Big data analytics for automated QoE management in mobile networks

Antonio J. García, Matías Toril, Pablo Oliver, Salvador Luna-Ramírez, and Rafael García

***Abstract*—Over the last years, there has been a significant increase in the number of services in mobile networks. Such a trend has forced operators to change their network management processes to ensure an adequate user Quality of Experience (QoE), instead of an adequate Quality of Service (QoS). As a result, Customer Experience Management (CEM) is now a critical task for mobile network operators, which demand tools for QoE monitoring on a user basis. With the latest advances in information technologies, the newest traffic monitoring and analysis (TMA) solutions can leverage the huge amount of information available from network elements and interfaces in mobile networks. However, data processing algorithms in these tools are still to be defined. In this work, we review the shortcomings and challenges in the use of TMA applications in mobile networks, and how these can be empowered by big data analytics (BDA). For this purpose, a methodology to validate a generic big-data driven TMA framework with user terminal agents in a real cellular network is outlined. Then, a use case is presented to show the potential and limitations of these applications for monitoring end-user QoE in a live Long-Term Evolution (LTE) network.**

***Index Terms*—Big data, mobile network, management, automation, QoE, traffic monitoring**

I. Introduction

In the last years, there has been an exponential growth in the demand of mobile services. In parallel, the success of smartphones and tablets has changed traffic patterns in mobile networks. Changes will continue in future 5G networks with the introduction of network virtualization and machine-type communications, making traffic management a very challenging task.

At the same time, the constant increase in users' expectation has forced operators to change the way they manage their networks. Legacy management processes, focused on network performance and Quality of Service (QoS), have been replaced by a more modern approach focused on user satisfaction, referred to as Quality of Experience (QoE). This new paradigm has become a key differentiating factor in a market where networks and services are quite similar between operators. As a result, Customer Experience Management (CEM) is now one of the most important tasks for mobile operators [1].

For simplicity, operators currently monitor QoE based on data provided by network equipment (e.g., counters or logs). From this data, key performance indicators related to network resources (Resource Key Performance Indicators, R-KPIs) are computed. Occasionally, this data is complemented with measurements from simple network protocol analyzers, limited in space (one or a few interfaces) and time (a short period). Thus, mobile operators discard a huge amount of information in the form of measurements and interaction registers generated by their networks [2]. Such data volume will be even larger in 5G with the addition of new network nodes in ultra-dense cell deployments and machine-type communications [3].

With recent advances in information technologies, it is now possible to process massive volumes of information with big data analytics (BDA) platforms [4]. 'Big Data' refers to data that cannot be processed by traditional means due to its volume, velocity and variety (e.g., connection traces or packet-level traffic). With BDA, operators can better explain network performance by discovering hidden relationships between system variables (self-awareness) and take corrective actions proactively by predicting trend changes (self-adaptiveness). For this reason, specifying BDA systems has become a priority for standardization bodies (e.g., ITU-T Study Group 13 [5]).

An area of great potential for BDA is Traffic Monitoring and Analysis (TMA). Legacy TMA tools [6] were focused on the overall network performance. In contrast, the latest TMA solutions are able to passively monitor all the traffic crossing the network at a very fine granularity. This is achieved by adding new network elements (deep packet inspectors) that store a full or partial copy of each frame from different protocol layers. Then, packet-level traffic analysis allows to build service-specific performance metrics with end-to-end network visibility (Service KPI, S-KPI), which can be mapped more easily into QoE figures. Such an approach is already used by the newest TMA solutions based on big networking data (e.g., [7]). Nonetheless, algorithms inside these platforms are still to be defined [2].

In this work, we revise the shortcomings and challenges in the use of these applications for mobile networks, and how these can be empowered by BDA. Architectural and implementation issues are not covered here. For clarity, a generic framework for big-data driven TMA in mobile networks is first introduced. Then, a methodology to validate these applications is described. Later, a real use case is presented to show the potential of BDA for analyzing QoE in a live mobile network. Finally, open issues are discussed.

II. Big-data driven TMA framework for QoE analysis

Fig. 1 shows the structure of a generic TMA application for QoE analysis in mobile networks. Most of its components are included in the newest TMA solutions (e.g., [7]). Three main layers can be differentiated.

Big data: Traffic monitoring at network layer is performed by sniffing transit packets from selected network interfaces.

Antonio J. García, Matías Toril, Pablo Oliver and Salvador Luna-Ramírez are with the University of Málaga.

Rafael García is with Ericsson Spain.

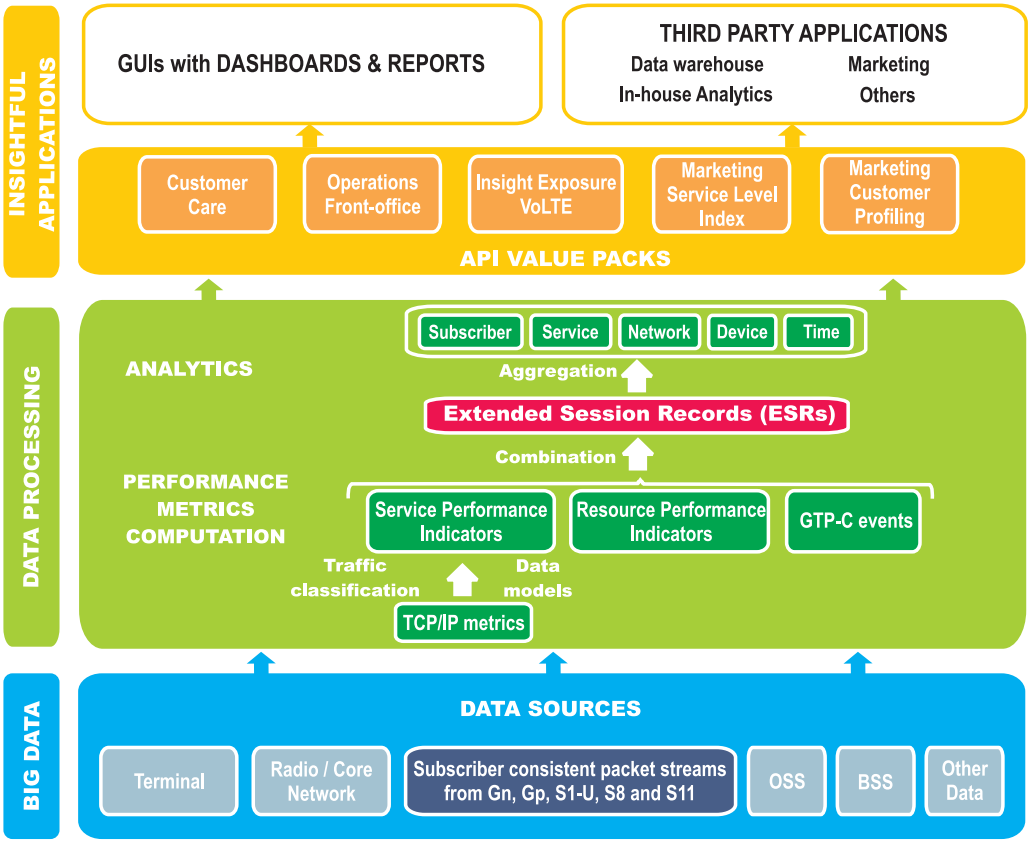

Fig. 1. Basic scheme of a TMA application for QoE monitoring.

The *pcap* (packet capture) Application Programming Interface (API) is often used for this purpose [8]. Traffic is classified on a session and subscriber basis by Internet Protocol (IP) address and International Mobile Subscriber Identity (IMSI). Then, information is enriched with data gathered by radio/core network equipment, Operating/Business Support Systems (OSS/BSS) or user terminals controlled by the operator. Context information (e.g., time of day, location, weather or device type) can be added to aid troubleshooting and optimization processes.

Data processing: The inputs of data processing are Transmission Control Protocol (TCP) performance metrics, such as IP session throughput, packet loss ratio or round-trip time (RTT), computed from packet-level analysis on a session basis. Such metrics can be aggregated per service by classifying packets depending on service class. Current service list includes real-time conversational (e.g., Voice-over-LTE, Skype, online games) and broadcast (e.g., Internet radio/television) services, and non-real time streaming (e.g., YouTube, Spotify), interactive (e.g., web browsing, social networking, app download) and background (e.g., email, file sharing, cloud services) services. Events of GPRS Tunneling Protocol for Control Plane (GTP-C) are also supplied, providing information in signaling flows in the core network.

Insightful Applications: Graphical Users Interfaces (GUIs) with dashboards are implemented to assess the QoE throughout the network. Third party applications may also use this structure for data warehousing, in-house analytics or marketing. Understanding customer behavior is key to create more effective promotions and pricing strategies, which ultimately improve user satisfaction and reduce churn [9].

The data processing workflow is as follows. Passive traffic monitoring is implemented at key network links, totally standardized, reducing as much as possible the interaction with production equipment. Such a monitoring point is referred to as a *probe* point. Ideally, the probe should be located in core network interfaces, where traffic from large geographical areas is aggregated. In 3G networks, the Gn link between the Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN), carrying both control and user plane data, is the preferred interface to monitor end-to-end performance. In 4G and first non-standalone 5G New Radio networks, monitored links are the S11 interface between the Mobility Management Entity (MME) and Serving Gateway (SGW) in the control plane, and S1-U between the eNodeB and SGW in the user plane. Cross-correlating the information from these interfaces is needed to identify the user. Additional interfaces, such as Gp in 3G and S8 in 4G, are monitored in roaming scenarios. Likewise, different interfaces of IP Multimedia Subsystem (IMS) may be monitored (e.g., Mm link) to collect messages exchanged between IMS core and external IP networks. All this information is combined with performance counters from network equipment or more refined indicators from connection traces.

Later, data processing derives insights for QoE management from the available big data, consisting of: a) subscriber packet streams from key network interfaces, and b) counters and trace

files with signaling events on a subscriber basis, generated by nodes in the radio and core network. The output of data processing consists of Extended Session Records (ESRs) [9]. These are periodic records generated per-service for each subscriber, combining S-KPIs with data from network elements involved in the session.

Data processing includes traffic classification, TCP/IP metrics calculation and S-KPI estimation. Once packet-level traces are stored, traffic classification is performed to determine the specific service for each traffic flow. At this point, TCP/IP (for short, TCP) performance metrics can easily be computed from raw data. This is possible because TCP flow control makes that traffic dynamics is correlated to end-to-end performance. Thus, a problem affecting a TCP flow at some point (e.g., radio interface) can be detected by observing traffic at a different location along the path (e.g., network probe at a core network interface). As IP address is included in monitored traffic, TCP metrics can be separated by link direction (downlink or uplink). Moreover, TCP metrics can also be segregated by network segment (probe-to-terminal or probe-to-Internet) by correlating payload and acknowledge messages in both directions of the interface. To reduce computational load, TCP metrics are computed only for selected services and sufficiently large data bursts.

After traffic classification, S-KPIs are calculated on a user and session basis. This is performed by searching for relevant actions in the service. For instance, web download time is calculated as the time gap between the first Hypertext Transfer Protocol (HTTP) GET message sent by the user and the last HTTP response message replied by service provider. S-KPI estimates are then combined with other KPIs obtained from different network segments (e.g., radio or core). After time alignment for data merging, output is included in ESRs, broken down by user, service, radio access technology or service provider. As a result, every ESR contains a vast amount of subscriber and session specific data, from which end-to-end performance can be estimated. Thus, it is possible to pinpoint user sessions with unacceptable S-KPI values. Finally, in order to make data more practicable, ESRs are aggregated on a per-period basis (generally, 1-5 minutes) by keeping the granularity indicated above.

III. Analytics

The data collected by TMA can be used offline for knowledge discovery by data analytics. Fig. 2 shows several interrelated disciplines involved in this process, from the simplest data visualization steps in Exploratory Data Analysis (EDA) to the most sophisticated Machine Learning (ML) algorithms. The aim here is to build models for classifying, characterizing and predicting the performance of each individual session, for which the most important features must be selected.

1) *Classification*: A model is needed to segregate sessions per service class. In the past, this was done by analyzing protocol messages, which is not possible anymore due to traffic encryption. Alternatively, user sessions can be grouped by checking traffic attributes. This can be done by heuristic rules or automatic clustering algorithms based on unsupervised learning (e.g., k-means) [10].
2) *Regression*: A model must be found to estimate S-KPIs from TCP metrics. Such a mapping can be derived from measurements taken with terminal agents in lab environments. Then, model construction can be performed by classical regression techniques (e.g., generalized linear regression) or more complex supervised learning algorithms (e.g., support vector machines, neural networks or ensemble algorithms) [11].
3) *Forecasting*: A model must be derived to predict QoE trends on different time scales, so that proactive control decisions can be taken driven by QoE criteria. This can be done by traditional time series analysis (e.g., ARIMA) or regression based on supervised learning (e.g., support vector machines, neural networks). The latter can take advantage of historical data from similar network elements to improve prediction accuracy.
4) *Feature selection*: The most significant TCP metrics for each S-KPI must be identified. Reducing the number of variables in the model reduces the computational load, speeds up the learning process, improves generalization capability and makes interpretation easier for the operator.

In these processes, statistical modeling can be used to isolate the effect of an important variable or ensure that models are interpretable. In contrast, ML is preferred when high-order interactions between predictors are expected or estimation accuracy is the main goal.

IV. Validation of a big-data driven TMA application

The accuracy of S-KPI estimates obtained with a TMA application needs to be assessed. In this section, a generic methodology to check the accuracy of S-KPI estimates with a *Terminal Agent* (TA) is described.

Equipment. TAs are software applications running on commercial user terminals that include a S-KPI analyzer together with automatic procedures that mimic user interactions. TAs have direct access to one side of the communication channel, so they can measure end-to-end performance (and, hence, S-KPIs) accurately.

Automation. Validation tests include data captures from both TA and probe (i.e., user and network interface). On the one hand, the TA automatically triggers service requests as the user would do. In parallel, the TMA application sniffs packets at network layer to draw estimates of the end-to-end performance experienced by the TA. Finally, S-KPI estimates from TMA are compared against real S-KPI values measured by the TA.

Experiments. Tests must cover a wide variety of scenarios (i.e., most demanded services, user profiles and network conditions, such as high/low interference, high/low spectral efficiency, handover ...) and be repeated for a sufficiently large time period (e.g., one day). In each loop, services have to be tested in a sequential order.

Time alignment. A correction is needed for the time offset between clocks in the terminal and network interfaces. For a precise alignment, three conditions must be fulfilled. First, service tests must be separated in time within each loop to

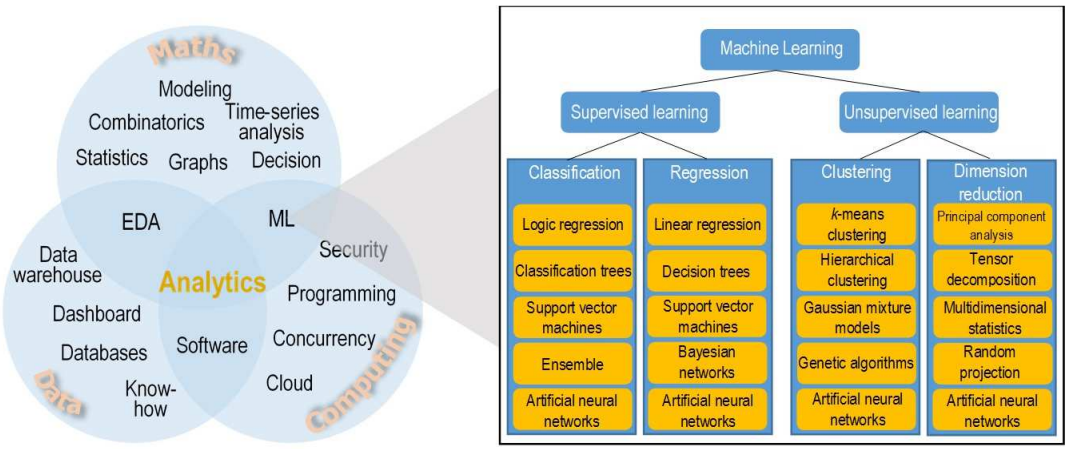

Fig. 2. Taxonomy of machine learning algorithms.

clearly isolate each service request in a single ESR. Second, every service has to be tested only once per loop. Note that information is aggregated in ESRs, and, thus, bad service performance in one session could be hidden with other good performing sessions of the same service. Third, to make alignment easier, a preliminary and fictitious service connection, labeled as *starting test* header, is established to identify the start of periodic service tests. For this purpose, a service with a large TCP data volume that stands out from the other services (e.g., a File Transfer Protocol, FTP, connection) is often selected.

Calibration. The purpose of validation is not only to assess the quality of S-KPI estimates, but also to tune internal parameters in the TMA application. Corrective alignment is always needed before deploying any TMA application in a commercial network.

An important consideration when validating a TMA application is how S-KPIs are defined. Too often it is taken for granted that both TA and network probe are making the same calculations. This is not always true, which might incorrectly lead to a non-conforming report. For instance, TA usually considers DNS resolution time and TCP handshake when computing web download time. However, a TMA application is only able to identify the web browsing service once the first HTTP message is sent, which generally takes place hundreds of milliseconds later. Likewise, the TMA application is not able to measure some of the processes executed in the TA that generally depend on many factors (e.g., device model, operative system ...) and cannot be monitored by TMA. To quantify these differences, traffic generated by the TA can be analyzed to check how TA computes S-KPIs. Then, possible differences with the network probe can be corrected by adding an offset term in the S-KPI estimation process. Another issue is the access to dynamic web pages, including dynamic objects (e.g., advertisements). One of those objects may be wrongly interpreted by the TMA application as the end of the web page. To circumvent this problem, the application can be adjusted with a sufficiently large time gap to consider two consecutive data bursts as different sessions. These corrective actions increase the accuracy of measurements in most cases.

V. Use Case of Big Data Analytics for TMA Application

In this section, a field trial of a big-data driven TMA application for QoE monitoring and optimization in LTE is presented. The offline construction of S-KPI estimation and forecasting models by ML is covered first. Then, several QoE network performance statistics from the live network are presented.

A. Offline stage - construction of S-KPI estimation model

Descriptive analytics can be used to unveil the relationship between S-KPI measurements from terminal agents and TCP metrics from probes. This process entails determining the most significant TCP metrics for each S-KPI (feature selection) and deriving the TCP metric-to-SKPI mapping (model construction).

Fig. 3 shows the result of automatic model construction for one of the most important S-KPIs, namely the initial buffering time for video streaming service. In the example, feature selection is based on a filtering method (chi-squared test) and model construction is based on polynomial multivariate regression [11]. To build the model, the dataset is split into two groups for training and testing purposes (80% for training and 20% for testing). The x-axis represents the total number of features (TCP metrics) used for estimating the S-KPI, whereas the y-axis represents the mean squared error (*MSE*) between S-KPI measurements and estimates for each combination of selected features. The four lines denote different models (polynomial degrees) to map TCP metrics into S-KPI values. The solid circle shows the optimal combination of number of selected features and polynomial degree, resulting in the minimum *MSE* with the testing dataset (*MSE* = 1.18 s^2 for 5 features and 5th order polynomial). Moreover, results show that the initial buffering time can also be derived accurately by 4^{th} order polynomial regression with 5 variables (specifically, the average RTT from terminal to probe, the average RTT from probe to Internet, the average downlink throughput of video session, the average downlink throughput without considering the initial TCP slow start stage and the average video bitrate). The resulting model is tested by using a terminal agent,

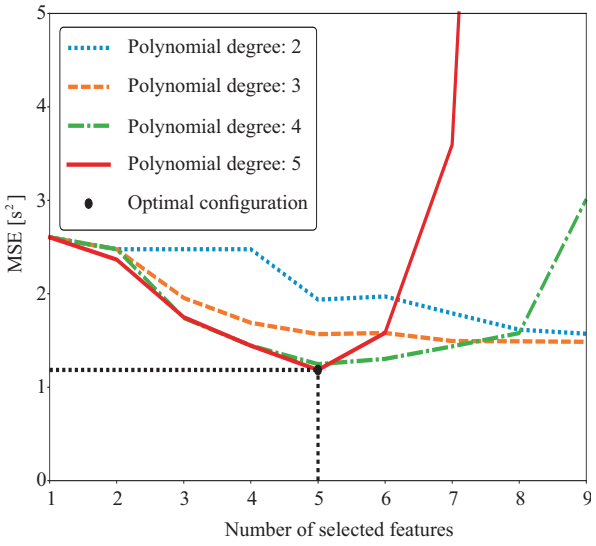

Fig. 3. Automatic model construction for the initial buffering time in video streaming service.

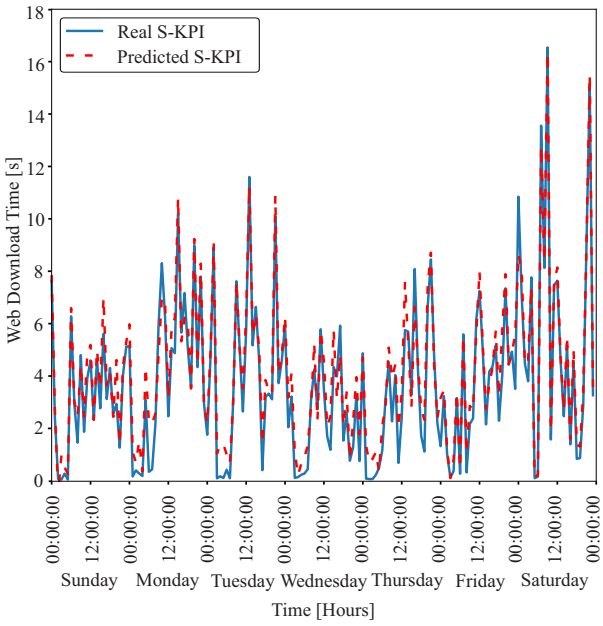

Fig. 4. Forecast of average web download time in a cell.

showing that the indicator ranges from 0 to 50 s depending on network conditions, with a 80-th percentile absolute error of just 0.16 s.

B. Offline stage - construction of S-KPI forecasting model

Predictive analytics can be used to foresee service performance degradation, so that corrective actions can be implemented. In our context, forecasting models may be used to predict QoE trends on different time scales on a service basis, so that proactive control decisions can be taken driven by QoE criteria. To this end, we can replicate legacy approaches based on time series analysis or apply ML algorithms to build more robust models.

Fig. 4 shows the results of a short-term forecasting model built with ML. The considered model predicts the average web download time of users in a cell for the next hour based on measurements from the previous 12 hours. To build the model, TMA was performed for several months in a large geographical area comprising many cells of a live LTE network. With a S-KPI model, the web download time is estimated on a session basis, and then aggregated on a cell and hourly basis. The resulting dataset is split into two groups for training and testing purposes (80% and 20%, respectively). Then, a forecasting model is derived by training a recurrent neural network (specifically, a long short-term memory network) [11] with the set of time series from cells in the training dataset. Then, the forecasting model is evaluated with a week-long time series from a randomly selected cell in the testing dataset. For this purpose, the model is executed 7*24 times, each taking the S-KPI value from the previous 12 hours as an input.

The figure shows real S-KPI measurements by a solid line and predicted values by a dashed line. It is observed that the predicted pattern is very close to the original one, which is confirmed by the small *MSE* value observed in the whole testing dataset (*MSE* = 0.95). Similar models can be constructed with finer time granularity (e.g., minutes) to reduce reaction times.

C. Online stage - QoE assessment

To check the tool, TMA is performed for a whole day in a live LTE network comprising 4828 cells. Based on operator demand, three different services are evaluated: video streaming, web browsing and mobile broadband, consisting of a speed test based on a FTP download.

The following S-KPIs are analyzed on a session basis: a) the video stalling ratio (i.e., total stall time divided by total time spent watching a video) and the video start delay for video streaming, b) the web page display success ratio and the average web page response delay for web browsing, and c) the average downlink throughput for mobile broadband. Each S-KPI is computed per cell by aggregating all sessions of a service in a cell.

Table I presents S-KPI values for cells in the network. Columns 4-6 show average, 95^{th} and 5^{th}-percentile values of S-KPIs computed on a cell basis to evaluate the performance of the worst cells. Column 3 shows the minimum performance threshold per S-KPI defined by the operator to label a cell as conforming or non-conforming. Results in Table I show that web browsing and mobile broadband services perform well, with at least 74.11% of cells fulfilling minimum S-KPI thresholds. However, video streaming shows worse results.

The previous statistics allow to identify bad service performance in a cell. Then, an automatic root cause analysis can be performed by evaluating all R-KPIs (i.e., radio access, transport and core KPIs) stored in ESRs. This is often done by checking performance differences against the rest of the network at the same time period. If this is the case, the specific issue is notified to the operator, who must take the required corrective actions. This kind of analysis also allows to identify bad performance from the service provider side, avoiding unnecessary network changes. In the field trial, a closer analysis of ESRs showed that bad user experience for video streaming was due to a large RTT value caused by the

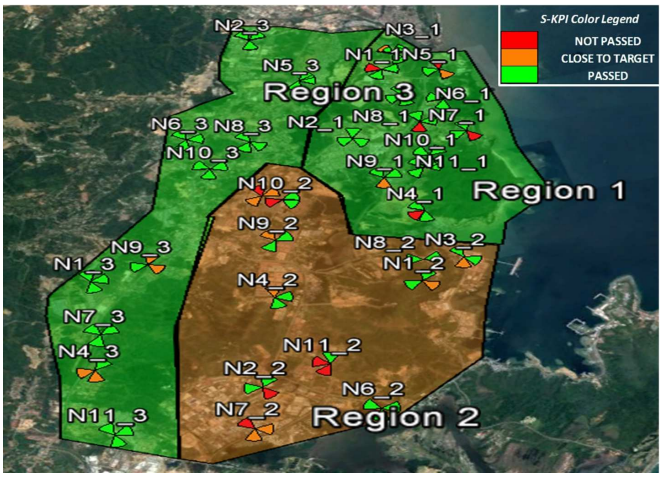

Fig. 5. S-KPI compliance map.

delay from intra-frequency handover, as well as a high packet loss ratio in the core network due to a service provider issue. From this analysis, the operator was able to isolate causes due only to its network and take corrective actions to improve the QoE of non-conforming cells and services.

The output of a TMA application can be processed by dashboards to graphically show the network status in terms of QoE. Dashboards complement S-KPI statistics by including QoE maps showing the degree of compliance by geographical areas. For this purpose, the degree of fulfillment of S-KPI thresholds for the three considered services is averaged per cell. Fig. 5 shows an example of S-KPI compliance map, where cells are represented by sectors. Every sector is represented with a different color depending on whether the target values are satisfied, close to be satisfied or not complying at all. In the example, the area is broken down into 3 regions according to the average S-KPI values. Similarly, QoE maps can be drawn by mapping S-KPI values into QoE figures with utility functions. Such maps make it easier for the operator to spot regions of bad end-user performance and trigger local actions.

VI. Open Issues

The following problems must be solved for a proper QoE monitoring in future 5G networks.

Metric calculation. To reduce computational load, some metrics are only calculated at very specific moments during a service session, so that the number of measurement samples per ESR may not be enough to ensure statistical reliability. For instance, RTT is typically estimated by the TMA application as the time difference between TCP SYNC and SYNC-ACK messages. These messages are only sent at the beginning of the connection, so only one RTT value is available per connection. Although validation can be successful (i.e., RTT values from both TA and network probe fit), this metric may not be relevant. For QoE management, RTT must be periodically measured to detect network performance fluctuations. This can only be done by checking other protocol layers.

New QoE models. User satisfaction is measured by mapping S-KPI values to QoE measures by means of utility functions. Such functions are derived from subjective tests with real users in lab environments, crowd-sourced user feedback or inferred from session times. Although a wide variety of QoE models are already available, the introduction of new services and the increase in customer expectations requires updating existing ones. An example is the introduction of Dynamic Adaptive Streaming over HTTP (DASH), for which image quality (and not stalling ratio) is now the most important S-KPI. Likewise, utility functions may consider contextual factors, such as device (e.g., smartphone vs computer), location (e.g., indoor vs outdoor) or subscription plan (e.g., normal vs premium), which information is available on a session basis. In this context, ML techniques can be used to keep track of the most relevant factors.

Data encryption. In the last years, most service providers have included encryption for privacy reasons, which makes traffic classification much more difficult. Even if encrypted traffic is correctly classified, protocol messages are not available, so that detection of triggering actions relevant for S-KPI estimates has to be done by other means. One approach consists of the use of statistical models that blindly relate TCP metrics and S-KPIs in controlled lab environments, which can then be used in live environments to estimate end-user experience from TCP metrics [12].

New transport mechanisms. Most TMA tools rely on TCP, whose flow control ensures that traffic dynamics is correlated to end-to-end performance. This is not the case for User Datagram Protocol (UDP). Likewise, multiplexing several streams in the same connection makes it very difficult to isolate the performance of each flow. This is the case of the new Quick UDP Internet Connection (QUIC) protocol.

Proxies. The introduction of Performance-Enhancing Proxies (PEP) negatively affects the accuracy of end-to-end performance estimations. In cellular networks, PEPs are introduced for some services (e.g., web browsing) to cope with large round-trip delay times by splitting TCP connections into multiple connections [13]. Unfortunately, PEPs modify some TCP metrics derived by network probes (e.g., RTT on the Internet side), which makes these measurements useless for estimating end-to-end performance.

Real time. Reactive QoE management requires: a) understanding how different consumer segments behave depending on perceived QoE, b) detecting what a customer does in near real time, c) identifying the consumer context, and d) providing a set of measures that is truly relevant.

Network slicing. Network slicing is one of the key virtualization technologies in 5G, allowing network operators to provide dedicated virtual networks with functionality specific to a service or customer over a common physical infrastructure. This requires storing network configuration parameters in ESRs when service degradation occurs.

VII. Conclusions

Passive network monitoring for QoE management purposes is gaining momentum in the industry. In this paper, a generic methodology to validate big-data driven TMA solutions for QoE monitoring in mobile networks has been proposed. Then,

TABLE I
S-KPIs IN SCENARIO.

Service	S-KPI	Threshold	Average value	95^{th} percentile	5^{th} percentile	No. fulfilling cells
Video streaming	Video stalling ratio	4 %	2.85 %	13.34 %		848 (17.15%)
	Start delay	2 s	3.51 s	9.96 s		2450 (17.15%)
Web browsing	Display success ratio	98 %	97.22 %		91.26 %	3578 (74.11%)
	Response delay	2 s	0.29 s	0.64 s		4747 (98.32%)
Mobile broadband	FTP DL Throughput	5 Mbps	13.83 Mbps		2.59 Mbps	4312 (89.31%)

a real use case has been presented to show the potential of big data analytics for automatic knowledge discovery related to QoE. Field trials have shown how these tools can detect QoE problems in a live LTE network effectively. The S-KPI forecasting models derived from analytics can then be integrated into Self-Organizing Network (SON) platforms to proactively change network parameters on a service basis [14][15]. Some open issues in the development of these applications have also been identified, which will be addressed by data scientists in the coming years.

ACKNOWLEDGMENT

This work was funded by the Spanish Ministry of Economy and Competitiveness (TEC2015-69982-R) and Ericsson Spain.

REFERENCES

[1] D. Soldani, S. Das, M. Hassan, J. Hassan, G. Mandyam, "Traffic management for mobile broadband networks," *IEEE Communications Magazine*, vol. 49, no. 10, pp. 98–100, 2011.
[2] N. Baldo, L. Giupponi, J. Mangues-Bafalluy, "Big Data Empowered Self Organized Networks," *20th European Wireless Conference*, pp. 1–8, 2014.
[3] A. Imran, A. Zoha, A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Network*, vol. 28, pp. 27–33, 2014.
[4] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, W. Xiang, "Big data-driven optimization for mobile networks toward 5G," *IEEE Network*, vol. 30, no. 1, pp. 44–51, 2016.
[5] ITU-T, "Big data standardization roadmap," International Telecommunication Union, Recommendation Y.3600, 2016.
[6] F. Ricciato, "Traffic monitoring and analysis for the optimization of a 3G network," *IEEE Wireless Communications*, vol. 13, no. 6, pp. 42–49, 2006.
[7] A. Baer, P. Casas, A. D'Alconzo, P. Fiadino, L. Golab, M. Mellia, E. Schikuta, "DBStream: A holistic approach to large-scale network traffic monitoring and analysis," *Computer Networks*, vol. 107, pp. 5–19, 2016.
[8] [Online]. Available: https://www.winpcap.org/ntar/draft/PCAP-DumpFileFormat.html, [accessed on 19.1.2018]
[9] D. Šipuš, "Big data analytics for communication service providers," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE 39th Int. Conv.*, 2016, pp. 513–517.
[10] T. T. Nguyen, G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, pp. 56–76, 2008.
[11] M. Mohammed, M. B. Khan, E. B. M. Bashier, *Machine learning: algorithms and applications*. CRC Press, 2016.
[12] P. Fiadino, P. Casas, A. D'Alconzo, M. Schiavone, A. Baer, "Grasping Popular Applications in Cellular Networks With Big Data Analytics Platforms," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 681–695, 2016.
[13] X. Xu, Y. Jiang, T. Flach, E. Katz-Bassett, D. Choffnes, R. Govindan, "Investigating Transparent Web Proxies in Cellular Networks," in *Proc. 16th Int. Conf. Passive and Active Measurement (PAM 2015)*, J. Mirkovic and Y. Liu, Ed., 2015, pp. 262–276.
[14] P. V. Klaine, M. A. Imran, O. Onireti, R. D. Souza, "A Survey of Machine Learning Techniques Applied to Self-Organizing Cellular Networks," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2392–2431, 2017.
[15] I Chih-Lin, Q. Sun, Z. Liu, S. Zhang, S. Han, "The Big-Data-Driven Intelligent Wireless Network: Architecture, Use Cases, Solutions, and Future Trends," *IEEE Vehicular Technology Magazine*, vol. 12, no. 4, pp. 20–29, 2017.

BIOGRAPHIES

ANTONIO J. GARCÍA received his M.S. degree in Telecommunication Engineering from the University of Málaga, Spain, in 2014. Since 2014, he joined the Communications Engineering Department, University of Málaga, where he is currently working toward the Ph.D. degree in Telecommunications Engineering in a collaborative project with Ericsson. His research interests are focused on planning and optimization of mobile radio access networks based on users' experience.

MATÍAS TORIL received his M.S. and Ph.D. degrees in Telecommunication Engineering from the University of Málaga, Spain, in 1995 and 2007, respectively. Since 1997, he is Lecturer in the Communications Engineering Department, University of Málaga, where he is currently Full Professor. He has co-authored more than 100 publications in leading conferences and journals and 3 patents owned by Nokia Corporation. His current research interests include self-organizing networks, radio resource management and data analytics.

PABLO OLIVER received his M.S. degree in Telecommunication Engineering from the University of Málaga, Spain, in 2013. Since 2013, he joined the Communications Engineering Department, University of Málaga, where he is currently working toward the Ph.D. degree in Telecommunications Engineering in a collaborative project with Ericsson. His research interests are focused on planning and optimization of mobile radio access networks based on users' experience.

SALVADOR LUNA received his M.S. and Ph.D. degrees in Telecommunication Engineering from the University of Málaga, Spain, in 2000 and 2010, respectively. Since 2000, he has been with the Communications Engineering Department, University of Málaga, where he is currently Associate Professor. His research interests include self-optimization of mobile radio access networks and radio resource management.

RAFAEL GARCÍA received his M.S. in Telecommunication Engineering from the University of Málaga, Spain, in 2004. From 2004 to 2010 he was with Optimi in Málaga, working in the R&D department. In 2011, he joined Ericsson as development leader, changing to researcher on 2013. He has lead several research projects related to mobile networks in the areas of capacity planning, network performance monitoring and troubleshooting and user experience monitoring. His current research interests include user experience on mobile networks and high scale data processing.