Computer Communications 116 (2018) 240–252

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

# Anomaly detection mechanisms to find social events using cellular traffic data

Rosario G. Garroppo*,[a], Saverio Niccolini[b]

[a] Dipartimento di Ingegneria dell'Informazione, University of Pisa, Pisa, Italy
[b] Network Research Division, NEC Laboratories Europe, Heidelberg, Germany

ARTICLE INFO

*Keywords:*
Discrete stationary wavelets transform
Anomaly detection
Cellular traffic
Social events
Spatial analysis
Time analysis

ABSTRACT

The design of new tools to detect on-the-fly traffic anomaly without scalability problems is a key point to exploit the cellular system for monitoring social activities. To this goal, the paper proposes two methods based on the wavelet analysis of the cumulative cellular traffic. The utilisation of the wavelets permits to easily filter "normal" traffic anomalies such as the periodic trends present in the cellular traffic. The two presented approaches, denoted as Spatial Analysis (SA) and Time Analysis (TA), differ on how they consider the spatial information of the traffic data. We examine the performance of the considered algorithms using cellular traffic data acquired from one the most important Italian Mobile Network Operator in the city of Milan throughout December 2013.

The results highlight the weak points of TA and some important features of SA. Both approaches overcome the performance of one reference algorithm present in literature. The strategy used in the SA emerges as the most suitable for exploiting the spatial correlation when we aim at the detection of the traffic anomaly focused on the localisation of social events.

## 1. Introduction

Identifying real-world phenomena through the analysis of network traffic has recently attracted the attention of many researchers. This topic is one of the most promising in the context of cellular networks. In these systems, in addition to the traffic patterns of users, we have rough information on their geographic location. Cellular traffic data can easily be extended and augmented with external information, such as the location of nearby events, etc. Some mobile network operators have launched many research challenges and made available their traffic data to explore the potentialities that the analysis of cellular network data offers [1–3]. Based on these data, researchers have carried out a lot of studies considering different problems. For example, in [4] the focus is on the paths of user mobility, while in [5] the authors define a method to determine the levels of poverty in different geographical areas. Other works have proposed methods to link the traffic patterns to external real-world events and observations, see the survey [6] for details.

In this work, we study how the aggregate cellular traffic data can be used to identify social events. We pose this as an online anomaly detection problem and to solve it we propose scalable algorithms, which can run in the operational network. Our mechanisms do not require detailed information on traffic such as user Id, per packet or per flow information as in [7,8]. Instead, they use only cumulative information on the observed traffic at different Base Stations (BSs). The algorithms assume the continuous monitoring of the amount of traffic in non-overlapping Time Slots (TSs) and for each BS. We use wavelet decompositions to analyse the traffic over different timescales. This approach permits to easily filter periodic trends in the observed signals and to highlight the variation of traffic patterns in the time. The basic idea of the proposed methods is the exploiting of the spatial correlation of these traffic variations. The two presented approaches, i.e. Time Analysis (TA) and Spatial Analysis (SA), differ on how they consider the spatial information. In particular, TA is based on the application of hard thresholding de–noising techniques, which consider separately the traffic behaviour for each BS. At a particular TS, TA confirms a detected alarm in a BS only if in the neighbouring area more than a selected number of other BSs have triggered alarms. This strategy permits to take into account the spatial information of the traffic. On the contrary, at a particular TS, SA evaluates if to trigger or not an alarm by means of the analysis of the traffic variations over the whole network.

The rest of the paper is organised as follows. The next Section discussed the related work and summarises the paper contribution. Section 3 presents the background on wavelets theory, and in particular on the Discrete Wavelets Transform (DWT) and discrete Stationary Wavelets Transform (SWT). Section 4 describes the proposed

* Corresponding author.
*E-mail addresses:* r.garroppo@iet.unipi.it (R.G. Garroppo), Saverio.Niccolini@neclab.eu (S. Niccolini).

https://doi.org/10.1016/j.comcom.2017.12.009
Received 6 December 2016; Received in revised form 17 November 2017; Accepted 15 December 2017
Available online 16 December 2017
0140-3664/ 

approaches, while Section 5 presents the dataset and some preliminary tests. Section 6 discusses the simulation analysis carried out with the actual data. The analysis compares the performance of the proposed approaches and of one reference algorithm present in literature. Section 7 shows the detection performance of the proposed solutions carried out adding random artificial anomalies to the original data. Section 8 draws the concluding remarks.

## 2. Related work

In the last years, the problem of detecting traffic anomalies in cellular systems has attracted the attention of a lot of researchers. The results are a set of methods, which differ in the required traffic information, the kind of detected anomalies, the used tools, the online or offline operation, etc. As an example, in [7] the authors present a graph–based anomaly detection to find anomalies that can aide in the security of users, their phones, their personal information, and the companies that provide them services. In [8], the authors address the problem of automatic network traffic anomaly detection and classification using Machine Learning (ML) based techniques.

Different works have considered the design of methods to run on-the-fly in commercial networks. For example, in [9] the authors propose a method for identifying deviations in timeseries distribution based on a statistical change detection algorithm. The study is based on a large dataset from an operational 3G mobile network. The proposed method is able to cope with the marked non-stationarity and daily/weekly seasonality that characterise the traffic mix in a large public network. In [10], the authors present a method based on the change point detection applied to two sets of features extracted from DNS data. The symptomatic features are defined such that their abrupt change directly relates to the presence of abnormal and potentially harmful events, while diagnostic features shall provide contextual details of the anomalies, pointing to their root causes. These methods consider per-user traffic information. This approach requires a large amount of data to be processed, with a higher complexity of the monitoring platform. On the contrary, our methods are based on the knowledge of the aggregated traffic in each BS. This assumption implies a lower amount of data to consider in the detection algorithm, leading to a lower complexity of the monitoring system.

Traffic anomaly techniques are based on different approaches, as shown in [11–13]. For example, we can mention the machine learning approach proposed in [14,15], the combination of filtering and statistical methods discussed in [16], the technique based on principal subspace tracking suggested in [17], the traffic feature distributions used in [18,19], the utilisation of big data analytics presented in [20], the method based on the variation in the entropy associated to the network traffic [21], and the kernel recursive least squares used in [22].

We have chosen to exploit the properties of wavelet decomposition for designing our scheme. In [23], the authors suggested the wavelet transform for the modelling and the synthetic generation of multifractal traffic. More recently, this tool has represented the base for the design of some online traffic anomaly detection methods. In [24], the authors proposed a tool exploiting the wavelet packets in order to detect network traffic attacks in real time. The detection mechanism of the tool considers the iterated cumulative sums of squares (ICSS) algorithm and the Schwarz information criterion (SIC) algorithm for the identification of multiple variance change points in sequence data. These algorithms are integrated with another approach aimed at detecting sharp jumps and cusps in the data. However, the proposed method does not consider spatial information. In [25], the authors present a technique for traffic anomaly detection based on the analysis of the correlation of destination IP addresses in outgoing traffic at an egress router. The address correlation data are transformed through discrete wavelet transform for effective detection of anomalies through statistical analysis. The technique requires the IP address information, which is expensive to acquire.

In [26], the authors propose a signal analysis technique for detecting network traffic anomalies. They analyse the applications of general wavelet filters to the traffic data representing the byte and packet counts, over five minute sampling intervals, from wide-area routing links. They show that wavelets provide a powerful tool for isolating characteristics of signals via a combined time-scale representation. With respect to [26], we add the spatial information associated to the BSs in order to locate the alarm and then the related social event.

Focused on the anomaly detection for DoS attack prevention or for network malfunctioning identification, other works proposed wavelet theory to improve well-known techniques, e.g. CUSUM in [27], or to design new algorithms that jointly use other techniques, e.g. PCA in [28]. This paper differs from these works because the anomaly detection is aimed at the triggering of alarms related to social events. Furthermore, we exploit the available spatial information.

### 2.1. Cellular traffic data and real-world phenomena

A lot of works have used cellular traffic data to relate network information and real-world phenomena. Most of them are based on the data made available in the framework of the Data For Development (D4D) challenges [1,2].

In [6], the authors provide an extensive review of results on the analysis of mobile phone datasets, characterised by detailed information. A large amount of data in these datasets permit to carry out research on social networks, mobility, geography, urban planning, help towards development, and security. Detailed information available in the D4D datasets permits to study the relation between cellular traffic and social environment and events. For example, in [4] the authors describe some strategies to identify the cities, the city population, the strength of social ties among cities, the urban mobility in the largest city of Abidjan, the residential districts and the work areas. In [29], the authors have monitored six million users of a mobile network in Wuxi (China) from October 24, 2013, to March 24, 2014 for recovering individuals' commute routes. The dataset includes a huge amount of information, but it is not publicly available.

In this work, we use the Open Data of Telecom Italia (ODTI) [3], which contains cumulative information on the traffic acquired in different BSs. Unlike D4D datasets, ODTI does not provide data on individual users. ODTI dataset does not permit to carry out analysis aimed at finding anomalies for aiding in the security of users or at evaluating the mobility pattern. However, ODTI dataset is complete and contains all the measured data for each BS in each TS of the observation period. In [30] the authors present early experiments in predicting land use and demographics considering jointly heterogeneous open data and an ODTI [3] dataset. These ODTI data provide phone activity information over time and also over space (due to the positioning of transceiver towers, i.e. BSs), which is a strong indicator of the presence of people and of the mobility in urban environments. They demonstrate that an approach leveraging diverse datasets can be effective in aiding smarter urban planning efforts. In [31], the authors demonstrate that ODTI data can be used to infer information about the behaviour of foreign people in Milan, using simple statistical tools.

Recent studies aim at the estimation of the road traffic starting from mobility data acquired from cellular networks. As an example, the authors of [32–34] present systems able to detect incidents and predict travel times on main traffic roads. These systems are based on the monitoring of the mobility paths of cellular network devices or on the analysis of signalling traffic in the core network. The proposed approaches require per-user traffic information.

The contribution of this paper significantly differs from these previous works because we aim at defining an on-the-fly and scalable algorithm that permits us to isolate and characterise patterns that substantially deviate from the "normal" behaviour of the network. The idea is to regard the traffic load of the different BSs in a large cellular