Machine learning, data mining and Big Data frameworks for network monitoring and troubleshooting

Alessandro D'Alconzo[a,*], Pere Barlet-Ros[b], Kensuke Fukuda[c], David Choffnes[d]

[a] *AIT, Austrian Institute of Technology, Vienna, Austria*
[b] *UPC BarcelonaTech, Barcelona, Spain,*
[c] *National Institute of Informatics, Tokyo, Japan*
[d] *Northeastern University, Boston, USA*

1. Introduction

The scale and the complexity of the Internet has dramatically increased in the last few years. At the same time, Internet services are becoming increasingly complex with the introduction of cloud infrastructures, Content Delivery Networks (CDNs) and mobile Internet usage. This complexity will continue to grow in the future with the rise of Machine-to-Machine communication and ubiquitous wearable devices. In this scenario it becomes even more compelling and challenging to design scalable network traffic monitoring and analysis tools able to shed light on the complex interplay between network infrastructure and the traffic profiles generated by a continuously growing number of applications.

The current and future network monitoring frameworks cannot rely only on information gathered at a single network interconnect, but must consolidate information from various vantage points distributed across the network. Many systems for the extraction of operational statistics from computer network interconnects have been designed and implemented in the last decades. Those systems generate huge amounts of data in various formats and granularity from several vantage points and devices, ranging from packet level data to statistics about whole flows and system logs. However, despite recent major advances of Big Data analysis frameworks, their application to the network traffic monitoring and analysis domain remains poorly understood and investigated.

Furthermore, critical applications such as detection of anomalies, network attacks and intrusions, require fast mechanisms for online analysis of thousands of events per second, as well as efficient techniques for offline analysis of massive historical data. Statistical modeling, data mining and machine learning-based techniques able to detect, characterize, and troubleshoot network anomalies and security incidents, promise to efficiently shed light on this enormous amount of data. Nonetheless, the unprecedented size of the data at hand poses new performance and scalability challenges to traditional methods and tools.

The purpose of this special issue is to bring together state-of-the-art studies proposing novel scalable techniques and frameworks capable of collecting and analyzing massive datasets of network traffic traces and performance measurements, along with novel methodologies and techniques able to extract information from the gathered data for the purpose of tackling typical networking problems such as performance characterization, security and troubleshooting.

Out of 40 submissions, 8 articles have been selected after at least two review rounds. All papers have received at least three reviews from experts in the different areas including network measurements, traffic analysis, data mining and machine learning. In addition the two best papers (on topics related to the Special Issue) of the 8th International Workshop Traffic Monitoring and Analysis (TMA'16)[1] have been accepted in their extended version, after peer reviewing. The Spe-

*Corresponding author
Email addresses: alessandro.dalconzo@ait.ac.at (Alessandro D'Alconzo), pbarlet@ac.upc.edu (Pere Barlet-Ros), kensuke@nii.ac.jp (Kensuke Fukuda), choffnes@ccs.neu.edu (David Choffnes)

[1] tma.ifip.org/2016/

cial Issue consists of 10 papers organized in four groups: (I) Frameworks and methodologies, (II) Security and troubleshooting, (III) Distributed measurements for network characterization, and (III) Network applications characterization.

The first part of the special issue is devoted to frameworks and methodologies to collect, process and analyze big datasets obtained from monitoring large-scale networks. In fact, when considering passively collecting and then processing network traffic traces, the need to analyze raw data at several Gbps and to extract higher level indexes from the stream of packets poses typical big data-like challenges.

In *DBStream: A Holistic Approach to Large-Scale Network Traffic Monitoring and Analysis*, **A. Bär et al.** present DBStream, a holistic approach to large-scale network monitoring and analysis applications. They show how its Continuous Execution Language (CEL) can be used to automate several data processing and analysis tasks typical for monitoring operational ISP networks. The paper discusses the performance of DBStream as compared to MapReduce processing engines and shows how intelligent job scheduling can increase its performance even further. Furthermore, the paper shows the versatility of DBStream by explaining how it has been integrated to import and process data from two passive network monitoring systems, namely METAWIN and Tstat. Finally, multiple examples of network monitoring applications are given, ranging from simple statistical analysis to more complex traffic classification tasks applying machine learning techniques using the Weka toolkit.

In *Statistical Network Monitoring: Methodology and Application to Carrier-Grade NAT*, **E. Bocchi et al.** engineer a methodology to extract, collect and process passive traffic traces. In particular, they design and implement analytics that, based on a filtering process and on the building of empirical distributions, enable the comparison between two generic collections, e.g., data gathered from two different vantage points, from different populations, or at different times. The ultimate goal is to highlight statistically significant differences that could be useful to flag incidents for the network manager. As a use-case the authors apply the proposed methodology to assess the impact of Carrier-Grade NAT (CGN), a technology that Internet Service Providers (ISPs) deploy to limit the usage of expensive public IP addresses. They process a large dataset of passive measurements collected from an ISP using CGN for part of its customers, extract detailed per-flow information, derive higher level statistics, and finally look for statistically significant differences in connectivity and performance figures of customers being offered public or private addresses.

The second part of the special issue addresses typical network management tasks such as troubleshooting and security.

Growing network complexity mandates automated tools and methodologies for troubleshooting. In *Framework, Models and Controlled Experiments for Network Troubleshooting*, **F. Espinet et al.** follow a crowd-sourcing approach and argue for the need to deploy measurement probes at the edge of the network, which can be either under the control of the users (e.g., end-user devices) or the ISP (e.g., home gateways), and that raises an interesting tradeoff. They define a framework for network troubleshooting, and its implementation as open source software named NetProbes. In data mining terms, depending on the amount of information available to the probes (e.g., ISP topology), they formalize the network troubleshooting task as either a clustering or a classification problem. In networking terms, these algorithms allow respectively end-users to assess the severity of the network performance degradation, and ISPs to precisely identify the faulty link. Both problems are solved with an algorithm that achieves perfect classification under the assumption of a strategic selection of probes (e.g., assisted by an ISP), and they assess its performance degradation under a naive random selection.

The increasing number of attacks against computing infrastructure, which is of critical importance for enterprises, drives the need to deploy progressively more sophisticated defense solutions to protect network assets. An essential component of the defense are Intrusion Detection Systems (IDS) searching for evidence of ongoing malicious activities (network attacks) in network traffic crossing the defense perimeter. Many intrusion detection systems are implemented as ensembles of relatively simple, yet heterogeneous detectors, where some of them can be specialized to particular types of intrusions, whereas others can be general anomaly detectors capable of detecting previously unseen attacks at the expense of higher false alarm rates. **M. Grill and T. Pevný** present in *Learning Combination of Anomaly Detectors for Security Domain*, a novel

technique of finding a convex combination of outputs of anomaly detectors maximizing the accuracy in τ-quantile of most anomalous samples. Such an approach better reflects the needs of the security domain in which subsequent analysis of alarms is costly and can be done only on a small number of alarms. An extensive experimental evaluation and comparison to prior art using real network data on two existing intrusion detection systems shows that the proposed method not only outperforms prior work, but it is also more robust to noise in training data labels, which is another important feature for deployment in practice.

Characterization of flows by temporal patterns enables traffic classification and filtering for network management and network security in situations where full packet data is not accessible (e.g., obfuscated or encrypted traffic) or cannot be analyzed due to privacy concerns or resource limitations. **F. Iglesias and T. Tseby** study the temporal behavior of communication flows in the Internet in *Time-Activity Footprints in IP Traffic*. They define a time activity feature vector that describes the temporal behavior of flows. Then cluster analysis is used to capture the most common time activity patterns in real Internet traffic using traces from the MAWI dataset. They find a set of 7 time-activity footprints and show that about 95% of the analyzed flows can be characterized based on such footprints, which represent different behaviors for the three main protocols, i.e., TCP, UDP and ICMP. In addition, they report that the majority of the observed flows consisted of short, one-time bursts, corresponding to a large number of scanning, probing, DoS attacks and back-scatter traffic in the network, whereas flows transmitting meaningful data became outliers among short, one-time bursts of unwanted traffic.

The third part of the special issue covers exploitation of distributed measurements for characterization of distributed large-scale networks and services.

In particular, *The Good, the Bad and the Implications of Profiling Mobile Broadband Coverage* from **A. Lutu et al.**, investigates the prevalent mobile network coverage profiles capturing the availability of the different access technology in an area. Indeed, given the increasing heterogeneity of technologies in the last mile of Mobile Broadband (MBB) networks, further support for seamless connectivity across multiple network types relies on understanding the prevalent network coverage profiles. Correlating these coverage profiles with network performance metrics is of great importance both for regulators and operators in order to forestall disturbances for applications running on top of MBB networks. For this purpose, they authors deploy custom measurement nodes on-board five Norwegian inter-city trains and collect a unique geo-tagged dataset along the train routes. Then they build a coverage mosaic, where they divide the routes into segments, and analyze the coverage of individual operators in each segment. The prevalent coverage profiles of MBB networks along the train routes, and the classification of each segment is obtained using hierarchical clustering. Finally, the areas classified within each profile are assessed in terms of the packet-loss and HTTP download performance.

Performance of CDNs is difficult to characterize because it depends on a number of factors such as scale and distributed nature. Furthermore, as the volume of content served by CDNs grows, these networks evolve over time to improve performance. In *DBit: Assessing Statistically Significant Differences in CDN Performance*, from **Z. Akhtar et al.** develop a methodology called DBit that can determine whether one CDNs user-perceived performance is statistically different from another. They validate DBit and demonstrate its usefulness on CDNs used for photo and video delivery. PlanetLab and and RIPE Atlas nodes hosted in end-user homes in about 1500 autonomous systems worldwide are used to obtain photo and video fetches from three Photo CDNs and two Video CDNs. The authors show that DBit can identify significant performance differences not just between CDNs, but also across time and location.

Similarly, *Multidimensional Cloud Latency Monitoring and Evaluation* from **O. Tomanek et al.**, deals with measurement and performance evaluation of a cloud service. In particular the work focuses on cloud-service latency from the user's point of view, using the multidimensional latency measurements obtained from CLAudit, an in-house designed active-probing platform, deployed across PlanetLab and Microsoft Azure data centers. The multiple geographic Vantage Points, multiple protocol layers and multiple data center locations of CLAudit measurements allow pinpointing with great precision if, where and what kind of a particular latency-generating event has happened. The authors demonstrate the utility of the multidimen-

sional approach and document the differences in the measured Cloud-services latency over time.

The last part of the special issue covers network applications characterization.

A typical problem is measurement of network Quality of Service (QoS) which has attracted considerable research effort over the last two decades. The recent trend towards Internet Service Providers (ISP's) offering application-specific QoS is creating possibilities for more sophisticated QoS metrics to be offered by ISP's in service level agreements. This in turn could be used for the purposes of improved network optimization and billing according to application specific QoS guarantees. **S. Middleton and S. Modafferi** report in *Scalable Classification of QoS for Real-Time Interactive Applications from IP Traffic Measurements* about a scalable near real-time approach using passively logged IP traffic data for classification of application latency and packet loss across a range of real-time interactive applications. They run experiments with popular online games, a video streaming application as well as a VoIP application, using a mixture of laboratory and real-world deployments, with run times ranging from hours to days, and observe a combination of real and simulated ISP latency and packet loss events. They show that their classifier achieves high accuracy, while retaining near real-time performance. With new business models between ISP's and application developers being actively considered this work represents a significant contribution to the debate by providing scientific evidence relating to a novel approach to scalable QoS measurement.

Social networks are among the most popular applications in the nowadays Internet, and community detection is one of the important methods for understanding the mechanism behind the function of social networks. The recently developed label propagation algorithm (LPA) has been gaining increasing attention because of its excellent characteristics, such as a succinct framework, linear time and space complexity, easy parallelization, etc. However, several limitations of the LPA algorithm, including random label initialization and greedy label updating, hinder its application to complex networks. In *A social community detection algorithm based on parallel grey label propagation*, **Q. Zhang et al.** propose a new parallel LPA algorithm. First, grey relational analysis is integrated into the label updating process, which is based on vertex similarity. Second, parallel propagation steps are comprehensively studied to utilize parallel computation power efficiently. Finally, randomness in label updating is significantly reduced via automatic label selection and label weight thresholding. Experiments conducted on artificial and real social networks demonstrate that the proposed algorithm is scalable and exhibits high clustering accuracy.

The Guest Editors would like to thank the authors, the Journal Editor-in-chief and especially the reviewers. The time and efforts they have devoted to provide detailed comments and suggestions has contributed to significantly improve the technical and scientific level, as well as the presentation quality of the accepted papers.

This special issue is the outcome of the dissemination activity of the European Union FP7 project mPlane (grant agreement n. 318627) at the IFIP TC6 2014/2 Strategic Review Meeting in Dagstuhl (event n. 14463). Participants of the mPlane project were actively involved as reviewers and authors of the papers. The work devoted to the special issue is partly supported by the Vienna Science and Technology Fund (WWTF) through project ICT15-129, Big-DAMA[2], by the Spanish Ministry of Economy and Competitiveness and EU FEDER, under grant TEC2014-59583-C2-2-R (SUNSET project).

List of reviewers

- Osamu Akashi
- Bahaa Al-Musawi
- Zied Aouini
- Zachary Bischof
- Enrico Bocchi
- Nevil Brownlee
- Fabian Bustamante
- Christian Callegari
- Pedro Casas
- Ignacio Castro
- Angelo Coluccia
- Luca Deri
- Giorgos Dimopoulos
- Benoit Donnet
- Maurizio Dusi

[2] `https://bigdama.ait.ac.at/`

- Jeroen Famaey
- Olivier Festor
- Pierdomenico Fiadino
- Alessandro Finamore
- Romain Fontugne
- Jerome Francois
- Yongquan Fu
- Kostas Giotis
- Eduard Glatz
- Felix Iglesias Vazquez
- Hyun-chul Kim
- Mirja Khlewind
- Minseok Kwon
- Guy Leduc
- Yeonhee Lee
- Fangfan Li
- Toru Mano
- Alessandro Margara
- Johan Mazel
- Marco Milanesio
- Daisuke Miyamoto
- Giovane C. M. Moura
- Jens Myrup Pedersen
- Ashkan Nikravesh
- Philippe Owezarski
- Maria Papadopouli
- Antonio Pescapè
- Michalis Polychronakis
- Javier Ramos
- Fabio Ricciato
- Philipp Svoboda
- Hajime Tazaki
- Brian Trammell
- Stefano Traverso
- Matteo Varvello
- Safaa Zeidan
- Liang Zhu
- Tanja Zseby

Guest Editors biographies

Alessandro D'Alconzo received the M.Sc. degree in Electronic Engineering with honors in 2003, and the Ph.D. in Information and Telecommunication Engineering in 2007, from Polytechnic of Bari, Italy. He is Scientist in the Digital Safety & Security department of AIT, Austrian Institute of Technology. From 2007 to 2015, he was Senior Researcher in the Communication Networks Area of the Telecommunications Research Center Vienna (FTW). From 2008 to 2013 he has been Management Committee representative for Austria and Secretary of the COST Action IC0703 *Traffic Monitoring and Analysis*. He has extensive experience in contributing and managing EU funded projects, as well as in applied research projects in the field of network traffic measurements in collaboration with national telecommunication operators. His research interests embrace network measurements and traffic monitoring, ranging from design and implementation of statistical based anomaly detection algorithms and root cause analysis, to Quality of Experience evaluation, and application of secure multiparty computation techniques to cross-domain network monitoring and troubleshooting.

Pere Barlet-Ros received the M.Sc. and Ph.D. degrees in Computer Science from the Universitat Politecnica de Catalunya (UPC) in 2003 and 2008, respectively. He is currently an Associate Professor with the Computer Architecture Department and a Senior Researcher with the Advanced Broadband Communications Center of the UPC. He is the Chairman and co-founder of Network Polygraph (`https://polygraph.io`), a start-up company that provides network visibility as a cloud service. He was also a visiting researcher with Endace, New Zealand (Winter 2004), Intel Research Cambridge, UK (Summer 2004) and Intel Labs Berkeley, California (Summer 2007). His research interests are in the fields of network monitoring, network data analysis, traffic classification, anomaly detection, SDN measurement and online privacy.

Kensuke Fukuda is an associate professor at the National Institute of Informatics (NII). He earned his Ph.D degree in computer science from Keio University in 1999. He worked in NTT laboratories from 1999 to 2005, and joined NII in 2006. He was a visiting scholar at

Boston University in 2002 and a visiting scholar at the University of Southern California / Information Sciences Institute in 2014-2015. He was also a researcher of PRESTO JST (Sakigake) in 2008-2012. His current research interests are Internet traffic measurement and analysis, intelligent network control architectures, and scientific aspects of networks.

David Choffnes received his M.S. and Ph.D. in Computer Science from Northwestern University in 2006 and 2010, respectively. He is currently an assistant professor in the College of Computer and Information Science at Northeastern University. Previously he was a postdoctoral research associate at the University of Washington. His research is primarily in the areas of distributed systems and networking, focusing on mobile systems and privacy.