Big Data Meets Telcos: A Proactive Caching Perspective

Ejder Bastug, Mehdi Bennis, Engin Zaydan, Manhal Abdel Kader, Alper Karatep, Ahmet Salih Er, Merouane Debbah

Big Data Meets Telcos: A Proactive Caching Perspective

Ejder Baştuğ, Mehdi Bennis, Engin Zeydan, Manhal Abdel Kader, Alper Karatepe, Ahmet Salih Er, and Mérouane Debbah

Abstract: **Mobile cellular networks are becoming increasingly complex to manage while classical deployment/optimization techniques and current solutions (i.e., cell densification, acquiring more spectrum, etc.) are cost-ineffective and thus seen as stopgaps. This calls for development of novel approaches that leverage recent advances in storage/memory, context-awareness, edge/cloud computing, and falls into framework of *big data*. However, the big data by itself is yet another complex phenomena to handle and comes with its notorious 4V: velocity, voracity, volume and variety. In this work, we address these issues in optimization of $5G$ wireless networks via the notion of proactive caching at the base stations. In particular, we investigate the gains of proactive caching in terms of backhaul offloadings and request satisfactions, while tackling the large-amount of available data for content popularity estimation. In order to estimate the content popularity, we first collect users' mobile traffic data from a Turkish telecom operator from several base stations in hours of time interval. Then, an analysis is carried out locally on a big data platform and the gains of proactive caching at the base stations are investigated via numerical simulations. It turns out that several gains are possible depending on the level of available information and storage size. For instance, with 10% of content ratings and 15.4 Gbyte of storage size (87% of total catalog size), proactive caching achieves 100% of request satisfaction and offloads 98% of the backhaul when considering 16 base stations.**

Index Terms: **Proactive caching, content popularity estimation, big data, machine learning, 5G cellular networks.**

I. INTRODUCTION

THE unprecedented increase in data traffic demand driven by mobile video, online social media and over-the-top (OTT) applications are compelling mobile operators to look for innovative ways to manage their increasingly complex networks. This explosion of traffic stemming from diverse domain (e.g., healthcare, machine-to-machine communication, connected cars, user-generated content, smart metering, to mention a few) have different characteristics (e.g., structured/non-structured) and is commonly referred to as *Big Data* [1]. While big data come with "big blessings" there are formidable challenges in dealing with large-scale data sets due to the sheer volume and dimensionality of the data. A fundamental challenge of big data analytics is to shift through large volumes of data in order to discover hidden patterns for actionable decision making. Indeed, the era of collecting and storing data in remote standalone servers where decision making is done offline has dawned. Rather, telecom operators are exploring decentralized and flexible network architectures whereby predictive resource management play a crucial role leveraging recent advances in storage/memory, context-awareness and edge/cloud computing [2–4]. In the realm of wireless, big data brings to network planning a variety of new information sets that can be interconnected to achieve a better understanding of users and networks (e.g., location, user velocity, social geodata, etc.). Moreover, public data from social networks such as Twitter and Facebook provides additional side information about the life of the network, which can be further exploited. The associated benefits are a higher accuracy of user location information or the ability to easily identify and predict user clustering, for example for special events. Undoubtedly, the huge potential associated with big data has sparked a flurry of research interest from industry, government and academics (see [5] for a recent survey), and will continue to do so in the coming years.

At the same time, mobile cellular networks are evolving towards the next generation of 5G wireless communication, in which ultra-dense networks, millimetre wave communications, massive multiple-input multiple-output (massive-MIMO), edge caching, device-to-device communications play a pivotal role (see [6] and references therein). Unlike the base station-centric architecture paradigm assuming *dumb* terminals and in which network optimization is carried out in a *reactive* way, 5G networks will be truly disruptive in terms of being user-centric, context-aware and proactive/anticipatory in nature. While continued evolution in spectral efficiency is expected, the maturity of air interfaces of current systems (LTE-Advanced) mean that no major improvements of spectral efficiency can be anticipated. Additional measures like the brute force expansion of wireless infrastructure (number of cells) and the licensing of more spectrum are prohibitively expensive. Thus, innovative solutions are called upon.

In this work, based on the motivations and issues above, we are intent to propose a proactive caching architecture for optimization of 5G wireless networks where we exploit large amount of available data with the help of big data analytics and machine learning tools. In other words, we investigate the gains

Manuscript received April 30, 2015; approved for publication by Zhu Han, Guest Editor, September 24, 2015.

This research has been supported by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering), the SHARING project under the Finland grant 128010 and TUBITAK TEYDEB 1509 project grant (numbered 9120067) and the project BESTCOM.

E. Baştuğ, M. Debbah, M. A. Kader are with Large Networks and Systems Group (LANEAS), CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France, emails: {ejder.bastug, merouane.debbah}@centralesupelec.fr and manhalak@gmail.com.

M. Bennis is with Centre for Wireless Communications, University of Oulu, Finland, email: bennis@ee.oulu.fi.

E. Zeydan, A. Karatepe and A. S. Er are with AveaLabs, Istanbul, Turkey, email: {engin.zeydan, alper.karatepe, ahmetsalih.er}@{avea.com.tr, turktelekom.com.tr}.

M. Debbah is also with Mathematical and Algorithmic Sciences Lab, Huawei France R&D, Paris, France.

1229-2370/15/$10.00

of proactive caching both in terms of backhaul offloadings and request satisfactions, where machine learning tools are used to model and predict the spatio-temporal user behaviour for proactive cache decision. By caching strategic contents at the edge of network, namely at the base stations, network resources are utilized more efficiently and users' experience is further improved. However, the estimation of content popularity tied with spatio-temporal behaviour of users is a very complex problem due to the high dimensional aspects of data, data sparsity and lack of measurements. In this regard, we present a platform to parallelize the computation and execution of the content prediction algorithms for cache decision at the base stations. As a real-world case study, a large amount of data collected from a Turkish telecom operator, one of the largest mobile operator in Turkey with 16.2 million of active subscribers, is examined for various caching scenarios. Particularly, the traces of mobile users' activities are collected from several base stations in hours of time interval and are analysed inside the network under the privacy concerns and regulations. The analysis is carried out on a big data platform and caching at the base stations has been investigated for further improvements of users' experience and backhaul offloadings.

A. Prior Work and Our Contribution

The use of big data in mobile computing research has been investigated recently such as in [7]. The idea of caching at the edge of wireless network has also been studied in various works [8–14], including proactive caching for 5G wireless networks [2]. In detail, a proactive caching procedure using perfect knowledge of content popularity is studied in [8]. A caching architecture (namely FemtoCaching) relying on cache-enabled user devices and small base stations is introduced in [9]. The caching problem as a many-to-many matching game is formulated in [11] and caching gains are characterized numerically. Deployment aspects of cache-enabled base stations via stochastic geometry tools is investigated in [10] where the outage probability is derived as a function of signal-to-interference-plus-noise ratio (SINR), base station density and storage size. For optimal cache allocations, an approximation framework based on a well-known facility location problem is given in [12]. The impact of unknown content popularity on cache decision is characterized in [14]. The advantage of multicast transmission together with caching at the base station is investigated in [13]. We refer our readers to [15] for a recent survey and more comprehensive details.

Compared to the works mentioned above, our main contribution in this work is to make tighter connections of big data phenomena with caching in $5G$ wireless networks, by proposing a proactive caching architecture where statistical machine learning tools are exploited for content popularity estimation. Combined with a large-scale real-world case study, this is perhaps the first attempt on this direction and highlights a huge potential of big data for 5G wireless networks.

The rest of paper is organized as follows. Our network model for proactive caching is detailed in Section II. A practical case study of content popularity estimation on a big data platform is presented in Section III, including a characterization of users' traffic pattern. Subsequently, numerical results for cache-enabled base stations and relevant discussions are carried out in Section IV. We finally conclude in Section V and draw our future directions in the same section.

II. NETWORK MODEL

Suppose a network deployment of M small base stations (SBSs) from the set $\mathcal{M} = \{1, \dots, M\}$ and N user terminals (UTs) from the set $\mathcal{N} = \{1, \dots, N\}$. Each SBS m has access to the broadband Internet connection via a wired backhaul link with capacity C_m Mbyte/s, and is able to provide this broadband service to its users via a wireless link with total capacity of C'_m Mbyte/s. Due to the motivation that the backhaul capacity is generally limited in densely deployed SBSs scenarios [6], we further consider that $C_m < C'_m$. Also, assume that each user $n \in \mathcal{N}$ is connected to only one SBS and is served via unicast sessions[1]. In particular, we assume that UTs request contents (i.e., videos, files, news, etc.) from a library $\mathcal{F} = \{1, \dots, F\}$, where each content f in this library has a size of $L(f)$ Mbyte and bitrate requirement of $B(f)$ Mbyte/s, with

$$L_{\min} = \min_{f \in \mathcal{F}} \{L(f)\} > 0 \tag{1}$$

$$L_{\max} = \max_{f \in \mathcal{F}} \{L(f)\} < \infty \tag{2}$$

and

$$B_{\min} = \min_{f \in \mathcal{F}} \{B(f)\} > 0 \tag{3}$$

$$B_{\max} = \max_{f \in \mathcal{F}} \{B(f)\} < \infty. \tag{4}$$

The users' content requests in fact follow a Zipf-like distribution $P_{\mathcal{F}}(f), \forall f \in \mathcal{F}$ given as [17]:

$$P_{\mathcal{F}}(f) = \frac{\Omega}{f^{\alpha}} \tag{5}$$

where

$$\Omega = \Big(\sum_{i=1}^{F} \frac{1}{i^{\alpha}} \Big)^{-1}.$$

The parameter α in (5) describes the steepness of the distribution. This kind of power laws is used to characterize many real-world phenomena, such as the distribution of files in the web-proxies [17] and the traffic dynamics of cellular devices [18]. Higher values of α corresponds to a steeper distribution, meaning that a small subset of contents are highly popular than the rest of the catalog (namely users have very similar interests). On the other hand, the lower values describe a more uniform behaviour with almost equal popularity of contents (namely users have more distinct interests). The parameter α can take different values depending on users' behaviour and SBSs deployment strategies (i.e., home, enterprise, urban and rural environments), and its practical value in our experimental setup will be given in the subsequent sections.

Given such a global content popularity in the decreasing ordered case, the content popularity matrix of the m-th SBS at time t is specifically described by $\mathbf{P}^m(t) \in \mathbb{R}^{\mathbb{N} \times \mathbb{F}}$ where each

[1]The unicast service model can also be extended to the multicast case. See [13, 16] for studies in this direction.

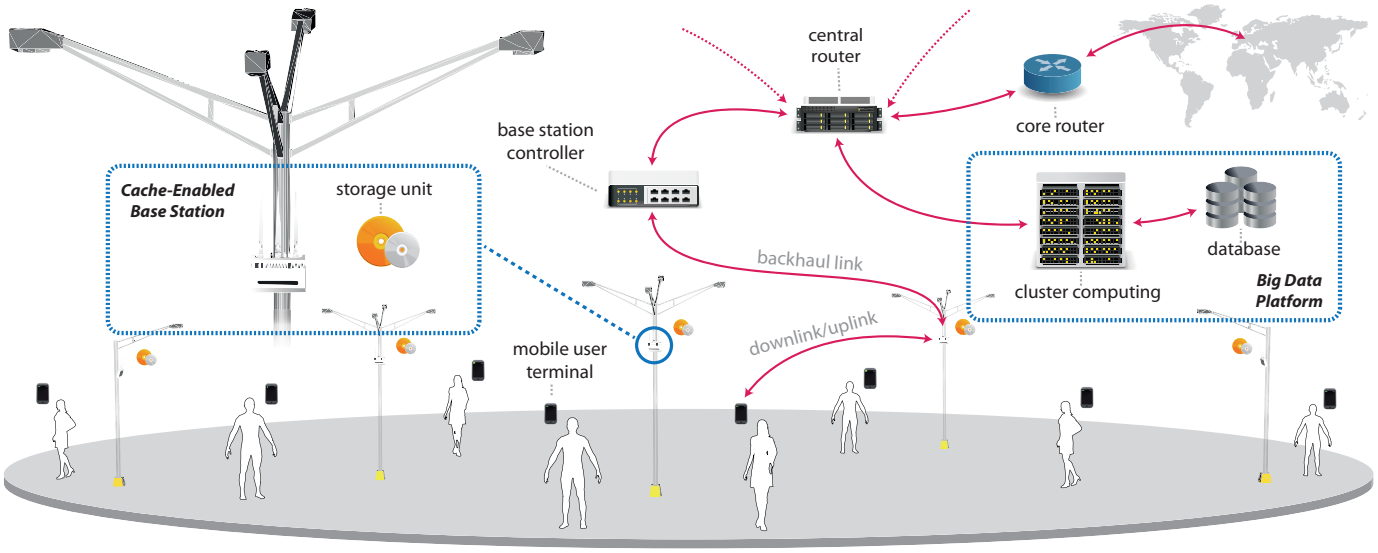


Fig. 1: An illustration of the network model. A *big data platform* is in charge of tracking/predicting users' demand, whereas *cache-enabled base stations* store the strategic contents predicted on the big data platform.

entry $P^m_{n,f}(t)$ corresponds to the probability that the n-th user requests the f-th content. In fact, the matrix $\mathbf{P}^m(t)$ is the local content popularity distribution observed at the base station m at time t, whereas the Zipf distribution $P_{\mathcal{F}}(f), \forall f \in \mathcal{F}$ is used to characterize the global content popularity distribution of all contents in (decreasing) sorted order.

In this scenario, we consider that each SBS has a finite storage capacity of S_m and proactively caches selected contents from the library $\mathcal{F}$ during peak-off hours. By doing so, the bottlenecks caused by the limited-backhaul are avoided during the delivery of users' content requests in peak hours. The amount of satisfied requests and backhaul load are of paramount importance and are defined as follows. Suppose that D number of contents are requested during the duration of T seconds, and are represented by the set $\mathcal{D} = \{1, ..., D\}$. Assume that the delivery of content is started immediately when the request $d \in \mathcal{D}$ arrives to the SBS. Then, the request d is called *satisfied* if the rate of content delivery is equal or higher than the bitrate of the content in the end of service, such as:

$$\frac{L(f_d)}{\tau'(f_d) - \tau(f_d)} \geq B(f_d) \tag{6}$$

where f_d describes the requested content, $L(f_d)$ and $B(f_d)$ are the size and bitrate of the content, $\tau(f_d)$ is the arrival time of the content request and $\tau'(f_d)$ the end time delivery.[2] Defining the condition in (6) stems from the fact that, if the delivery rate is not equal nor higher than the bitrate of the requested content, the interruption during the playback (or download) occurs thus users would have less quality-of-experience (QoE)[3]. Therefore, the situations where this condition holds are more desirable for better QoE. In (6), note also that the end time of delivery for request d, denoted by $\tau'(d)$, highly depends on the load of the system, capacities of the backhaul and wireless links as well as availability of contents at the base stations. Given this definition of satisfied requests and related explanations, the users' average request *satisfaction ratio* is then defined for the set of all requests, that is:

$$\eta(\mathcal{D}) = \frac{1}{D} \sum_{d \in \mathcal{D}} \mathbb{1}\left\{ \frac{L(f_d)}{\tau'(f_d) - \tau(f_d)} \geq B(f_d) \right\} \tag{7}$$

where $\mathbb{1}\{...\}$ is the indicator function which takes 1 if the statement holds and 0 otherwise. Now, denoting $R_d(t)$ Mbyte/s as the instantaneous rate of backhaul for the request d at time t, with $R_d(t) \leq C_m, \forall m \in \mathcal{M}$, the average *backhaul load* is then expressed as:

$$\rho(\mathcal{D}) = \frac{1}{D} \sum_{d \in \mathcal{D}} \frac{1}{L(f_d)} \sum_{t=\tau(f_d)}^{\tau'(f_d)} R_d(t). \tag{8}$$

Here, the outer sum is over the set of all requests whereas the inner sum gives the total amount of information passed over the backhul for request d which is at most equal to the length of requested file $L(f_d)$. The instantaneous rate of backhul for request d, denoted by $R_d(t)$, heavily depends on the load of the system, capacity of the backhaul link and cached contents at the base stations.

In fact, by pre-fetching the contents at the SBSs, the access delays to the contents are minimized especially during the peak hours, thus yielding higher satisfaction ratio and less backhaul load. To elaborate this, now consider the cache decision matrix of SBSs as $\mathbf{X}(t) \in \{0,1\}^{M \times F}$, where the entry $x_{m,f}(t)$ takes 1 if the f-th content is cached at the m-th SBS at time t, and 0 otherwise. Then, the backhaul offloading problem under a specific

[2]One can also consider/exploit future information (i.e., start time of requests, end time of content delivery) in the context of proactive resource allocation (see [19] for instance).

[3]In practice, a video content has typically a bitrate requirement ranging from 1.5 to 68 Mbit/s [20].

request satisfaction constraint is formally given as follows:

$$\underset{\mathbf{X}(t),\mathbf{P}^m(t)}{\text{minimize}} \quad \rho(\mathcal{D}) \tag{9}$$

$$\text{subject to} \quad L_{\min} \leq L(f_d) \leq L_{\max}, \qquad \forall d \in \mathcal{D}, \tag{9a}$$

$$B_{\min} \leq B(f_d) \leq B_{\max}, \qquad \forall d \in \mathcal{D}, \tag{9b}$$

$$R_d(t) \leq C_m, \qquad \forall t, \forall d \in \mathcal{D}, \forall m \in \mathcal{M}, \tag{9c}$$

$$R'_d(t) \leq C'_m, \qquad \forall t, \forall d \in \mathcal{D}, \forall m \in \mathcal{M}, \tag{9d}$$

$$\sum_{f \in \mathcal{F}} L(f) x_{m,f}(t) \leq S_m, \qquad \forall t, \forall m \in \mathcal{M}, \tag{9e}$$

$$\sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} P^m_{n,f}(t) = 1, \qquad \forall t, \forall m \in \mathcal{M}, \tag{9f}$$

$$x_{m,f}(t) \in \{0,1\}, \qquad \forall t, \forall f \in \mathcal{F}, \forall m \in \mathcal{M}, \tag{9g}$$

$$\eta_{\min} \leq \eta(\mathcal{D}), \tag{9h}$$

where $R'_d(t)$ Mbyte/s describes the instantaneous rate of wireless link for request d and $\eta_{\min}$ represents the minimum target satisfaction ratio. In particular, the constraints (9a) and (9b) are to bound the length and bitrate of contents in the catalog for feasible solution, the constraints (9c) and (9d) are the backhaul and wireless link capacity constraints, (9e) holds for storage capacity for caching, (9f) is to ensure the content popularity matrix as a probability measure, (9g) denotes the binary decision variables of caching, and finally the expression in (9h) is the satisfaction ratio constraint for QoE.

In order to tackle this problem, the cache decision matrix $\mathbf{X}(t)$ and the content popularity matrix estimation $\mathbf{P}^m(t)$ have to be optimized jointly. However, solving the problem (9) is very challenging as:

i) the storage capacity of SBSs, the backhaul and wireless link capacities are limited.
ii) the catalog size and number of users with unknown ratings[4] are very large in practice.
iii) the optimal uncoded[5] cache decision for a given demand is non-tractable [8, 9, 12].
iv) the SBSs have to track, learn and estimate the sparse content popularity/rating matrix SBSs $\mathbf{P}^m(t)$ while making the cache decision.

In order to overcome these issues, we restrict ourselves to the fact that cache decision is made during peak-off hours, thus $\mathbf{X}(t)$ remains static during the content delivery in peak hours and is represented by $\mathbf{X}$. Additionally, the content popularity matrix is stationary during T time slots and identical among the base stations, thus $\mathbf{P}^m(t)$ is represented by $\mathbf{P}$.

After these considerations, we now suppose that the problem can be decomposed into two parts in which the content popularity matrix $\mathbf{P}$ is first estimated, then is used in the caching decision $\mathbf{X}$ accordingly. In fact, if sufficient amount of users' ratings are available at the SBSs, we can construct a k-rank approximate popularity matrix $\mathbf{P} \approx \mathbf{N}^T\mathbf{F}$, by jointly learning the factor matrices $\mathbf{N} \in \mathbb{R}^{\urcorner \times \mathbb{N}}$ and $\mathbf{F} \in \mathbb{R}^{\urcorner \times \mathbb{F}}$ that minimizes the following cost function:

$$\underset{\mathbf{P}}{\text{minimize}} \sum_{P_{ij} \in \mathcal{P}} \Big(\mathbf{n}_i^T \mathbf{f}_j - P_{ij}\Big)^2 + \mu\Big(||\mathbf{N}||_F^2 + ||\mathbf{F}||_F^2\Big) \tag{10}$$

where the summation is done over the corresponding user/content rating pairs P_{ij} in the training set $\mathcal{P}$. The vectors $\mathbf{n}_i$ and $\mathbf{f}_j$ here describe the i-th and j-th columns of $\mathbf{N}$ and $\mathbf{F}$ matrices respectively, and $||.||_F^2$ represents the Frobenius norm. The parameter μ is used to provide a balance between the regularization and fitting the training data. Therein, high correspondence between the user factor matrix $\mathbf{N}$ and content factor matrix $\mathbf{F}$ leads to a better estimate of $\mathbf{P}$. In fact, the problem (10) is a regularized least square problem where the matrix factorization is embedded in the formulation. Despite various approaches, the matrix factorization methods are commonly used to solve this kind of problems and has many applications such as in recommendation systems (i.e., Netflix video recommendation). In our case detailed in the following sections, we have used regularized sparse singular value decomposition (SVD) to solve the problem algorithmically which exploits the least square nature of the problem. The overview of these approaches, sometimes called collaborative filtering (CF) tools, can be found in [22,23]. When the estimation of content popularity matrix $\mathbf{P}$ is obtained, the caching decision $\mathbf{X}$ can be made in this scenario accordingly.

In practice, the estimation of $\mathbf{P}$ in (10) can be done by collecting/analysing large amount of available data on a *big-data platform* of the network operator, and strategic/popular contents from this estimation can be stored at the *cache-enabled base stations* whose cache decisions are represented by $\mathbf{X}$. By doing this, the backhaul offloading problem in (9) is minimized and higher satisfactions are achieved. Our network model including such an infrastructure is illustrated in Fig. 1. In the following, as a case study, we detail our big data platform and present users' traffic characteristics by analysing large amount of data on this platform. The processed data will be used to estimate the content popularity matrix $\mathbf{P}$ which is essentially required for the cache decision $\mathbf{X}$ and will be detailed in the upcoming sections.

III. BIG DATA PLATFORM

The big data platform used in this work runs in the operator's core network. As mentioned before, the purpose of this platform is to store users' traffic data and extract useful information which are going to be used for content popularity estimation. In a nutshell, the operator's network consists of several districts with more than 10 regional core areas throughout Turkey. The average total traffic over all regional areas consists of approximately over 15 billion packets in uplink direction and over 20 billion packets in the downlink direction daily. This corresponds to approximately over 80 TByte of total data flowing in uplink and downlink daily in a mobile operator's core network. The data usage behaviour results in exponential increase in data traffic of a mobile operator. For example, in 2012, the approximate total data traffic was over 7 TByte in both uplink and downlink daily traffic.

The streaming traces which will be detailed in the sequel, are obtained from one of the operator's core network region, includes the mobile traffic from many base stations, and are captured by a server on a high speed link of 200 Mbit/sec at peak

[4]The term "rating" refers to the empirical value of content popularity/probability and is interchangeable throughout the paper.

[5]In the information theoretical sense, the caching decision can be categorized into "coding" and "uncoded" groups (see [21] for example).

hours. In order to capture Internet traffic data by the server in this platform, a procedure is initialized by mirroring real-world Gn interface data.[6] After mirroring stage of Gn interface, network traffic is transferred into the server on the platform. For our analysis, we have collected traffic of approximately 7 hours starting from 12 pm to 7 pm on Saturday 21'st of March 2015. This traffic is processed on the big data platform which is essentially based on Hadoop.

A. Hadoop platform

Among the available platforms, Hadoop stands out as the most notable one as it is an open source solution [24]. It is made up of a storage module, namely Hadoop Distributed File System (HDFS) and a computation module, namely MapReduce. Whereas HDFS can have centralized or distributed implementations, MapReduce inherently has a distributed structure that enables it to execute jobs in parallel on multiple nodes.

As stated in previous subsection, the accuracy and precision of the proposed mechanism was tested in operator's network. A data processing platform was implemented through using Cloudera's Distribution Including Apache Hadoop (CDH4) [25] version on four nodes including one cluster name node, with computations powers corresponding to each node with INTEL Xeon CPU E5-2670 running @2.6 GHz, 32 Core CPU, 132 GByte RAM, 20 TByte hard disk. This platform is used to extract the useful information from raw data which is described as follows.

B. Data extraction process

First, the raw data is parsed using Wireshark command line utility *tshark* [26] in order to extract the relevant fields of CELL-ID (or service area code (SAC) in our case, in order to uniquely identify a *service area* within a *location area*[7]), LAC, Hypertext Transfer Protocol (HTTP) request-uniform resource identifier (URI), tunnel endpoint identifier (TEID)[8] and TEID-DATA for data and control plane packets respectively, and FRAME TIME indicating arrival time of packets. The HTTP Request-URI is a Uniform Resource Identifier that identifies the resource upon which to apply the request. The *control* packets contain the information elements that carry the information required for future data packets. It contains cell identification ID (CELL-ID), LAC and TEID-DATA fields. The *data* packets contain HTTP-URI and TEID fields.

In the next step, after obtaining those relevant fields from both control and data packets, the extracted data is transferred into HDFS for further analysis. In HDFS, there can be done many data analytics performed over the collected data using Hive Query language (QL) [27]. For example, in order to calculate the HTTP Request-URIs at specific location, the HTTP-URI can be joined with CELL-ID-LAC fields over the same TEID and TEID-DATA fields for data and control packets respectively. In our analysis, due to the limitations on observable number of rows of HTTP-URI fields with a corresponding CELL-ID-LAC fields after mapping, we have proceeded with HTTP Request-URIs and TEID mappings.

From HDFS, a temporary table named *traces-table-temp* is constructed using Hive QL. The *traces-table-temp* has HTTP Request-URI, FRAME TIME and TEID fields. After constructing this table, the sizes of each HTTP Request-URI request is calculated using a separate URI-*size calculator* program that uses HTTPClient API [28] in order to obtain the final table called *traces-table* with fields of SIZE, HTTP Request-URIs, FRAME TIME and TEID. This table has approximately over 420.000 of 4 millions HTTP Request-URI's with SIZE field returned as not zero or null due to unavailability of HTTP response for some requests. Note that in a given session with a specific TEID, there can be multiple HTTP Request-URIs. Each TEID belongs to specific user. Each user can also have multiple TEIDs with multiple HTTP Request-URIs. The steps of data extraction process on the platform is summarized in Fig. 2. Note that the data extraction process is specific to our scenario for proactive caching. However, similar studies in terms of usage of big data platform and exploitation of big data analytics for telecom operators can be found in [29–34].

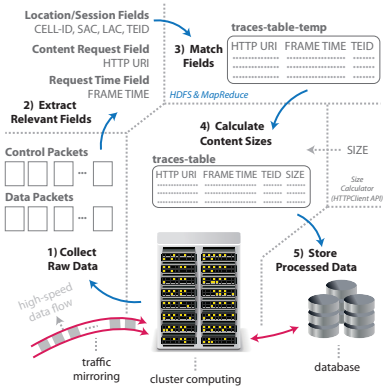


Fig. 2: An overview of the data extraction process on the big data platform.

C. Traffic Characteristics

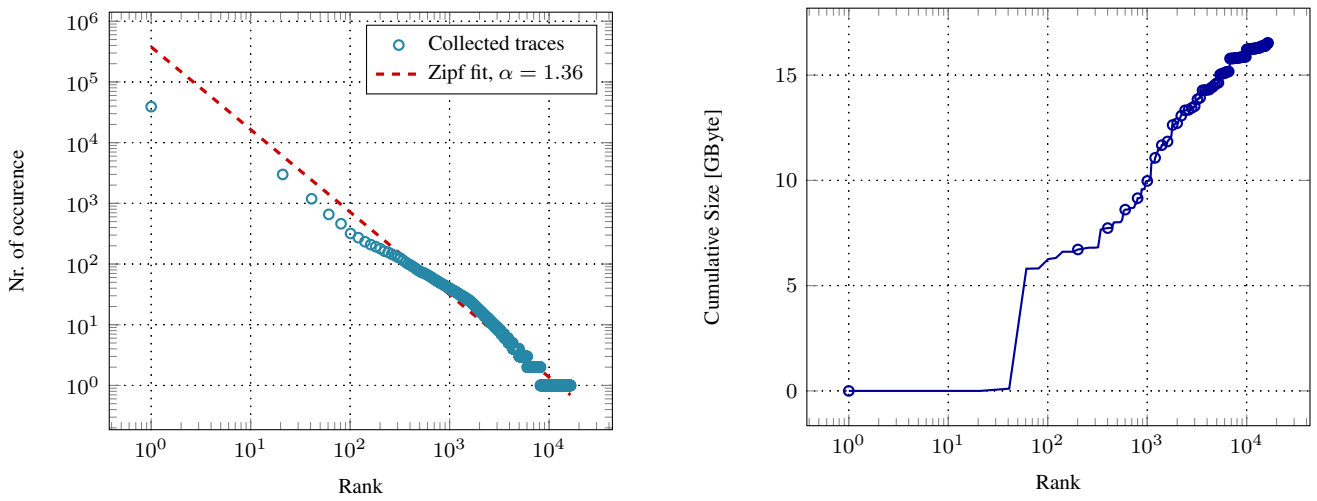
Based on information available in *traces-table*, the global content popularity distribution (namely HTTP-URI popularity distribution) in a decreasing ranked order is plotted in Fig. 3a. According to this available experimental data, we observe that the popularity behaviour of contents follows a Zipf law with steepness parameter $\alpha = 1.36$.[9] Therein, the Zipf curve is calculated in the least square sense from the collected traces and the parameter α is then found by evaluating the slope of the curve. On the other hand, cumulative size of ranked contents is given in

[6] Gn is an interface between Serving GPRS Support Node (SGCN) and Gateway GPRS Support Node (GGSN). Network packets sent from a user terminal to the packet data network (PDN), e.g. internet, pass through SGCN and GGSN where GPRS Tunneling Protocol (GTP) constitutes the main protocol in network packets flowing through Gn interface.

[7] The service area identified by SAC is an area of one or more base stations, and belongs to a location area which is uniquely identified by location area code (LAC). Typically, tens or even hundreds of base stations operates in a given location area.

[8] A TEID uniquely identifies a tunnel endpoint on the receiving end of the GTP tunnel. A local TEID value is assigned at the receiving end of a GTP tunnel in order to send messages through the tunnel.

[9] The value of steepness parameter α can change depending on the scenario. For instance, the steepness parameter of content popularities in YouTube catalog varies from 1.5 to 2.5 [35, 36].



(a) Global content popularity distribution.

(b) Cumulative size distribution.

Fig. 3: Behaviour of content popularity distribution.

Fig. 3b. The cumulative size up to 41-th most-popular contents has 0.1 GByte of size, whereas a dramatical increase appears afterwards. This basically shows that most of the requested contents in our traces has low content sizes and contents with larger sizes are relatively less requested.

We would like to note that a detailed characterization of the traffic for caching is left for future work. Indeed, characterization of the traffic in web proxies which are placed in the intermediate level of network [17], a specific video content catalog in a campus network [37], mobile traffic of users in Mexico [38] can be found in the literature. Compared to these works, we focus on the characterization traffic of mobile users collected from base stations in a large regional area and exploit this information for proactive caching (i.e., content popularity distribution, cumulative size distribution). Based on information available in *traces-table*, we in the following simulate a scenario of cache-enabled base stations.

IV. NUMERICAL RESULTS AND DISCUSSIONS

The list of parameters for numerical setup is given in Table 1. For ease of analysis, the storage, backhaul, and wireless link capacities of small cells are assumed to be identical within each other.

Table 1: List of simulation parameters.

Parameter	Description	Value
T	Time slots	6 hours 47 minutes
D	Number of requests	422529
F	Number of contents	16419
M	Number of small cells	16
$L_{\min}$	Min. size of a content	1 Byte
$L_{\max}$	Max. size of a content	6.024 GByte
$B(f)$	Bitrate of content f	4 Mbyte/s
$\sum_m C_m$	Total backhaul link capacity	3.8 Mbyte/s
$\sum_m \sum_n C'_m$	Total wireless link capacity	120 Mbyte/s

In the simulations, all of D number of requests are taken from the processed data (namely *traces-table*), spanning over a time duration of 6 hours 47 minutes. The arrival times of each request (FRAME TIME), requested content (HTTP-URI) and content size (SIZE) are taken from the same table. Then, these requests are associated to M base stations pseudo-randomly. In order to solve the backhaul offloading problem in (9), the content popularity matrix $\mathbf{P}$ and caching strategy $\mathbf{X}$ are evaluated separately. In particular, the following two methods are used for constructing the content popularity matrix $\mathbf{P}$:

- *Ground Truth*: The content popularity matrix $\mathbf{P}$ is constructed from all available information in *traces-table* instead of solving the problem in (10). Note that the rows of $\mathbf{P}$ represent base stations and columns are contents. The rating density of this matrix is 6.42%.
- *Collaborative Filtering*: For the estimation of content popularity matrix $\mathbf{P}$, the problem in (10) is attempted by first choosing 10% of ratings from *traces-table* uniformly at random. Then, these ratings are used in the training stage of the algorithm and missing entries/ratings of $\mathbf{P}$ are estimated. Particularly, the regularized SVD from the CF methods [23, 39] is used in the algorithmic part.

After constructing the content popularity matrix $\mathbf{P}$ based on these above methods, the cache decision (modelled by the matrix $\mathbf{X}$) is made by storing the most-popular contents greedily at the SBSs until no storage space remains (see [8] for the details). Having these contents cached proactively at the SBSs at $t = 0$, the requests are then served until all of the contents are delivered. The performance metrics request satisfaction and backhaul load are calculated accordingly.

The evolution of users' request satisfaction with respect to the storage size is given Fig. 4a. The storage size is given in terms of percentage where 100% of storage size represents the sum of all size of contents in the catalog (17.7 GByte). From zero storage (0%) to full storage (100%), we can seen that the users' request satisfaction increases monotonically and goes up to 100%, both in ground truth and collaborative filtering approaches. However,

there is a performance gap between the ground truth and CF until 87% of storage size, which is due to the estimation errors. For instance, with 40% of storage size, the ground truth achieves 92% of satisfaction whereas the CF has value of 69%.

The evolution of backhaul load/usage with respect to the storage size of SBSs is given in Fig. 4b. As the storage size of SBSs increases, we see that both approaches reduces backhaul usage (namely higher offloading gains). For example, with 87% of storage size for caching, both approaches offload 98% of backhaul usage. The performance of ground truth is evidently higher than the CF as all of the available information is taken into consideration for caching. We also note that there is a dramatical decrease of backhaul usage in both approaches after a specific storage size. In fact, most of the previous works on caching assume a content catalog with identical content sizes. In our case, we are dealing with real traces in the numerical setup where the size of contents differs from content to content, as discussed in the previous section (see Fig. 3b). According to this scenario, on the one hand, caching a highly popular content with very small size might not reduce the backhaul usage dramatically. On the other hand, caching a popular content with very high size can dramatically reduce the backhaul usage. Therefore, as the CF approach used here is solely based on content popularity, it fails to capture these content size aspects on the backhaul usage, which in turn results in higher storage requirements to achieve the same performance as in the ground truth. This shows the importance of size distribution of popular contents.

We have so far compared the performance gains of these approaches with 10% of rating density in CF. In fact, as the rating density of CF for training increases, we expect to have less estimation error, thus resulting closer satisfaction gains to the ground truth. To show this, the change of root-mean-square error (RMSE) with respect to the training rating density is given in Fig. 5. Therein, we define the error as the root-mean-square of difference between users' content satisfaction of the ground truth and CF approaches over all possible storage sizes. Clearly, as observed in Fig. 5, the performance of CF is improved by increasing the rating density, thus confirming our intuitions.

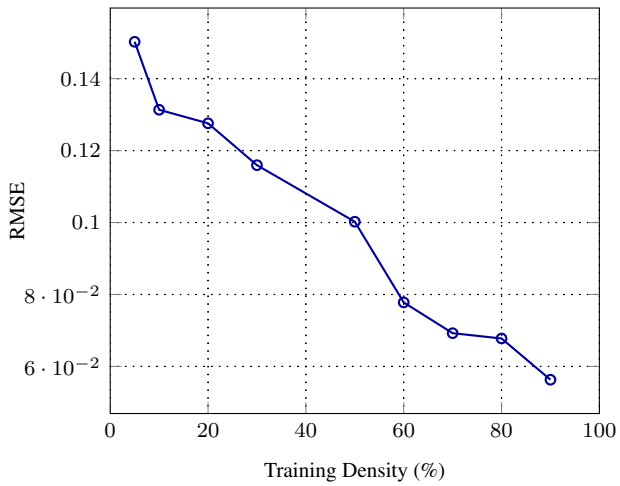


Fig. 5: Evolution of RMSE with respect to the training density.

V. CONCLUSIONS

In this work, we have studied a proactive caching approach for 5G wireless networks by exploiting large amount of available data and employing machine learning tools. In particular, an experimental setup for data collection/extraction process has been demonstrated on a big data platform and machine learning tools (CF in particular) have been applied to predict the content popularity distribution. Depending on the rating density and storage size, the numerical results showed that several caching gains are possible in terms of users' request satisfactions and backhaul offloadings. An interesting future direction of this work is to conduct a more *detailed characterization of the traffic* which captures different spatio-temporal content access patterns. In order to estimate the content access patterns for cache decision, the development of *novel machine learning algorithms* is yet another interesting direction. Finally, design of new *deterministic/randomized cache decision algorithms* are required and should not be purely based on content popularity and storing most popular contents, so that higher backhaul offloading can be achieved while satisfying users' requests.

REFERENCES

[1] C. Lynch, "Big data: How do your data grow?," *Nature*, vol. 455, no. 7209, pp. 28–29, 2008.

[2] E. Baştuğ, M. Bennis, and M. Debbah, "Living on the Edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, pp. 82 – 89, August 2014.

[3] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, pp. 13–16, ACM, 2012.

[4] T. H. Luan, L. Gao, Z. Li, Y. Xiang, and L. Sun, "Fog computing: Focusing on mobile users at the edge," *[Online] arXiv:1502.01815*, 2015.

[5] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[6] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?," *IEEE Journal on Selected Areas in Communications*, vol. 32, pp. 1065–1082, June 2014.

[7] J. K. Laurila, D. Gatica-Perez, I. Aad, B. J., O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen, "The Mobile Data Challenge: Big Data for Mobile Computing Research," in *Pervasive Computing*, 2012.

[8] E. Baştuğ, J.-L. Guénégo, and M. Debbah, "Proactive small cell networks," in *20th International Conference on Telecommunications (ICT'13)*, (Casablanca, Morocco), May 2013.

[9] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, 2013.

[10] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP Journal on Wireless Communications and Networking*, p. 41, February 2015.

[11] K. Hamidouche, W. Saad, and M. Debbah, "Many-to-many matching games for proactive social-caching in wireless small cell networks," in *12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 569–574, May 2014.

[12] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Transactions on Communications*, vol. 62, pp. 3665–3677, October 2014.

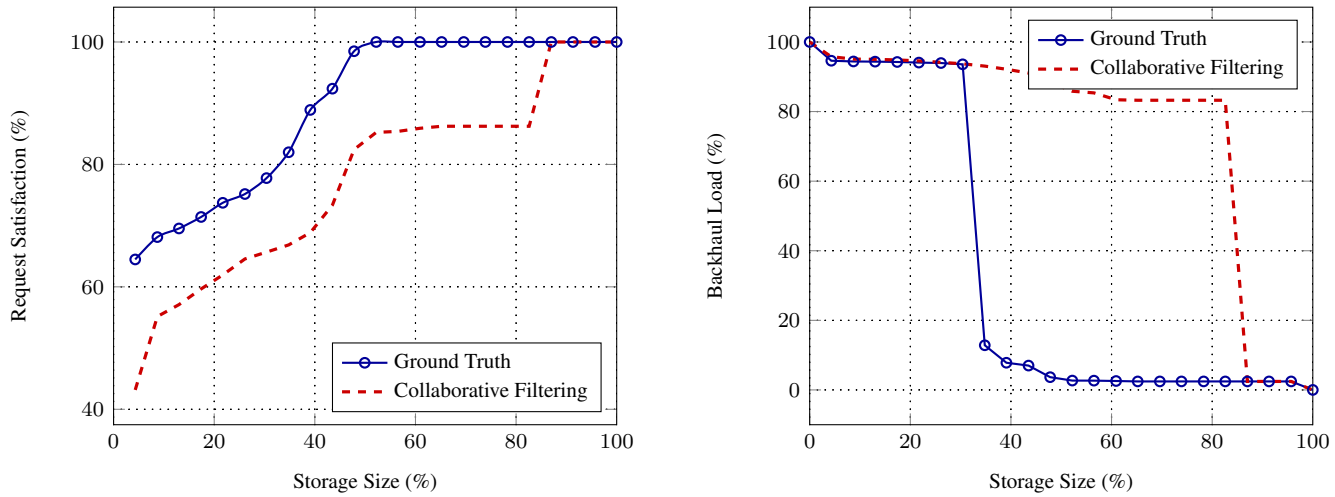
[13] B. Zhou, Y. Cui, and M. Tao, "Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks," *[Online] arXiv:1504.04428*, 2015.

[14] B. B. Nagaraja and K. G. Nagananda, "Caching with unknown popularity profiles in small cell networks," *[Online] arXiv:1504.03632*, 2015.

[15] E. Baştuğ, M. Bennis, and M. Debbah, *Proactive Caching in 5G Small Cell Networks*. Wiley, [In Minor Revision] 2015.

[16] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," in *IEEE Wireless Communications and Networking Conference (WCNC'2014)*, pp. 2300–2305, IEEE, 2014.

[17] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *IEEE Eighteenth*



(a) Evolution of satisfaction with respect to the storage size.

(b) Evolution of backhaul usage with respect to the storage size.

Fig. 4: Simulation results of proactive caching at the base stations.

Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'99), vol. 1, pp. 126–134, IEEE, 1999.

[18] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pp. 305–316, ACM, 2011.

[19] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive data download and user demand shaping for data networks," *submitted to IEEE Transactions on Information Theory [Online] arXiv: 1304.5745*, 2014.

[20] Google, "Recommended upload encoding settings (Advanced)." `https://goo.gl/KJXfhh`, 2015. [Online; accessed 30-August-2015].

[21] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, pp. 2856–2867, May 2014.

[22] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, pp. 30–37, August 2009.

[23] J. Lee, M. Sun, and G. Lebanon, "A comparative study of collaborative filtering algorithms," *[Online] arXiv: 1205.3193*, 2012.

[24] "Apache Hadoop." `http://hadoop.apache.org/`, 2015. [Online; accessed 02-April-2015].

[25] "Cloudera." `http://www.cloudera.com/content/cloudera/en/documentation.html`, 2015. [Online; accessed 02-April-2015].

[26] "The Wireshark Network Analyzer 1.12.2." `https://www.wireshark.org/docs/man-pages/tshark.html`, 2015. [Online; accessed 02-April-2015].

[27] "Apache Hive TM." `https://hive.apache.org/`, 2015. [Online; accessed 02-April-2015].

[28] Apache, "HttpClient API Tutorial." `https://hc.apache.org/httpcomponents-client-ga/tutorial/pdf/httpclient-tutorial.pdf`, 2015. [Online; accessed 25-April-2015].

[29] Y. Dong, Q. Ke, Y. Cai, B. Wu, and B. Wang, "Teledata: data mining, social network analysis and statistics analysis system based on cloud computing in telecommunication industry," in *Proceedings of the third international workshop on Cloud data management*, pp. 41–48, ACM, 2011.

[30] H.-D. J. Jeong, W. Hyun, J. Lim, and I. You, "Anomaly teletraffic intrusion detection systems on hadoop-based platforms: A survey of some problems and solutions," in *15th International Conference on Network-Based Information Systems (NBiS)*, pp. 766–770, IEEE, 2012.

[31] J. Magnusson and T. Kvernvik, "Subscriber classification within telecom networks utilizing big data technologies and machine learning," in *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, pp. 77–84, ACM, 2012.

[32] W. Indyk, T. Kajdanowicz, P. Kazienko, and S. Plamowski, "Mapreduce approach to collective classification for networks," in *Artificial Intelligence and Soft Computing*, pp. 656–663, Springer, 2012.

[33] O. F. Celebi, E. Zeydan, O. F. Kurt, O. Dedeoglu, O. Ileri, B. A. Sungur, A. Akan, and S. Ergut, "On use of big data for enhancing network coverage analysis," in *20th International Conference on Telecommunications (ICT'13)*, (Casablanca, Morocco), May 2013.

[34] I. A. Karatepe and E. Zeydan, "Anomaly detection in cellular network data using big data analytics," in *Proceedings of European Wireless 2014*, pp. 1–5, VDE, 2014.

[35] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 1–14, ACM, 2007.

[36] D. Rossi, G. Rossini, *et al.*, "On sizing ccn content stores by exploiting topological information.," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 280–285, 2012.

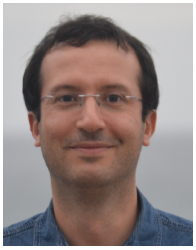
[37] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network–measurements, models, and implications," *Computer Networks*, vol. 53, no. 4, pp. 501–514, 2009.

[38] E. Mucelli Rezende Oliveira, A. Carneiro Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area," Research Report RR-8613, INRIA, October 2014.

[39] A. Paterek, "Improving regularized singular value decomposition for collaborative filtering," in *Proceedings of KDD cup and workshop*, vol. 2007, pp. 5–8, 2007.

Ejder Baştuğ was born in Mardin, Turkey. He received his M.Sc. in Electronic Engineering from Fatih University, Istanbul, in 2012. Since then, he is pursuing his PhD at Centrale-Supélec, under the guidance of Prof. Mérouane Debbah and Prof. Jean-Claude Belfiore. His Ph.D. topic is on distributed caching methods in small cell networks, whereas his main research interests are related to stochastic geometry and machine learning tools for wireless communications. He has been in the executive committee of IEEE WCNC 2014, organizing committee of IEEE BlackSeaCom 2015, technical programme committee of EuCNC 2015, and chair of several international workshops. He is in the organizing committee of one of the two flagship conferences of IEEE Communications Society: IEEE ICC 2017, to be held in Paris.

Mehdi Bennis received his M.Sc. degree from the Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland and the Eurecom Institute, France in 2002. He obtained his Ph.D. degree in electrical engineering December 2009 on spectrum sharing for future mobile cellular systems. His main research interests are in radio resource management, heterogeneous networks, game theory and machine learning. Dr. Bennis was the co-PI of the Broadband Evolved FEMTO (FP7-BeFEMTO) project, and is currently the PI of the European CELTIC project SHARING (2013-2016). He has published more than 100 research papers in international conferences, journals, book chapters and patents. . Dr. Bennis is the winner of the 2015 IEEE Communications Society Fred W. Ellersick Prize.



Engin Zeydan received the PhD degree in February 2011 from the Department of Electrical and Computer Engineering at Stevens Institute of Technology, Hoboken, NJ, USA. Previously, he received his M.S. and B.S. degrees from the Department of Electrical and Electronics Engineering at Middle East Technical University, Ankara, Turkey, in 2006 and 2004, respectively. Dr. Engin Zeydan is working as an R&D engineer for Avea, a mobile operator in Turkey since August 2011. His research interests are in the area of wireless communication networks. Dr. Zeydan is a recipient of the Exemplary Reviewer for IEEE Communications Letters Award, 2011. Engin Zeydan is primarily responsible for carrying out European Commission and nationally funded research activities at Avea İletişim Hizmetleri A.Ş..

Manhal Abdel Kader is from Syr-El-Dannieh, Lebanon. He studied a double degree program, where he had obtained (i) his Engineering diploma in Telecommunication & Computer Sciences from Lebanese University and (ii) his Master's degree in Advanced Wireless Communications Systems (SAR) from Central-Supélec, Paris-Sud 11 & ENS Cachan after being selected for a scholarship program "Bourse Master Ile-de-France" in 2014. He was an intern researcher at Large Networks and Systems Group (LANEAS), focusing on "Big Data and proactive caching algorithms in Small Cell Networks". Since then, he is working as Implementation & Coordination Software Engineer consultant for Amadeus IT Group in the R&D department, managing implementation projects like Database Migrations, building servers, etc. His main research interests are in wireless communications, Big Data, Data Science and machine learning.

Ilyas Alper Karatepe received his B.S. degree in Electrical & Electronics Engineering from Hacettepe University, Ankara, Turkey in 2004. He is studying MSc in Software Engineering at Bogazici University, Istanbul, Turkey since September 2013. İlyas Alper Karatepe currently works in AveaLabs division of Avea, a mobile operator in Turkey, as a senior software engineer since August 2013. Previously, he was working as Middleware Architect for Avea, since April 2010. Before Avea, he worked for Alcatel Lucent, a global telecommunications equipment company, between September 2007 - April 2010 and for Vodafone Turkey, a global mobile operator, between January 2005 - August 2006. His expertise is in Software Development, SOA, System Integration, and Femto&ADSL Modem Management Systems. His research interests are big data analytics and data mining.

Ahmet Salih Er received the B.S. degree in Computer Engineering from Istanbul Technical University, Turkey, in 2008. From December 2008 to August 2011, he was with Vodafone Turkey working as a software development engineer in Service Development Platform. In August 2011, he joined to Turkcell as a software development engineer for Campaign Management System. In December 2013, he joined Avealabs of AVEA as a senior software engineer. His current research interests include SDN, Big Data Analytics, Social Media Analysis, Service Oriented Architectures. He is also studying PhD degree in Kadir Has University Management Information Systems program since September 2014.

Mérouane Debbah entered the Ecole Normale Supérieure de Cachan (France) in 1996 where he received his M.Sc and Ph.D. degrees respectively. He worked for Motorola Labs (Saclay, France) from 1999-2002 and the Vienna Research Center for Telecommunications (Vienna, Austria) until 2003. From 2003 to 2007, he joined the Mobile Communications department of the Institut Eurecom (Sophia Antipolis, France) as an Assistant Professor. Since 2007, he is a Full Professor at CentraleSupelec (Gif-sur-Yvette, France). From 2007 to 2014, he was the director of the Alcatel-Lucent Chair on Flexible Radio. Since 2014, he is Vice-President of the Huawei France R&D center and director of the Mathematical and Algorithmic Sciences Lab. His research interests lie in fundamental mathematics, algorithms, statistics, information & communication sciences research. He is an Associate Editor in Chief of the journal Random Matrix: Theory and Applications and was an associate and senior area editor for IEEE Transactions on Signal Processing respectively in 2011-2013 and 2013-2014. Mérouane Debbah is a recipient of the ERC grant MORE (Advanced Mathematical Tools for Complex Network Engineering). He is a IEEE Fellow, a WWRF Fellow and a member of the academic senate of Paris-Saclay. He has managed 8 EU projects and more than 24 national and international projects. He received 14 best paper awards, among which the 2007 IEEE GLOBECOM best paper award, the Wi-Opt 2009 best paper award, the 2010 Newcom++ best paper award, the WUN CogCom Best Paper 2012 and 2013 Award, the 2014 WCNC best paper award, the 2015 ICC best paper award, the 2015 IEEE Communications Society Leonard G. Abraham Prize and 2015 IEEE Communications Society Fred W. Ellersick Prize as well as the Valuetools 2007, Valuetools 2008, CrownCom2009, Valuetools 2012 and SAM 2014 best student paper awards. He is the recipient of the Mario Boella award in 2005, the IEEE Glavieux Prize Award in 2011 and the Qualcomm Innovation Prize Award in 2012.