Final publication can be found in: http://ieeexplore.ieee.org/document/7600929/

Knowledge-based 5G Radio Access Network Planning and Optimization

J. Pérez-Romero, O. Sallent, R. Ferrús, R. Agustí
Dept. of Signal Theory and Communications
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
e-mail:[jorperez, sallent, ferrus, ramon]@tsc.upc.edu

***Abstract*— Self-Organizing Network (SON) functionalities in forthcoming 5G systems stand as a promising field for the fertilization of Artificial Intelligence (AI) mechanisms. In this respect, this paper analyzes how SON functionalities for radio access network planning and optimization in future 5G systems can be built upon AI concepts. A framework is presented that processes input data from multiple sources and extracts, through learning-based classification, prediction and clustering models, relevant knowledge to drive the decisions made by 5G SON functionalities. Different candidate AI-based tools for knowledge discovery are identified together with the associated knowledge models that can be extracted. On this basis, the applicability of these models to a comprehensive range of 5G SON functions across the categories of self-planning, self-optimization and self-healing is analyzed. Finally, the paper identifies the research directions deriving from the proposed framework.**

Keywords—5G; Knowledge discovery; Artificial Intelligence; Self-Organizing Networks

I. INTRODUCTION

As a next step in the evolution of mobile communication systems, research carried out by industry and academia is nowadays focused on the development of the new generation of mobile and wireless systems, known as the 5th Generation (5G). Requirements for future 5G system have already been identified and discussed in different fora [1][2]. Examples of challenging requirements for 5G radio access include 1000 times higher mobile data volume per area, 10 to 100 times higher number of connected devices, 10 to 100 times higher typical user data rate, 10 times longer battery life for low power devices and 5 times reduced end-to-end latency.

In addition to extending the performance envelope of radio access capabilities, 5G is also envisioned to include unprecedented network flexibility and highly efficient/adaptive network resource usage thanks to the advances in technologies such as Software Defined Networking (SDN) and Network Function Virtualization (NFV). This will allow supporting current and new business models involving different types of customers and partnerships [2]. Specifically, 5G is expected to allow Mobile Network Operators (MNOs) to better support customers from a number of vertical industries (e.g., e-health, automotive, energy). Challenging performance indicators in terms of network flexibility and network resource efficiency include a 1000-fold reduction in service deployment time (reaching a complete deployment time in the order of hours) and 5 times reduced network management OPEX (Operational Expenditures) [3].

Furthermore, the management and operation of 5G networks will be fueled by the advent of big data analytics [4], which are already essential technologies in different sectors (e.g. entertainment, financial services, automotive, logistics). Therefore, with the huge amount of data generated by mobile networks, it can be envisaged that big data technologies will play a central role in 5G to extract the highest possible value of the available data, particularly enabling the extraction of valuable knowledge about the experience of individual users, and exploiting it for enhancing the efficiency in mobile service delivery.

The 5G challenge for MNOs is very much about how to balance investments, user experience and profitability. In this respect, and sustained on the pillars of network flexibility and data volume handling as key technology enablers, two main areas of improvement are identified: (1) The methodologies used for Radio Access Network (RAN) planning, which can result in significant CAPEX (Capital Expenditures) savings and (2) The methodologies used for RAN optimization, which can result in significant OPEX savings. This paper claims that there is the need and the opportunity to revisit the actual methodologies for RAN planning and optimization in light of the challenging network efficiency requirements for 5G, forging the vision of an ultra-efficient RAN fully exploiting cognitive capabilities that embrace knowledge and intelligence, increasing the degree of automation, making the network more self-autonomous and enabling a personalized user experience across the MNO's RAN.

In this context, this paper supports the idea that Artificial Intelligence (AI) mechanisms, which intend to develop intelligent systems able to perceive and analyze the environment and take the appropriate actions, will fertilize in the 5G ecosystem. Indeed, while many seeds can be found in the literature both from an academic/theoretical perspective (e.g., connected to the so-called Cognitive Networks) and from a practical perspective in current 3G/4G networks (e.g., connected to the so-called Self-Organizing Networks (SON)), more ambitious objectives can be targeted and the 5G era is the proper time for AI-based RAN management to happen.

In particular, the paper discusses in Section II the evolution envisaged in SON to deal with 5G scenarios, and justifies the use of AI-based knowledge discovery mechanisms relying on machine learning as a means to extract models that reflect the user and network behaviors. Specifically, the paper identifies candidate AI-based tools for knowledge discovery and

This work has been supported by the EU funded H2020 5G-PPP project SESAME under the grant agreement 671596 and by the Spanish Research Council and FEDER funds under RAMSES grant (ref. TEC2013-41698-R).

enumerates in Section III different knowledge models that can be extracted through these tools, together with their applicability to specific SON functions. The paper concludes in Section IV by pointing out some research directions.

II. AI-Based 5G SON

The vision of the future 5G corresponds to a highly heterogeneous network at different levels, including multiple Radio Access Technologies (RATs), multiple cell layers, multiple spectrum bands, multiple types of devices and services, etc. Consequently, the overall RAN planning and optimization processes that constitute a key point for the success of the 5G concept will exhibit tremendous complexity. In this direction, legacy systems such as 2G/3G/4G already started the path towards a higher degree of automation in the planning and optimization processes through the introduction of SON functionalities.

SON refers to a set of features and capabilities designed to reduce or remove the need for manual activities in the lifecycle of the network. With the introduction of SON, classical manual planning, deployment, optimization and maintenance activities of the network can be replaced and/or supported by more autonomous and automated processes [5], operating costs can be reduced and human errors minimized. In the context of this paper, SON functions that can benefit from the application of AI concepts are organized in the following categories:

- Self-planning: Automatization of the process of deciding the need to roll out new network resources (e.g., new cells, new component carriers in an existing cell) in specific areas and determining the adequate configurations and settings for these resources. Examples of functions include the planning of a new cell and the spectrum planning.
- Self-optimization: Once the network is in operational state, the self-optimization includes the set of processes intended to improve or maintain the network performance in terms of coverage, capacity and service quality by tuning the different network settings. Examples functions include Mobility Load Balancing, Mobility Robustness Optimisation, Automated Neighbour Relation, Coverage and Capacity Optimization, optimization of admission control and packet scheduling, intercell interference coordination and energy saving.
- Self-healing: Automation of the processes related to fault management and fault correction, usually associated to hardware and/or software problems, in order to keep the network operational while awaiting a more permanent solution to fix it and/or prevent disruptive problems from arising. Examples of self-healing functions include Cell Outage Detection and Cell Outage Compensation.

Given that RAN's management processes are rather technology-agnostic, it is expected that the bulk of 5G SON functionalities will be similar to those encountered in legacy systems. Nevertheless, additional functionalities may arise as a result of advanced mechanisms. For example, in light of the more advanced spectrum management models, which in general comprise licensed, light licensed and unlicensed components, the provisioning of the spectrum resources to be exploited at a given time and location in 5G should be considered from a wider perspective and be conceived as a dynamic self-planning mechanism.

While the current SON vision exhibits an intrinsic reactive design approach and a lack of end-to-end knowledge of the network [8], the inclusion of AI-based tools enables to shift the evolution of the SON paradigm in 5G towards a more proactive approach able to exploit the huge amount of data available by the MNO and to incorporate additional dimensions coming from the characterization of end-user experience and end-user behavior.

Fig. 1 illustrates the three main stages envisaged for AI-based 5G SON, which are described in the following.

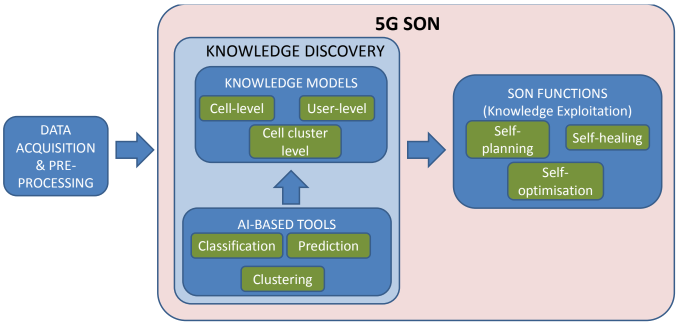

Fig. 1. AI-based 5G SON

- The acquisition and pre-processing of input data exploiting the wide variety of available data sources. In the 5G and big data era, it is feasible going far beyond the network data (e.g. performance measurements, network counters, etc.) that has been traditionally used in legacy systems, and consider also other dimensions such as the user-specific data, the content associated to applications or even external data from outside the MNO domain (e.g. planned events, weather forecasts, etc.).
- The Knowledge Discovery stage supported by AI-based tools that will smartly process the gathered input data to come up with exploitable knowledge models that represent the network/user behavior in a way that can be directly used to make smart network planning and optimization decisions. AI-based tools rely on machine learning to carry out the mining of the input data and extract relevant knowledge models at different levels: cell level (contains the characterization of the conditions on a per cell basis), cell cluster level (characterization of groups of cells built according to their similarities) and user level (contains the characterization of the conditions experienced by individual users). The general goal of machine learning is to build computer systems that can adapt and learn from their experience [7]. Specific machine learning functions that are relevant here are [6]:
    - Classification: It is the process of finding a model or function that describes and distinguishes data classes or concepts. The obtained model (i.e. the classifier) is then used to determine the class to which an object belongs. The object is the entity to be classified and it is usually represented by a tuple that includes a set of attribute values (e.g. an object could be a cell and the attributes could be different performance measurements associated to that cell). Classification process assumes that the possible classes are predefined in advance. Then, the classifier model is usually obtained from a supervised learning algorithm that analyses a set of training tuples associated with known classes. Table I briefly describes different classification tools.

o Prediction: It intends to find models to anticipate future values of a certain parameter. Prediction models usually exploit the trend analysis of input data in terms of four major components, namely long-term movements that indicate the general direction in which a time-series graph is moving over a long interval of time, cyclic movements that refer to oscillations about a trend line, seasonal movements that are systematic or calendar related, and irregular or random movements that characterize the sporadic motion of time series due to random or chance events. Table II summarizes some relevant prediction tools.

o Clustering: It consists in grouping a set of objects in a way that objects within the same cluster are similar to one another and dissimilar to the objects in other clusters. Clustering mechanisms do not rely on a set of predefined classes but these are obtained through an unsupervised learning process. Clustering tools are described in Table III.

- The Knowledge Exploitation stage will apply the obtained knowledge models to drive the decision-making of the actions associated to the SON functionalities.

TABLE I. AI-BASED TOOLS FOR CLASSIFICATION

Classification tools	
Decision Tree Induction	Classification uses a flow-chart structure where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes. The flow-chart structure is built during the supervised learning stage through a top-down recursive divide-and-conquer manner, starting from the training set which is recursively partitioned into smaller subsets.
Bayesian classification	Classification process evaluates the probability that a given tuple belongs to a class based on its attributes and selects the class with the highest probability. The probability computation is done using Bayes' theorem whose terms are obtained from the analysis of the training set.
Rule-based classification	Classification uses a set of if/then rules obtained from decision trees or directly from the training data.
Fuzzy Logic	Classification is similar to rule-based classification but allowing "fuzzy" thresholds to be defined for each class. Then, the classification is given in terms of a value between 0 and 1 that represents the degree of membership that a certain attribute value has in a given category.
Neural networks	Classification uses a feed-forward neural network that consists of an input layer, one or more hidden layers and an output layer. Each layer is made up of processing units called neurons. The input attributes of the object to classify are fed into the neurons making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer. The process is repeated until reaching the output layer, whose neurons provide the selected class. The weights of the connections between neurons are learnt during the supervised learning phase using a back propagation algorithm.
Support Vector Machines (SVM)	Classification uses the optimal boundary that separates the input tuples of the training set in their corresponding classes. This boundary is found during the training stage through a nonlinear mapping to transform the original training data into a higher dimension so that the optimal boundary becomes a hyperplane. SVM classifier is originally intended to do a binary classification, but multi-class SVM classifiers can be built by hierarchically combining multiple binary SVM classifiers.
K-nearest neighbour (k-NN)	Classification is done by comparing the test tuple to classify with the set of training tuples and finding the k training tuples with shortest distance to the test tuple. The assigned class is the most common class among these k training tuples.

In summary, given the ultra-high level of efficiency associated to the design of future 5G systems, the target is to gain in-depth and detailed knowledge about the whole ecosystem, understanding hidden patterns, data structures and relationships, which in turn will enable ultra-efficient management and optimization. In this respect, the higher level of knowledge about the network and its users constitutes a key differential factor between SON in 5G and in legacy systems.

TABLE II. AI-BASED TOOLS FOR PREDICTION

Prediction tools	
Seasonal Auto-Regressive Integrated Moving Average (ARIMA)	In ARIMA the time series is modelled as a linear combination of its past values and the past values of an error series. Seasonal ARIMA extends this with the inclusion of one or multiple seasonal factors to capture e.g. weekly variations, monthly variations, etc. The setting of the model parameters can be done automatically based on past input data applying algorithms such as the minimization of Akaike Information Criterion (AIC).
Holt-Winters Exponential Smoothing with seasonal patterns	The time series is modelled by a local mean, a local trend and a local seasonal factor that are updated by exponential smoothing. Either one or multiple seasonalities can be considered. The setting of the model parameters based on the past observations can be done automatically through the minimization of the AIC.
Support Vector Regression (SVR)	The prediction function is obtained by solving a convex optimization problem based on the SVM concept [9]. It does not need to assume linear relationship of the predicted parameter with respect to previous values, so it can work in linear, non-linear, stationary and non-stationary systems.

TABLE III. AI-BASED TOOLS FOR CLUSTERING

Clustering tools	
Partitioning methods	Clusters are formed to optimize an objective partitioning criterion such as a dissimilarity function based on distance. Existing partitioning algorithms include k-means, Partitioning Around Medoids (PAM) or Clustering LARge Applications (CLARANS) [6]. The number of clusters has to be known in advance.
Hierarchical methods	Objects are grouped into a tree of clusters that is iteratively modified in an agglomerative (i.e. starting with clusters of one object and then merging them according to some similarity criterion) or a divisive way (i.e. starting with a single cluster that contains all the objects and progressively subdividing it into smaller clusters). Hierarchical algorithms do not require the number of clusters to be known in advance but it can be dynamically found by the process.
Density-based methods	Clusters are considered as dense regions of objects in the data space that are separated by regions of low density.
Grid-based methods	Clustering uses a multiresolution grid that quantizes the object space into a finite number of cells.
Self-Organizing Map (SOM)	Clustering relies on a neural network model. Each neuron has an associated weight with as many components as the number of attributes of the input objects. An unsupervised learning process updates the values of the weights applying the Kohonen's algorithm over each input object [10]. At the end, the weight of each neuron captures the attribute values of a cluster of input objects.
Geospatial clustering	Clustered objects are associated to geographical coordinates and clustering is done by grouping points based on their connectivity, density and/or reachability in geographical space. Specific tools include Kernel Density Estimation (KDE) [11], which is based on estimating the density of points, or the Geo-ProZones (GPZ) algorithm, which applies an adaptive partitioning method to detect regions with a high concentration of points [12].

III. KNOWLEDGE MODELS TO SUPPORT 5G SON

Different knowledge models derived from the AI-based tools in the knowledge discovery stage are discussed in the following for the three abovementioned cell, user and cell cluster levels. In turn, Table IV analyzes the potential applicability of these knowledge models for different SON functions.

A. Cell-level models

This level includes the knowledge that characterizes the existing conditions in a cell. This encompasses the following models:

1) Traffic characterization. From a time-domain standpoint, this defines how the traffic of a cell varies as a function of time. The traffic can be measured in different ways, such as the load factor, the total number of users, the total data rate, etc., and it can be aggregated or split among Quality of Service (QoS) classes. From a space-domain perspective, the traffic characterization can be done in different terms, such as the geographical distribution of the users, traffic load, services/applications or QoS class. In general, given the proliferation of multiple small cells that can be located indoors and deployed in tall buildings, a 3D characterization can be required. Then, AI-based techniques will be applied over past observations of traffic in order to provide:

- Classification of the time domain traffic pattern: Time correlations in the traffic evolution of a given cell should be detected to identify existing seasonalities at different levels (e.g. intra-day variations, variations during the week between working days and weekend, variations in the traffic between winter and summer, etc.) and classify the cell accordingly. The classification tools of Table I can be used to extract these models.

- Learning the traffic behavior in the time domain: This refers to the identification of a model that captures the cell traffic at different periods of time (e.g. hours, days of the week, etc.) and allows identifying time periods exhibiting similar traffic levels. Candidate tools to extract this knowledge include some of the clustering techniques listed in Table III, such as partitioning, hierarchical methods or SOM.

- Prediction of the future traffic: A prediction model can be extracted to anticipate future values of the traffic evolution in a cell. This can feed various decision-making processes regarding planning (e.g., in order to anticipate the need to deploy additional network nodes) and optimization (e.g., in order to tune handover parameters in neighboring cells to absorb traffic if the cell is anticipated to be overloaded), mainly depending on the time scale at which the prediction is conducted. Candidate tools are the techniques listed in Table II.

- Clustering spatial traffic (hot-spots): It targets the identification of concentrations of users in limited geographical areas. Candidate tools include the geo-spatial clustering techniques of Table III.

- Learning mobility patterns: This intends to identify if the traffic follows some specific mobility patterns inside the cell that can be characterized in terms of prototype or representative trajectories followed by many of the users (e.g. trajectories directed towards specific points such as a metro station, etc.). Candidate tools to extract this knowledge include some of the clustering mechanisms of Table III like the partitioning methods or the SOM.

2) Performance characterization. The assessment of a cell's performance involves multiple measurements and Key Performance Indicators (KPIs) that can be organized under different categories depending on the specific performance criterion. Typically, it can be distinguished among accessibility KPIs (e.g., success rate in the set-up of new calls), retainability (i.e., how often an end-user abnormally loses a call or a session), mobility (e.g., number of handovers to different target cells, handover types or handover causes), QoS related KPIs and associated measurements (e.g., cell bit rate, throughput per user, latency and packet loss rate), resource utilization related measurements (e.g., transmit power per carrier, percentage of time that all the resource blocks devoted to traffic have been used), and RF measurements (e.g., distributions of the Channel Quality Indicators (CQIs), of the received power levels, etc.). Traditionally, measurements such as accessibility rate or dropping rate are aggregated on a cell basis and averaged along relatively long-term periods (e.g. days or weeks). These averaged values are used to trigger various optimization processes. Instead, the AI-based approach presented in this paper intends to attain much deeper exploitation of these KPIs and extract additional knowledge based on the time and spatial domain analysis of the abovementioned indicators. Some possibilities are listed in the following:

- Learning the time domain pattern of a performance indicator: This characterizes the time evolution of a given performance indicator, with the objective of identifying existing hidden patterns that would remain undetected if only aggregated measurements along several days/weeks were considered. For example, it can be automatically detected if the dropping rate in a cell exceeds certain thresholds during some specific hours, and if this situation exhibits some regularity, meaning that actions should be triggered to optimize the performance for those specific hours. Clustering tools listed in Table III like partitioning, hierarchical methods or SOM are candidates to extract this knowledge.

- Prediction of a performance indicator: This involves the definition of a prediction model to anticipate future values of a performance indicator at different time scales. Prediction should be based on past observations of the indicator, but it can also consider past observations of other related indicators as additional features. For example, the prediction model of the throughput per user can take as inputs the observations of the signal to noise and interference conditions seen by the users and the resource usage. Prediction models of Table II can be used to extract this knowledge.

- Learning space-domain black-spots: The characterization of the performance indicators in the spatial dimension allows the identification of specific areas where the desired performance limits are not met, such as black-spots where there is a high concentration of dropped calls, reduced throughput levels or low signal strength. Candidate tools include the geo-spatial clustering tools of Table III.

- Classification of the general performance status of the cell: Given the high number of existing performance metrics, and the fact that hidden correlations between them can exist, it can be useful to combine these metrics into a simpler indicator that reflects the overall performance behavior of the cell. Supervised classification tools like those of Table I can be useful here.

B. User-level models

The capabilities offered by big data and big data analytics technologies for processing unprecedented amounts of data will enable the exploitation of the user-data dimension in the different management processes of future 5G networks. Besides supporting relevant business processes, such as customer experience management, valuable knowledge extracted from the characterisation of the network usage made by the individual users can be exploited, not only for increasing the efficiency of the network optimization decision making processes, but also for a better personalization of the network services offered to the different users. Clearly, the extraction of this knowledge involves a trade-off between the achievable degree of user personalization and the complexity associated to the huge volume of data that needs to be processed to achieve it. In general, all the cell-level models described above admit desaggregation and analysis at user level. Nevertheless, in particular, the following components are envisaged to have high interest and applicability for the user-level characterization:

1) Time-domain traffic pattern characterization: This should reflect the behavioral patterns of an individual user when experiencing mobile service in terms of the type of services consumed at different periods of time or the traffic volume generated. Examples include the classification of the time domain traffic pattern of the user (e.g. primarily use of streaming services at night, primarily use of web navigation at noon, etc.), learning the time traffic pattern of a user in order to identify regularities, the extraction of a prediction model at user level to anticipate the service demand of individual users, etc.

2) Spatial-domain traffic pattern characterization: This captures the behavioral patterns of an individual user from the spatial perspective, reflecting which cells the user has been connected to, together with the type of services/applications and the traffic volume generated by the user in each cell. Besides, the analysis of the order in which the user connects to the different cells along a service session can provide information about the trajectories followed by the user (e.g. travelling from home to the office, going from home to the gym, etc.) and this can be used to predict the next cells that a user will be connected to, detect the prototype trajectories that the user follows along the day, etc.

3) Performance characterization: This identifies the performance experienced by an individual user in terms of the different KPIs discussed above, such as accessibility, dropping rate, throughput, latency, etc. Besides, the comparison between the performance experienced by the user and the performance at the cell level can be a useful indicator to decide if specific actions need to be carried out, since it may happen that the overall cell performance (i.e. aggregated or averaged for all the users) is adequate but a specific user is repetitively affected by bad performance (e.g. because the user is very often located in an area of the cell with poor coverage, etc.) so that actions can be triggered (e.g., increase priority level in packet scheduling mechanisms) to enhance user's satisfaction.

C. Cell clustering models

Cell clustering refers to the process of identifying groups of cells that exhibit certain similarities, so that a more efficient management can be carried out by considering the group of cells as a whole rather than on considering each cell individually. The cell clustering model can capture different perspectives:

1) Clustering cells that exhibit a certain degree of mutual interaction, e.g. in the form of inter-cell interference, coverage overlapping, neighboring relationships, etc. In this case, cell clustering will be closely linked to the geographical proximity between cells. The adequate cluster size or the specification of the cluster borders need to be considered. Acquiring knowledge about commonalities affecting a certain group of cells can lead to more efficient decisions taken at area level rather than at cell level both in terms of planning (e.g., high load levels in several neighboring cells may advice the addition of a new site in the area) and operations (e.g., high dropping affecting a number of neighboring cells can be associated to an external source of interference affecting a wide area).

2) Clustering cells that exhibit similar characteristics or a similar behavior, regardless of whether they are closely located or not. The rationality of this approach is that the knowledge learnt for one cell can be valid for the rest of cells of the same cluster. In other words, SON functions can benefit from the knowledge of cells that exhibit a similar behavior, because the actions that are learnt to be good (bad) for one cell can also be good (bad) for other cells of the same cluster. The clustering process according to the cell similarities can be done based on different dimensions: (i) performance (e.g. group cells that offer similar accessibility, retainability, QoS KPIs, etc.), (ii) traffic characteristics (e.g. group cells that exhibit similar traffic volume, service distributions, etc.), (iii) RF characteristics (e.g. group cells with similar received power distributions, interference, etc.). While the performance-based clustering may lead to a highly aggregated vision, since it will be the result of multiple different effects, the clustering based on e.g. traffic or RF may provide a more detailed vision of the cell behavior. The identification of cells that are similar from the RF perspective and with similar traffic types may allow extrapolating effects from one cell to the other. For example the behavior of a cell in front of a given traffic level, can be extrapolated to other cells that have similar characteristics, so it is possible to predict how these cells will behave when they reach the traffic volume as the initial cell. When doing this clustering, each cell is considered as an object characterized by a number of features including aspects such as the RF measurements, traffic patterns, static attributes, user-level characteristics, etc. The clustering algorithm will then analyze these features and will group the cells according to their similarity, usually measured in terms of a distance function. Candidate tools to perform this process include the techniques discussed in Table III. Specifically, hierarchical methods can be particularly suitable as long as the number of possible clusters is not known a priori but it has to be derived from the observed behaviors of the involved cells.

TABLE IV. APPLICABILITY OF THE KNOWLEDGE MODELS IN DIFFERENT SON FUNCTIONS

SON functions \ Knowledge Models		Cell-level									User-level			Cell clustering	
		Classification of the traffic pattern	Learning the traffic behavior in time domain	Traffic prediction	Clustering spatial traffic (Hot-spots)	Learning mobility patterns	Learning time pattern of a performance	Prediction of a performance indicator	Learning space-domain black-spots	General performance status of a cell	User time-domain traffic pattern	User spatial-domain traffic pattern	User performance	Cell clustering based on interaction	Cell clustering based on similarities
Self-planning	Planning a new cell			•	•	•		•	•	•					
	RF planning of a new cell				•	•			•						
	Spectrum planning	•	•	•			•	•		•	•			•	•
Self-optimization	Coverage and Capacity Optimization				•	•			•					•	•
	Automatic Neighbor Relations					•	•		•					•	
	Intercell Interference Coordination		•	•	•		•	•	•			•	•	•	
	Mobility Robustness Optimization				•	•	•	•	•			•	•	•	
	Mobility Load Balancing	•	•	•	•	•	•	•	•					•	
	Optimization of admission control / congestion		•	•			•	•			•	•	•		•
	Energy Saving	•	•	•	•		•	•	•		•	•	•	•	
	SON coordination	•	•	•	•	•	•	•	•		•	•	•	•	•
Self-healing	Cell Outage Detection	•	•				•		•	•					
	Cell Outage Compensation	•	•	•	•		•	•	•					•	

IV. RESEARCH DIRECTIONS AND CONCLUDING REMARKS

This paper has presented a vision of an AI-based 5G SON framework that processes input data from very different sources and extracts, through learning-based classification, prediction and clustering models, relevant knowledge models used to drive the SON decisions. Following a taxonomy of SON functions and a detailed list of AI-based tools that could empower the 5G SON framework, a number of potential knowledge models and their applicability to the different SON functions have been identified.

The full introduction of AI-based support for supporting 5G SON presents a number of challenges and associated research directions, as discussed in the following.

a) Data acquisition and pre-processing: The integration of big data analytics tools (e.g. Hadoop) into the 5G SON ecosystem opens the door for implementing both the pre-processing and the knowledge discovery at a larger scale in terms of the amount of data (e.g. using larger training sets sizes, numbers of attributes, data sources, etc.), thanks to enhanced acquisition, storage and processing capabilities. Big data analytics are essential to exploit the user-level dimension, which inherently requires managing massive amounts of data. In this respect, the suitability of existing big data technologies should be analyzed in the context of 5G SON requirements in terms of data volume, processing times and variety of data.

b) AI-based tools for knowledge discovery: The application of the AI-based tools for classification, prediction and clustering in 5G SON involves many aspects to be traded off, such as the accuracy of the models, the computational requirements, the amount of required training data and the time needed to execute the training stage. The wide variety of tools, as reflected in Tables I, II and III, and their diverse characteristics invites to a systematic analysis to delimitate strengths and weaknesses of such tools when applied to a wireless mobile environment. Besides, a robust design of knowledge discovery needs to include monitoring mechanisms to check the validity and reliability of the learnt models and to identify the need for modifying these models to deal with changes in the environment, in the network configuration or in the users' behavior.

c) Knowledge-based 5G SON solutions: Table IV has pointed out, qualitatively, which knowledge models can provide valuable support to the various SON functions. Therefore, the materialization of SON solutions would involve an in-depth analysis of the components captured in Table IV. For some SON functions, it can be envisaged that a coherent combination of several knowledge models would be of interest, particularly by combining proactive capabilities (e.g., obtained from prediction models) with reactive capabilities (e.g., obtained from spatio-temporal models characterizing past behaviours). Matching the time scale of operation of 5G SON solutions with the time scale inherent to the different knowledge models is another challenge that deserves careful attention. Knowledge models will typically characterize the network/user behavior at a relatively long-term basis, after having observed a sufficient amount of different situations to extract representative models. Then, for 5G SON solutions operating in the long term (e.g., self-planning), these models can directly be exploited. In turn, for those functions that are typically more dynamic and involve decision making at shorter time scales (e.g., self-optimization), combining and integrating the knowledge models with other more reactive AI-based techniques, such as reinforcement learning tools, can lead to a more efficient and robust operation.

d) Implementation considerations: It is necessary to devise which AI-based 5G SON solutions will be implemented following a centralized approach, i.e. residing at the management systems such as the Element Manager (EM) or the Network Manager (NM), a distributed approach, i.e. residing at the network elements such as the cells, or a hybrid approach that includes some components at the management systems and some others at the cells [13]. Then, the design of a 5G SON solution needs to adequately map the knowledge discovery and exploitation stages of Fig. 1 onto these centralized/decentralized/hybrid components. The choice should trade off aspects such as: (1) availability of data (e.g. solutions running at the NM will be typically constrained by the data made available through the open interface between EM and NM), (2) computational capabilities (e.g. centralized approaches will typically be able to support more computationally demanding operations than decentralized approaches, although this limitation can be overcome with the introduction of Mobile Edge Computing technologies that enhance the computational capabilities at the RAN), (3) signalling requirements to transfer the data between the involved entities.

e) SDN/NFV enabled SON: The inclusion of softwarization technologies such as SDN and NFV will substantially change the way how the 5G networks will be managed in relation to legacy systems, thus having a direct impact on the design and development of 5G SON solutions. The applicability of SDN to mobile networks, generically referred to as Software Defined Mobile Networks (SDMN) [14] will bring a systematic abstraction and modularity of the functions within the RAN, enabling a hierarchical control architecture in which the high control layer controls lower layers through defining behaviors without the need to know their specific implementation. Network views at the high control layer have to be built on the proper abstraction of lower layers through defined open control interfaces and primitives. This will be essential to facilitate the implementation of programmable SON functions and the support of AI-based knowledge discovery depending on the type of data that is made available at the different layers. Besides, combining the programmable control with the possibility to run the control programs on general purpose computing/storage resources, as facilitated by NFV, enables the deployment of very flexible control functions for different groups of network entities or end users, as required. Whatever SON function is considered, if proper open control interfaces are established, it can be implemented as a Virtual Network Function (VNF) fitting into the NFV-MANO (Network Function Virtualization - Management and Orchestration) framework of [15]. This provides an inherent flexibility through easy instantiation, modification and termination procedures, an inherent efficiency in hardware utilization, since VNFs are executed on a pool of shared resources, and an inherent capability to add new functionalities and/or extend/upgrade/evolve existing VNFs. Therefore, this facilitates a more open market, where third-parties can provide SON solutions that can be easily and flexibly implemented in any NFV-enabled network.

REFERENCES

[1] METIS 2020 project, http://www.metis2020.com

[2] R. El Hattachi, J. Erfanian (editors) "NGMN 5G White Paper", NGMN Alliance, February, 2015

[3] 5G-PPP, "5G Vision. The 5G Infrastructure Public Private Partnership: the next generation of communication networks and services", https://5g-ppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf

[4] C-L I., Y. Liu, S. Han, S. Wang, G. Liu, "On Big Data Analytics for Greener and Softer RAN", IEEE Access, August, 2015.

[5] J. Ramiro, K. Hamied, Self-Organizing Networks. Self-planning, self-optimization and self-healing for GSM, UMTS and LTE, John Wiley & Sons, 2012.

[6] J. Han, M. Kamber, "Data Mining Concepts and Techniques", 2nd edition, Elsevier, 2006.

[7] R.A.Wilson, F.C.Keil, The MIT Encyclopedia of the Cognitive Sciences, MIT Press, 1999.

[8] A. Imran, A. Zoha, A. Abu-Dayya, "Challenges in 5G: How to Empower SON with Big Data for Enabling 5G", IEEE Network, November/December, 2014, pp. 27-33.

[9] N. I. Sapankevych, R. Sankar, "Time Series Prediction: Using Support Vector Machines: A Survey", IEEE Computational Intelligence Magazine, May, 2009.

[10] T. Kohonen, "Essentials of the self-organizing map", Neural Networks, Vol.37, pp. 52-65, 2013.

[11] A. C. Gatrell, T. C. Bailey, P. J. Diggle, B. S. Rowlingson, "Spatial point pattern analysis and its application in geographical epidemiology", Transactions of the Institute of British Geographers, Vol. 21, No. 1 (1996), pp. 256-274.

[12] A.J. Brimicombe, "A dual approach to cluster discovery in point event data sets", Computers Environment and Urban Systems, 31(1), pp. 4-18, 2007.

[13] 3GPP 32.500 v12.1.0, "Self-Organizing Networks (SON); Concepts and requirements (Release 12)", September, 2014.

[14] T. Chen, et al. "Software Defined Mobile Networks: Concept, Survey, and Research Directions", IEEE Comm. Mag., Nov., 2015, pp. 126-133.

[15] ETSI GS NFV-MAN 001 (V1.1.1) "Network Function Virtualisation (NFV); Management and Orchestration", December, 2014