Computer Networks 57 (2013) 3446–3462

Contents lists available at ScienceDirect

Computer Networks

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

Distribution-based anomaly detection via generalized likelihood ratio test: A general Maximum Entropy approach

A. Coluccia [a,*], A. D'Alconzo [b], F. Ricciato [a]

[a] University of Salento, via Monteroni, 73100 Lecce, Italy
[b] Forschungszentrum Telekommunikation Wien (FTW), Vienna, Austria

ARTICLE INFO

Article history:
Received 4 March 2013
Received in revised form 20 June 2013
Accepted 29 July 2013
Available online 11 August 2013

Keywords:
Anomaly detection
Maximum Entropy (ME)
Network traffic
Generalized Likelihood Ratio Test (GLRT)
Maximum Likelihood (ML)
3G cellular networks

ABSTRACT

We address the problem of detecting "anomalies" in the network traffic produced by a large population of end-users following a distribution-based change detection approach. In the considered scenario, different traffic variables are monitored at different levels of temporal aggregation (timescales), resulting in a grid of variable/timescale nodes. For every node, a set of per-user traffic counters is maintained and then summarized into histograms for every time bin, obtaining a timeseries of empirical (discrete) distributions for every variable/timescale node. Within this framework, we tackle the problem of designing a formal Distribution-based Change Detector (DCD) able to identify statistically-significant deviations from the past behavior of each individual timeseries.

For the detection task we propose a novel methodology based on a Maximum Entropy (ME) modeling approach. Each empirical distribution (sample observation) is mapped to a set of ME model parameters, called "characteristic vector", via closed-form Maximum Likelihood (ML) estimation. This allows to derive a detection rule based on a formal hypothesis test (Generalized Likelihood Ratio Test, GLRT) to measure the coherence of the current observation, i.e., its characteristic vector, to the given reference. The latter is dynamically identified taking into account the typical non-stationarity displayed by real network traffic. Numerical results on synthetic data demonstrates the robustness of our detector, while the evaluation on a labeled dataset from an operational 3G cellular network confirms the capability of the proposed method to identify real traffic anomalies.

1. Introduction

Modern data and communication networks are exposed to many types of problems and security threats. In order to respond quickly and minimize service degradation, network operators require tools capable to promptly detect "abnormal" traffic conditions, i.e. *anomalies*. This is even more demanding in third-generation (3G) cellular networks, which are highly heterogeneous, complex and constantly evolving systems, and as such are exposed to unanticipated types of problems and threats [1–3]. Anomaly Detection (AD) in network traffic is a well-explored field, and several different techniques have been proposed (see e.g. [4,5] and references therein). Generally speaking, the statistical-based AD approach seeks to identify a *reference* representative of the "normal" behavior and then look for any "significant" deviation from it. In other words, *anomaly* is defined as anything deviating from the expected behavior—*expectation* is a key concept here [6]. Thus, a complete AD scheme consists logically of a *reference identification* method followed by a *detection rule* for testing consistency between the observed data and the reference. As the state of the network and the behavior of its users change (e.g. following daily and weekly cycles, and long-term trends) so do the notions of "normal" behavior

[*] Corresponding author. Tel.: +39 0832297206.
E-mail addresses: angelo.coluccia@unisalento.it (A. Coluccia), dalconzo@ftw.at (A. D'Alconzo), fabio.ricciato@unisalento.it (F. Ricciato).

1389-1286/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comnet.2013.07.028

and “significant” deviation. Therefore the AD system should be adaptive: the reference identification as well as the detection rule must be dynamically updated so as to track the physiological changes in the traffic patterns.

Statistical-based AD can be applied to virtually any type of temporally-structured traffic data, or *traffic representation*, from coarse scalar time-series (e.g. of total volume or entropy) to finer-grain multidimensional representations (e.g. vectors, sketchs, histograms) of the underlying traffic process, extracted by some more or less involved procedure that typically requires feature selection, aggregation and tracking of per-flow states [7]. Moreover, in order to detect anomalies occurring at different timescales, the AD system should consider traffic data at different levels of temporal aggregation (*multi-resolution*). Operators of access networks are particularly concerned with revealing macro-anomalies, i.e. events that affect many network users (i.e., their “customers”) rather than micro-anomalies with impact limited to one or a few users, since the former more likely point to a problem in the shared network or service infrastructure. This motivates us to consider a *distribution-based approach*, where the network traffic is represented by (a set of) traffic distributions across users. In this way, we aim at profiling the aggregate behavior of the whole user population, rather than of individual users, which is in line with the goal of capturing macro-anomalies. More specifically, we consider a reference scenario where a passive monitoring system measures multiple *traffic variables*—e.g. number of packets of a certain type, such as “number of TCP SYN packets sent in uplink to port 80” or “number of distinct IP addresses contacted” or “volume of traffic on port 25” and so on (we will detail the formulation in Section 3)– for each individual user and at different temporal aggregation scales, from 1 min up to 1 day. For every variable and timescale, the data observed in each time-bin is summarized into a binned histogram—where bins are intervals that partition the span of the variable—that represents the empirical distribution of that variable across users. Therefore, we obtain a set of distribution timeseries, each referring to a different traffic variable and timescale. Each timeseries is then processed by a separate instance of a Distribution Change Detector (DCD) that learns adaptively the “normal” reference profile and detects if the current observation deviates “significantly” from the reference.

In the given reference scenario the number of variable/timescale combinations is large, and each follows a specific profile and temporal pattern different from the others. It would be practically infeasible to tailor the design and parametrization of the DCD module to each and every timeseries, therefore a suitable DCD should fulfill the following requirements:

- *Versatility*: to model different traffic variables at different timescales and aggregation, without manual tuning.
- *Adaptiveness*: to adjust the reference identification and the detection rule to the physiological changes in the traffic composition.
- *Low-complexity*: to allow on-line implementation for a sufficiently large number of variables.

The goal of this paper is to derive a DCD with such properties by following a theoretically-grounded methodology.

The statistical-based AD approaches present in the literature can be grouped into two main categories: model-based (parametric) and model-free (non-parametric). The former (e.g. the popular CUSUM [8,9] and other classical hypothesis testing methods) are based on strict *a priori* assumptions on the statistical characteristics of the data, which allow formal tractation and better control over the achievable performances (e.g. the Probability of False Alarm, PFA) but lack versatility. On the other hand, model-free methods based on general non-parametric techniques (e.g., PCA [10] to name just one) or simple heuristics (e.g. [7]) are more flexible but, lacking a formal hypothesis testing framework, must resort to heuristic detection rules which are difficult to control. In the present contribution *we aim at breaking such a tension between the need for a tractable statistical model and the elusive peculiarities of empirical data*. The proposed methodology is somehow a “third way” in between the model-based and model-free avenues. We leverage a Maximum Entropy (ME) approach to learn from empirical data the statistical model with the highest probability of being close to the underlying distribution out of a broad set of distributions, i.e. the Gibbs family. The set of ME model parameters, estimated via a Maximum Likeihoood (ML), represent the “characteristic vector” associated to the empirical distribution. Based on the latter, we derive a formal hypothesis test to establish whether the current sample is “compatible” with a reference extracted from selected past observations.

The formal test is obtained by applying the Generalized Likelihood Ratio Test (GLRT) theory to the problem at hand. This is a general approach for hypothesis testing in multidimensional data, which yields a formal and rigorous test provided that an *a priori* statistical model for the data is available and that the Maximum Likelihood (ML) estimation of its parameters can be obtained, possibly in closed-form to grant reasonable computational complexity. This makes the GLRT approach very powerful but difficult to derive, unless very tractable models are postulated (e.g., Gaussian). Indeed, real network traffic does not exhibit a simple statistical behavior and is non-stationary, i.e. it is difficult to model in a simple way. We introduced the general idea of GLRT for anomaly detection in network traffic in [11], without providing however any particular algorithm for the modeling. To the best of our knowledge, the only other attempt to apply a GLRT to anomaly detection in network traffic is [12], limited to data that can be modeled by α-stable distributions. An additional drawback of this approach is that such a model has no closed-form for the distributions, therefore powerful numerical methods are needed for parameter estimation. Furthermore, the lack of an analytical form impedes the derivation of a low-complexity GLRT detector. Conversely, our approach is general since does not make any assumption on the data at hand. The key idea is to use a Maximum Entropy (ME) approach for obtaining a general parametric model, which opens the door to formal hypothesis testing, hence to GLRT. Furthermore, thanks to the derivation in closed-form of the ML estimator of the characteristic vector, the