



Advanced spatial network metrics for cognitive management of 5G networks

Ignacio Sanchez-Navarro¹ · Jorge Bernal Bernabe^{1,2} · Jose M. Alcaraz-Calero¹ · Qi Wang¹

© The Author(s) 2020

Abstract

The emerging fifth-generation (5G) mobile networks are empowered by softwarization and programmability, leading to the huge potentials of unprecedented flexibility and capability in cognitive network management such as self-reconfiguration and self-optimization. To help unlock such potentials, this paper proposes a novel framework that is able to monitor and calculate 5G network topological information in terms of advanced spatial metrics. These metrics, together with enabling and optimization algorithms, are purposely designed to address the complexity of 5G network topologies introduced by network virtualization and infrastructure sharing among operators (multi-tenancy). Consequently, this new framework, centred on a topology monitoring agent (TMA), enables on-demand 5G networks' spatial knowledge and topological awareness required by 5G cognitive network management in making smart decisions in various autonomous network management tasks including but not limited to virtual network function placement strategies. The paper describes several technical use cases enabled by the proposed framework, including proactive cache allocation, computation offloading, node overloading alerting, and load balancing. Finally, a realistic 5G testbed is deployed with the central component TMA, together with the new spatial metrics and associated algorithms, implemented. Experimental results empirically validate the proposed approach and demonstrate the scalability and performance of the TMA component.

Keywords 5G networks · Topology management · Spatial network metrics · Cognitive management

1 Introduction

Network management in the forthcoming fifth-generation (5G) mobile networks is notably influenced by the softwarization of network infrastructures where several hardware components are virtualized and by the multi-tenancy of the network infrastructures where hardware components are shared by different mobile operators. The main motivation of these 5G capabilities is the reduction of both capital and operational costs. 5G virtual network functions (VNFs) can now be deployed automatically and on-demand on the Edge and the Core segments of the 5G network and can be migrated between the computers that belong to the same net-

work segment or across network segments as required. These characteristics, along with the mobility of 5G users across the different antennas of 5G networks, make 5G network topologies highly dynamic and dependent of the status of the network and the dynamics of the human behaviour. This complicates the optimal management of resources and services in the 5G networks.

Unlike previous 3G/4G networks, the dynamic nature of the novel 5G topologies imposes continuous decisions about the concrete location where to perform VNF placements and migrations to maximize the effectiveness of the services delivered by such VNFs (e.g., cache, load balancer, and baseband unit). For example, a cache service will reduce bandwidth consumption in the network links located after such a cache service due to the hits into the cached content and will reduce the latency in the delivery of cached content to the final users. Thus, the placement strategy of such a cache service will play a vital role in deciding where is the current most effective location to maximize latency reduction and to minimize bandwidth consumption.

Communicated by V. Loia.

✉ Ignacio Sanchez-Navarro
ignacio.sanchez@uws.ac.uk

¹ School of Computing, Engineering and Physical Sciences, University of West of Scotland, Paisley, UK

² Departamento de Ingenieria de la Informacion y las Comunicaciones, University of Murcia, Murcia, Spain

A way to address the placement strategy of services in 5G networks is to make use of spatial network metrics to identify critical locations in the 5G network topology, represented as evolving graphs with hardware and software resources. Spatial network metrics are indices that provide insights about the centrality of the vertices of a graph and can be interpreted as a level of effectiveness of a given resource due to the spatial location with respect to the whole topological structure of the network.

Spatial network metrics can be combined with other key 5G performance metrics to create composed metrics used to enhance the dynamic placement of services. A cognitive network management framework that monitors, calculates and aggregates in real time such metrics, can make use of them to perform autonomous decisions, such as dynamic VNF placements, load balancing, caching, network communication dynamics and VNF migration, among others.

In this research work, we present a new cognitive network management framework, focusing on topology monitoring based on new spatial metrics in 5G networks. There are several use cases that can benefit from our cognitive framework. For instance, composed spatial metrics can be used to infer the optimal placement for key 5G architectural components such as Centralized Unit (CU), Distributed Unit (DU) or User Plane Function (UPF) for providing support to different 5G use cases such as ultra Reliable, Low-Latency Communications (uRLLC) and enhanced Mobile Broad-Band (eMBB) where latency and throughput need to be minimized or maximized, respectively. Massive Machine-Type Communications (mMTC) refer to Internet of Things (IoT) networks where millions of 5G devices create complex topologies and where IoT services can be offloaded to the Mobile/Multi-access Edge Computing (MEC) platform, to perform computation offloading of the resource-constrained IoT devices to the network edges, thereby overcoming the resource constraints and achieving energy consumption reduction in the IoT devices. Spatial metrics can help determine the closest geographical areas to perform such offloading. Analogously, the optimal placement of caches and load balancers in the network will further enhance the effectiveness of such uses cases. In case the reader is interested, Kim et al. (2017) provide a comprehensive description of the 5G architecture and its components.

Recent research work such as Salva-Garcia et al. (2018) and Neves et al. (2016) rely on diverse performance metrics to trigger autonomous behaviours in the cognitive network management framework according to the current status of the network. These metrics encompass resource metrics (e.g., memory, disk, CPU), network metrics (e.g., throughput, latency, delay, capacity), wireless metrics (e.g., Received Signal Strength Indicator or RSSI, and Received Signal Code Power or RSCP) and service metrics (e.g., cost, revenue, request/second). However, 5G cognitive management frame-

works should consider not only those traditional performance and capacity metrics, as it has been studied so far, but also spatial network metrics that can empower the cognitive network management framework to make more meaningful management decisions by leveraging the spatial context-awareness in determining optimal operation locations. To the best of our knowledge, there is not yet any framework able to calculate or make use of 5G spatial network metrics combined with traditional ones to make valuable network management decisions. A cause that has contributed to this lack of support has been the complexity associated to the calculation of this type of metrics, which increases exponentially according to the size of the topology, and thus it is impractical, even for small topologies (hundreds of devices), to calculate these metrics in a useful time scale, i.e. seconds, to enable real-time cognitive network monitoring and management.

This research work demonstrates how spatial network metrics can be monitored and calculated for large-scale 5G topologies, and how they can be combined with traditional performance metrics to serve as an enabler for our 5G cognitive management framework, to make valuable decisions that optimize the allocation and management of 5G VNFs and traffic. Specifically, our new composed spatial metrics would allow higher 5G network resources utilization efficiency, reduced latency, increased availability, optimal VNF allocation, migrations of VNFs, balancing the network traffic, and balancing the computational load among edge nodes. Nonetheless, the computation of spatial metrics on 5G network topology with millions of resources (switches, physical machines, virtual machines, users, etc.), is challenging. To meet this challenge, we propose an efficient 5G topology monitoring agent (TMA) to quantify those new spatial metrics. This agent has been significantly optimized over the traditional ways to calculate spatial centrality metrics in graphs. Our TMA makes computationally feasible the on-demand quantification of combined 5G metrics in large topologies with hundreds of thousands of nodes.

The contributions of this paper are manifold:

- This paper defines a new 5G cognitive network management framework for continuous monitoring and calculating spatial metrics in 5G networks.
- The paper proposes new 5G-tailored, efficient and scalable spatial network metrics over evolving 5G network topologies to meet 5G multi-tenant and virtualized networking requirements.
- The paper proposes new composed metrics useful for the cognitive management framework to make allocations decisions such as cache allocation, load balancing, function computation offloading in 5G networks.
- Finally, the proposed approach has been successfully designed, implemented and deployed in a realistic 5G testbed. The paper demonstrates the feasibility and scal-

Table 1 Analysis of the state of the art with respect to our contribution against the challenges required to achieve Advanced Spatial Network Metrics for 5G Cognitive Network Management

	Exploits Spatial information	Applies closeness centrality	Metric based	Optimized for 5G topologies	ScalableSolution
Kourtellis et al. (2015)	✓	×	×	×	×
Dolev et al. (2010)	✓	×	✓	×	×
Kchiche and Kamoun (2010)	✓	✓	×	×	×
Wang et al. (2017)	✓	✓	×	×	×
Musumeci et al. (2016)	×	×	✓	✓	×
Baštuđ et al. (2014) and Baštuđ et al. (2016)	✓	×	×	×	×
Tulu et al. (2018)	✓	×	×	×	×
Li et al. (2017)	✓	×	✓	✓	×
Zhao et al. (2018)	✓	✓	×	×	×
Dourado et al. (2013)	✓	×	✓	×	×
Bajpai and Schönwälder (2015)	×	×	✓	×	×
Goel et al. (2015)	×	×	✓	×	×
Sheu et al. (2017)	✓	✓	×	×	×
Tizghadam and Leon-Garcia (2010)	✓	✓	×	×	×
Our contribution	✓	✓	✓	✓	✓

ability through empirical performance evaluation in this testbed.

The rest of this paper is structured as follows. Section 2 reviews related work. Section 3 introduces the 5G cognitive network management framework. Section 4 delves into the design of the novel 5G spatial metrics, followed by Sect. 5, which describes concrete use cases for the usage of the proposed metrics. Section 6 describes the most relevant implementation details of the proposed framework focusing on the TMA. Subsequently, Sect. 7 is devoted to the performance evaluation in a realistic 5G testbed. Finally, Sect. 8 concludes the paper highlighting the main achievements.

2 Related work

Table 1 provides a comparative analysis of the current state of the art against the challenges proposed in this research work and how our contribution compares with them to allow the reader to clearly identify our contribution. Each column in the table is a key enabler to achieve the proposed challenges.

Spatial network metrics such as Closeness Centrality proposed in Sabidussi (1966) and Betweenness Centrality in Freeman (1977) can identify influential nodes in a graph, but, as it is known, they are difficult to be applied in large-scale networks due to the computational complexity as reported by Chen et al. (2012), even more in evolved graphs such as those resultant of 5G monitored networks as the algorithm runs in $O(n^2)$ time.

Calculating centrality metrics over large graphs is a challenging problem. The problem of efficiency in centrality metrics calculation has been studied in the past and recently. Kourtellis et al. (2015) propose a scalable algorithm for betweenness centrality in large graphs using MapReduce that scales by adding more compute nodes. Even with this kind of parallelization optimization, it might take 10^5 s with a 100k graph size, which makes this unfeasible for 5G network real-time management. Therefore, in our approach, instead of betweenness centrality, we leverage closeness centrality metric that provides similar insights about network graphs, but with less complexity time $O(nm)$ to make practicable real-time and online network management.

In Dolev et al. (2010), the authors present a Routing Betweenness Centrality metric for communication networks, by considering network flows created by arbitrary loop-free routing strategies, where routing decisions depend on the packet target alone and considering the source-target of the packet.

In Kchiche and Kamoun (2010), the proposed scheme relies on centrality index to calculate the best placement of access points for vehicular networks, or to deploy Road Side Units (RSUs) proposed in Wang et al. (2017). However, in both cases, the results are based on simulation environments and they considered small network topologies (20–200 nodes) that are not actually affected by the complexity of spatial metrics.

Musumeci et al. (2016) propose an optimization algorithm of DU placement optimization in 5G for Cloud-based Radio Access Networks (C-RAN), considering diverse net-

work metrics such as traffic requests, fibre per link, fronthaul latency, network metrics, but they do not consider centrality metrics.

In Baştuğ et al. (2014) and Baştuğ et al. (2016), the authors analyze the spatial structure of networks to perform effectively a proactive caching in 5G small cells networks. Unfortunately, they do not deal with large-scale graphs and do not consider optimal centrality indexes to maintain up-to-date metrics in the evolving 5G graphs.

The work in Tulu et al. (2018) resorts to spatial network metrics to calculate influence nodes to serve as content-centric mobile 5G networks. However, again, they focus their experiments on small network graphs (85 nodes).

Li et al. (2017) propose a “Caching-as-a-Service” (CaaS) for 5G, either cloud-based radio access networks and virtualized Evolved Packet Core (EPC). They discuss the allocation performance evaluation according to diverse metrics, but they just mention the possible usage of shortest path betweenness centrality as a potential metric, without delving into details.

Recently, Zhao et al. (2018) study the optimal placement of cloudlets in IoT access points (MEC nodes) for access delay minimization in large scale IoT networks. Their fitness function considers several attributes, including the evaluation of closeness centrality of the access points. However, they have validated the approach in a simulated environment with only 1000 IoT devices and 40 access points.

Network metrics aggregation and composition are still not being broadly adopted in currently available monitoring tools, even less on inter-domain and complex communications networks. In this regard, some software implementations such as Dourado et al. (2013) allow metric composition considering intra-domain performance measurements in communications networks. They analyse metrics for quantifying end-to-end minimum-mean delay in spatial composition, but they do not consider spatial centrality metrics. Indeed, given the current state of the art surveyed in Bajpai and Schönwälder (2015) and Goel et al. (2015), there is not any implemented model or monitoring tool for computing aggregated spatial metrics for 5G networks, and they focus only on network metrics end-to-end latency, last-mile latency, latency-under-load, mean latency, end-to-end packet loss or jitter, upstream-downstream throughput-goodput, network availability, etc.

Similarly, the work in Sheu et al. (2017) leads to a routing algorithm for Software-Defined Networking (SDN), evaluated at small scale that considers Centrality in the weight functions to prevent the bottleneck in the multicast routing and choose the nodes or links that can balance the loads across the network.

Spatial metrics can be used to enrich the graphical representation of network topologies, thereby helping on human-centric network management, where administrators can directly govern the network resources and flows.

In Tizghadam and Leon-Garcia (2010), the authors discuss the usage of centrality indexes for control algorithms in communications networks, including allocation capacities, design and dynamic control of physical/virtual network topologies.

To the best of our knowledge, this is the first proposal of monitoring and calculate efficient and optimized spatial centrality metrics, used as a baseline for cognitive management at the scale of evolving 5G networks.

3 Cognitive 5G network management framework

Figure 1 introduces the 5G multi-tenant network deployment and the proposed cognitive management framework to manage such networks, according to the contextual information, including our proposed spatial metrics to be defined in Sect. 4. The following subsections provide details of the framework.

3.1 5G multi-tenant network

The bottom part of Fig. 1 depicts the deployment of a 5G multi-tenant network where both physical and virtual layers are presented. The physical layer is composed of the hardware available in the infrastructure. The antennas provide coverage to the 5G users to gain radio access to the network. These antennas are connected by means of the front-haul to a DU. A DU controls the new 5G radio interface, sending the radio signals through the mid-haul network segment to the edge of the network. The edge is the closest location where a VNF can be deployed close to the final users, in order to achieve computational off-loading and low-latency capabilities. Usually, a CU can be deployed in this network segment to do not insert delays in the control of the radio interface, to allow quick handovers. The edges of the network are connected by means of the back-haul to the core network where the main 5G VNF Core services are deployed. The UPF VNF is the mobile anchor where all the mobile traffic is received to get access to the Internet. In the control plane, An Access and Mobility Function (AMF) VNF provides mobile equipment authentication, authorization and mobility management. A Session Management Function (SMF) VNF is responsible for session management and allocates IP addresses to UE devices. It is noted how all the different 5G architectural components are virtualized with the only exception of the DU and that multiple instances of such components are instantiated inside the same physical machine in multi-tenancy scenarios. The Core network is finally connected to the Internet through the backbone network segment.

In production-ready deployments, the idea is to maximize the number of 5G subscribers that can be supported with

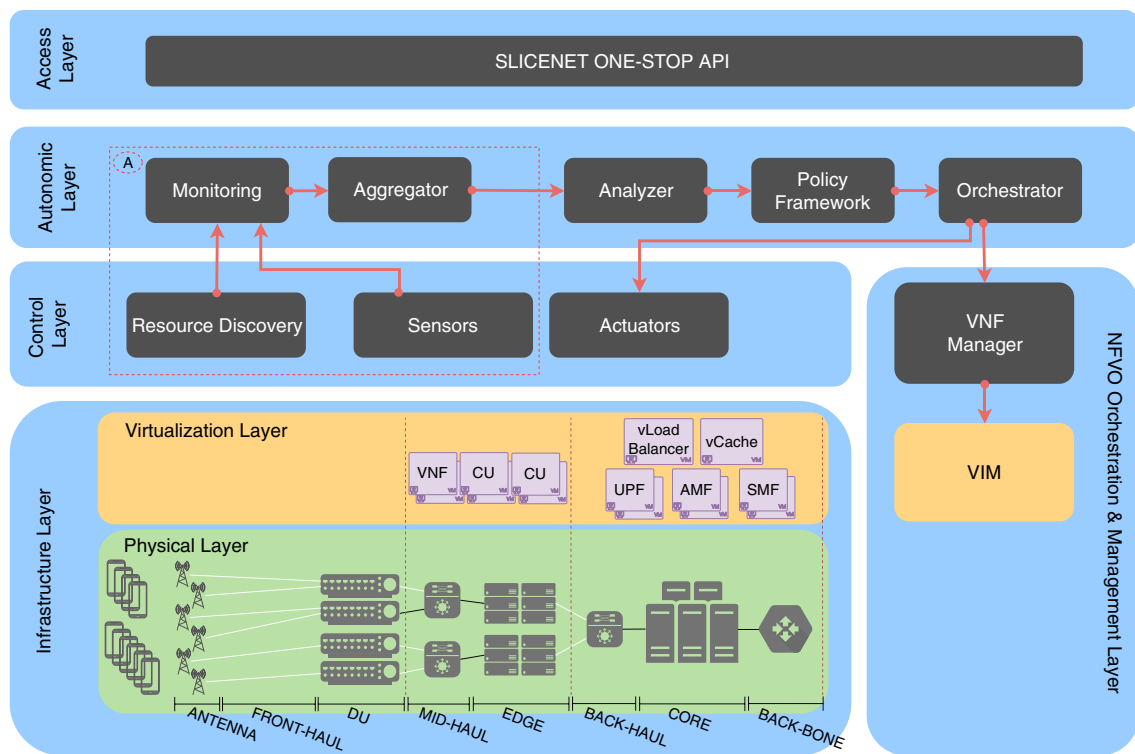


Fig. 1 5G cognitive management framework

the minimal set of hardware resources while providing the expected quality of service.

3.2 5G cognitive management framework

The upper part of Fig. 1 illustrates the 5G cognitive management framework proposed. The architecture has been aligned with the H2020 5G SliceNet project, which is a high-profile EU project on cognitive network management for 5G networks and whose consortium comprises 16 partners including leading telecommunication operators, vendors and other 5G stakeholders. The 5G multi-tenant network has been instrumented with sensors, actuators and discovery agents in order to enable the management of the infrastructure. They are depicted in the Control layer of Fig. 1. The sensors provide metrics to the management framework. The discovery agents provide topological information about the 5G multi-tenant infrastructure. The actuators are in charge of controlling resources, communications and services of the infrastructure.

The autonomic layer depicted in Fig. 1 contains all the components that provide cognitive management capabilities. First, the monitoring component collects both topological information and sensed metrics to make them homogeneous and processable over a common format. The aggregator component produces composed metrics that are defined as a temporal or spatial aggregation of the elemental metrics

sensed. The analyzer component is in charge of identifying alerts in the 5G network. An alert is an alarm that requires attention by the administrator. A sub-optimal device configuration, hardware failure or bad configuration in devices or services or an external attack are good examples of alerts. The analyzer correlates the information received from the monitoring component in order to determine existing or emerging issues. The policy framework component then performs decisions about actions to deal with such alerts. This decision is an implementation plan that indicates how to overcome the issues expressed in the alerts. The orchestrator component is in charge of executing, step by step, in an orderly manner the plan provided by the policy framework. This plan is executed by interacting with actuators and VNF Management. The architecture is fully aligned with the ETSI NFV Management and Orchestration (MANO) architecture. The VNF Manager deals with the life cycle and controls services, while the Virtual Infrastructure Manager (VIM) manages the life cycle of virtual resources in a multi-tenant environment.

This architecture enables a closed control loop where the system can react autonomously through executing a mitigation plan against several alerts that have been defined in the system without any human intervention. Finally, the SliceNet one-stop Application Program Interface (API) allows onboarding new autonomic capabilities and controlling the exposure of information to the final users about the status of the network, among others.

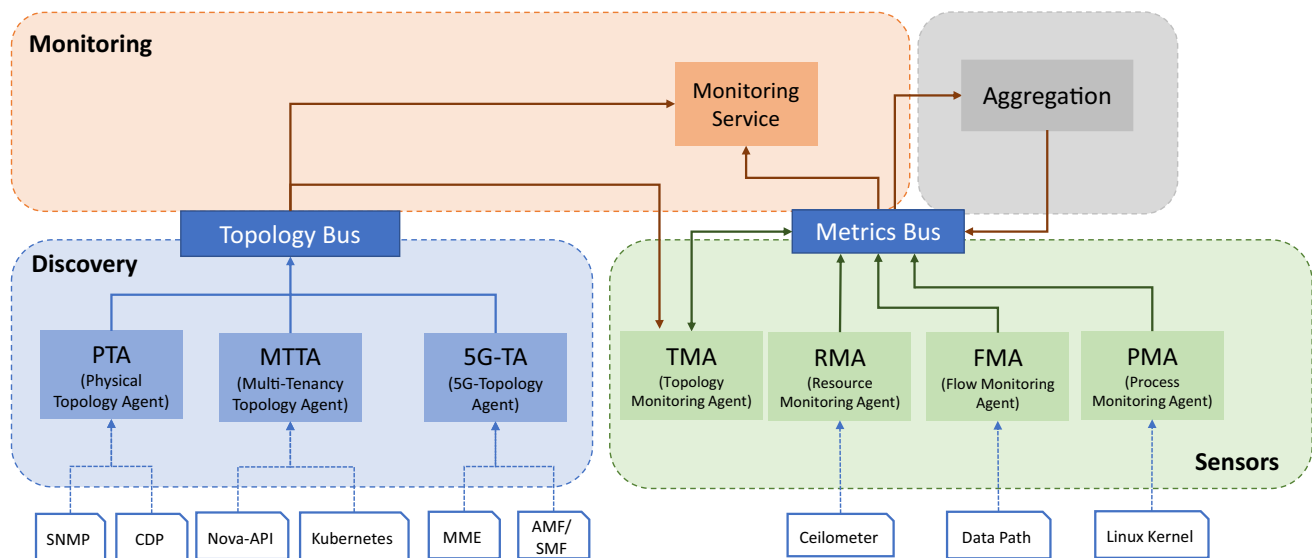


Fig. 2 Continuous 5G monitoring and discovery framework

3.3 Monitoring and discovery of 5G network topologies

Figure 2 shows a detailed description of the part of the cognitive management plane dealing with monitoring and topology discovery. This is a zoom of the components under the dotted square labelled as A in Fig. 1. On the topology discovery side, our architecture proposes the following three components to provide updated information about the topology.

- The *Physical Topology Agent (PTA)* relies on traditional technologies such as the Simple Network Management Protocol (SNMP) to provide information about the physical topology of the 5G network. There is only one PTA deployed in the infrastructure. This component performs a periodic SNMP walk from different entry points defined by the administrator. This SNMP walk interrogates recursively the status of the LLDP learned neighbours of each node and then their neighbours.
- The *Multi-Tenant Topology Agent (MTTA)* utilizes the cloud stack APIs to retrieve information about the virtual topologies available in the 5G network and about the tenants that own such topologies. To this end, some key management APIs such as OpenStack and Kubernetes have been integrated. It retrieves periodically the list of containers and virtual machine allocated in each physical machine in order to generate the virtual topological information. There is only one MTTA deployed in the infrastructure.
- The *5G-Topology Agent (5G-TA)* provides information about the attachment of the 5G mobile users to the specific 5G DUs and keeps tracking the user mobility across antennas. To achieve this, it has been integrated with the

control plane of both 4G and 5G networks by retrieving such information from the 4G Mobility Management Element (MME) and 5G AMF/SMF, respectively. To be concrete, this integration has been performed by creating an ad-hoc API in the MME implementation of the Mosaic 5G project as described by Nikaein et al. (2018). Marco Alaez et al. (2017) provide a comprehensive explanation of this API.

These three monitoring components pull topological information from the infrastructure at periodic intervals. This periodic interrogation is used to maintain in memory the updated network topology. Then, they only report the topological information when there are changes with respect to the previously reported state to deliver only the topological changes. The three components report the topological information using a common topological model to a common message bus. The key is the separation of responsibilities of each of the components with respect to the devices being reported by them. To be concrete, the PTA reports on physical devices and their connections, including switches, routers, and DUs. The MTTA reports on virtual machines and containers and their connections with the physical machines. And finally, the 5G-TA reports on UE devices and their connection to the DUs.

On the sensing side, our architecture employs four components. Firstly, the following three provide performance metrics about resources, flows and operating system processes (services) respectively.

- The *Resource Monitoring Agent (RMA)* reports metrics about the following: i) the physical device where it is installed, such as CPU Usage, Memory Size, Memory

I/O, Disk Size, Disk I/O, Cache faults, and kWatt/Hour; ii) the virtual machines allocated in such physical machine, such as Memory I/O, Disk I/O, and Memory Used; iii) both physical and virtual network interfaces, such as Bandwidth Consumed, Bandwidth Capacity or Bandwidth Negotiated. This component is installed in each physical machine of the 5G infrastructure.

- The *Flow Monitoring Agent (FMA)* reports metrics about the flows of the interfaces that are being monitored. These metrics include Experienced Data Rate (Downlink), Experienced Data Rate (Uplink), Area Traffic Capacity (Downlink), Area Traffic Capacity (Uplink), Overall User Density and End-to-End Latency. This component is installed in each physical machine of the 5G infrastructure.
- The *Process Monitoring Agent (PMA)* is installed in each physical machine of the 5G infrastructure. It reports metrics about each of the services/processes that are running in such machines. These metrics include process I/O, memory I/O, disk I/O, context switches, sleep time, idle time, and so on.

In addition, a *Topology Monitoring Agent (TMA)* is proposed as the cornerstone component of the monitoring framework. It is in charge of computing in real time the 5G spatial metrics (defined later in the next section, Sect. 4), according to both sensing and topological information. Unlike the rest of the sensor's components, the TMA receives both topological information from all the topology discovery sensors and metrics from the other sensing components as inputs. The TMA keeps the 5G topology updated in a graph structure using the add/remove primitives described in the next section. The TMA performs periodic calculation of the 5G spatial metrics. The calculation of the spatial metrics is detailed in the next section and it may require the usage of performance metrics, and this is the reason why this component also receives such information. The TMA computes dynamically diverse metrics over the 5G network, which is represented as an evolving graph. The graph changes continuously according to the reports coming from the topology discovery sensors. The framework computes continuously the metrics for each node, as the nodes and resources evolve along the time. Then, it provides periodic reports of the spatial metrics associated with such nodes of the 5G topology. Those metrics are then exploited by the cognitive layer of the framework to make advanced context-aware decisions for 5G network management.

Both the spatial and performance metrics are received by the aggregator, which is able to compute composed metrics based on spatial or temporal aggregation of the elemental metrics. Thus, the aggregator combines diverse network metrics, resource metrics, and topology information, along with spatial metrics, to come up with complex 5G spatial network

metrics. Such composed metrics are also reported back again to allow the cognitive system to utilize them.

4 Spatial 5G closeness centrality network metric

All the segments of the 5G network including Radio Access Network (RAN), Edge, Core network as well as mobile users are continuously monitored by the topology discovery sensors, which allow modelling dynamically the 5G network as a connected directed graph $G = (N, E)$, consisting of a set N nodes or vertices, corresponding to the 5G network devices, and set E of edges corresponding to the network connections between devices.

Our graph G changes continuously as the 5G network evolves, considering physical devices holding virtual network functions (VNFs), and their connections. The network links are directed edges with associated metrics, which are calculated continuously and reported by the sensing components per node, such as for instance, latency $l(v, v')$, bandwidth utilized $b(v, v')$ and bandwidth available $ba(v, v')$.

Figure 3 represents graphically the graph of the 5G network topology, where nodes N are represented in circles and edges E are represented in lines. As explained in Sect. 3.1, UE represents the users mobile terminal, DU and CU represent Distributed Units and Centralized Units, respectively. It is noted that some nodes have a “v” prefix, representing that this is a virtual machine where such a service is running. The VNFs are the set of virtual network functions deployed in the 5G network to perform computational off-loading to the edges. The vSMF, vAMF, vPCF, vUDM, vAUSF are all the architectural components of the 5G network running in virtualized machines. In the figure, nodes with heptagonal shape represent Physical Machines P located in both the Edge and the Core Networks that are potential candidates for deploying virtual machines V for implementing virtual network functions. Thus, $V = \{v \in N \mid v.type == VirtualMachine\}$ and $P = \{p \in N \mid p.type == PhysicalMachine\}$.

Spatial networks metrics rely on networks graphs, obtained dynamically from the topology discovery agents to calculate meaningful information that is ultimately used by the cognitive network management framework to make wise control decisions to manage the 5G network. However, the current state of the art in sensing tools, topology discovery tools, and inventory systems are not yet capable of providing spatial network metrics of 5G multi-tenant infrastructure. So far, 4G networks did not have to deal with virtualization of the network functions and the multi-tenancy capabilities needed to enable concurrent operations over the 5G infrastructure. These 5G characteristics combined with the dense deployments foreseen in 5G networks where thousands of UEs can be connected to the same DU lead to an explosion of the num-

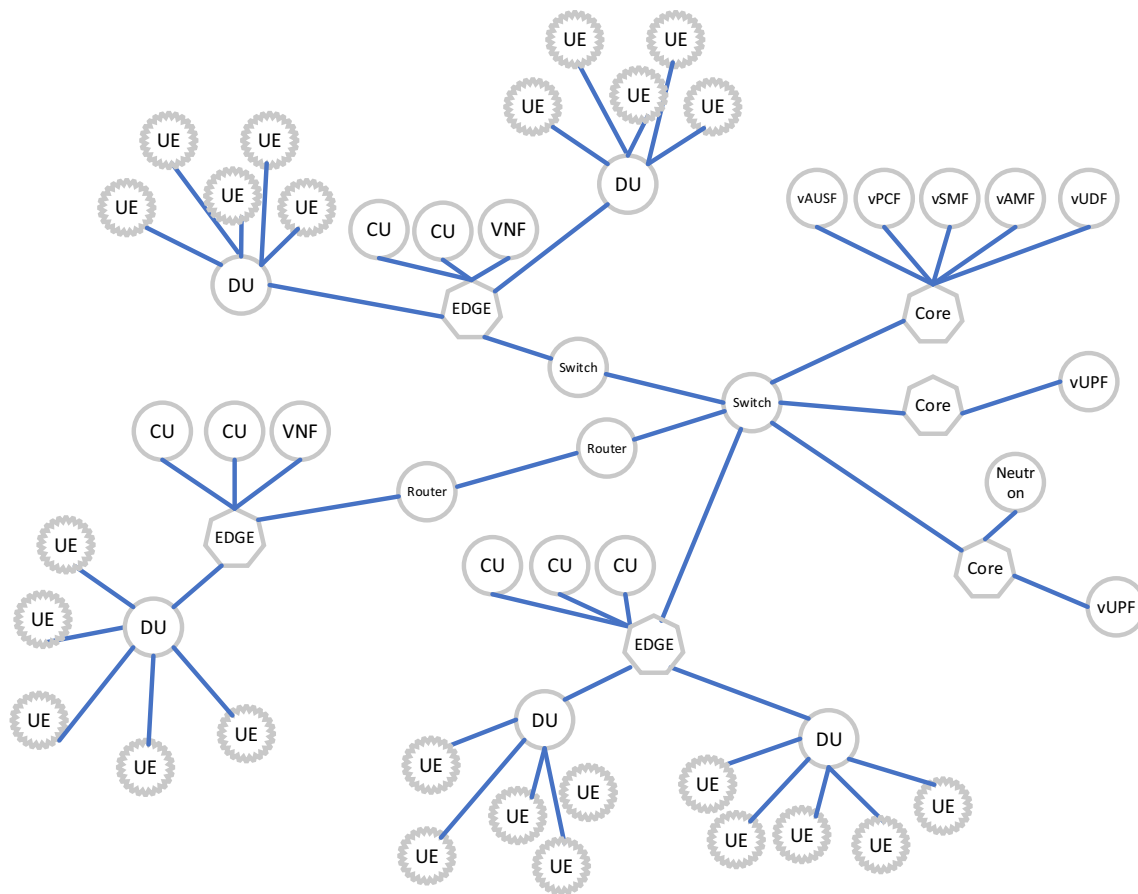


Fig. 3 5G network graph representation

ber of nodes (physical/virtualized and 5G users) and their connections in the network topology.

It can result in topologies with hundreds of thousands of Vertices and Edges, which makes the management of 5G network topology graphs impracticable, above all when computing those metrics in real time and considering the whole 5G Network topology.

Certain network management decisions require the whole network picture to make valuable decisions, for instance, allocation of caches functions in a proper MEC node depends on the number of connected nodes (including users) and their spatial locations and demands. The load balancing function allocation is also influenced by the combination of the current flows, and spatial network representation, since such a network function should be allocated on-demand wisely in central and strategic locations to perform balancing to the network traffic wherever and whenever is needed.

4.1 Tailored spatial metric for multi-tenant 5G networks

Our tailored spatial metric for multi-tenant 5G network features several optimizations over traditional spatial metrics

based on centrality. Such optimizations make the calculation of the metrics treatable and computationally feasible for real-time management of the 5G network.

$$CC(v) = \frac{1}{\sum_{n \in N} d_G(v, n)} \quad (1)$$

Our proposal extends and adapts the Closeness Centrality metric by Sabidussi (1966), defined as how close a node is to all the other nodes of the network. Nodes are more central as they are closer to the majority of the other nodes. Closeness Centrality ($CC(v)$) is mathematically defined in Eq. 1, where $d_G(v, n)$ is the function that calculates the distance of the shortest path from node v to node n in the Graph G . This distance is the minimum length of any path between both nodes in the graph. Being the path the sequence of vertices and edges beginning in v and ending in n , where each edge connects its previous with its subsequent vertex. The distance is the sum of the weights of its edges, where the weight measures the strength of a link. In our case, the weight will be embodied by additional 5G network performance metrics calculated by our sensing components, such as latency or bandwidth, depending on the specific

use case being addressed (e.g. cache allocation, load balancing, computation offloading, etc., as shown in the next section). The distance function requires an algorithm supporting weighted graphs able to calculate the shortest-path to other vertices, such as Dijkstra’s algorithm, which runs in time $O(m + n \log n)$ (being m the number of edges and n the number of nodes) and as a result, the complexity of traditional Closeness Centrality (Sabidussi 1966) is defined as $O(nm + n^2 \log n)$.

Our newly proposed 5G closeness centrality (5GCC) metric, presented in the following subsections, performs two main optimizations over the traditional way of calculating closeness centrality in a whole graph: node selection optimization which defined the set x as a subset of N as nodes to calculate the metric ($O(xm + xn \log n)$ where $x < n$). This complexity is further reduced by the node prune optimization which decreases the set of destination nodes y , defined as a significantly reduced subset of N , leading to a decreased complexity when calculating the shortest path defined as $O(xm + xy \log y)$ where $y < n$.

4.1.1 Node selection optimization

This optimization aims to create a subset of nodes $P = \{p \in N \mid p.type == PhysicalServer\}$ which includes only those nodes that are subject to serve as possible candidates v to allocate VNFs, i.e. only physical nodes p in the graph $G(N, E)$ will be considered for calculating their $CC(v)$. It significantly reduces the number of nodes v required to calculate the $5GCC(v)$, and therefore the number of pairs to which quantifying the distance functions $d_G(v, n)$, and consequently the weighted shortest-paths. It should be noted that reducing the number of nodes needed for calculating the $5GCC(v)$ does not imply reducing the number of target vertexes n of the distance function $d_G(v, n)$ that will still consider the whole Graph $G(N, E)$. Therefore, it requires calculating the shortest path between the candidate physical node v to every single node n in the graph (either physical, virtual or mobile-UE nodes).

4.1.2 Node prune optimization

The second optimization is the pruning of the leaves nodes of the 5G graph, i.e. the user mobiles attached to DU nodes while maintaining the same resultant closeness centrality index values when computing $CC(v)$ for each $n \in P$. This optimization reduces significantly the number of nodes in the graph, as the mobiles represent the vast majority of nodes in the 5G network topology. It minimizes the time needed to quantify the distance function $d_G(v, n)$ (which implies quantifying the shortest-path).

$$D = \{d \in N \mid d.type \neq Ue \text{ AND } d.type \neq Du\},$$

$$R = \{r \in N \mid r.type == Du\}$$

$$5GCC(v) = \frac{1}{\sum_{d \in D} d_G(v, d) + \sum_{r \in R} 5Gd_G(v, r)} \quad (2)$$

Thus, our 5G closeness centrality metric, defined in Eq. 2, splits the closeness centrality quantification in two spaces. For those 5G nodes n in set D , which are not part of UE mobiles or DU sets, it computes the distance function as it is done traditionally in $CC(v)$ metric, i.e. using function $d_G(v, n)$, whereas for those nodes r of type DU (set R), the distance is calculated with our new function $5Gd_G(v, r)$ defined in Algorithm 1.

Function $5Gd_G(v, r)$ requires two structures that are kept in cache memory. The first one $UEInDu(DU)$ holds up-to-date the actual UE mobiles belonging to each DU. The second one $UEsWeightInDu(DU)$ maintains the sum of the weights for each edge connected to the DU. These two structures are needed to allow the system to get the shortest path to every UE. Basically, when the shortest path to a DU is calculated from a node n , the shortest path from this node n to every UE connected to this DU is the same plus the additional weight for every UE connected to the DU. These structures are managed by our topology manager that performs the network graph management and quantifies the metrics, adding and removing continuously the nodes in these structures and in the graph, as the 5G network evolves.

```

Data:  $G(N, E), UEInDu(DU), UEsWeightInDu(DU)$ 
1 Function  $5Gd_G(v, r)$ :
2    $out \leftarrow 0;$ 
3    $duPathWeight \leftarrow d_G(v, r);$ 
4    $mobilesPath \leftarrow$ 
    $duPathWeight * UEInDu[r] + UEsWeightInDu[r];$ 
5    $out \leftarrow out + duPathWeight + mobilesPath;$ 
6   return  $out;$ 
7 End Function
    
```

Algorithm 1: 5G distance function

As can be seen the function $5Gd_G(v, r)$ defined in Algorithm 1, adds to the total summation of $d_G(v, r)$, the weights from the DU towards each UE connected, taking those weights from the $UEsWeightInDu$ structure.

As a result, our 5GCC metric, using as baseline $CC(v)$, and after applying the optimizations, is defined in Algorithm 2. This algorithm requires access to the whole graph $G(N, E)$. Also, it requires access to a set of Physical Servers (PS) containing the candidate nodes selected to calculate their 5G Closeness Centrality. The set of all DUs used in the calculation of the shortest path for the optimized 5GCC. UE set comprises the UEs available in the graph. And finally, a set D containing the rest of destination nodes in the graph for which the shortest path to them must be calculated. The algorithm returns the vector with the 5G Closeness Centrality

values for each node that is worth calculating the 5G spatial metrics, i.e. those physical nodes that might allocate a VNF.

```

Data:  $G(N, E)$ ,
         $PS = \{ps \in N \mid ps.type == PhysicalServer\}$ ,
         $DU = \{du \in N \mid du.type == Du\}$ ,
         $UE = \{ue \in N \mid ue.type == Ue\}$ ,
         $D = \{d \in N \mid d \notin UE \text{ AND } d \notin DU\}$ 
Result:  $5GCC[]$ 
1 for  $ps \in PS$  do
2    $out \leftarrow 0$ ;
3   for  $du \in DU$  do
4      $out \leftarrow out + 5Gd_G(ps, du)$ ;
5   end
6   for  $d \in D$  do
7      $out \leftarrow out + d_G(ps, d)$ ;
8   end
9    $5GCC[ps] \leftarrow \frac{1}{out}$ ;
10 end
11 return  $5GCC[]$ ;

```

Algorithm 2: 5G closeness centrality algorithm

4.2 5G network graph management

5G networks are subject to continuous topology changes that require autonomous and dynamic management of the associated evolving network graph. The topology changes are motivated by different factors. On one hand, due to 5G mobility, users are continuously performing handovers from one DU to another. Besides, UE devices are continuously connecting/disconnecting from the network, e.g. user devices restart or simply lose their coverage. These kind of changes are minimal for the topology but sufficient to force a recalculation of the closeness centrality in the whole graph. On the other hand, the virtualization and on-demand provisioning of VNFs lead to continuous and dynamic changes in the network topology beyond the leaves nodes in the graph that requires recalculating our 5GCC indexes, as this quantification depends on the overall status of the graph. This section defines the procedure followed by our topology manager to maintain in real time the graph and associated structures, as well as calculate efficiently the 5GCC metrics.

4.2.1 Add and remove nodes/edges in 5G graph

The 5GCC is calculated by the TMA of the framework, updating the 5G graph according to those dynamic changes, and recalculating the 5GCC for each node in the graph. To this aim, it relies on two main functions defined in Algorithm 3.

The *addEdge()* function updates the graph when a new connection whose source (or destination) is a physical or virtual 5G node. However, when the change is minor, i.e. a connection related to a UE node has been

added, it only updates the cache structures *UEInDu* and *UEsWeightInDu*. The *addNode()* function basically adds the node in the graph, but only if it is not UE, thereby avoiding changes that might trigger a recalculation of the 5GCC in the whole 5G graph. The *removeNode()* and *removeEdge()* functions are intended to detach the UE from a DU in the event of a handover or disconnection, and they are defined in a similar way as the add ones, but subtracting the weights from the *UEsWeightInDu* and removing the node and edges whenever necessary in the graph (i.e. if it is not UE). When an *addEdge* implies any node beyond UE devices, then it triggers the execution of Algorithm 2.

```

Data:  $G(N, E)$ ,  $UEInDu(DU)$ ,  $UEsWeightInDu(DU)$ 
1 Function addEdge ( $e$ ) :
2   if ( $e.getSource() \neq UE$ ) then
3      $UEInDu(e.getDest()) \leftarrow$ 
4        $UEInDu(e.getDest()) + 1$ ;
5      $UEsWeightInDu(e.getDest()) \leftarrow$ 
6        $UEsWeightInDu(e.getDest()) + e.weight$ ;
7   else
8      $G(N, E) \leftarrow G(N, E) + e$ ;
9   end
10 End Function
11 Function addNode ( $n$ ) :
12   if ( $n.type \neq UE$ ) then
13      $G(N, E) \leftarrow G(N, E) + n$ ;
14   end
15 End Function

```

Algorithm 3: Graph Management. Add nodes/edges functions

4.3 5GCC retrieval optimization

This optimization is intended to minimize the number of times needed to recalculate the entire 5GCC when a topology change is produced in the 5G network. To this aim, we exploit the memory structures presented previously so that when there is only a minor change i.e. a variation related to UE nodes, the 5GCC index for the rest of the nodes does not need to be recalculated entirely as it is traditionally done, which avoids executing fully Algorithm 2.

Thus this optimization avoids computing the shortest-path $d_G(ps, d)$ each time a UE node is added to the graph, reducing to the minimum the complexity to keep up-to-date the 5GCC metric in the entire 5G network topology. Thus, for these cases, which might represent more than 95% of changes in the 5G network, Algorithm 2 is given as input the set D empty, meaning the loop of lines number 6–8, which calculates distance function $d_G(ps, d)$ are not executed. In addition, for this case, we have defined another 5G distance function $5Gd'_G(v, r)$ similar to the one defined in Algorithm 1, but slightly changed, without code of line num-

ber 3, meaning the distance function $d_G(v, r)$ is obtained from another structure *DistanceVR* kept in memory-cache that holds the distance values for each pair physical server and DU. The rest of the function description is kept, where lines 4–5 in Algorithm 1 update the weights, and therefore the resultant 5GCC, according to the added/removed UE in the graph. It is noted that when this function is executed, the *UESWeightInDu* structure is up-to-date since the add/remove functions managed by the TMA have already updated those structures.

4.4 Network function placement optimization based on 5G-spatial metrics

When a service request reaches the cognitive management framework, the policy framework component must decide how to allocate optimally certain VNFs in the substrate physical nodes, depending on the service request necessities and the current status of the entire 5G management network, including physical and virtual appliances.

$$\forall p \in P, \forall r \in R_v : \sum_{v \in V_p} C_r^d(v) \leq C_r^a(p) - C_r^s(v') \quad (3)$$

We formulate the placement problem as an optimization procedure that starts firstly by checking hard constraints, either, placement constraint (e.g. nodes resource restrictions and current status in terms of computing, memory, etc.), as well as network constraints (e.g. resource's connections bandwidth, latency constraints, or flows). This hard constraints phase ends up with a subset of potential candidates to allocate the VNF, as defined in Eq. 3, where V_p denotes the subset of V that are already allocated in a physical p machine and R_v is the set of resource requirements for a given node v . Thus, an $r \in R$, might refer, for instance, to resource storage requirement or a memory requirement of the VNF v .

$C_r^d(v)$ refers to the constraint function that gets the requirement value r , deployed d for a given virtual node v . Similarly, $C_r^a(p)$ denotes the constraint function value for that requirement r , currently available a in a given physical node p . Similarly, $C_r^s(v)$ denotes the constraint function value of the requested of the requirement r , of the virtual service s to be allocated, in a given v' virtualized node.

When the inequality is met $\forall r \in R_v$, the node p satisfies the constrains requirements and becomes potential candidate to allocate the VNF in the node v . Thus, the set $PS = \{p_1..p_n\}$ that satisfies the requirements is used as input for Algorithm 2, which calculates the 5GCC for each $p \in PS$.

The objective function for the allocation problem in those candidates depends on the use case being faced by the cognitive framework. In the case of the cache allocation, the objective function minimizes the latency, selecting less loaded physical edge in the graph, which is closer to the final

UE terminals, and at the same time, is central to the maximum number of UE terminals to maximize the number of cache hits.

For this second phase, the 5G spatial metric defined above, combined with the actual network conditions, such as consumed bandwidth or latency retrieved from the FMA, RMA and PMA components of our framework, can help to determine optimally those allocations. The following section delves into complex and novel spatial metrics that combine both our 5GCC metric and other network metrics to address the particular objective of certain 5G use cases.

5 Spatial-based network management: 5G use cases

This section outlines a number of representative technical 5G use cases, where the proposed spatial metric can be further augmented to create new composed metrics that meet the specific requirements of the services for the purpose of smart 5G network monitoring/management.

5.1 Proactive cache allocation for 5G ultra-reliable and low-latency communication (URLLC)

Content distribution networks (CDN) are composed of data centre and a set of cache proxy servers deployed in strategic geographic locations to maximize the availability and performance in the network while reducing the latency of the content delivery applications such as video streaming and software downloads. The CDNs can be deployed over the 5G networks, exploiting the capabilities of the 5G infrastructure, thereby deploying on-demand the proxy caches as VNFs in Edge nodes.

The proper place to allocate those virtual caches (vCaches) varies dynamically as the network topology changes due to the user's mobility, and the geographic video demands, among others. Spatial metrics are perfect allies in this scenario to quantify the most influential physical edges to increase the content cache hits, and they can be combined with the actual end-to-end latency metrics to quantify locations with the lowest latency from the end-user to the cache through diverse network hops.

Thus, the proposed 5G Closeness Centrality metric, as defined in Sect. 4 and calculated by the TMA component of our framework, can be weighted with the latency measurements calculated in real time by the RMA component, resulting in a new metric called *5G-Cache Closeness Centrality*, which computes the indexes for each of the best physical nodes in the network topology represented in the graph to deploy a vCache VNF, thereby minimizing the delay of the user towards the cache.

Thus, the resultant metric is directly influenced by the network graph, the number of mobile users attached to the set of DUs connected to the Edge, along with the physical network conditions and the actual network conditions, network latency from the UE towards the edge node that deploys the vCache. The low mean UE-Edge latency values assigned to the Edge nodes will prioritize them in the spatial metric against physical machines in the core.

5.2 Computation offloading in mMTC

The *5G-Cache Closeness Centrality* metric can also be employed in the MEC scenario to assist in the Computation offloading in mMTC.

The metric can be employed to find out the best Edge node in the 5G network to allocate certain IoT computing tasks, performing computation offloading of IoT devices towards those nodes. Distributed MEC nodes (e.g. Fog Nodes as VNFs) allows reducing the data traffic at the Core network.

In addition to 5GCC and latency, the decision-making considers other indicators such as the number of available resources in the MEC node in terms of hardware/software capacities, during the constraint satisfaction defined in 4.4.

5.3 Node over-provisioning

Network operators need to quantify continuously network dimensions to satisfy committed network service guarantees specified in the Service-Level Agreements (SLAs). Over-provisioning considers, on one hand, the contracted SLAs e.g., in terms of bandwidth and latency, and, on the other hand, what is currently demanded in real time, considering the network's topology, and on-going measurements and metrics such as the consumed bandwidth.

The degree centrality metric is the simplest centrality index and is defined as the number of connections a node has in a graph. The degree centrality can be calculated using weights, and concretely for this case, the weights can refer to the bandwidth consumed in each of the network interfaces (connections) by a node.

$E_v =$ Edges of V

$$\text{OPC}(v) = \text{Capacity}_b(v) - \sum_{e \in E_v} \text{Bandwidth}_c(e) \quad (4)$$

Our proposed Node Over-provisioning Centrality $\text{OPC}(v)$ metric, combines the degree centrality weighted with the current bandwidth consumed. The resultant metric is defined in Eq. 4. This metric allows the operator to identify proactively those nodes that are near to be overloaded and could lead to a breach in the contracted SLA.

According to this metric's values, when the node is about to be overloaded $\text{OPC}(v) < \text{threshold}$ under a certain

threshold, the *planner* module of our cognitive management framework is notified, and it can infer certain actions to mitigate the over-provisioning, such as disconnecting certain UE terminals from the DU to force a reconnection to a less overloaded DU (it is noted that the RMA has means to monitor the on-going DU's reachability by the UE). In another example, if the $\text{OPC}(v)$ affects a physical node in the Core network, the *planner* can migrate a particular VNF to another physical node with a higher $\text{OPC}(v)$ value.

5.4 Optimal location of load balancers

Load balancers address scalability in distributed systems by balancing the workload among different servers. The optimal allocation of such VNFs can have a significant impact on the global performance of the network. Different criteria can be used to perform load balancing. It can be done based on real-time bandwidth, latency, the number of requests, etc. Let us consider bandwidth since it is widely adopted nowadays to take load balancing decisions. Our 5GCC can be combined with such metrics to determine the critical point in the network where more bandwidth is concentrated and, at the same time, where this selected node is more in the middle of the paths between servers and clients. To achieve this, $d_G^b(s, d)$ is defined as the shortest path weighted by bandwidth between the source (s) and the destination (d). Thus, replacing this function in equation 2, we can define the metric for the Optimal Location of Load Balancers (OLLB).

Therefore, human dynamics such as sleeping, working, relaxing, and moving to different locations will have an impact on the spatial aggregation of the real-time consumed bandwidths along the time, causing a recalculation of the OLLB. This OLLB metric can then be utilized to determine the location where to migrate or deploy a load balancer in real time to keep optimum the placement to maximize the balancing of the workloads in the network. OLLB considers available resources in MEC nodes, thereby it allows balancing the load among MEC and Core virtual and physical network nodes.

6 Implementation details

The proposed cognitive network management framework has been prototyped by focusing on implementing the core contribution of this paper, i.e., the TMA, the 5G spatial metrics and the enabling and optimization algorithms, all in a realistic 5G infrastructure testbed.

Specifically, the TMA has been implemented in Java 8 using RabbitMQ v3.7.16¹ as a message broker to receive both topology information and performance metrics, and

¹ RabbitMQ is available at <https://www.rabbitmq.com/>.

also to publish the calculated spatial metrics into the publication/subscription message bus. The TMA uses as an in-memory graph database, Apache Tinkerpop v3.4.2 backend.²

The TMA has been implemented using three different execution threads. Firstly, a reception thread is in charge of receiving information from both topology and metrics and updating the graph database accordingly. This thread maintains the cache structures required for the calculation of our 5GCC metric. Secondly, a calculation thread is running at periodic intervals and it performs the update of the values of the spatial network metrics for all the devices of the graph. To do so, the definition of the metric is implemented in the graph engine using the Gremlin query language. Gremlin is a functional, data-flow language that enables users to express queries over their graph. The execution of the Gremlin query makes use of the cached structure or triggers the complete recalculation of the metrics depending on the status of the cache. To allow this optimization to happen, we have combined Java code with Gremlin queries in order to create the optimized iteration of the graph. As a result, this thread keeps in memory the updated values of the spatial metrics. The third and last thread is the sender, in charge of sending periodically the metrics of the devices that have changed with respect to previous values (due to a recalculation). This is a way to optimize the overhead of the cognitive management framework which needs to deal with several metrics for thousands of devices and thus it improves scalability.

The TMA allows on-boarding different metrics from a configuration file where the metrics are defined based on their gremlin query associated. Each metric runs in a different calculation thread. The node selection optimization has been implemented directly as a filter of the Gremlin language whereas the node prune optimization has been implemented by a wrapper of the management functions of the graph (add/remove vertex/edge) to decide if the node needs to be added to or removed from the graph and if the cache structure needs to be updated. The implementation of the algorithm matches exactly the approach presented in the previous section. For the cache, the TMA has implemented the *UEInDu* structure as a *HashMap<String,Set<String>>* and the *UEsWeigthInDu* structure as a *HashMap <String, Long>* as a way to make the most efficient acceleration in the calculation of the spatial metrics.

7 Empirical validation

All the empirical executions performed in this research work have been carried out in a Cyberserve XE5-308S v4 computer, with Dual E5-2660 v4 Intel Xeon, 14 Cores, 2.00 GHz,

35M Cache, 105 W with hyperthreading activated, 128GiB DIMM DDR4 Synchronous 2400 MHz, 1.6 TB Intel SSD PCIe, and Ubuntu 18.04.2 LTS. That computer runs the TMA software component for the calculation of the spatial metrics. Our 5G infrastructure is composed of nine Cyberserve computers with the same specifications indicated. They run OpenStack Newton³ where one computer is the cloud controller and the other eight are computes employed for a 5G mobile edge computing network. These computes have been deployed inside the Mosaic 5G stack.⁴ The infrastructure is deployed with Ethus USRP X310 as DU. All the switches of the infrastructure are Netgear GS724T at 1 Gbps ethernet. Three different testbeds have been designed and executed to validate the suitability of the approach proposed. They are explained in the following subsections together with the results achieved.

7.1 5G topology scalability results

Firstly, a scalability test has been conducted with the main aim to show how the proposed metrics can be calculated on large scale infrastructures. To emulate large-scale infrastructures with up to 1 million UE devices, we have collected our own topological information of the 40 UE devices and 3 DUs we have in our premises and then created an augmented dataset for different sizes and topology shapes. The dataset created triggers real 5G events such as adding or removal of nodes/edges in the network. For this purpose, we use exactly the same APIs as those used in our 5G deployment in production in order to make sure we gather the same results in terms of performance as if running in such large-scale infrastructure. Our 5G deployment matches that described in Sect. 3.2.

The different topologies analyzed are depicted in Table 2 where the number of UE devices connected to a DU is ranged exponentially from 1 to 1024, the number of physical machines in the edge of the network is ranged exponentially from 1 to 256, and the number of physical machines in the core of the network is ranged from 16 to 64. The rest of the values have been kept constant and set to realistic numbers. Four DUs have been allocated per edge physical machine and Eight tenants share each of these physical machines in both edge and core. In 5G operations, it implies that Eight different telecommunication operators share the same edge, which is a very realistic scenario. For instance, Edotco which is a tower company owned by Asian telecommunications group has at least six subsidiary mobile operators sharing the same infrastructure (GSMA 2019). Also, it is expected an increment on the number of tenants sharing the same edges due to the accessibility of 5G networks for start-ups network companies (Gabriel 2020). In addition, it has been decided to

² Apache TinkerPop is available at <http://tinkerpop.apache.org/>.

³ OpenStack is available at <https://www.openstack.org/>.

⁴ Mosaic 5G is available at <http://mosaic-5g.io/>.

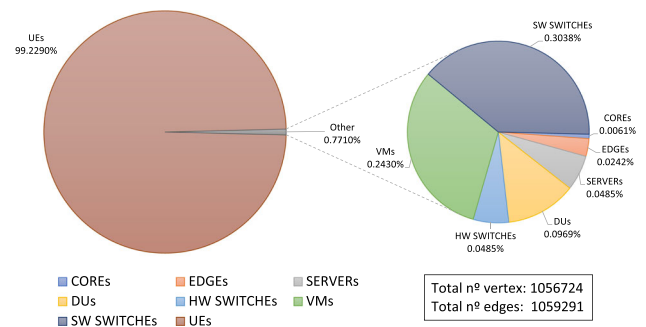
Table 2 Ranging of generated scenarios for empirical validation experiments

Property	Min value	Max value	Function
Number of $UE \times DU$	1	1024	$\sum_{i=0}^{10} 2^i$
Number of $DU \times Edge$	4	4	$CONST(4)$
Number of $Tenants \times Edge$	8	8	$CONST(8)$
Number of $Edge$	1	256	$\sum_{i=0}^8 2^i$
Number of $Servers \times Tenant$	1	1	$CONST(1)$
Number of $Tenants \times Core$	8	8	$CONST(8)$
Number of $Core$	16	64	$\sum_{i=4}^6 2^i$
Bandwidth $RAN - Edge$	10Gbps	10Gbps	$CONST(10\ Gbps)$
Bandwidth $Edge - Core$	100Gbps	100Gbps	$CONST(100\ Gbps)$
Global latency	30 μ s	30 μ s	$CONST(30\ \mu s)$

connect only one server to each of the physical machines of the core network. Such computer represents an access point to the Internet. Finally, the bandwidth available and latency values have been fixed to specific values to represent real scenarios. However, these values don't affect the performance of TMA but only the result value of the metric. The bandwidth has been fixed to 10Gbps between the RAN and the Edge and 100Gbps between Edge and Core. These values are expected values in 5G networks due to the maximum capacity of network cards used in these segments. The latency has been fixed to 30 μ s for all the segments which are a typical average value.

It is noted that the largest scenario represented in Table 2 to be executed corresponds to a 5G topology with $1024 * 4 * 256 = 1,048,576$ mobile users connected which is the size of a very realistic medium-size city. For that scenario, Fig. 4 shows the description of the population of nodes and edges that conform that topology in order to allow a reader to understand the nature of the graph. As shown in Fig. 4, 99.21% of the nodes and connections represent UE devices. It is also worth remarking that only the 0.03% represented physical machines in edge and core network segments are the candidates to allocated VNFs. This largest scenario is composed by a graph with 1,056,724 vertices and 1,059,291 edges.

To facilitate a 3D representation of the spatial-temporal relationship, we represent each dimension of the three in different figures. Thus, Figs. 5, 6 and 7 represent analogous graphs when it exponentially increases the number of physical machines used in the core network segment. Each of these figures shows an exponential increase in both the number of UE (x -axis) and the number of edges (y -axis). The time shown in the z -axis follows a linear increase. From the three figures, in all the cases, it can be seen how the increase in the number of UE does not affect the time to calculate the 5GCC metric. This is a significant achievement since they are the vast majority of nodes connected to the 5G network. It can be seen as well in all the figures that when the number

**Fig. 4** Distribution analysis of the population of the 5G network topology

of edges is increased, the calculation time increases accordingly. Similarly, comparing the same data series in different plots, it can be seen how increasing the number of cores, the calculation time also increases, accordingly. It means that the internal topology of the 5G network has an influence on the calculation time, mainly due to the fact that the algorithms calculate the shortest paths. The overall execution time increases almost linearly regardless of the workload in the number of edges.

Let us focus on the largest scenario, with 1 million 5G UE connected to the 5G network. The reader can see how each of Figs. 5, 6 and 7 represents three series plotted in different colours. They represent different scenarios over the same 5G topology. Specifically, they represent the levels of stress of the 5G topology. If the infrastructure is not running any service inside, all the physical machines can be candidates to host services due to the fact that the constraint satisfaction will not discard any unsuitable candidates. However, if the infrastructure is running several services and close to be out of remaining free capacity, the number of physical machines that can be candidates to be selected to allocated services is significantly reduced. The different series represent the percentage of physical machines that are suitable to allocate the new service. They are 100% candidates (theoretical—

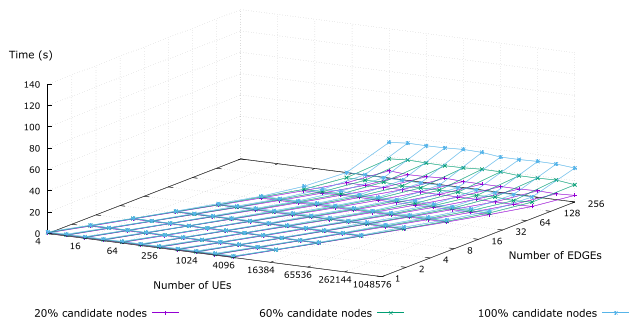


Fig. 5 Time required by TMA to calculate 5GCC without cached data for a topology with 16 Core physical machines when ranging the number of edge machines and the number of UE devices connected to the DU of such edges

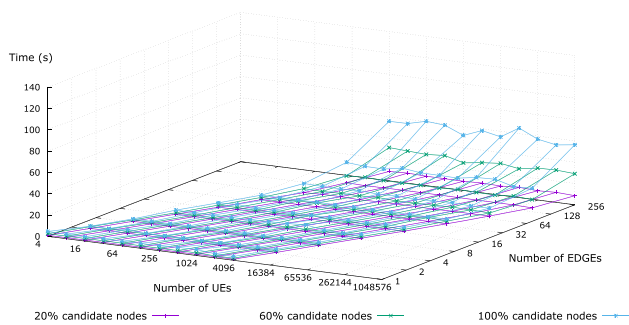


Fig. 6 Time required by TMA to calculate 5GCC without cached data for a topology with 32 Core physical machines when ranging the number of edge machines and the number of UE devices connected to the DU of such edges

no stress), 20% of candidates (realistic—highly stressed scenario) and 80% of candidates (realistic—low stressed scenario). Other intervals have been calculated and they follow the expected trend. It is noted that when the infrastructure is more stressed, the time to calculate the metric is significantly reduced. It helps on the scalability of the calculation of the metric. For 256 edges, 16 core machines and 1 million UE devices in Fig. 5, the metric is calculated in fewer than 20s for very realistic scenarios with 80% of suitable candidates, thereby allowing decision-making related to the structure of the network in a very short time, e.g., allocating a virtual Cache or a virtual Load Balancer. For a similar scenario but with a much more complex core network of 64 core machines, the time to calculate the metric is 60s, which is still within a very acceptable boundary in practical terms. It is very important to remark that all the results presented herein do not make use of the 5GCC retrieval optimization and they are the time to calculate the metric without any cached data (i.e., worst case scenario).

Figure 8 shows how the number of core physical machines affects the calculation time of the metric when fixing the number of edges to 128. It can be seen how the impact on ranging the edges is more dominant than ranging core machines.

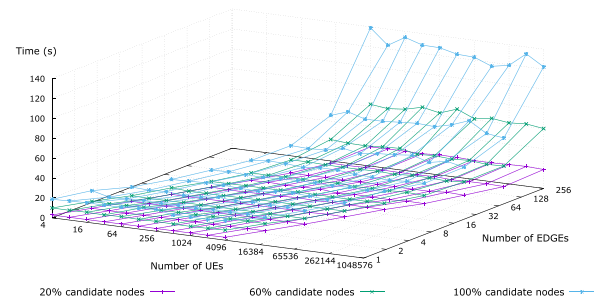


Fig. 7 Time required by TMA to calculate 5GCC without cached data for a topology with 64 Core physical machines when ranging the number of edge machines and the number of UE devices connected to the DU of such edges

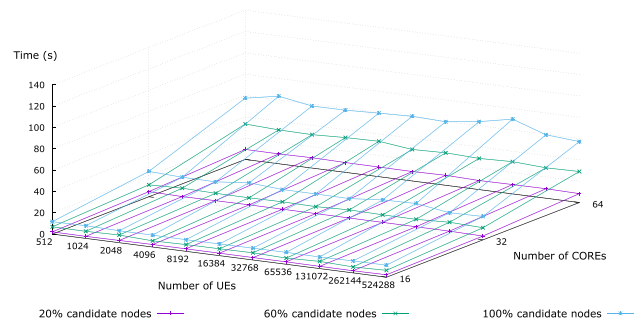


Fig. 8 Time required by TMA to calculate 5GCC without cached data for a topology with 128 Edge physical machines when ranging the number of core machines and the number of UE devices connected to the DU of such edges

Moreover, in all the cases analyzed, the metric shows a linear trend despite the exponential aspect of the axis. This is a clear sign of the high scalability results achieved with our proposal, and thus the proposed approach can be applied in large-scale networking scenarios such as the Load balancing case in mMTC. These results make the metric usable and feasible and can bring significant improvements over the decision-making in the cognitive framework.

7.2 Algorithm computational efficiency results

Secondly, a comparison of performance has been undertaken to show the improvement in computational complexity achieved with the different optimization schemes proposed in the calculation of the spatial metrics over the traditional reference *CC* metric.

Figure 9 shows the time to calculate the spatial metric for a small-size scenario with 4 DUs, 256 UE devices per DU, 4 Edge machines and 4 Core machines where all of the machines are potential candidates to allocate a VNF. It is noted that this small-size scenario is to establish a comparison with the traditional Closeness Centrality (*CC*) metric shown in the “No optimization” series of Fig. 9. The reader can see that it follows an exponential trend in complexity. The points

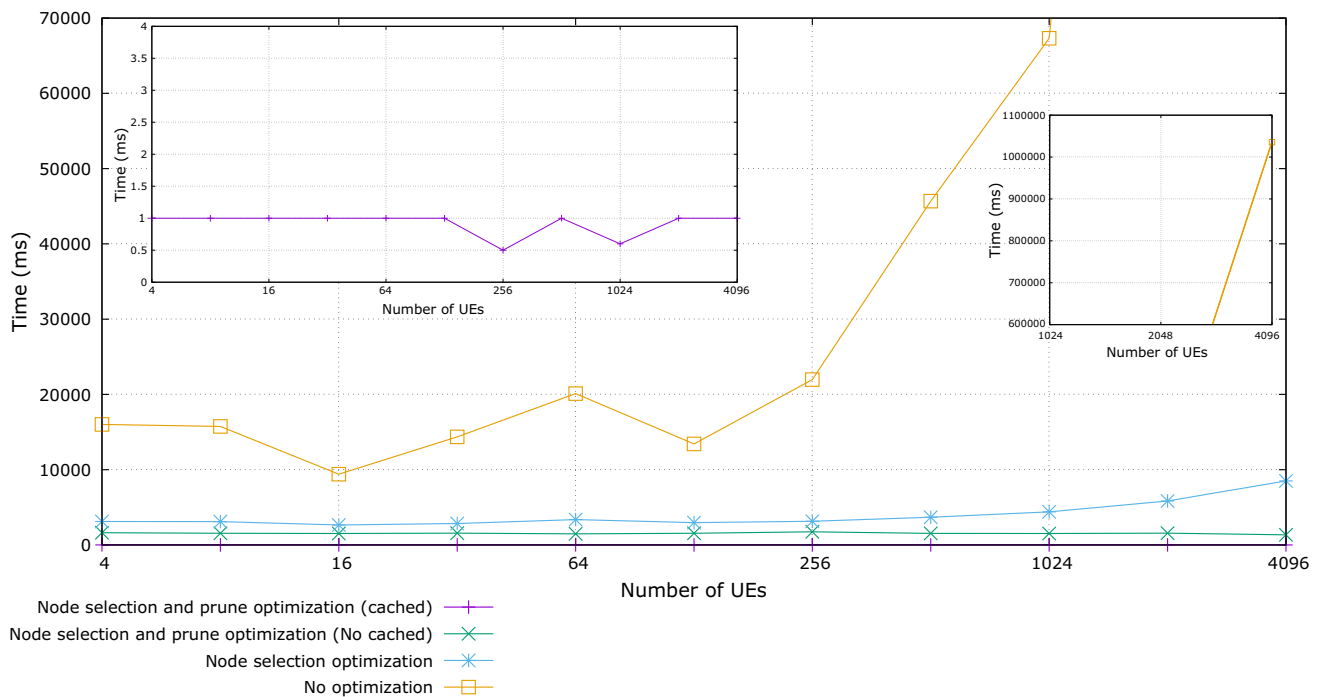


Fig. 9 Comparison of computational complexity in the calculation of 5GCC metric over the different optimization schemes

of the graph for 2048 and 4096 UE devices are plotted out of the scale, and this is why the small sub-figure embedded in the right side has been embedded to allow the reader to see the impractical nature of this *CC* metric due to the complexity associated to its calculation. It takes about 1500.000 ms to compute the largest scenario. The complexity is significantly reduced by two levels of magnitude when the node selection optimization is executed, leading to about 10,000 ms. It is further optimized by another level of magnitude when the prune optimization is also applied over the previous optimization, resulting in about 1000 ms. These results are plotted in Fig. 9. When the 5GCC retrieval optimization is further utilized and the calculation of the metric is cached, the execution time is reduced by yet another extra three levels of magnitude. The y-axis in Fig. 9 does not show enough resolution, and this is why it has been decided to create the sub-graph located in the left-side of Fig. 9 with a zoom in the scale. It can be seen that the calculation time is remarkably reduced to only around 1 ms for the largest scenario analyzed. This excellent result clearly indicates that the proposed approach is able to support real-time use cases such as the proactive cache allocation case and also the computation offloading case. It is noted that it is foreseen that in the vast majority of the time the metric will be cached and only when new VNFs are deployed or migrated it will need to invalidate the cache. These results clearly validate the scalability and practicality of the results achieved in this contribution and the feasibility of the practical application into the 5G networks.

7.3 Graph management results

Lastly, Fig. 10 depicts the analysis of the performance of the TMA component in terms of the management of topological changes in the infrastructure. This experiment has monitored the time required to insert each of the elements of the largest graph utilized in this research work, i.e. the scenario composed by 64 Cores, 256 Edges, 8 tenants which implies 8 VMs in every Core and Edge, 1 server per Core VM, 4 DUs per Edge and 1024 UEs per DU, in total 1,048,576 mobile users. The graph is composed of 1,056,724 vertices and 1,059,291 edges, representing devices and connections between devices respectively. This times plot includes both the addition of the components to the Gremlin graph framework and the updates for all the different cache structures used in the TMA for optimizing the calculation of the metrics (described in previous sections). Figure 10 shows a whisker plot with the analysis of such behaviours. It is worth mentioning that the four quartiles of the distribution are concentrated in one point, representing almost 100% of the cases, yielding around 0.05753782 and 0.05263078 ms, respectively for the insertion of vertices/devices and edges/connections. These results show very similar behaviour between the management of topological changes for both devices and connections. Moreover, these results are at the microsecond scale, which clearly demonstrates the stability on the behaviour of the proposed system and the efficiency in dealing with constant topological changes. It is noted how there are some outliers shown in both plots, which may be due to the fact that the

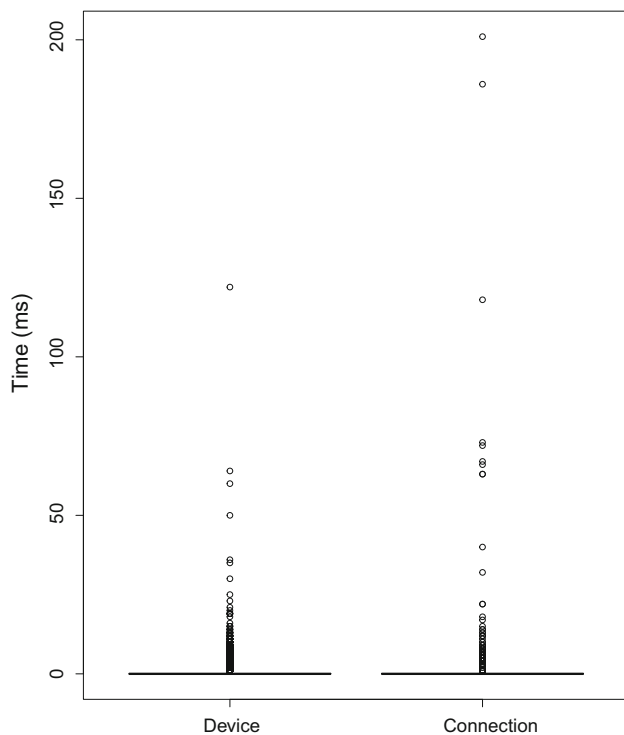


Fig. 10 Whisker plot showing insert time to the database for vertices (Devices) and edges (Connections)

calculation of the metrics maybe work in parallel in another thread while the topological change has arrived, thus causing some sharing of resources of the same machine. They represent the 0.00001% of the cases and the worst-case scenario is around 200 ms, which is still acceptable in use cases that do not have a strict time constraint.

8 Discussion

This section presents a discussion about other alternatives that could be considered and developed in future work. In this paper, four different use cases for the novel multi-tenant 5G networks have been analyzed. However, the solution proposed in this work is not restricted to these use cases since it is based on the universal definition of closeness centrality.

Our proposal has optimized the number of nodes where 5GCC has been calculated by selecting a smaller subset. The reader may think of a possible optimization where the metric is tailored for specific purposes and the destinations used for the calculation of the shortest path are also restricted to a smaller subset, shaping the graph according to a specific problem faced. E.g. to calculate the closeness to respect to a concrete type of VNF, rather than the closeness to respect to all the other nodes in the network. However, we would lose generality achieved in our proposed 5GCC metric. Moreover, the algorithm would need to iterate anyway to all the

nodes available to determine their type of node (e.g. type of VNF) and, in consequence, it will not be any improvement in terms of performance. The alternative would be to have an in-memory graph for each type of metric, which its associated consumption in memory and scalability limitations are associated with.

Another possible optimization could be based on the extension of the support in the framework for additional kind of spatial metrics. For instance, *Degree Centrality*, a much simpler metric in computational cost, can provide the number of neighbours that every node has connected. This metric can be applied when deciding where to deploy a new router in the infrastructure depending on whether a router can route the maximum number of different destinations. In addition, *Betweenness Centrality* metric that measures how many times a node appears in the shortest path between any other nodes in the graph might be applied in similar 5G scenarios as the ones discussed in this paper, giving meaningful information about the centrality. However, such a metric requires higher computation cost on average, and cannot be easily optimized simplifying the graph, as it is done in our proposal.

As future work, authors would like to parallelize the calculation of the metrics using a map-reduce or other alternative distributed framework to further increase both efficiency and scalability. In addition, authors will look into the adaptation and applicability of the proposed model to efficiently manage beyond 5G networks and IoT complex network topologies.

9 Conclusions

5G cognitive network management requires advanced network topological knowledge in a new complicated networking paradigm featured with virtualization and multi-tenancy. This paper has proposed a novel framework that is capable of meeting such challenging requirements for various use cases such as cache allocation, computation offloading and load balancing. In particular, within this framework, new spatial metrics especially a 5G Closeness Centrality metric has been designed, and a set of enabling or optimization algorithms have been proposed. Furthermore, a new 5G architectural component TMA is introduced to monitor such spatial metrics over virtualized, multi-tenanted 5G networks. Moreover, a set of optimization schemes have explored to accelerate the calculation of the new 5G spatial metrics. A realistic testbed has been deployed to test, validate and evaluate the proposed approach. Empirical results have demonstrated the high scalability of the approach over varied network topologies of large scales (with over 1 million mobile user devices and over 1 million connections), and the real-time performance in calculating the spatial metrics to allow timely cognition (only about 1 ms when optimization schemes are applied) and in

handling constant topology changes in 5G networks (in the order of 10s of ms).

Acknowledgements This work was funded by the European Commission Horizon 2020 5G-PPP Program under Grant Agreement Number H2020-ICT-2016-2/761913 (“SliceNet: End-to-End Cognitive Network Slicing and Slice Management Framework in Visualized Multi-Domain, Multi-Tenant 5G Networks”), and by AXA Postdoctoral Scholarship awarded by the AXA Research Fund (Cyber-SecIoT project). The project has also received funding by post-doctoral international mobility fellowship “CERU - On the Move” from University of Murcia, and by the 5G Video Lab project supported by UWS.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bajpai V, Schönwälder J (2015) A survey on internet performance measurement platforms and related standardization efforts. *IEEE Commun Surv Tutor* 17(3):1313–1341
- Baştuğ E, Bennis M, Debbah M (2014) Living on the edge: the role of proactive caching in 5g wireless networks. *IEEE Commun Mag* 52(8):82–89
- Baştuğ E, Bennis M, Debbah M (2016) Proactive caching in 5g small cell networks. In: *Towards 5G: applications, requirements and candidate technologies*. Wiley Online Library, pp 78–98
- Chen D, Lü L, Shang M-S, Zhang Y-C, Zhou T (2012) Identifying influential nodes in complex networks. *Phys A* 391(4):1777–1787
- Dolev S, Elovici Y, Puzis R (2010) Routing betweenness centrality. *J ACM (JACM)* 57(4):25
- Dourado RA, Sampaio LN, Suruagy Monteiro JA (2013) On the composition of performance metrics in multi-domain networks. *IEEE Commun Mag* 51(11):72–77
- Freeman LC (1977) A set of measures of centrality based on betweenness. *Sociometry* 40:35–41
- Gabriel C (2020) Sharing telco infrastructure will accelerate edge startups’ expansion. Accessed 8 Apr 2020
- Goel U, Wittie MP, Claffy KC, Le A (2015) Survey of end-to-end mobile network measurement testbeds, tools, and services. *IEEE Commun Surv Tutor* 18(1):105–123
- GSMA (2019) Future networks. Infrastructure sharing: an overview. Accessed 8 Apr 2020
- Kchiche A, Kamoun F (2010) Centrality-based access-points deployment for vehicular networks. In: *2010 17th international conference on telecommunications*, pp 700–706
- Kim J, Kim D, Choi S (2017) 3GPP SA2 architecture and functions for 5G mobile communication. *ICT Express* 3(1):1–8
- Kourtellis N, Morales GDF, Bonchi F (2015) Scalable online betweenness centrality in evolving graphs. *IEEE Trans Knowl Data Eng* 27(9):2494–2506
- Li X, Wang X, Li K, Leung VCM (2017) Caas: Caching as a service for 5g networks. *IEEE Access* 5:5982–5993
- Marco Alaez R, Alcaraz Calero JM, Wang Q, Belqasmi F, El Barachi M, Badra M, Alfandi O (2017) Open-source based testbed for multioperator 4g/5g infrastructure sharing in virtual environments. *Wirel Commun Mobile Comput* 2017:1–11
- Musumeci F, Bellanzon C, Carapellese N, Tornatore M, Pattavina A, Gosselin S (2016) Optimal BBU placement for 5G C-RAN deployment over WDM aggregation networks. *J Lightwave Technol* 34(8):1963–1970
- Neves P, Calé R, Costa MR, Parada C, Parreira B, Alcaraz-Calero J, Wang Q, Nightingale J, Chirivella-Perez E, Jiang W et al (2016) The selfnet approach for autonomic management in an NFV/SDN networking paradigm. *Int J Distrib Sens Netw* 12(2):2897479
- Nikaein N, Chang C-Y, Alexandris K (2018) Mosaic5g: agile and flexible service platforms for 5g research. *ACM SIGCOMM Comput Commun Rev* 48(3):29–34
- Sabidussi G (1966) The centrality index of a graph. *Psychometrika* 31(4):581–603
- Salva-Garcia P, Alcaraz-Calero JM, Alaez RM, Chirivella-Perez E, Nightingale J, Wang Q (2018) 5g-uhd: design, prototyping and empirical evaluation of adaptive ultra-high-definition video streaming based on scalable h. 265 in virtualised 5g networks. *Comput Commun* 118:171–184
- Sheu J-P, Chen Y-C (2017) A scalable and bandwidth-efficient multicast algorithm based on segment routing in software-defined networking. In: *2017 IEEE international conference on communications (ICC)*. IEEE, pp 1–6
- Tizghadam A, Leon-Garcia A (2010) Betweenness centrality and resistance distance in communication networks. *IEEE Netw* 24(6):10–16
- Tulu MM, Hou R, Li C, Amentie MD (2018) Cluster head selection method for content-centric mobile social network in 5g. *IET Commun* 12(6):402–408
- Wang Z, Zheng J, Wu Y, Mitton N (2017) A centrality-based RSU deployment approach for vehicular ad hoc networks. In: *2017 IEEE international conference on communications (ICC)*. IEEE, pp 1–5
- Zhao L, Sun W, Shi Y, Liu J (2018) Optimal placement of cloudlets for access delay minimization in sdn-based internet of things networks. *IEEE Internet Things J* 5(2):1334–1344

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.