This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2892767, IEEE Transactions on Industrial Informatics

Joint Computation Offloading, Power Allocation, and Channel Assignment for 5G-enabled Traffic Management Systems

Zhaolong Ning, Xiaojie Wang, Joel J. P. C. Rodrigues, Feng Xia

***Abstract*—Due to the ever-increasing requirements of delay-sensitive and mission-critical applications in 5G, Mobile Edge Computing (MEC) is promising to react and support real-time interactive systems. However, it is still challenging to construct a 5G-enabled traffic management system, owing to the qualification of ultra-low latency and ubiquitous connectivity. Furthermore, the computing resources and storage capacities of edge nodes are limited, thus computation offloading is a fundamental issue for real-time traffic management. This paper puts forward a hybrid computation offloading framework for real-time traffic management in 5G networks. Specially, we consider both Non-Orthogonal Multiple Access (NOMA) enabled and Vehicle-to-Vehicle (V2V) based traffic offloading. The investigated problem is formulated as a joint task distribution, sub-channel assignment, and power allocation problem, with the objective of maximizing the sum offloading rate. After that, we prove its NP-hardness and decompose it into three subproblems, which can be solved iteratively. Performance evaluations illustrate the effectiveness of our framework.**

***Index Terms*—Computation offloading, edge computing, 5G, traffic management, power allocation, channel assignment.**

I. Introduction

THe generated data by mobile devices are estimated to increase over 50% annually from now to 2020 [1]. With the development of smart vehicles and sensors, various kinds and huge amounts of data can be collected. It is estimated that over 1.5 billion vehicles will be connected by 2020, and each vehicle will produce approximately 30 TB data one day, posing a great challenge for the real-time traffic management [2]. The unprecedented growth of mobile data makes the bandwidth and spectrum provided by cellular networks increasingly saturated. Furthermore, how to process and manage the generated traffic data in a real-time manner is rather challenging. This is because real-time traffic management in vehicular networks calls for prompt responses, a high quality of user experiences and massive network connectivities, especially for delay-sensitive applications [3]. Internet of Vehicles (IoV) is acknowledged as a key research field for the development of industrial applications, such as traffic management and road safety [4]. The research and testbed systems on IoV-based industrial automation have been conducted by many worldwide automakers, such as BMW, Volvo, and Toyota.

Z. Ning is with Chongqing Key Laboratory of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing, China, and also with School of Software, Dalian University of Technology, Dalian, China, and also with State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China. (Email: zhaolongning@dlut.edu.cn)

X. Wang and F. Xia (Co-Corresponding author) are with School of Software, Dalian University of Technology, Dalian, China. (Email: wangxj1988@mail.dlut.edu.cn, f.xia@ieee.org)

J. Rodrigues (Co-Corresponding author) is with National Institute of Telecommunications (Inatel), Santa Rita do Sapucaí-MG, Brazil, and also with Instituto de Telecomunicações, Portugal, and also with University of Fortaleza (UNIFOR), Fortaleza-CE, Brazil. (Email: joeljr@ieee.org)

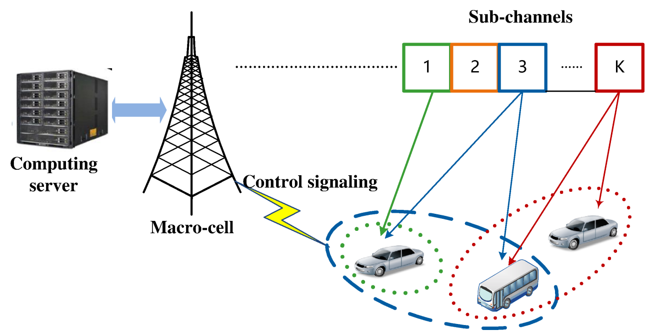


Fig. 1: NOMA for MEC-enabled vehicular networks.

With the objective of taking full advantage of wireless spectrum and increasing the utilization of bandwidths, Non-Orthogonal Multiple Access (NOMA) is a promising technology in 5G networks, which can enable non-orthogonal channel accesses through power domain or code domain multiplexing [5], [6]. Different from the traditional Orthogonal Frequency Division Multiple Access (OFDMA), multiple users can share the same sub-channel in NOMA to gain multiplexing benefits. An illustrative framework for NOMA-based communications between a 5G macro-cell and vehicular users is demonstrated in Fig. 1, where the macro-cell separates the bandwidth to multiple sub-channels and transmits signals to different vehicular users. Sub-channels can be reused by vehicular users for communications and traffic offloading. It should be noticed that the co-channel interference in NOMA is unavoidable, and multi-user detection techniques can be leveraged to decode the received signals and alleviate the inter-user interference in NOMA.

In order to jointly consider sub-channel assignment and power allocation, the authors in [7] concentrate on resource allocation in a downlink NOMA-based wireless network. A low-complexity matching algorithm is presented, where a two-sided exchange-stable matching is constructed by considering the interaction of both users and sub-channels. Due to the heterogeneous access of V2X networks, NOMA is integrated with cellular-based vehicular networks to provide massive

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2892767, IEEE Transactions on Industrial Informatics

connectivities and reduce resource collision, so that the low latency and high reliability requirements in 5G can be satisfied. With the ever-increasing number of devices in Internet of Things (IoT), NOMA is promising to enable the communications of massive devices in cellular networks with constrained radio resources.

Since around 45% of the generated data by IoTs will be processed by edge devices, Mobile Edge Computing (MEC) is with great promise for real-time traffic management, which can largely shorten the processing time compared with the centralized Mobile Cloud Computing (MCC) by transferring the computing capability from the central cloud to the edge. Compared with MCC, the following three advantages of MEC can be drawn: 1) Closeness: Data generation and processing are close to data sources and in the proximity of users; 2) Diversity: Various kinds of edge devices with distinct computing capabilities coexist, such as RoadSide Units (RSUs), vehicles and WiFi hotspots; 3) Constrained resources: Edge nodes are generally with lower computing capabilities than those of cloud servers.

In order to fulfill latency-sensitive applications in 5G, edge or fog computing is promising to decrease the transmission time during computation offloading, by taking advantage of nearby nodes with available resources [8]. A NOMA-based cooperative computation offloading algorithm is presented in [9] to minimize the network latency. By integrating with social trust, link stability can also be strengthened. In [10], the authors construct a small-cell network framework for task offloading in 5G, and they concentrate on the energy consumption of offloading from the aspects of both communication and computation.

Heterogeneous physical resources and network access patterns coexist in the current intelligent transportation system. Fog-Radio Access Network (F-RAN) and NOMA are acknowledged to satisfy the heterogeneous requirements in 5G networks. For the sake of maximizing the network utility, the authors in [11] allocate power and sub-channel resources in NOMA-based F-RANs by taking the co-channel interference into consideration. With the objective of promoting the caching capability of edge devices, the authors in [12] design a vehicle-aided edge caching framework for 5G-enabled IoVs by joint scheduling the caching and computing resources of edge devices. In order to minimize the total network cost of energy and user delay, an optimization problem, by jointly considering scheduling and resource allocation between cloud and fog devices, is formulated in [13]. The cooperation between cloud and fog computing is investigated in [14]. A joint offloading decision, resource allocation, and power control scheme is presented to minimize the maximal weighted cost of energy consumption and delay.

In order to manage network traffic in a real-time manner, this paper puts forward a hybrid computation offloading scheme for a 5G-enabled traffic management system by taking advantage of NOMA and MEC technologies. Our objective is to maximize the link rate by enabling various offloading schemes, and the studied problem is formulated as a joint task distribution, sub-channel assignment, and power allocation problem. We further demonstrate its NP-hardness. Due to its high computational complexity, the formulated problem is decomposed into three subproblems. To the best of our knowledge, our work is a prior attempt to construct a real-time traffic management system. It supports heterogeneous network accesses and offloading in 5G, empowering by the integration of NOMA and MEC technologies. The main contribution of this paper can be summarized as follow:

1) We investigate a hybrid computation offloading scheme in 5G, and formulate an optimization problem by jointly considering task distribution, sub-channel assignment and power allocation. By considering both NOMA and MEC technologies, our objective is to maximize the achievable sum rate.
2) Because of the NP-hardness of the formulated problem, we decompose it into three subproblems, i.e., dynamic offloading decision, power allocation, and sub-channel assignment, so that the formulated problem can be solved in an iterative manner.
3) For the offloading decision between cellular and Vehicle-to-Vehicle (V2V) offloading, we theoretically derive the maximum tolerable transmission delay by stochastic theory. If the estimated delay is above the threshold, cellular network based offloading is required. Otherwise, a V2V based predictive offloading can be leveraged to decrease the network cost.
4) In order to minimize network interference caused by concurrent transmission links, we study a joint power control and channel allocation scheme, and solve it iteratively by power allocation and channel assignment.

Performance evaluations based on a city-road map demonstrate the effectiveness of our method comparing with some existing solutions. The rest of this paper is organized as follows. In Section II, we state the system model. Problem formulation is specified in Section III. After that, we present a heuristic solution to solve the optimization problem in a distributed manner. Performance evaluations in Section V demonstrate the effectiveness of our solution before concluding our work in Section VI.

II. System Model

The system model is illustrated in Fig. 2. The real-time traffic management system in 5G has distinct network access patterns, including a macro-cell equipped with a computing server for computing-intensive task implementation, RSUs and vehicles [15]. Specially, RSUs can be equipped with MEC servers for traffic offloading. The main components are specified as follows:

1) Macro-cell: It is regulated by mobile operators, and can almost cover wireless communications among vehicles in urban areas. However, message offloading via macro-cells charges additional costs, and their bandwidth is becoming ever-saturated. Generally, one macro-cell in 5G can play the role of a router for global communication, an MEC server, and a local controller. A computing server is necessary to be deployed for the macro-cell to offer computing and offloading services.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2892767, IEEE Transactions on Industrial Informatics

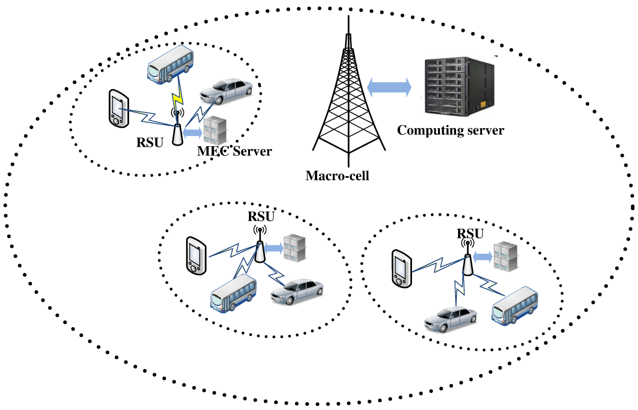


Fig. 2: An illustration of MEC-enabled vehicular network.

2) RSUs: They are installed along roads to connect vehicles by providing wireless communications. RSUs upload the generated messages by vehicles to the traffic management server. However, deploying a huge number of RSUs is expensive for network accesses. In our work, merely a limited number of RSUs exist and they can obtain the location information of neighboring RSUs by information synchronization. In order to satisfy the ever-increasing computation requirements of vehicles, an MEC server can be integrated with an RSU for storage and offloading. Due to the high speed of vehicles, high-efficient communications between MEC servers and vehicles are significant.

3) Vehicles: Vehicles equipped with On Board Units (OBUs) can be leveraged for V2V communications by bluetooth, Wi-Fi and other short distance communication technologies. Thanks to the development of sensors and wireless communication technologies, vehicles close to each other can exchange information in a device-to-device manner.

A specific pre-installed software inside a vehicle can be leveraged by the driver or passengers to record the occurred event, such as traffic jams, traffic accidents, or road surface damages [16]. The occurred event can be recorded by pictures, texts or short videos, when the vehicle detects the event during its route. After that, the vehicle packages the recorded information into messages before offloading. Although an MEC server is equipped with an RSU, its computing resources may not be strong enough to meet the demands of vehicles. Therefore, a computing server is equipped with the macro-cell to provide computing resources for RSUs, so that the MEC server can also offload its computation tasks to the macro-cell. The objective of our 5G-enabled traffic management system is to provide high-efficient communication and computing services, so that the tasks of vehicles can be fulfilled timely.

For the sake of promoting routing efficiency while reducing the communication overhead, our work leverages the position-based clustering scheme in [17] to enable V2V communications. The cluster structure can be determined by the traffic information and geographic positions. If a vehicle is selected as the cluster head, the maximum distance between the cluster head and cluster members can be predefined for cluster size control. After that, the cluster head generates a message to describe the happened event by feature extraction. In order to increase the transmission rate and make full use of wireless spectrum, the cluster head can upload the generated messages by either macro-cell or RSUs.

If the macro-cell is selected, the caused delay for message forwarding is almost negligible while additional costs are added, such as some charged fees by the mobile operator. Otherwise, a geographical message forwarding scheme deserves to be designed, so that the message can be delivered to a suitable RSU. Although message offloading by RSUs is free, it will cause extra delay.

For the considered network supporting NOMA, it can cover a group of edge devices (such as RSUs), denoted by $\mathcal{M} = \{1, ..., M\}$, to offload network traffic. In our work, macro-cells and RSUs are two main infrastructures to provide network access for vehicular users. The available bandwidth of the macro-cell is separated into a group of sub-channels, denoted by $k \in \mathcal{K}$. There are N users with J computation tasks to be offloaded. For one user within the coverage of the macro-cell, its task can be expressed by $\tau_{n,j} = \{d_{n,j}, c_{n,j}, T_{n,j}, n \in \mathcal{N}, j \in \mathcal{J}\}$. Herein, $d_{n,j}$ is the size of task j to be offloaded by user n, and $c_{n,j}$ denotes the required number of CPU cycles for user n to fulfill task j, and $T_{n,j}$ is the maximum tolerable latency of task j for user n. Similar with many existing researches (such as [7] and [13]), we consider a quasi-static network circumstance, and assume the channel condition is stable during the periods of transmission and offloading, while it varies independently from one to another. The macro-cell is assumed to have the full knowledge of the Channel Side Information (CSI), based on which the non-overlapping sub-channels and transmission power can be allocated to users.

A. Communication Model

According to NOMA, one user can receive information from edge devices via multiple sub-channels, and one sub-channel can be assigned to distinct users. We denote $p_{m,n}$ as the allocated power of the mth edge device to the nth user, and the received signal $y_{m,n}$ can be illustrated as:

$$y_{m,n} = \sqrt{p_{m,n}} h_{m,n} x_{m,n} + \sum_{i \neq n, i \in \mathcal{N}} \sqrt{p_{m,i} h_{m,i} x_{m,i}} + z_{m,n}, \tag{1}$$

where the channel gain between the nth user and the mth edge device is $h_{m,n}$, and $x_{m,n}$ is the transmitted signal. The first item is the desired signal from the nth user, while the second item is the intra-cell interference from other users on the same frequency band. Symbol $z_{m,n}$ is the additive white Gaussian noise for the nth user with variance δ^2.

Since the sub-channel can be reused by different users in the MEC-based offloading networks with NOMA, the signal transmission of the nth user causes interference to other users. In order to demodulate the target message, each user should apply Successive Interference Cancellation (SIC) [18]. Generally speaking, low-power level is allocated to the user with a high channel gain after all the users with high-power levels are decoded by the SIC. High-power level is assigned to the user with a low channel gain to recover the signals by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2892767, IEEE Transactions on Industrial Informatics

regarding the low-power level signals as the noise during SIC decoding. Consequently, the user with the highest channel gain suffers from the interference from all the other users, while the signal transmitted by the user with the lowest channel gain is almost interference-free. Specifically, if the channel gain satisfying $h_{m,n}^2/z_{m,n} > h_{m,n'}^2/z_{m,n'}$, the receiver of the nth user can cancel the interference from the n'th user. For the users with higher channel gains than those of the n'th user, their signals are viewed as noise to decode $x_{m,n'}$. The Signal-to-Interference-plus-Noise Ratio (SINR) of the nth user served by the mth edge device can be calculated by:

$$\Gamma_{m,n} = \frac{p_{m,n} h_{m,n}^2}{\sum_{\frac{h_{m,n}^2}{z_{m,n}} > \frac{h_{m,n'}^2}{z_{m,n'}}} p_{m,n'} h_{m,n'}^2 + \delta^2}. \tag{2}$$

The date rate of the nth user that transmitting to the mth edge device within one time slot can be illustrated as:

$$R_{m,n} = \omega_{m,n} log_2(1 + \Gamma_{m,n}), \tag{3}$$

where $\omega_{m,n}$ is the occupied bandwidth of user n served by edge device m. The total achievable rate is the rate summation of all the concurrent transmission links.

Similarly, if the nth user transmits signals directly to the macro-cell over sub-channel k, the achievable rate is:

$$R_{k,n} = \sum_{k \in \mathcal{K}} \omega_{k,n} b_{k,n} log_2(1 + \Gamma_{k,n}), \tag{4}$$

where $\omega_{k,n}$ and $\Gamma_{k,n}$ are the bandwidth and SINR value between user n and the macro-cell over the kth sub-channel. Binary variable $b_{k,n}$ represents whether sub-channel k is allocated to user n.

B. Mobile Edge Computing Model

If the mobile edge device is selected by users to offload network traffic, an extra delay is caused by transmitting the computation data to the edge devices through wireless links.

If user n sends computation task $\tau_{n,j}$ to edge device m, the transmission latency for the jth offloading task is:

$$T_{m,n,j}^T = \frac{d_{n,j}}{R_{m,n}}. \tag{5}$$

We denote $T_{m,s,j}$ as the computation time of task $\tau_{n,j}$ offloaded to edge device m by the sth user. If edge device m is selected by the sth user, $b_{m,s} = 1$; otherwise, $b_{m,s} = 0$. The delay of user n to process task j by the mth edge device is:

$$T_{m,n,j}^D = \sum_{\frac{h_{m,s}^2}{z_{m,s}} > \frac{h_{m,n}^2}{z_{m,n}}} b_{m,s} T_{m,s,j}. \tag{6}$$

Due to the limited computation ability of edge devices, the execution time of $T_{m,s,j}^E$ can be calculated by:

$$T_{m,s,j}^E = \frac{Q_{s,j}}{C_{m,s}}, \tag{7}$$

where $Q_{s,j}$ is the queue of task j to be processed by user s, and $C_{m,s}$ is the computing capability of edge device m allocated to user s.

Therefore, the total latency of edge computing for user n to offload task j to edge device m can be worked out by:

$$T_{m,n,j} = T_{m,n,j}^T + T_{m,n,j}^D + T_{m,n,j}^E. \tag{8}$$

III. Problem Formulation

For the sake of constructing a real-time traffic management system in 5G, this section formulates the investigated problem as a sum-rate maximization problem. We consider the constraints of offloading decision, latency, task assignment, power allocation, channel assignment and so on. The total sum rate is the summation achieved by both macro-cell based and edge device based transmission, which is proportional to the network total utility, i.e.,

$$\mathcal{U}_{total} = \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{J}} (\alpha_{n,j} R_{m,n} + \beta_{n,j} R_{k,n}), \tag{9}$$

where $\alpha_{n,j}$ and $\beta_{n,j}$ are binary variables to denote whether the jth task of the nth user is handled by the edge device or macro-cell, respectively.

The optimization problem is formulated as follows:

$$\max \mathcal{U}_{total}, \tag{10}$$

s.t.

$$\alpha_{n,j} + \beta_{n,j} = 1, \forall n \in \mathcal{N}, j \in \mathcal{J}, \tag{11}$$

$$\sum_{n \in \mathcal{N}} b_{k,n} \leq \eta_u, \forall k \in \mathcal{K}, \tag{12}$$

$$\sum_{k \in \mathcal{K}} b_{k,n} \leq \eta_s, \forall n \in \mathcal{N}, \tag{13}$$

$$\sum_{m \in \mathcal{M}} b_{m,n} \leq 1, \forall n \in \mathcal{N}, \tag{14}$$

$$\sum_{k \in \mathcal{K}} b_{k,n} \omega_{k,n} \leq W_M, \forall n \in \mathcal{N}, \tag{15}$$

$$\sum_{m \in \mathcal{M}} b_{m,n} \omega_{m,n} \leq W_E, \forall n \in \mathcal{N}, \tag{16}$$

$$\sum_{k \in \mathcal{K}} b_{k,n} p_{k,n} \leq P_M, \forall k \in \mathcal{K}, \tag{17}$$

$$\sum_{m \in \mathcal{M}} b_{m,n} p_{m,n} \leq P_E, \forall n \in \mathcal{N}, \tag{18}$$

$$\sum_{n \in \mathcal{N}} b_{m,n} C_{m,n} \leq C_m, \forall m \in \mathcal{M}, \tag{19}$$

$$\alpha_{n,j} \sum_{m \in \mathcal{M}} b_{m,n} T_{m,n,j} \leq T_{n,j}, \forall n \in \mathcal{N}, j \in \mathcal{J}, \tag{20}$$

$$\alpha_{n,j}, \beta_{n,j}, b_{k,n}, b_{m,n} \in \{0, 1\}. \tag{21}$$

Constrain (11) shows each task can be offloaded by either the macro-cell or edge devices. Constraints (12) and (13) indicate each sub-channel can merely be allocated to at most η_u users, and each user can occupy η_s sub-channels at most. As illustrated in (14), each user can choose one edge device to offload tasks at most. The constraints for bandwidth allocation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2892767, IEEE Transactions on Industrial Informatics

of both the macro-cell and edge devices are respectively demonstrated in (15) and (16), where W_M and W_E are the available bandwidths of both macro-cell and edge devices, respectively. The power consumption of both macro-cell and edge devices should be no more than their limited transmission power (P_M and P_E respectively) as demonstrated in (17) and (18). In (19), the computing resources allocated to users should not exceed the maximum computing resource of the edge device (i.e., C_m). The tolerated delay of task offloading by the edge device is illustrated in (20).

Proposition 1: The sum-rate maximization problem in equation (10) is NP-hard.

Proof: The joint problem of sub-channel assignment and power allocation for NOMA in [7] has demonstrated to be NP-hard. Beyond that work, we investigate task offloading and user selection between the macro-cell and edge devices. Since a simplified case of equation (10) is NP-hard, Proposition 1 holds. ■

IV. Joint Computation Offloading, Power Allocation and Channel Assignment

Due to the computational complexity of the formulated problem, this section presents a joint Computation Offloading, Power Allocation and Channel Assignment (COPACA) scheme to solve the formulated sum-rate maximization problem in a distributed manner. An iterative solution is put forward, where we solve the offloading decision problem before considering the power allocation and channel assignment problems.

A. Computation Offloading Decision

This subsection puts forward a hybrid computation offloading framework for real-time traffic management by taking advantage of heterogeneous network accesses in 5G.

When an event happens, a message can be generated by one vehicle before forwarding it to the cluster head. After that, the cluster head needs to decide how to upload the generated message. If the message is uploaded before the traffic management server takes actions, all the vehicles within the cluster are rewarded. The utility of cluster l (U_l) has relationships with the server reward, size and cost of the uploaded message, and the transmission delay to the server, which are the main factors to determine server response timeliness and vehicle benefit. Similar with the definition of utility function in [19], the utility function of cluster l is:

$$U_l = y_i(t,s)(1 - e^{-l_m c_m \Delta t}\theta), \tag{22}$$

where the payment of the traffic management server is $y_i(t,s)$ according to time t and report quality s. The size and cost of an uploaded message through the macro-cell are represented by l_m and c_m, respectively. The delay between message offloading through the macro-cell and an RSU is denoted by Δt, equaling to the message transmission time from the cluster head to the nearest RSU. Binary variable θ denotes whether RSU is employed for message offloading. Two main parts are contained in Δt, i.e., message transmission time from its current position to the nearest RSU and message offloading time from an RSU to the traffic management server. Compared with the former part, message offloading time can be ignored. If the computed utility of cluster l by the macro-cell is higher than that calculated by RSU forwarding, the macro-cell is selected for message offloading; Otherwise, RSUs are preferred. It should be noticed that the estimation of Δt affects the message offloading strategy. Next, we specify the estimation of Δt. When a message is received by the traffic management system, its accuracy, such as the time, location and specified description of the recorded event, will be validated. Our work integrates the data trustworthiness assurance solution in [20] to validate the message. If the received message is confirmed, a notification message will be formed and broadcasted to the passing-by vehicles through RSUs. Otherwise, the traffic management system will wait for other messages.

The real map can be represented by a directed graph G=(O, L), including node set O to illustrate road intersections and link set L to represent road segments among road intersections. According to [21] and [22], traffic flow entering a road segment can be modeled as a Poisson process. We define the average arrival rate as $\lambda_{p,q}$, where $p,q \in O$ and $l_{p,q} \in L$. Similar with [23], we also assume each vehicle's speed ($v_{p,q}$) keeps constant during its period on the route. The traffic density can be calculated by $\rho_{p,q} = \lambda_{p,q}/E(v_{p,q})$, where $E(v_{p,q})$ is the average vehicle speed on road segment $l_{p,q}$.

The locations of neighboring RSUs can be aware by the local RSUs. RSUs can exchange the information of their local maps and the latest updated time to the vehicles passing by, thus the cluster head can obtain the local information of the nearest RSU. The K-shortest paths from the location of the cluster head to the nearest RSU can be worked out by some existing methods (such as $Dijkstra$ algorithm).

If one cluster with the generated messages encounters another cluster, the encountered one is asked to send the information of both traveling possibility along a predefined path and the estimated time to the nearest RSU. For the selection of the next-hop cluster, the one with a large traveling probability and short estimated time is preferred. When a cluster approaches to a road section, its turning and encounter possibilities are calculated by the cluster head according to its travel plan. If another road segment is suitable for transmission, messages will be taken into this road segment along a predefined path by the cluster. Otherwise, they will be forwarded to a following cluster, so that other vehicles can take messages to the suitable direction at the road interaction. The transmission delay of all the paths estimated by the cluster head is demonstrated in the following theorem.

Theorem 1: The expected path delay from the cluster head to the nearest RSU along with intersections $o_1 \rightarrow o_2 ... \rightarrow o_k$, can be worked out by:

$$\sum_{i=1}^{k-1} E(t_{i,i+1}) = \sum_{i=1}^{k-1} Y - \sum_{i=1}^{k-1} (Y+t) e^{-\lambda_{i,i+1} t}, \tag{23}$$

where $Y = \frac{1}{\lambda_{i,i+1}} + t^E_{i,i+1}$, and $t^E_{i,i+1}$ is the average travel time

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2892767, IEEE Transactions on Industrial Informatics

of a vehicle along road segment $l_{i,i+1}$. $t^E_{i,i+1} = \frac{L_{i,i+1}}{E(v_{i,i+1})}$, where $L_{i,i+1}$ is the length of road segment $l_{i,i+1}$, and $E(v_{i,i+1})$ is the average speed of vehicles on $l_{i,i+1}$.

Proof: Denote the maximum waiting time in one road intersection by t, and the message will be dropped if it cannot be delivered to the right road segment within time t. For one road segment, the expected delay can be calculated by:

$$E(t_{i,i+1}) = \int_0^t (\tau + t^E_{i,i+1}) f_{i,i+1}(\tau) d\tau. \tag{24}$$

The probability density function of the message waiting time at road intersections i to $i+1$ is expressed by $f_{i,i+1}(\tau)$. Because the traffic flow entering a road segment can be modeled as a Poisson process with an average arrival rate $\lambda_{i,i+1}$, $f_{i,i+1}(\tau) = \lambda_{i,i+1} e^{-\lambda_{i,i+1}\tau}$ holds. The expected delivery delay from road segments i to $i+1$ is:

$$E(t_{i,i+1}) = \frac{1}{\lambda_{i,i+1}} - (t + \frac{1}{\lambda_{i,i+1}} + t^E_{i,i+1}) e^{-\lambda_{i,i+1}t} + t^E_{i,i+1}. \tag{25}$$

The expected delivery delay of a path can be calculated by:

$$\begin{aligned} \sum_{i=1}^{k-1} E(t_{i,i+1}) &= \sum_{i=1}^{k-1} (\frac{1}{\lambda_{i,i+1}} + t^E_{i,i+1}) \\ &- \sum_{i=1}^{k-1} (\frac{1}{\lambda_{i,i+1}} + t^E_{i,i+1} + t) e^{-\lambda_{i,i+1}t}. \end{aligned} \tag{26}$$

Therefore, Theorem 1 can be proved. ■

The expected message delivery delay is compared with Δt by calculating the ratio $P(\sum_{i=1}^{k-1} E(t_{i,i+1})) \leq \Delta t$. If this ratio is above a predefined threshold, RSU is suitable for offloading; otherwise, macro-cell is selected for offloading.

B. A Joint Power Allocation and Channel Assignment Heuristic Algorithm

It can be observed that the high computational complexity in equation (10) is mainly from three parts, i.e., offloading decision, power allocation and channel assignment. After solving the offloading decision constraint in equation (11), this subsection presents a joint power control and channel allocation heuristic algorithm to solve the formulated problem. Although some existing schemes have been investigated to solve the joint power control and channel allocation problem, the number of link combinations with distinct transmission power is huge, not to mention the channel assignment. The presented joint power control and channel allocation heuristic algorithm in this subsection contains the following three main steps:

- Prune edges: Remove some extra-essential links from the link graph to reduce the searching space.
- Power allocation: Allocate transmission power to activate multi-link transmissions of both edge devices and the macro-cell.
- Channel assignment: Reuse the spectrum frequency to maximize the total number of users accessing into sub-channels.

1) Prune edges: Generally, the nearby vehicular users can receive the forwarded messages with a high transmission rate, while a low rate for farther neighbors. Therefore, a great number of message transmission paths with distinct link rates exist. It is obvious that the computational complexity of the formulated problem in equation (10) increases rapidly as the matching number of users and edge devices enhances. Thus, we first calculate the K-shortest paths between users and the cluster head according to Yen's algorithm [24]. The link cost is also set as the reciprocal of the maximum possible transmission link as in [25]. The tradeoff between the computational speed and the network performance has direct relationship with parameter K, and we set $K = 3$.

2) Power allocation: With the objective of relaxing power constraints in the formulated problem, we present a distributed power allocation scheme to determine $b_{m,n}$ and $p_{m,n}$.

Although one link with large transmission power can acquire a high rate, it will cause strong interference and limit the spectrum reuse of other links. NOMA is a promising technology for spectrum reuse in 5G, whereas, the sub-channel cannot be reused without limitation due to the hardware constraint. Furthermore, it is demonstrated in [26] that merely three or four sub-bands can be divided from the whole bandwidth, if the system does not have any delay constraint. If a strict delay constraint has to be satisfied, the whole bandwidth should be reused by merely one sub-band.

For the sake of increasing the sum-rate of 5G-enabled vehicular networks, a trade-off has to be made between the maximum achievable transmission rate and the number of concurrent transmission links. One intuitive approach for the balance of link rate and spatial reuse is the allocation of transmission power, by which the interference of concurrent transmission links can be controlled to a limited area [27] [28]. The core concept of our power allocation scheme is to supplement links into the link group iteratively, so that the generated interference by concurrent transmission links can be controlled to an appropriate level.

We define configuration s as a link set, where links inside can be activated concurrently. Initially, it is null. In order to take the advantage of wireless spectrum, links joining into the configuration start from the one with the best channel state, which can be reflected in the achievable SINR value and the ability to resist network interference. Interference range is generally leveraged to denote the ability for interference resistance, within which the SINR threshold can be satisfied for link transmission even though other concurrent transmission links exist. It is clear that one transmission link with a large SINR value has a larger interference range and strong ability to resist the interference from other concurrent transmission links. The procedure to select links for configuration construction is illustrated as follows:

$$L_{m,n} \rightarrow \arg\{\max_{m\in\mathcal{M},n\in\mathcal{N}} (\frac{\Gamma_{R_{m,n}} h^2_{m,n}}{\sum_{h\in\mathcal{M}-\{m\}} h^2_{h,n}})\}, \tag{27}$$

where $\Gamma_{R_{m,n}}$ is the SINR threshold for transmission with rate $R_{m,n}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2892767, IEEE Transactions on Industrial Informatics

After the first transmission link is chosen, the minimum required power is:

$$P_{m,n} \geq \frac{\Gamma_{R_{m,n}} \delta^2}{h_{m,n}^2}. \tag{28}$$

Next, other links are supplemented into the configuration with an iterative manner. If another link $L_{m',n'}$ joins into configuration s with transmission power $P_{m',n'}$ ($m' \in \mathcal{M} - m, n' \in \mathcal{N} - n$), its generated interference to the existing links is:

$$\Delta I_n \geq P_{m',n'} h_{m',n}^2. \tag{29}$$

With the objective of overcoming the interference brought by the new added link to other links, their transmission power should be added by at least:

$$\Delta P_{m,n} = \frac{\Delta I_n \Gamma_{R_{m,n}}}{h_{m,n}^2} \geq \frac{P_{m',n'} h_{m',n}^2 \Gamma_{R_{m,n}}}{h_{m,n}^2}. \tag{30}$$

The constraint in equation (30) puts a threshold on $P_{m',n'}$:

$$P_{m',n'} \leq \frac{\Delta P_{m,n} h_{m,n}^2}{h_{m',n}^2 \Gamma_{R_{m,n}}} = \frac{(P_{max} - P_{m,n}) h_{m,n}^2}{h_{m',n}^2 \Gamma_{R_{m,n}}}, \tag{31}$$

where P_{max} equals to either P_M or P_E according to offloading decision.

Therefore, the problem of link selection becomes to maximize the SINR value at the destination node, i.e.,

$$l_{m,n} \rightarrow \arg\{\max_{m',n'}\{\min_{m,n} \frac{(P_{max} - P_{m,n}) h_{m,n}^2}{h_{m',n}^2 \Gamma_{R_{m,n}} I_n}\}\}. \tag{32}$$

The link is selected to join into the configuration until its allocated transmission power is larger than P_{max} or no existing link needs to be activated. By our distributed power allocation method, one configuration containing multiple links can be activated to reuse the wireless spectrum simultaneously.

3) Sub-channel assignment: Through the worked out offloading decision and power allocation scheme, the transmission links for traffic offloading can be activated simultaneously for orthogonal multiple access. In this subsection, we develop a heuristic sub-channel allocation scheme for 5G-enabled vehicular network with NOMA.

At the initial stage of each scheduling period, vehicular users update their positions and network states before the macro-cell allocates frequency resources. Because co-channel interference exists for vehicular users with NOMA, how to maximize the utility function in equation (10) can be transferred to maximize $\varphi_{n,k}$, which is a binary variable to represent whether sub-channel k can be accessed by vehicular user n. Therefore, our objective is to maximize the total number of $\varphi_{n,k}$, indicating as many vehicular users as possible should be assigned to sub-channels.

Before illustrating the channel assignment method, we define some binary variables. If sub-channel k is assigned to configuration s, $y_k^s = 1$. If configuration s contains node n, $a_n^s = 1$. If sub-channel k is leveraged by node n in configuration s, $\nu_{n,k}^s = 1$. The channel assignment problem can be formulated as:

$$\max\ \varphi_{n,k}, \tag{33}$$

s.t.

$$\sum_{k \in \mathcal{K}} y_k^s = 1, \tag{34}$$

$$\sum_{k \in \mathcal{K}} \sum_{s \in S} y_k^s \leq |\eta_s|, \tag{35}$$

$$\sum_{k \in \mathcal{K}} \sum_{s \in S} \nu_{n,k}^s \leq |\eta_u|, \tag{36}$$

$$\nu_{n,k}^s \geq y_k^s a_n^s. \tag{37}$$

Constraint (34) ensures each configuration is assigned to one time slot. Constraints (35) and (36) guarantee each vehicular user can take up η_s sub-channels at most, and one sub-channel can be assigned to η_u users at most. The relationship between the sub-channel and the node is demonstrated in constraint (37). It indicates that sub-channel k should be assigned to configuration s and node n is in configuration s.

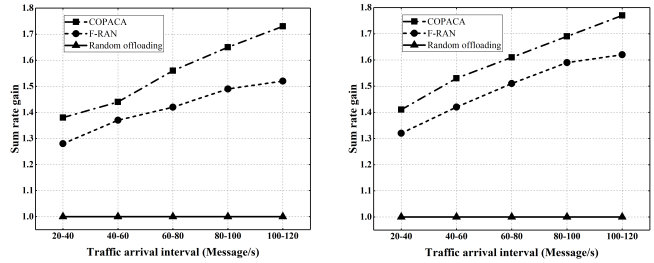
Since the formulated sub-channel problem is with a high computational complexity, we put forward a heuristic method to assign sub-channels for vehicular users. We first define the minimum and maximum numbers of users to access into the sub-channel as $\psi/\max(\eta_s, \eta_u)$ and $\psi/\min(\eta_s, \eta_u)$, respectively, where ψ is the total number of users accessed in the sub-channel. The initial value of $\varphi_{n,k}$ is set to $\psi/\max(\eta_s, \eta_u)$. If we can find a suitable allocation satisfying network constraints (e.g., the SINR threshold, and transmission power sub-channel assignment constraints), $\varphi_{n,k}$ increases by 1. Otherwise, the output is the number of sub-channel k accessed by user n.

V. Performance Evaluation

This section evaluates the performance of our presented CO-PACA. In order to validate network performance, we leverage a real-world trajectory of taxis in Shanghai, China in April 2015, including the information of GPS locations, speed and record time. According to the administrative division of Shanghai [29], its downtown has seven regions. Putuo and Huangpu districts are selected as examples, where RSUs are deployed randomly. The communication ranges of RSU and vehicle are set to 300m and 50m, respectively. The size of message is set between 40 MB and 500 MB. The CPU frequency of the vehicular users ranges between 1.5GHz and 2.5GHz, while that of the RSU is between 5GHz and 10GHz. The CPU frequency of the macro-cell is set to 30GHz. According to the AT&T's prepaid plan (https://www.att.com/prepaid/plans.html), we set the fee caused by macro-cell is $0.007 per MB. The bandwidth of edge devices is 10 Mb. Each simulation is run 100 times to obtain the average performance value.

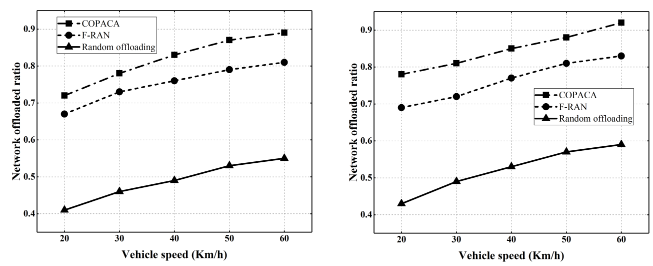
Two existing schemes are utilized for comparison, NOMA-based F-RAN [11] and random offloading. For F-RAN, it jointly considers the obtained transmission rate, interference and edge cache. However, the interaction between edge nodes and the macro-cell has not well been investigated, especially the V2V based predictive offloading has not been considered.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2892767, IEEE Transactions on Industrial Informatics



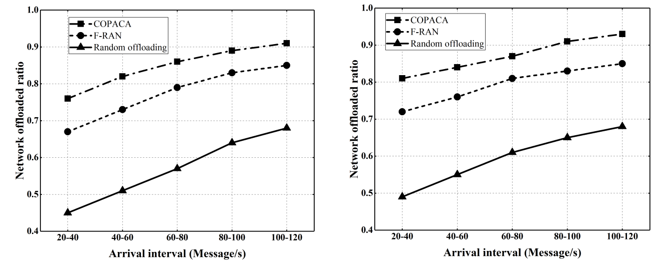
(a) Putuo district (b) Huangpu district

Fig. 3: Sum rate gain with different traffic arrival intervals.



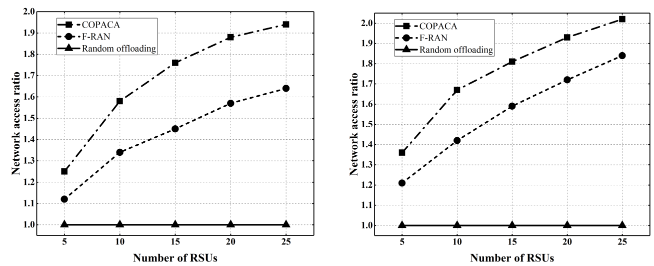
(a) Putuo district (b) Huangpu district

Fig. 4: Network offloaded ratio with different vehicle speeds.



(a) Putuo district (b) Huangpu district

Fig. 5: Network offloaded ratio with different arrival intervals.



(a) Putuo district (b) Huangpu district

Fig. 6: Network access ratio with different number of RSUs.

The following performance metrics are evaluated:

- Sum rate gain: the ratio between the achieved sum rate obtained by different schemes and that achieved by the random offloading scheme.
- Network offloaded ratio: percentage of traffic that can be offloaded by either the macro-cell or RSUs.
- Network access ratio: the ratio between messages uploaded by RSUs and those uploaded by the random offloading scheme.

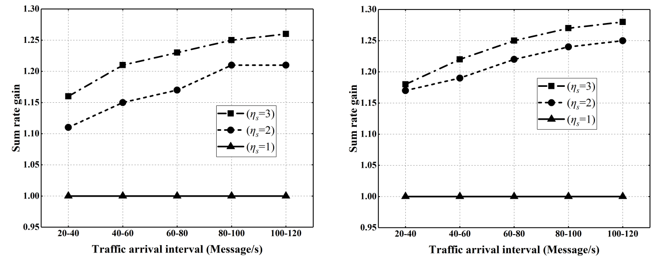
Four major network factors, i.e., traffic arrival intervals, number of sub-channels, the vehicle speed, and the number of RSUs are taken into consideration. We first fix the number of sub-channels to 2, the vehicle speed to 40Km/h, and the number of RSUs to 15. The sum rate gain with different traffic arrival intervals in Putuo and Huangpu districts is illustrated in Fig. 3. It is distinct that our COPACA outperforms F-RAN scheme. For example, when the traffic arrival rate is 80-100 message/s, the achieved sum-rate gain by COPACA is around 10% higher than that in F-RAN scheme in Fig. 3(a). The reason is that V2V-based offloading in COPACA is considered by leveraging the clustering-based traffic management scheme. However, vehicles in F-RAN scheme do not upload messages cooperatively, reducing the potential sum rate gain.

Then we fix the number of sub-channels to 2, the traffic arrival rate to 60-80 message/s, and the number of RSUs to 15. Network offloaded ratio with different vehicle speeds is shown in Fig. 4. Compared with the random offloading scheme, more traffic can be offloaded by both COPACA and F-RAN schemes. This is because these two schemes jointly consider computation offloading and power allocation, so that the traffic can be offloaded in a suitable manner. As the vehicle speed increases, the network offloaded ratio of these three methods increase constantly. The reason is that a high speed results in little time spent for vehicle traveling, vehicles can access distant RSUs with short time, and the delay constraint of message uploading is likely to be satisfied. If a vehicle travels with a low speed, merely few edge devices can be connected.

After that, we fix the number of sub-channels to 2, the vehicle speed to 40Km/h, and the number of RSUs to 15. Network offloaded ratio with different arrival intervals is demonstrated in Fig. 5. We can see the trends in Fig. 5 are similar with those in Fig. 4 when the number of arrival intervals increases. We can also observe that the performance gained in Huangpu district is slightly better than that in Putuo district. This is mainly due to the statistics of traffic flows and vehicles in distinct areas.

Next, we fix the number of sub-channels to 2, the vehicle speed to 40Km/h, and the arrival interval to 60-80 message/s. Network access ratio with different number of RSUs is shown in Fig. 6. Obviously, more traffic tends to be offloaded by the RSUs instead of the macro-cell as the number of RSUs increases. Furthermore, as the coverage of RSUs becomes large, more traffic can be uploaded to the RSUs with a low cost, which can alleviate the bandwidth of the macro-cell effectively.

Finally, we fix the vehicle speed to 40Km/h and the number of RSUs to 15. The sum rate gain obtained by our COPACA with different number of sub-channels is illustrated in Fig. 7, where $\eta_s = 1$ corresponds to the OFDMA-based solution. As the number of traffic arrival interval increases, the sum rate of COPACA with different number of sub-channel increases,



(a) Putuo district (b) Huangpu district

Fig. 7: Sum rate gain with different number of sub-channels.

illustrating our scheme can provide good network performance with dense network traffic. It is noted that the gap between $\eta_s = 1$ and $\eta_s = 2$ is large, while the counterpart between $\eta_s = 2$ and $\eta_s = 3$ is not such large. Due to the constraints of hardware and delay, η_s is generally not set larger than 3.

VI. Conclusion

IoVs is a fundamental research field for industrial IoTs, and is promising to pave the way for traffic management by industrial automation. This paper investigates a joint computation offloading, power allocation and channel assignment scheme for 5G-enabled traffic management systems, with the purpose of maximizing the achievable sum rate. Since the formulated problem is NP-hard, it is separated into three subproblems. For the dynamic offloading decision between the macro-cell and V2V, the maximum tolerable transmission delay is theoretically derived. After that, distributed power allocation and channel assignment schemes are specified. To our best knowledge, this is a prior effort for establishing a 5G-enabled traffic management system, which supports heterogeneous network access and offloading selection. Real-world dataset based performance demonstrates the effectiveness of our scheme from the aspects of the sum rate gain, the offloaded ratio and the access ratio.

VII. Acknowledgments

This work has been partially supported by the Brazilian National Council for Research and Development (CNPq) via Grant No. 309335/2017-5, by National Funding from the FCT-Fundação para a Ciência e a Tecnologia through the UID/EEA/500008/2013 Project, by Finep, with resources from Funttel, grant No. 01.14.0231.00 under the Radiocommunication Reference Center (Centro de Referência em Radiocomunicações CRR) project of the National Institute of Telecommunications (Instituto Nacional de Telecomunicações Inatel), Brazil, by the National Natural Science Foundation of China under grant 61502075 and 61572106, by China Postdoctoral Science Foundation under grant 2018T110210, by State Key Laboratory for Novel Software Technology, Nanjing University (KFKT2018B04), and by Dalian Science and Technology Innovation Fund under Grant No. 2018J12GX048.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020," *Cisco White paper*, 2016.

[2] Z. Ning, J. Huang, and X. Wang, "Vehicular fog computing: Enabling real-time traffic management for smart cities," *IEEE Wireless Communications, Doi: 10.1109/MWC.2018.1700441*, 2018.

[3] Z. Zhou, F. Xiong, C. Xu, Y. He, and S. Mumtaz, "Energy-efficient vehicular heterogeneous networks for green cities," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1522–1531, 2018.

[4] X. Wang, Z. Ning, and L. Wang, "Offloading in Internet of vehicles: A fog-enabled real-time traffic management system," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4568–4578, 2018.

[5] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Bjrnson, K. Yang, and I. Chih-Lin, "Millimeter wave communications for future mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1909–1935, 2017.

[6] Z. Ning, X. Wang, and J. Huang, "Non-orthogonal multiple access for mobile edge computing enabled vehicular networks," *IEEE Vehicular Technology Magazine, Doi: 10.1109/MVT.2018.2882873*, 2018.

[7] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7686–7698, 2016.

[8] S. Mumtaz, B. Ai, A. Al-Dulaimi, and K. F. Tsang, "5G and beyond mobile technologies and applications for industrial IoT (IIoT)," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 6, pp. 2588–2591, 2018.

[9] Y. Ai, L. Wang, B. Jiao, and K. C. Chen, "Exploiting NOMA into socially enabled computation offloading," in *International Conference on Wireless Communications and Signal Processing*, pp. 1–6, 2017.

[10] L. Yang, H. Zhang, M. Li, J. Guo, and H. Ji, "Mobile edge computing empowered energy efficient task offloading in 5G," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6398–6409, 2018.

[11] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource allocation in NOMA-based fog radio access networks," *IEEE Wireless Communications*, vol. 25, no. 3, 2018.

[12] K. Zhang, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Cooperative content caching in 5G networks with mobile edge computing," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 80–87, 2018.

[13] Y. Liu, R. Yu, X. Li, H. Ji, and V. Leung, "Hybrid computation offloading in fog and cloud networks with non-orthogonal multiple access," in *IEEE INFOCOM WKSHPS*, pp. 154–159, 2018.

[14] J. Du, L. Zhao, J. Feng, and X. Chu, "Computation offloading and resource allocation in mixed fog/cloud computing systems with min-max fairness guarantee," *IEEE Transactions on Communications*, vol. 66, no. 4, pp. 1594–1608, 2018.

[15] X. Wang, Z. Ning, X. Hu, L. Wang, J. Cheng, and V. Leung, "Optimizing content dissemination for real-time traffic management in large-scale internet of vehicle systems," *IEEE Transactions on Vehicular Technology, Doi: 10.1109/TVT.2018.2886010*, 2018.

[16] Z. Zhou, H. Yu, C. Xu, Y. Zhang, S. Mumtaz, and J. Rodriguez, "Dependable content distribution in D2D-based cooperative vehicular networks: A big data-integrated coalition game approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 953–964, 2018.

[17] C. Wu, T. Yoshinaga, X. Chen, L. Zhang, and Y. Ji, "Cluster-based content distribution integrating LTE and IEEE 802.11p with fuzzy logic and Q-learning," *IEEE Computational Intelligence Magazine*, vol. 13, no. 1, pp. 41–50, 2018.

[18] J. C. Liberti, S. Moshavi, and P. G. Zablocky, *Successive interference cancellation*. U.S. Patent 8,670,418, 2014.

[19] T. Ng and W. Yu, "Joint optimization of relay strategies and resource allocations in cooperative cellular networks," *IEEE Journal on Selected areas in Communications*, vol. 25, no. 2, pp. 328–339, 2007.

[20] A. Dua, N. Bulusu, W. C. Feng, and W. Hu, "Towards trustworthy participatory sensing," in *Usenix Conference on Hot Topics in Security*, pp. 8–8, 2009.

[21] J. He, L. Cai, P. Cheng, and J. Pan, "Delay minimization for data dissemination in large-scale vanets with buses and taxis," *IEEE Transactions on Mobile Computing*, vol. 15, no. 8, pp. 1939–1950, 2016.

[22] L.-W. Chen and C.-C. Chang, "Cooperative traffic control with green wave coordination for multiple intersections based on the internet of vehicles," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 7, pp. 1321–1335, 2017.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2892767, IEEE Transactions on Industrial Informatics

[23] R. Atallah, M. Khabbaz, and C. Assi, "Multi-hop V2I communications: A feasibility study, modelling and performance analysis," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2801–2810, 2017.

[24] J. Y. Yen, "An algorithm for finding shortest routes from all source nodes to a given destination in general networks," *Quarterly of Applied Mathematics*, vol. 27, no. 4, p. 526, 1970.

[25] J. J. Glvez and P. M. Ruiz, "Joint link rate allocation, routing and channel assignment in multi-rate multi-channel wireless networks," *Ad Hoc Networks*, vol. 29, pp. 78–98, 2015.

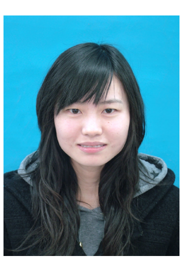
[26] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2238–2252, 2017.

[27] Z. Ning, F. Xia, X. Hu, Z. Chen, and M. S. Obaidat, "Social-oriented adaptive transmission in opportunistic Internet of smartphones," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 810–820, 2017.

[28] S. Mumtaz, K. M. Saidul Huq, and J. Rodriguez, "Direct mobile-to-mobile communication: Paradigm for 5G," *Wireless Communications IEEE*, vol. 21, no. 5, pp. 14–23, 2014.

[29] "List of administrative divisions of Shanghai." https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Shanghai.

Zhaolong Ning (M'14-SM'18) received the M.S. and PhD degrees in 2011 and 2014 respectively from Northeastern University, Shenyang, China. He was a Research Fellow at Kyushu University from 2013 to 2014, Japan. He is an associate professor in the School of Software, Dalian University of Technology, China. Dr. Ning has published 2 books and over 80 scientific papers in international journals and conferences. His research interests include mobile edge computing, vehicular network, and network optimization.



Xiaojie Wang received the M.S. degree from Northeastern University, China, in 2011. From 2011 to 2015, she was a software engineer in NeuSoft Corporation, China. Currently, she is pursuing the Ph.D. degree with the School of Software, Dalian University of Technology, Dalian, China. Her research interests are vehicular social networks, edge computing and network optimization.

Joel J. P. C. Rodrigues (S'01-M'06-SM'06) is currently a Professor and a Senior Researcher with the National Institute of Telecommunications, Brazil, and a Senior Researcher with the Instituto de Telecomunicaes, Portugal. He has been a Professor with the University of Beira Interior, Portugal, and a Visiting Professor with the University of Fortaleza, Brazil. He is also the Leader of the Internet of Things Research Group, CNPq. He has authored or co-authored over 550 papers in refereed international journals and conferences, three books, and two patents. He is a member of the IEEE ComSoc Board of Governors as the Director for Conference Development, an IEEE ComSoc Distinguished Lecturer, the President of the Scientific Council at ParkUrbis, Covilh Science and Technology Park, the past Chair of the IEEE ComSoc Technical Committee on eHealth, the past Chair of the IEEE ComSoc Technical Committee on Communications Software, a Steering Committee Member of the IEEE Life Sciences Technical Community and the Publications Co-Chair, and a Member Representative of the IEEE Communications Society on the IEEE Biometrics Council. He is a member of many international TPCs and participated in several international conferences organization. He is an Editorial Board Member of several high-reputed journals. He had been awarded several Outstanding Leadership and Outstanding Service Awards by the IEEE Communications Society and several best papers awards. He is a Licensed Professional Engineer (as a Senior Member). He is the Editor-in Chief of three international journals.

Feng Xia (M'07-SM'12) received the BSc and PhD degrees from Zhejiang University, Hangzhou, China. He was a Research Fellow at Queensland University of Technology, Australia. He is currently a Full Professor in School of Software, Dalian University of Technology, China. He is the (Guest) Editor of several international journals. He serves as General Chair, PC Chair, Workshop Chair, or Publicity Chair of a number of conferences. Dr. Xia has published 2 books and over 200 scientific papers in international journals and conferences. His research interests include computational social science, big data, and mobile social networks.