Data Analytics Based Origin-Destination Core Traffic Modelling

F. Morales*, M. Ruiz, and L. Velasco
Optical Communications Group (GCO), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain.
**e-mail:fmorales@ac.upc.edu*

ABSTRACT
Traffic monitoring is an essential task for network operators since it allows evaluating network performance. Monitoring data from origin-destination (OD) traffic in core virtual network topologies can be collected from packet nodes and stored in a repository for further analysis, e.g., to detect anomalies or to create predicted traffic matrices for the near future. In this paper we propose a set of modules to support data analytics-based algorithms along with a machine learning procedure based on artificial neural networks (ANN) that provides robust and adaptive traffic models.
Keywords: traffic monitoring, machine learning, traffic prediction, multilayer networks.

1. INTRODUCTION

Data analytics is the interesting and challenging task of inspecting, transforming and modelling heterogeneous data coming from different sources, aiming at discovering useful information and supporting decision making. It comprises a wide range of approaches and algorithms that are successfully applied in many business, science and social fields. In networking, data analytics can be applied by monitoring the data plane, e.g., received power, impairments, and errors in the optical layer or service traffic in the MPLS layer.

Traffic monitoring is becoming an essential task for network operators since it allows evaluating the network performance. Monitoring data can be collected in a repository for further analysis, e.g., to localize failures and identify their cause or to create predicted traffic matrices for the near future. According to the ITU-T [1], performance events are counted second by second over every 15-minute period. At the end of a period, they are stored in historical registers usually in the network manager.

Applying data analytics to monitoring data can increase the knowledge about the network traffic, moving towards a cognitive network operation. The output of data analytics can be used to improve network functioning either by planning the network off-line or even by automating the decision making process while in-operation, by using machine learning and data stream mining algorithms feeding from monitoring traffic that can trigger further actions in the network automatically. This way of doing network management is collectively known as the observe-analyse-act (OAA) loop since it links together monitoring, data analytics and operation.

In multilayer core networks, virtual network topologies (VNT) are created by connecting MPLS routers through virtual links (vlinks) supported by lightpaths in the optical layer; traffic flows are conveyed through MPLS paths along vlinks on the VNT. The introduction of new services requiring large and dynamic bitrate connectivity is leading to high traffic dynamicity, with directional and volumetric changes within short times. To avoid the large overprovisioning given by statically managing the VNT to cope with this dynamic traffic scenario, network operators are looking for more efficient architectures to dynamically adapt the VNT in a cost-efficient manner. In that regard, including data analytics to enable the OAA loop in the network is presented as a promising option.

In this paper, we propose a series of modules to bring data analytics to the core VNT and allow OD traffic modelling. Specifically, a machine learning procedure is proposed to automatically fit artificial neural network (ANN) models to predict OD traffic.

2. GENERIC MODULES TO BRING DATA ANALYTICS TO THE NETWORK

Among the possibilities offered by data analytics, traffic modelling is an interesting one because of its capability to characterize traffic profiles and produce accurate predictions. Based on OD traffic predictions, the VNT can be reconfigured to anticipate near future traffic or cope with traffic anomalies. To obtain predictions, we propose a series of modules to bring data analytics to the network while in operation, which allow the operator to enable the observe-analyse-act loop in the network. Specifically, these modules allow monitoring traffic, modelling the monitored data to obtain accurate traffic predictions which are finally used to guide a decision making process responsible for VNT reconfiguration.

We assume that traffic monitoring data is collected at the edge IP routers at regular intervals, e.g., every 15 minutes. Every edge router collects a set of samples for the traffic to every other destination router, which is stored in a *collected data repository* (Fig. 1). Note that since we focus on OD traffic monitoring, $|N|\cdot(|N|-1)$ traffic samples need to be stored at every monitoring interval, where $|N|$ is the number of routers.

Following a predefined time period, e.g., every hour, a time series from the collected data repository is retrieved for each OD pair and pre-processed applying data stream mining *sketches* to conveniently summarize collected data thus producing modelled data representing the OD pair that is stored in a *modelled data* *repository*. Modelled data includes, among others, for every OD the minimum, maximum, average, and last collected bitrate measurement within the elapsed hour.

The set of modelled variables for the current period t is stored in a repository together with variables belonging to previous periods. A *prediction module* based on machine learning techniques generates

OD traffic predictions for the next period; these predictions are used by a *decision maker* module to enable data analytics-based decision making. Based on the predicted traffic, the decision maker might trigger a reconfiguration action of the VNT, in which case the current and predicted traffic matrices are provided to the *VNT optimizer* to adapt the VNT.

Let us now introduce two VNT adaptability use cases triggered by data analytics –based decision making. The first use case consists of periodically reconfiguring the VNT to near future traffic. Specifically, predicted OD traffic matrices for the next hour are used to adapt the capacity of the VNT. To find such reconfiguration, the VENTURE problem was proposed and solved [1]. The second use case is related with traffic anomaly detection. By comparing current traffic samples against traffic predictions based on normal traffic conditions, it is possible to detect an anomalous capacity increment in the VNT and trigger the needed reconfiguration to adapt the capacity and minimize traffic losses. To that end, the ODEON problem was proposed and solved [3].

Last but not least, note that all the proposed modules can be implemented in practice by extending ABNO's Operations, Administration, and Maintenance (OAM) handler, as experimentally demonstrated by authors in [4].

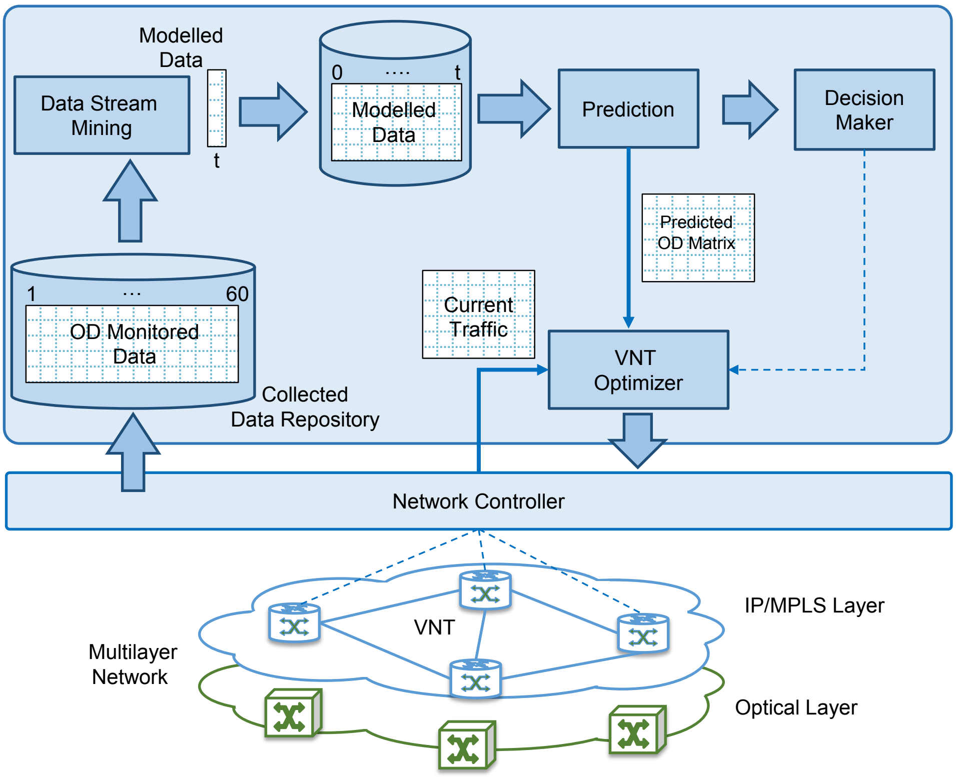

Figure 1. Applying the observe-analyse-act loop for VNT reconfiguration.

3. MACHINE LEARNING PROCEDURE FOR TRAFFIC PREDICTION

The previous prediction module consists of ANN-based models [5], selected because of its inherent capability of adapting to traffic changes in a non-supervised manner; we consider different ANNs to separately predict the traffic of each OD pair. To obtain the maximum traffic prediction at time t, each ANN is evaluated with p previous maximum bitrate measurements of the corresponding OD pair from the modelled data repository. Note that considering the maximum instead of average bitrate allows adapting the VNT to the maximum expected traffic hence, ensuring a better grade of service.

The size of an ANN depends on the number of inputs, hidden layers and neurons. Therefore, we consider ANN models with p inputs, s neurons in a single hidden layer and a single output. Consequently, $s\cdot(p+1)$ coefficients need to be found to specify every ANN. Aiming at keeping the number of coefficients small, we devised the algorithm in Fig. 2 that is triggered every time an ANN model needs to be refitted. It consists in three phases: *i*) input data pre-processing, *ii*) selection of the significant inputs and *iii*) dimensioning of the hidden layer.

In the first phase, a time series X with the maximum bitrate measurements for the selected OD pair is retrieved from the modelled data repository. The *auto-correlation function* (ACF) is applied to X and a list of *lags* is returned, where the i-th lag contains the average correlation between every value in the time series and its i-th previous value. Based on the lags analysis, a method is triggered to detect whether a periodic repetitive (seasonal) pattern is observable in X [6]. The resulting period *per* defines the number of inputs of the ANN; in case of non-seasonal data without observable periodical behaviour, we assume *per*=24 (i.e. one day) for convenience. Once *per* is obtained, X is transformed into a dataset D used for ANN fitting. Every row in D corresponds to a time t within the time series and every column corresponds to a lag within *per*.

The second phase is an iterative procedure that finds the ANN with the best trade-off between accuracy and number of inputs. This trade-off is captured numerically by the *Akaike Information Criterion* (AIC) [5]. Starting with p=*per*, the ANN routine fits an ANN from dataset D and returns the corresponding AIC value. While the obtained AIC value improves the minimum obtained so far, the best ANN is stored and p is decremented by effectively removing one input. Aiming at reducing the complexity of selecting the input to be removed, we select the lag with the lowest ACF. When the minimum AIC is reached, the third phase is executed to increase even more the accuracy of the model by adding hidden neurons until the AIC does not improve. The best ANN is eventually returned.

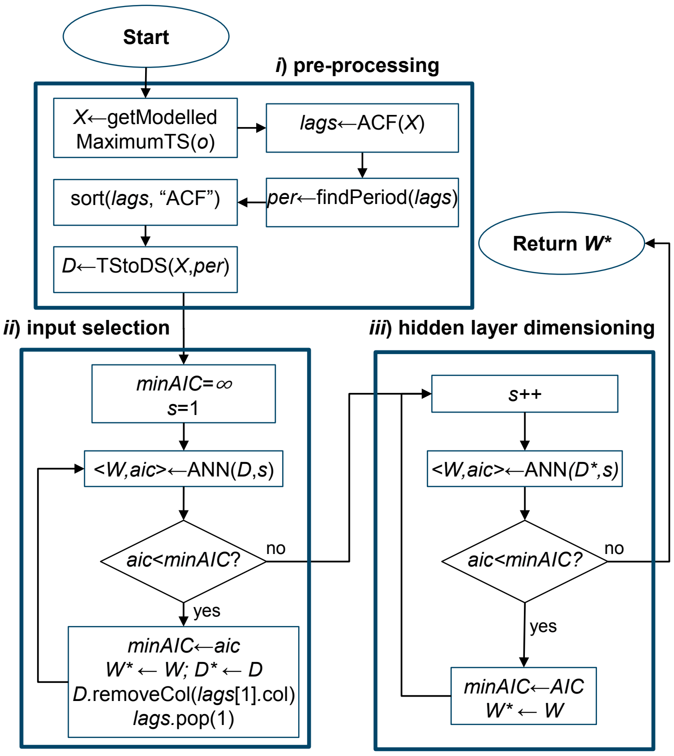

Figure 2. Self-learning ANN fitting algorithm.

4. SIMULATION RESULTS

For evaluation purposes, we implemented an event-driven simulator in OMNeT++ containing the modules described in Fig. 1. To generate monitored traffic measurements for different OD traffic profiles we implemented generators that inject traffic following three pre-defined traffic profiles named as *Users, Business* and Datacenter-to-Datacenter (*DC2DC*) (see average daily evolution in Fig. 3a). This generation was based in a traffic generation framework for telecom cloud –based simulation [7]. The Users traffic profile represents thousands of aggregated multimedia-like connections such as video streaming, with high variability and a traffic peak at noon hours. The Business traffic profile is similar to Users', but with the traffic peak shifted toward midday, at central business hours. Finally, the DC2DC traffic represents tenths of inter-datacenter connections performing scheduled tasks such as database synchronization; for this reason, it presents a lower variability and predominates at night.

The ANN models are trained applying the fitting algorithm in Fig. 2 on a training dataset with modelled data belonging to the last weeks. Results in Fig. 3b illustrate the average size and goodness-of-fit of the trained ANN models. Recall that during the input selection phase, the number of inputs p is decreased aiming at minimizing the AIC value. We observe in the figure that the minimum AIC is on average reached at p=4, being mainly selected those inputs corresponding to lags $t-1$ to $t-4$. Results from the hidden layer dimensioning phase are shown in the table embedded in Fig. 3b, for a number of hidden neurons ranging from 1 to 3. Note that the minimum AIC is obtained for $s = 2$, which results in an ANN model with 10 coefficients that accurately predicts the output variable with a good trade-off between average and maximum relative errors (2.64% and 9.55%, respectively).

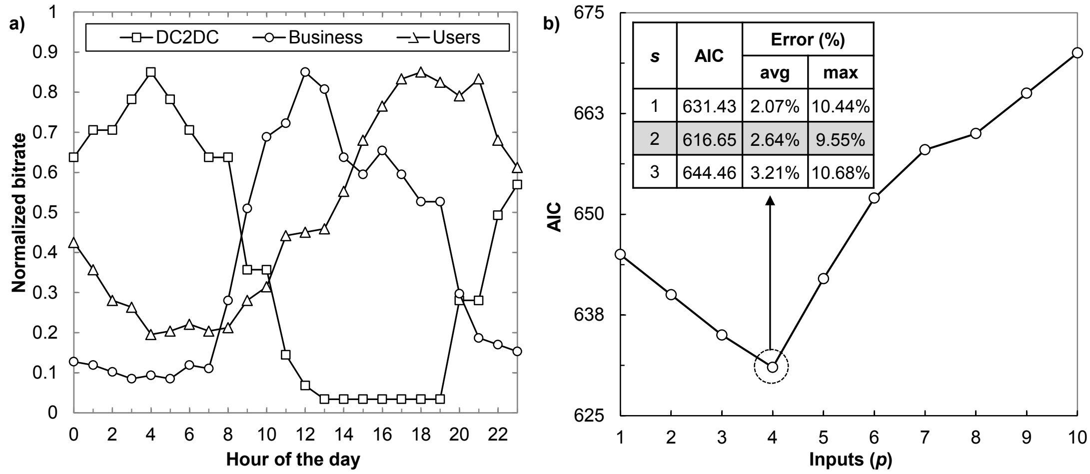

Figure 3. Average daily traffic profiles used in the simulation (a) and ANN goodness of fit (b).

5. CONCLUSIONS

OD traffic in the core VNT can be monitored and analysed to extract knowledge about the traffic that facilitates operating the network. To apply data analytics to monitored traffic in-operation, we propose a series of modules that allow storing monitoring data in a big data repository. By applying machine learning techniques, raw data is analysed and meaningful features about the traffic extracted. This richer, modelled data is processed by a prediction module which finally produces OD traffic predictions. Predicted OD traffic can be used to guide a decision maker in the process of VNT adaptability, adapting the VNT to future traffic matrices or traffic anomalies.

In order to obtain accurate traffic prediction, we propose ANN models inside the prediction module. In order to automatically fit these ANN models, we devise a self-learning ANN fitting algorithm that targets at providing accurate predictive models at a low parameter cost. Last, the quality of the proposed algorithm was demonstrated through exhaustive simulation.

ACKNOWLEDGEMENTS

This work was partially supported by the EC through the METRO-HAUL project (G.A. nº 761727), from the Spanish MINECO SYNERGY project (TEC2014-59995-R), and from the Catalan Institution for Research and Advanced Studies (ICREA).

REFERENCES

[1] ITU-T Rec. M.2120, 2002.
[2] F. Morales, M. Ruiz, and L. Velasco, "Virtual network topology reconfiguration based on big data analytics for traffic prediction," in *Proc. IEEE/OSA OFC*, 2016.
[3] L. Velasco, A. P. Vela, F. Morales, and M. Ruiz, "Designing, operating and re-optimizing elastic optical networks [Invited Tutorial]," *IEEE/OSA Journal of Lightwave Technology*, 2016.
[4] Ll. Gifre, F. Morales, L Velasco, and M. Ruiz, "Big data analytics for the virtual network topology reconfiguration use case," in *Proc. IEEE ICTON,* 2016.
[5] F. Emmert-Streib and M. Dehmer, "*Information Theory and Statistical Learning*," Springer Science & Business Media, 2008.
[6] J.D. Hamilton, "*Time Series Analysis*," Princeton University Press, 1994.
[7] A. P. Vela, A. Via, F. Morales, M. Ruiz, and L. Velasco, "Traffic generation for telecom cloud -based simulation," in *Proc. IEEE ICTON*, 2016.