IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON WIRELESS CACHIING TECHNIQUE FOR 5G

Received February 7, 2017, accepted April 7, 2017, date of publication April 24, 2017, date of current version June 7, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2693440

Social-Aware Edge Caching in Fog Radio Access Networks

XIANG WANG[1], SUPENG LENG[1], (Member, IEEE), and KUN YANG[2], (Senior Member, IEEE)
[1]School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K.

Corresponding author: Supeng Leng (spleng@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61374189, in part by the Fundamental Research Funds for the Central Universities, China, under Grant ZYGX2016Z011, in part by the EU FP7 Project CLIMBER under Grant PIRSES-GA-2012-318939, in part by the EU FP7 Project CROWN under Grant GA-2013-610524, in part by the UK EPSRC Project NIRVANA under Grant EP/L026031/1, in part by the UK EPSRC Project DANCER under Grant EP/K002643/1, and in part by the Program of the China Scholarship Council under Grant 201506070049.

ABSTRACT Fog radio access networks (F-RANs) are becoming an emerging and promising paradigm for fifth generation cellular communication systems. In F-RANs, distributed edge caching techniques among remote radio heads (RRHs) and user equipment (UE) can effectively alleviate the burdens on the fronthaul toward the base band unit pool and the bandwidth of the RANs. However, it is still not clear as to how social relationships affect the performance of edge caching schemes. This paper attempts to analyze the impact of mobile social networks (MSNs) on the performance of edge caching in F-RANs. We propose a Markov-chain-based model to analyze edge caching among edge nodes (i.e., RRHs and MSNs), as well as data sharing among the potential MSNs from the viewpoint of content diffusion in the F-RANs. Moreover, we analyze the edge caching schemes among UE to minimize the bandwidth consumption in the RANs. Finally, the optimal edge caching strategies among RRHs in terms of caching locations and time are introduced to minimize the bandwidth consumption of fronthaul and storage costs in the F-RANs. Simulation results show that the proposed edge caching schemes among UE and RRHs are able to reduce the bandwidth consumption of RANs and fronthaul effectively.

INDEX TERMS Edge caching, fog radio access network, mobile social network, Markov chain, network intervention.

I. INTRODUCTION

With the rapid advancement of wireless network technologies and the wide use of mobile phones, many efforts have been made in both industry and academia to design 5th-generation cellular communication systems. Among these research activities, F-RANs have been proposed as a promising paradigm for improving spectral efficiency (SE) for 5G systems by fully utilizing fog computing and cloud radio access networks (C-RANs) [1]. In F-RANs, the edge nodes, such as remote radio heads (RRHs) and user equipments (UEs), are capable of local signal processing, cooperative radio resource management, and distributed caching. It is observed that the emergence of distributed edge caching and delivery techniques through the edge nodes can effectively alleviate the burdens on the fronthaul toward the BBU pool and the bandwidth of the radio access networks [2]. Since a large amount of mobile media traffic is caused by the sharing of popular content (e.g., popular music and videos) through the social socialities or among the same interest groups [3]–[5], it is essential to analyze mobile user behaviors in terms of data sharing from social perspectives. The interests and the social ties of mobile users can be used to predict the data sharing among them [6]–[8], which is helpful in determining the optimal edge caching schemes among UEs. In addition, the user clustering characteristics in certain popular locations [9]–[11], such as shopping centers, public transportation hubs, and workplaces, can be utilized to decide the caching locations among RRHs, from which most UEs tend to download popular content. More specifically, as shown in Fig. 1, if popular content is cached in the edge nodes (e.g., UEs and RRHs), the demands from different users for the same content can be satisfied easily without duplicate transmissions from the remote content center (RCC) in the cloud. According to the interests and the social ties of mobile users, the content can be stored in proper UEs and shared to their neighbors in a proactive or reactive

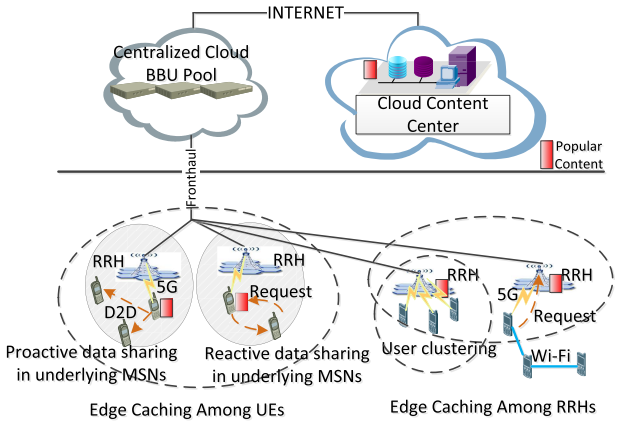

FIGURE 1. System architecture of F-RAN including edge caching among UEs and RRHs.

manner. In addition, the content can be stored among the RRHs in certain popular locations where users tend to cluster.

However, existing research works ignore the impact of social relationships on the performance of edge caching schemes in terms of content diffusion. We lack a model to analyze the content diffusion characteristics using edge caching in a 5G-based MSN. Moreover, network operators usually determine their edge caching scheme according to the popularity of the content. Nonetheless, the potential content sharing can actually be conducted in the potential MSNs. For instance, users are usually willing to share content with their friends via inexpensive wireless links (e.g., Wi-Fi and Bluetooth), which reduces the traffic over cellular networks. Consequently, the impacts of social ties and behaviors of UEs are considered in the design of edge caching schemes for efficient content diffusion in an F-RAN.

This paper attempts to model the behaviors of mobile users and content sharing patterns in an F-RAN. We study the impact of social relationships on the performance of edge caching schemes for the diffusion of popular content. In addition, with the cooperation of UEs in content sharing and caching, an edge caching scheme among UEs is proposed to minimize the bandwidth consumption in an F-RAN. Moreover, considering the content sharing in the potential MSNs, an edge caching scheme among RRHs is proposed to determine the optimal edge caching location and time for each RRH. The main contributions of this paper can be summarized as follows.

1) By modeling the process of content diffusion, we analyze the impact of edge caching among edge nodes (i.e., RRHs and UEs) and MSN-based data sharing on the performance of content diffusion in an F-RAN. In contrast to existing work, the proposed model can be used to compute the expectation of bandwidth consumption of an RAN and fronthaul with edge caching, as well as the corresponding content diffusion ratio in complicated scenarios in social-aware F-RANs.
2) Based on the proposed model, we formulate the optimal edge caching scheme among UEs to minimize the bandwidth consumption in an RAN. Suboptimal solutions with low complexity are proposed to address the optimization problem, and they are verified through simulation experiments.
3) Considering content sharing among UEs with social ties and the storage cost induced by edge caching among RRHs, the edge caching strategies among RRHs in terms of caching locations and time are introduced to minimize the bandwidth consumption of fronthaul induced by content diffusion in an F-RAN.

The remainder of this paper is organized as follows. Section 2 reviews some related works about edge caching. Section 3 presents an analysis model for edge caching in a 5G network. Section 4 presents the edge caching schemes among UEs to minimize the bandwidth consumption of an RAN. Section 5 provides an edge caching scheme among RRHs to minimize the bandwidth consumption of fronthaul and the storage cost. In section 6, the simulation results demonstrate the performance of the proposed solutions. Finally, Section 7 concludes the paper.

Notations: All uppercase boldface letters represent sets and matrices. Let tr(X), det(X), X^{-1} and X^H denote the trace, determinant, inverse and hermitian of a symmetric matrix X, respectively. Let $\mathbb{C}$ denote the set of complex and real matrices of size $x \times 1$. All letters at the right of different variables can be explained as follows: k represents the k-th slot, and m represents the different users.

II. RELATED WORKS

The core idea of F-RANs is to fully utilize local radio signal processing, data sharing, and storing capabilities in edge devices, in which the edge caching among edge nodes (such as UEs and RRHs) in radio access networks can efficiently decrease the heavy burden on RANs and fronthaul.

There are certain works focusing on designing edge caching schemes or algorithms to enhance the performance of F-RANs, in terms of the bandwidth consumption in content diffusion, the power consumption and energy efficiency of the network, and the security of transmission. In this paper, we focus on analyzing the bandwidth consumption of content diffusion in F-RANs from social perspectives. The paper in [2] summarized the recent advances in the performance analysis of F-RANs, where advanced edge cache and model selection schemes were introduced to improve the spectral efficiency and energy efficiency of the F-RANs. The scheme in [12] utilized social information and edge computing to efficiently decrease the end-to-end latency, where the cache of Internet contents, mobility management, and radio access control were studied. The work in [13] presented an information-theoretic framework to character the main trade-offs between the performance of F-RANs, in terms of delivery latency, and its resources, including caching and fronthaul capacities. The paper in [14] studied the joint design of cloud and edge processing for the downlink of F-RANs, where popular content caching strategies among RRHs were designed to maximize the delivery rate under fronthaul capacity and

RRH power constraints. In addition, the work in [15] proposed a collaborative strategy to implement caching in infrastructure and in mobile devices with D2D communication simultaneously. The paper [16] investigated the techniques related to caching in current mobile networks and discussed potential techniques for caching in 5G mobile networks, including RAN caching and evolved packet core (EPC) network caching. A novel edge caching scheme based on content-centric networking for information-centric networking has been proposed. The paper in [18] presented a content-centric-based network architecture consisting of UEs, communities, content centric nodes, small cells, and macro cells, and a caching scheme was presented to store replicas of mobile content.

However, to the best of our knowledge, none of the existing works provides an analysis model for edge caching in social-aware F-RANs. In addition, the advantage and interference of MSNs on the edge caching protocol design are usually neglected when reducing the bandwidth consumption of RANs and fronthaul in an F-RAN.

III. SYSTEM MODEL

In this section, we attempt to model the total content diffusion process by integrating edge caching and content sharing in social-aware F-RANs. With the proposed model, we are able to compute the expectation of the bandwidth consumption of RANs and fronthaul with edge caching in an F-RAN, as well as the corresponding content diffusion ratio prior to a previously defined deadline.

We consider an F-RAN that is composed of n UEs, $U = \{1, 2, \ldots, n\}$, and L RRHs, $R = \{1, 2, \ldots, L\}$. In the F-RAN, the popular content can be cached in the edge nodes (i.e., UEs and RRHs) or the remote cloud content center, as shown in Fig. 1. Each UE caching popular content will share the content with its neighbors who are interested in the content. In addition, the UEs can also access the content from the RRHs or the RCC.

In the system model, as in previous works, we assume that the arrival time t_r of the content access request follows an independent exponential distribution for each UE i ($i \in U$) [19], [20], denoted by $t_r \sim Exp(\lambda_{iR})$, where λ_{iR} indicates the frequency that the user i accesses the popular content. This can be calculated as the reciprocal of the statistic average time interval between two continuous accesses to the same type of content in history. In addition, we assume that encounters between node pairs follow the Poisson process according to the social tie strength between the nodes. In other words, the inter-contact time t_c between each pair of UEs follows an exponential distribution, denoted by $t_c \sim Exp(\lambda_{ij})$, where $i \neq j$ and $i, j \in U$.

Based on those assumptions, we model the process of content diffusion in social-aware F-RANs as a continuous Markov chain. A simple example of the state transition graph is shown in Fig. 2. In the Markov chain, each state m is denoted by $s_m : (s_m^1, \ldots, s_m^n)$, where $s_m^i \in \{0, 1\}$ indicates whether user i has already accessed the content in state s_m.

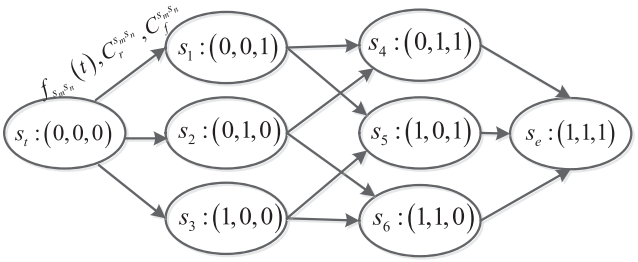

FIGURE 2. An example of the state transition graph of content diffusion when $n = 3$.

The state space is denoted by S, the size of which can be calculated by $|S| = 2^n$. State transitions are caused by the increase in the number of UEs who have already received the content. Then, the one-step state transition matrix A can be denoted by

$$A = \begin{bmatrix} p_{00}^m & \cdots & p_{0(L-1)}^m \\ \vdots & \ddots & \vdots \\ p_{(L-1)0}^m & \cdots & p_{(L-1)(L-1)}^m \end{bmatrix} \tag{1}$$

where element $p_{s_m s_n}$ is the probability that user i transmits from the state s_m to its next state s_n and can be calculated by **Theorem 1**.

Theorem 1: Given the state space S, the arrival time distribution of content access requests $t_r \sim Exp(\lambda_{iR})$ for each UE $i \in U$, and the inter-contact time $t_c \sim Exp(\lambda_{ij})$ between two UEs, the probability that user i transmits from the state s_m to its next state s_n can be calculated by

$$p_{s_m s_n} = \begin{cases} \dfrac{\left(\lambda_{jR} + \sum\limits_{i \in U_1(s_m)} \lambda_{ij}\right)\Big|_{j \in O_{s_m s_n}}}{\sum\limits_{k \in U_2(s_m)} \left(\lambda_{kR} + \sum\limits_{i \in U_1(s_m)} \lambda_{ik}\right)}, & if \ \left|O_{s_m s_n}\right| = 1 \\ 0, & otherwise \end{cases} \tag{2}$$

where

$$O_{s_m s_n} = \left\{ i \,\middle|\, s_m^i \neq s_n^i, i \in U \right\} \tag{3}$$

$$U_1(s_m) = \left\{ i \,\middle|\, s_m^i = 1, i \in U \right\} \tag{4}$$

$$U_2(s_m) = \left\{ i \,\middle|\, s_m^i = 0, i \in U \right\} \tag{5}$$

The $O_{s_m s_n}$ is the set of UEs that received the content from the edge nodes or the remote content center by transmitting the state from s_m to s_n. $U_1(s_m)$ indicates the set of UEs that have already received the content in state s_m, while $U_2(s_m)$ indicates the set of UEs that have still not received the content in state s_m.

Proof: See Appendix A. □

Based on Theorem 1, the probability density function (pdf) of each state transition, denoted by $f_{s_m s_n}$ ($s_m, s_n \in S$), can be calculated by

$$f_{s_m s_n}(t) = d_{s_m}(t) \cdot p_{s_m s_n} \tag{6}$$

where $d_{s_m}(t)$ is the pdf of the dwelling time t that the state will sojourn in its current state s_m, i.e., none of UEs will receive or access the content until time t, and can be derived by

$$d_{s_m}(t) = \lambda'_{s_m} \cdot e^{-\lambda'_{s_m} t} \tag{7}$$

where

$$\lambda'_{s_m} = \sum_{k \in U_2(s_m)} \left(\lambda_{kR} + \sum_{i \in U_1(s_m)} \lambda_{ik} \right) \tag{8}$$

Consequently, according to Eq.(2), Eq.(7), and Eq.(8), the function $f_{s_m s_n}(t)$ in Eq.(6) can be expressed by

$$f_{s_m s_n}(t) = \begin{cases} e^{-\lambda'_{s_m} t} \cdot \left(\lambda_{jR} + \sum_{i \in U_1(s_m)} \lambda_{ij} \right) \Big|_{j = O_{s_m s_n}}, & if \ \left| O_{s_m s_n} \right| = 1 \\ 0, & otherwise \end{cases} \tag{9}$$

There are three possible paths for each UE to access content: UE-UE, RRH-UE, and RCC-RRH-UE. As shown in Fig. 1, the path UE-UE indicates that the UE can receive the content from its neighbors in a proactive content sharing or reactive content sharing manner, RRH-UE indicates that the UE can receive the content from RRHs with RAN, and RCC-RRH-UE indicates that the UE can access the content stored in the remote content center. It is assumed that there exist priorities among the three possible paths, the relationship characterizing their priorities is UE-UE>RRH-UE>RCC-RRH-UE. In other words, to access the content, the UE will select the UE-UE path with the highest priority if there exists at least one UE caching the content among its neighbors; otherwise, it will select the RRH-UE path if at least one of its neighbor RRHs are caching the content. If none of its neighbor edge nodes, i.e., UEs and RRHs, are caching the content, the UE will select the RCC-RRH-UE path to access to the remote content center.

Then, we derive the probability of each possible path that the UE will select to receive the content. For each possible state transition from s_m to s_n (i.e., $p_{s_m s_n} > 0$), assume that the UE i received the content at time t and results in the state transition (i.e., $i = O_{s_m s_n}$). The probability that the received content originates from the other UEs, denoted by $p^{ue-ue}_{s_m s_n}$, can be calculated by

$$p^{ue-ue}_{s_m s_n} = \int_0^\infty \sum_{j \in U_1(s_m)} \lambda_{ij} \cdot e^{-\left(\lambda_{iR} + \sum_{j \in U_1(sm)} \lambda_{ij} \right) t} \tag{10}$$

Consequently, its complementary event is that the received content is received from RRHs or the RCC, denoted by $p^{rrh-ue}_{s_m s_n}$ and $p^{rcc-ue}_{s_m s_n}$, respectively. Their probabilities depend on the location of UE i and the content edge caching strategy (i.e., the caching locations among RRHs). We employ the steady-state distribution of the location of each UE i among L RRHs, denoted by $\boldsymbol{\pi}^i = \left\{ \pi^i_0, \pi^i_1, \ldots, \pi^i_L \right\}$, to predict its location among RRHs according to its carrier's clustering in the RRHs. The derivation of $\boldsymbol{\pi}^i$ can be found in [21].

In addition, the expectation of the bandwidth consumption of the RANs and fronthaul in each state transition from s_m to s_n can be denoted by $C_r^{s_m s_n}$ and $C_f^{s_m s_n}$, respectively, which are calculated in **Definition 1**.

Definition 1: Given the RRH set when caching the content R_c ($R_c \subseteq R$) and the steady-state distribution of each UE's location among L RRHs $\boldsymbol{\pi}^i$ ($i \in U$), the corresponding expectation of the bandwidth consumption of RANs and fronthaul in the possible state transition from s_m to s_n (i.e., $p_{s_m s_n} > 0$) can be denoted by

$$C_r^{s_m s_n} = D \cdot u_{D2D} \cdot p^{ue-ue}_{s_m s_n} + D \cdot u_r \cdot \left(1 - p^{ue-ue}_{s_m s_n} \right) \tag{11}$$

$$C_f^{s_m s_n} = D \cdot u_f \cdot \left(1 - p^{ue-ue}_{s_m s_n} \right) \cdot \sum_{l \in \{R - R_c\}} \pi_l^{O_{s_m s_n}} \tag{12}$$

where D denotes the size of the popular content (in bytes) and u_{D2D}, u_r and u_f denote the unit bandwidth consumption for transmitting one byte of data over D2D, RANs and backhaul (in Hz per byte).

After the derivation of the bandwidth consumptions of each state transition, we attempt to calculate the expectation of the bandwidth consumption for the overall content diffusion process. We first derive the probability of the possible sequence of the Markov chain, the corresponding bandwidth consumption and the probability density function of delay in **Lemma 2**.

Lemma 2: Given an arbitrary possible sequence of a Markov chain in the overall content diffusion process, denoted by $s = \{s_1, \ldots, s_k\}$, and the corresponding state transition time $\{t_1, \ldots, t_k\}$, where $k = \left| O_{s_t s_e} \right|$, s_t is the initial state and s_e is the destination state, the probability of the sequence of the Markov chain p_{seq}, the bandwidth consumption, and its corresponding probability density function of the delay $f(T)$ can be respectively calculated as follows:

$$p_s = \prod_{i=1}^{k-1} p_{s_i s_{i+1}} \tag{13}$$

$$C_r^s = \sum_{i=1}^{k-1} C_r^{s_i s_{i+1}} \tag{14}$$

$$C_f^s = \sum_{i=1}^{k-1} C_f^{s_i s_{i+1}} \tag{15}$$

$$f_s(T) = \left[\prod_{i=1}^{k-1} \lambda'_{s_i} \right] \sum_{j=1}^{k-1} \frac{e^{-\lambda'_{s_j} T}}{\prod_{q \neq j, q=1}^{k-1} \left(\lambda'_{s_q} - \lambda'_{s_j} \right)} \tag{16}$$

Proof: See Appendix B. □

Consequently, the expectation of the bandwidth consumption of RANs and fronthaul for the overall content diffusion process as well as the diffusion ratio before a previously specified deadline T_{th} can be derived using **Theorem 3**.

Theorem 3: Based on Lemma 2, given the state transition graph g and both the arbitrary initial state s_t and the destination state s_e : $(1, 1, \ldots, 1)$, the expectation of the bandwidth consumption of RANs and fronthaul and the diffusion ratio before a previously specified deadline T_{th} for content diffusion from the state s_t to s_e, denoted by $E(C_r)$, $E(C_f)$, and $p(T_{th})$, respectively, can be derived as follows:

$$E_{s_t s_e}(C_r) = \sum_{s \in M} p_s \cdot C_r^s \tag{17}$$

$$E_{s_t s_e}(C_f) = \sum_{s \in M} p_s \cdot C_f^s \tag{18}$$

$$p_{s_t s_e}(T_{th}) = \sum_{s \in M} p_s \cdot \int_0^{T_{th}} f_s(t)\, dt \tag{19}$$

where M is the consequence set that consists of all the paths between the initial node s_t and the destination node s_e in the state transition graph g.

Proof: See Appendix C. □

Consequently, based on the proposed model, given an arbitrary initial state and an arbitrary destination state of content diffusion in an F-RAN, we can compute the expectation of the bandwidth consumption of RANs and fronthaul with edge caching as well as the corresponding content diffusion ratio before a deadline. In the following two sections, the edge caching schemes among edge nodes (i.e., UEs and RRHs) are proposed based on the above proposed model.

IV. EDGE CACHING AMONG UEs

In this section, we analyze the edge caching among UEs with their cooperation for content caching and sharing to minimize the bandwidth consumption of RANs. Note that the bandwidth consumption of an RAN is only related to the edge caching scheme among UEs, while the bandwidth consumption of fronthaul is related to the edge caching scheme among UEs and RRHs. Consequently, in the remainder of this section, we first analyze the optimal edge caching schemes among UEs, based on which the optimal edge caching among RRHs will be analyzed later.

The cellular network operator usually takes advantage of the edge caching among UEs to reduce the bandwidth consumption of RANs by initially transmitting the content to a set of UEs and entrusting them to cache and share the content to the other UEs. However, the initial edge caching among UEs can also increase the bandwidth consumption of the RANs. Consider the following two extreme cases:

Case 1: The cellular network directly sends and caches the content to only one UE and entrusts him to disseminate the content to the other UEs. In this case, the initial bandwidth consumption for edge caching among UEs can be minimized, but the performance in terms of the delivery ratio and bandwidth reduction is poor.

Case 2: The cellular network directly sends and caches the content to all the UEs who are interested in the content. In this case, the bandwidth consumption is the highest, even though the performance can be ensured.

Both of the above cases are sub-optimal in terms of minimizing the RAN's bandwidth consumption and ensuring the performance of content diffusion. Consequently, we seek a tradeoff between the bandwidth consumption of RANs and the performance of content diffusion in an F-RAN. Based on the above proposed model in Section 3, we formulate an optimization problem to determine the optimal initial state s_t^* in the proposed model (i.e., the operator's best strategy for edge caching among UEs) to minimize the bandwidth consumption of an RAN during content diffusion under the constraint that the content diffusion ratio before the deadline T_{th} is larger than a previously defined threshold p_{th}. The optimization problem can be expressed as

$$\begin{aligned} s_t^* = & \underbrace{\max}_{s_t \in S} E_{s_t s_e}(C_r) \\ & s.t. \quad p_{s_t s_e}(T_{th}) > p_{th} \end{aligned} \tag{20}$$

where $s_e = (1, 1, \ldots, 1)$ indicates that all UEs have successfully received the content. However, the time complexity and space complexity of traversing all the paths from s_t to s_m in graph g are tremendous. Consequently, we attempt to derive the suboptimal solution of Eq.(20). We partition the optimization problem in Eq.(20) into two sub-problems to find its suboptimal solution (as shown in Algorithm 1). In the first sub-problem, given an initial bandwidth consumption of an RAN, we attempt to determine the optimal edge caching among UEs with the highest summation of their capacities for content diffusion. In the second sub-problem, we derive the optimal initial bandwidth consumption of an RAN.

Algorithm 1 Determine Optimal Edge Caching Among UEs

Require: T_{th}, **S**, D, u_r λ_{ij}, $i, j \in U$

Ensure: **The suboptimal edge caching among UEs $s_t^{*'}$.**

Step 1: Calculate the capacity of content diffusion for each UE according to Definition 1;

Step 2: Derive the suboptimal solution of the initial state $s_t^{*'}(B)$ according to Eq.(22) given an initial bandwidth consumption of RAN B;

Step 3: Calculate the optimal value of the initial bandwidth consumption of RAN B according to Eq.(23), where $B \in \{0, Du_r, 2Du_r, \ldots, nDu_r\}$;

return $s_t^{*'} = s_t^{*'}(B^*)$**;**

Since the capacities for content diffusion of each UE are different, to minimize the bandwidth consumption of an RAN, the network operator tends to select the UEs with the highest capacity for content diffusion when the initial bandwidth consumption is limited. To solve the first sub-problem, we first define the capacity for content diffusion for each UE, as shown in **Definition 2**.

Definition 2: Given the deadline of the content T_{th}, the capacity for content diffusion for each UE i is defined by

$$b_i = \sum_{j \in \{U-i\}} \int_0^{T_{th}} \lambda_{ij} e^{-\lambda_{ij} t} \cdot dt \tag{21}$$

Then, the **edge caching rule among UEs** can be defined as follows: given the initial bandwidth consumption of RAN B, $B \in \{0, Du_r, 2Du_r, \ldots, nDu_r\}$, the suboptimal edge caching among UEs with the minimum bandwidth consumption of an RAN in the overall content diffusion process can be denoted by $s_t^{*'}$ and calculated as follows:

$$s_t^{*'}(B) = \arg \underbrace{\max}_{s_t \in S} \sum_{i \in U_1(s_t)} b_i \cdot s_t^i$$
$$s.t.\ |U_1(s_t)| = \frac{B}{D \cdot u_r} \tag{22}$$

where $|U_1(s_t)|$ in the restrictive condition indicates the number of UEs that the network operator can initially transmit to with the given initial bandwidth consumption of the RAN.

Next, we attempt to determine the optimal initial bandwidth consumption of RAN B by solving the second subproblem, which can be derived as

$$B^* = \arg \underbrace{\max}_{B} E_{s_t^{*'}(B) s_e}(C_r)$$
$$s.t.\ B \in \{0, Du_r, 2Du_r, \ldots, nDu_r\} \tag{23}$$

where D indicates the size of the popular content (in bytes) and u_r indicates the unit bandwidth consumption for transmitting one byte of data in an RAN (in Hz per byte).

Consequently, the suboptimal solution $s_t^{*'}$ for the problem in Eq.(20) can be derived as

$$s_t^{*'} = s_t^{*'}(B^*) \tag{24}$$

Complexity Analysis: In Algorithm 1, the computational complexity is limited by the number of UEs n, the size of the state space $|S|$, and the number of edges in the state transition graph. Then, the computational complexity of Algorithm 1 is $O(n^2 + nVE)$, where $V = 2^n$ and $E = 4^n$. The computational complexity of the traversal global optimal algorithm is $O(2^n VE)$, which is much higher than that of our proposed algorithm. This is mainly because the optimization problem is partitioned into two sub-problem with lower computational complexity in our proposed algorithm 1.

V. EDGE CACHING AMONG RRHs

As mentioned in Section 3, there are three possible paths with different priorities for UE to access the content. The UE attempt to select the RRH-UE path to access to the content stored in the RRHs, if none of its neighbor UEs are caching the content. Although the edge caching among RRHs can reduce the bandwidth consumption of fronthaul, the content caching also results in storage resource consumption among RRHs. In this section, we analyze the edge caching among RRHs to minimize the bandwidth consumption of fronthaul and the storage cost of edge caching among RRHs.

To calculate the caching cost of each RRH, we need to first investigate how long the content should be cached in each RRH. We first define a utility function to calculate the difference between the reduced bandwidth by edge caching and the storage cost for each RRH l during unit time, which is defined in **Definition 3**.

Definition 3: Given an arbitrary state s_m, the edge caching utility at the RRH l during time $[t, t + \Delta T]$ can be defined by

$$Y_{s_m}^l(\Delta T, t) = -\alpha \cdot D \cdot \Delta T + \sum_{i \in U_2(s_m)} D \cdot u_f \cdot p_l(i, t) \tag{25}$$

where

$$p_l(i, t) = \pi_l^i \frac{\int_t^{t+T_{D2D}} f_i(t) dt}{1 - \int_0^t f_i(t) dt} \tag{26}$$
$$f_i(t) = e^{-\sum_{j \in U_1(s_m)} \lambda_{ji} t} \cdot \lambda_{iR} e^{-\lambda_{iR} t} \cdot \pi_l^i \tag{27}$$
$$\Delta T = T_{D2D} + T_{RAN} \tag{28}$$

The right part of Eq.(25) indicates the expectation of the reduced bandwidth of fronthaul by edge caching at RRH l during a time ΔT. The left part indicates the storage cost during ΔT, and α ($0 < \alpha < 1$) is a factor used to normalize the utility of the reduced fronthaul bandwidth consumption and the storage cost. P_l is the probability that UE i receives the content from RRH l instead of the other UEs during $[t, t + \Delta T]$, the pdf of which is denoted by $f_i(t)$. ΔT is the summation of the content transmission time of D2D and RRH-UE. In addition, it is assumed that the network operator can dynamically observe the state transition during the overall content diffusion process.

Next, we prove that $Y_{s_m}^l(\Delta T, t)$ is a strictly decreasing function with time and over a state transition, as proved in **Corollary 1**.

Corollary 1: For each RRH, given an arbitrary state s_m, the edge caching utility function $Y_{s_m}^i(\Delta T, t)$ is strictly decreasing with increasing time t before leaving from the state s_m. In addition, it is strictly decreasing over the state transition during content diffusion.

Proof: See Appendix D. □

Consequently, RRH l can determine the caching time of the content according to the edge caching rule in **Definition 4**.

*Definition 4: Edge Caching Time of RRHs: Based on **Corollary 1**, for each RRH l, $l \in R$, popular content should be dropped from the RRH's storage devices when $Y_{s_m}^i(\Delta T, t)$ becomes a negative value or when the deadline T_{th} is reached over time and over the state transition.*

Consequently, since the caching rule for each RRH was determined and the suboptimal initial edge caching among UEs (i.e., $s_t^{*'}$) was derived in Section 4, the edge caching locations among RRHs can be determined based on **Definition 5**.

Definition 5: Edge Caching Locations Among RRHs: Considering the edge caching and data sharing among UEs, popular content should be cached in RRH l, $l \in R$ if $Y_{s_t^{*'}}^l(\Delta T, 0) > 0$.

The condition $Y_{s_t^{*'}}^l(\Delta T, 0) > 0$ indicates that RRH l can achieve a positive utility by caching the content, considering the reduced bandwidth of fronthaul and the storage cost.

VI. PERFORMANCE EVALUATION

In this section, MATLAB-based simulations are conducted to study the performance of edge caching for content diffusion in an F-RAN. Moreover, simulation results are analyzed in terms of the bandwidth consumption of an RAN and the fronthaul toward the BBU pool, as well as the impact of edge caching on the performance of content diffusion, i.e., the average content diffusion ratio and delay.

We consider a scenario with 14 mobile users and 4 RRHs. According to their mutual contact frequencies, the average inter-contact time between any two users follows a uniform distribution. We consider the edge caching of one popular piece of content, which can be stored in the edge nodes among UEs and RRHs. Each UE can either access the content with an RAN or receive it from the other UEs via data sharing. The main parameter settings are listed in Table I.

TABLE 1. Simulation parameters.

Parameters	Values
The number of UEs	$M = 7$
The number of hotspots	$L = 4$
The deadline of the popular content	$T_{th} = 30$
The data size of the popular content	$D = 100$ bytes
The ratio of fronthaul bandwidth consumption to storage cost	$\alpha = 10$
The threshold of content diffusion ratio	$p_{th} = 0.95$
The arrival time of the content access request of each UE	$1/\lambda_{iR} \sim U(5, 25)$
Average inter-contact time between two UEs	$1/\lambda_{ij} \sim U(10, 50)$

A. AVERAGE CONTENT DIFFUSION DELAY

In this part, we investigate the impact of the bandwidth consumption in an RAN on the average content diffusion delay. Since each UE can access the content from an RAN or receive it from the other UEs via data sharing, the network operator can control the edge caching among UEs. A large value of the traffic cost in an RAN indicates that the network operator transmits and stores content in a greater number of UEs. It also means that the content will be rapidly diffused via the data sharing among UEs, which can result in a lower content diffusion delay.

Indeed, as shown in Fig. 3, the average content diffusion delay can be decreased with increasing traffic cost in an RAN. In addition, our proposed suboptimal solution, i.e., the UE's capacity-based edge caching among UEs, can approach the traversal global optimal solution while achieving lower complexity in the complexity analysis in Section 4. This is mainly because we partition the optimization problem in Section 4 into two sub-problems. In the edge caching among UEs, each UE's capacity for content diffusion is derived from our proposed suboptimal solution, thereby achieving low complexity.

B. AVERAGE DELIVERY RATIO

In this part, we review the relationship between the bandwidth consumption in an RAN and the content diffusion ratio.

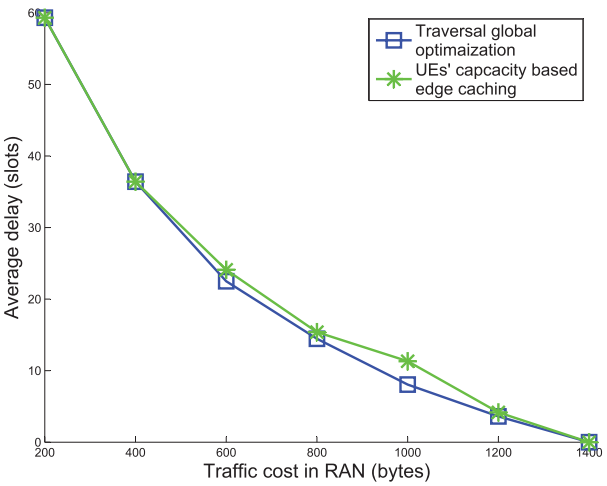

FIGURE 3. Average content diffusion delay vs traffic cost in RAN.

Considering the deadline of the popular content, the delivery ratio can be increased with increasing bandwidth consumption of direct transmission from RRHs to UEs. In addition, the data sharing among UEs depends on the mobility of UEs. This means that the network operator should select the optimal set of UEs to deliver the content according to their capacity for content diffusion.

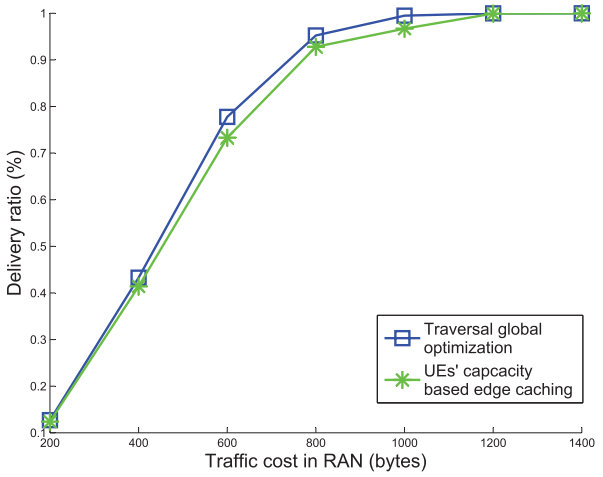

FIGURE 4. Average delivery ratio vs traffic cost in RAN.

Indeed, as shown in Fig. 4, the performance of content diffusion in terms of delivery ratio can be improved with increased traffic cost in an RAN. In addition, it also shows that our proposed suboptimal solution can approach the traversal global optimal solution in terms of delivery ratio while achieving lower complexity in the complexity analysis in Section 4.

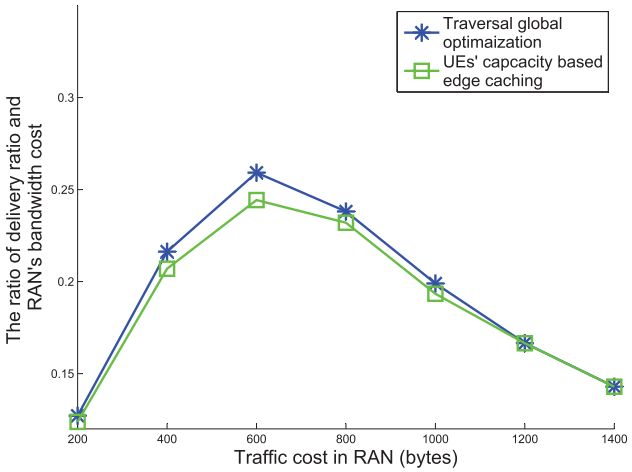

FIGURE 5. The ratio of delivery ratio and RAN's bandwidth consumption vs traffic cost in RAN.

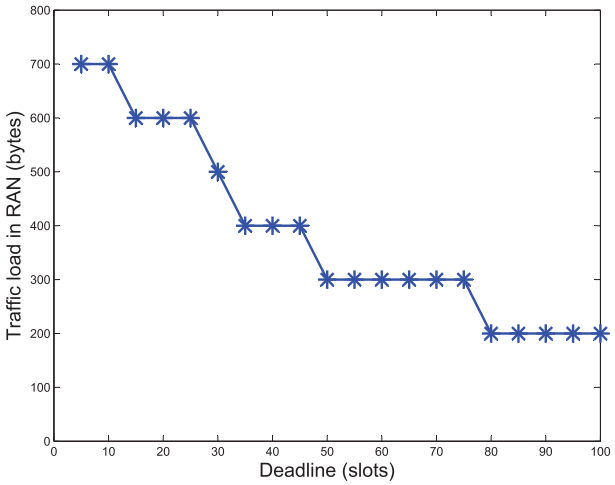

FIGURE 6. Traffic load in RAN vs the deadline of the popular content.

C. DELIVERY RATIO VERSUS BANDWIDTH CONSUMPTION

In this part, we attempt to characterize the efficiency of edge caching among UEs in terms of the ratio of delivery ratio to the RAN bandwidth consumption. There are two extreme cases. One case is where the cellular network directly sends the content to all the UEs, i.e., the traffic cost equals 1400 bytes in Fig. 5. In this case, the delivery ratio is the highest, while the corresponding bandwidth consumption is also the highest. Another case is where the network operator only directly sends to one UE, which can minimize the bandwidth consumption, but the delivery ratio cannot be ensured. Consequently, we attempt to find a tradeoff between the bandwidth consumption of an RAN and the performance of edge caching among UEs. As shown in Fig. 4, the ratio of the delivery ratio to the RAN's bandwidth consumption can be maximized when the traffic cost in an RAN equals 800 bytes.

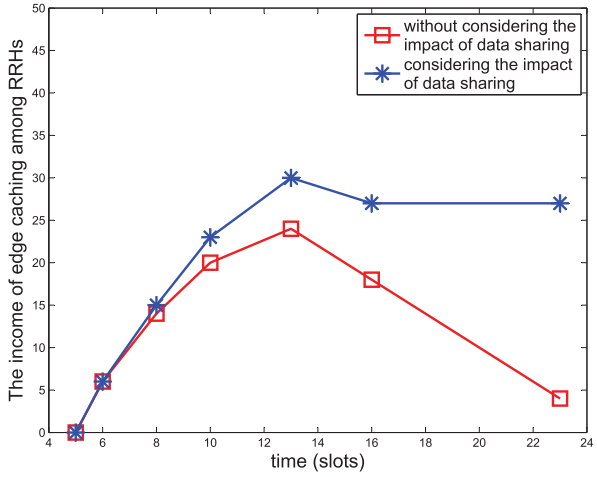

FIGURE 7. The income of edge caching among RRHs (reduced fronthaul bandwidth minus the storage cost).

D. TRAFFIC LOAD IN AN RAN

In this part, the traffic load of an RAN is analyzed under the constraint of the content's deadline. To ensure that the content can be delivered to the UEs before the deadline, the network operator should determine the minimal traffic load with edge caching among UEs. For instance, when the value of the deadline is small, such as $T_{th} = 5$, then the network operator should directly transmit the content to all the UEs. With increase deadline, edge caching among UEs can be employed to reduce the bandwidth consumption in an RAN by diffusing the content among the potential MSNs. As shown in Fig. 6, the traffic load in an RAN can be decreased with increasing deadline requirement by caching the data among the UEs in the MSNs.

E. THE INCOME OF EDGE CACHING AMONG RRHs:

This part determines the utilities of edge caching among RRHs by integrating the bandwidth reducing of the fronthaul and the storage cost of edge caching. Despite reducing the fronthaul's bandwidth, the edge caching among RRHs can also increase storage costs. Consequently, the network should find a tradeoff between edge caching among RRHs and the corresponding storage cost. We investigate the network operator's incomes, i.e., the utility of the reduced fronthaul bandwidth minus the storage cost with and without considering data sharing among UEs. As shown in Fig. 7, the utility of edge caching among RRHs increases with time. This is mainly because the UEs can download the content from the RRHs instead of the remote content center. However, the utility will be decreased when the time becomes excessive. This is mainly because the remaining UEs decrease in number over time, while the storage cost continually increases. In addition, the result in Fig. 7 also shows that our proposed solution can achieve high utility while considering data sharing. This is because the rules for edge caching among RRHs in terms of caching locations and time have been introduced. With the proposed edge caching rules among RRHs, the network can

dynamically determine where the content should be cached and when the contend should be drop, which can avoid unnecessary storage costs induced by edge caching among RRHs. needed in second column of first page if using

VII. CONCLUSION

This paper studied the social-aware edge caching techniques in F-RANs. We attempted to model the impact of edge caching schemes on the performance of content diffusion and sharing in a social-aware F-RAN. Based on the proposed analysis model, the edge caching schemes among UEs and RRHs were proposed to minimize the bandwidth consumption of RANs and the fronthaul toward BBU pools, as well as the storage cost of edge caching among RRHs. Simulation results validated by our theoretical model indicated that the proposed schemes can reduce the bandwidth consumption within the F-RANs effectively. Thus, the proposed social-aware edge caching schemes in F-RANs are helpful in accommodating the rapidly increasing mobile content traffic in the era of 5G networks.

APPENDIX A
PROOF OF THEOREM 1

We first prove the probability of a one-step state transition is zero when $\left|O_{s_m s_n}\right| \neq 1$. The transition probability is zero if more than two components of s_m, s_n are different, i.e., $\left|O_{s_m s_n}\right| > 1$, because the probability that two UEs receive the content simultaneously is zero. In addition, if $\left|O_{s_m s_n}\right| \neq 0$, then there is no state transition.

Then, we prove Eq.(2) on the condition that $\left|O_{s_m s_n}\right| = 1$. Assume that UE i receives the content at time t and results in the state transition from s_m to s_n, i.e., $O_{s_m s_n} = \{i\}$. We first calculate the probability that the other UEs, except UE i ($j \in \{U_2(s_m) - i\}$), have not received the content from the UEs or RRHs before the time t as follows:

$$P_1 = \prod_{k \in \{U_2(s_m) - i\}} e^{-\left(\lambda_{kR} + \sum\limits_{i \in U_1(s_m)} \lambda_{kR}\right) t} \tag{29}$$

In addition, the probability density function of user i receiving the content from the UEs or RRHs at time t can be calculated as

$$f(t) = \left(\lambda_{iR} + \sum_{k \in U_1(s_m)} \lambda_{ki}\right) \cdot e^{-\left(\lambda_{iR} + \sum\limits_{k \in U_1(s_m)} \lambda_{ki}\right) t} \tag{30}$$

Consequently, the probability that the user i receives the content from the UEs or RRHs before other UEs can be calculated as

$$p_{s_m s_n} = \int_0^\infty p_1 \cdot f(t) dt = \frac{\left(\lambda_{jR} + \sum\limits_{i \in U_1(s_m)} \lambda_{ij}\right)\Big|_{j \in O_{s_m s_n}}}{\sum\limits_{k \in U_2(s_m)} \left(\lambda_{kR} + \sum\limits_{i \in U_1(s_m)} \lambda_{ik}\right)} \tag{31}$$

APPENDIX B
PROOF OF LEMMA 2

We first prove the state transmit from s_t to s_e with $k = \left|O_{s_t s_e}\right|$ steps. Because each state transition will only result in a change in one component, the different component number of s_t and s_e indicates the steps required for the state to transmit from s_t to s_e.

In addition, because the state transitions in the sequence are independent of each other, only being dependent on the current state, the probability of a possible sequence of the Markov chain can be derived as the product of each state transition in the sequence. Moreover, the total time from s_t to s_e can be expressed as $T = \sum_{i=1}^{k-1} (t_{i+1} - t_t)$, where $t_{i+1} - t_i$ indicates the time that the state will dwell at the state s_i, which is an exponential distribution for all $i \in [1, \ldots, k]$. Then, the total time from s_t to s_e is a random variable T that equals the summation of $k-1$ independent exponentials, which can be calculated according to **Lemma 1** in [22].

APPENDIX C
PROOF OF THEOREM 3

There is a set of consequences of a Markov chain, denoted by M, in which the state transmits from s_t to s_e. Thus, to calculate the expectation of the bandwidth consumption of an RAN and fronthaul, we need to calculate this as the summation of the product of each possible consequence's ($s \in M$) probability and its corresponding bandwidth consumption of the RAN and fronthaul. In addition, given a content deadline $T_{th} > 0$, we need to derive the probability of each Markov chain's consequence occurring and its corresponding probability that the state will transmit from s_t to s_e before the deadline $T_t h$.

APPENDIX D
PROOF OF COROLLARY 1

We first prove that, given an arbitrary state s_m, the utility function $Y^i_{s_m}(\Delta T, t)$ is strictly decreasing with increasing time t before leaving from state s_m. Letting $\lambda_1 = \lambda_{iR}$ and $\lambda_2 = \sum\limits_{j \in U_1(s_m)} \lambda_{ji} + \lambda_{iR}$, Eq.(27) can be rewritten as

$$f_i(t) = \pi^i_l \lambda_1 e^{-(\lambda_2) t} \tag{32}$$

Then, we prove $p_1(i,t) < p_1(i,0)$ for $t > 0$.

$$p_l(i,0) = \pi^i_l \frac{\lambda_1}{\lambda_2} \left(1 - e^{-\lambda_2 T_{D2D}}\right) \tag{33}$$

$$p_l(i,t) = \pi^i_l \frac{\lambda_1}{\lambda_2} \frac{\left(1 - e^{-\lambda_2 T_{D2D}}\right)}{\frac{\lambda_2 - \lambda_1}{\lambda_2 e^{-\lambda_2 t}} + \frac{\lambda_1}{\lambda_2}} \tag{34}$$

Because we have the following equation for $t > 0$,

$$\frac{\lambda_2 - \lambda_1}{\lambda_2 e^{-\lambda_2 t}} + \frac{\lambda_1}{\lambda_2} > \frac{\lambda_2 - \lambda_1}{\lambda_2} + \frac{\lambda_1}{\lambda_2} = 1 \tag{35}$$

$p_1(i,t) < p_1(i,0)$ for $t > 0$. Consequently, $Y^i_{s_m}(\Delta T, t) < Y^i_{s_m}(\Delta T, 0)$ is proved for $t > 0$.

Then, we prove that the utility function $Y^l_{s_m}(\Delta T, t)$ is strictly decreasing with the state transition in content diffusion. Assume that the sequence of the Markov chain is denoted by $s = \{s_1, \ldots, s_k\}$. For arbitrary $m < n$, $m, n \in [1, \ldots, k]$, the number of UEs that have received the content in state s_m must be less than that in state s_n because the state transition is caused by an increase in the number of UEs that have received the content. Consequently, the $f_i(t)$ in state s_m is larger than it is in state s_n. In addition, the size of $U_2(s_m)$ (i.e., $|U_2(s_m)|$) in state s_m is larger than that in state s_n. Consequently, $Y^l_{s_m}(\Delta T, t)$ is strictly decreasing over the state transition.

REFERENCES

[1] X. Huang *et al.*, "Joint scheduling and beamforming coordination in cloud radio access networks with QoS guarantees," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5449–5460, Jul. 2016.

[2] M. Peng and K. Zhang, "Recent advances in fog radio access networks: Performance analysis and radio resource allocation," *IEEE Access*, vol. 4, pp. 5003–5009, Aug. 2016.

[3] Ericsson. *Ericsson Mobility Report*, accessed on 2015. [Online]. Available: http://www.ericsson.com/mobility-report

[4] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck, "To cache or not to cache: The 3G case," *IEEE Internet Comput.*, vol. 15, no. 2, pp. 27–34, Mar./Apr. 2011.

[5] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of caching strategies in modern cellular backhaul networks," in *Proc. ACM MobiSys*, Jun. 2013, pp. 319–332.

[6] B. Fan, S. Leng, and K. Yang, "A dynamic bandwidth allocation algorithm in mobile networks with big data of users and networks," *IEEE Netw.*, vol. 30, no. 1, pp. 6–10, Jan. 2016.

[7] N. Kayastha, D. Niyato, P. Wang, and E. Hossain, "Applications, architectures, and protocol design issues for mobile social networks: A survey," *Proc. IEEE*, vol. 99, no. 12, pp. 2130–2158, Dec. 2011.

[8] R. Akhtar, S. Leng, I. Memon, M. Ali, and L. Zhang, "Architecture of hybrid mobile social networks for efficient content delivery," *Wireless Pers. Commun.*, vol. 80, no. 1, pp. 85–96, Jan. 2015.

[9] J. Fan, J. Chen, Y. Du, W. Gao, J. Wu, and Y. Sun, "Geocommunity-based broadcasting for data dissemination in mobile social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 4, pp. 734–743, Apr. 2013.

[10] N. Vastardis, K. Yang, and S. Leng, "Socially-aware multi-phase opportunistic routing for distributed mobile social networks," *Wireless Pers. Commun.*, vol. 79, no. 2, pp. 1343–1368, Nov. 2014.

[11] B. Fan, S. Leng, K. Yang, and J. He, "Gathering point-aided viral marketing in decentralized mobile social networks," *IEEE Syst. J.*, to be published.

[12] S.-C. Hung, H. Hsu, S.-Y. Lien, and K.-C. Chen, "Architecture harmonization between cloud radio access networks and fog networks," *IEEE Access*, vol. 3, pp. 3019–3034, Dec. 2015.

[13] R. Tandon and O. Simeone, "Harnessing cloud and edge synergies: Toward an information theory of fog radio access networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 44–50, Aug. 2016.

[14] S. Park, O. Simeone, and S. S. Shitz, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.

[15] H. Hsu and K. Chen, "A resource allocation perspective on caching to achieve low latency," *IEEE Commun. Letters*, vol. 20, no. 1, pp. 145–148, Nov. 2015.

[16] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[17] B. Bai, L. Wang, Z. Han, W. Chen, and T. Svensson, "Caching based socially-aware D2D communications in wireless content delivery networks: A hypergraph framework," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 74–81, Aug. 2016.

[18] Z. Su and Q. Xu, "Content distribution over content centric mobile social networks in 5G," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 66–72, Jun. 2015.

[19] J. Wu, M. J. Xiao, and L. S. Huang, "Homing spread: Community home-based multi-copy routing in mobile social networks," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 2319–2327.

[20] M. Xiao, J. Wu, and L. Huang, "Community-aware opportunistic routing in mobile social networks," *IEEE Trans. Comput.*, vol. 63, no. 7, pp. 1682–1695, Jul. 2014.

[21] B. Fan, S. Leng, K. Yang, and Y. Zhang, "Optimal storage allocation on throwboxes in mobile social networks," *Comput. Netw.*, vol. 91, pp. 90–100, Nov. 2015.

[22] M. Balzs. *Sum of Independent Exponentials*, accessed on 2005. [Online]. Available: https://people.maths.bris.ac.uk/~mb13434/sumexp.pdf

XIANG WANG received the B.Eng. degree from Hunan University. He is currently pursuing the Ph.D. degree with the University of Electronic Science and Technology of China, Chengdu, China. His research interests include resource allocation and data delivery in mobile social networks.

SUPENG LENG (M'06) received the Ph.D. degree from Nanyang Technological University (NTU), Singapore. He is currently a Professor with the School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu, China. He has been a Research Fellow with the Network Technology Research Center, NTU. He has authored over 150 research papers. His research interests include resource, spectrum, energy, routing and networking in broadband wireless access networks, vehicular networks, Internet of Things, next-generation mobile networks, and smart grids. He currently serves as an Organizing Committee Chair and a Technical Program Committee Member for many international conferences and a Reviewer for over ten international research journals.

KUN YANG (SM'08) received the Ph.D. degree from the Department of Electronic and Electrical Engineering, University College London (UCL), U.K. He was at UCL, where he was involved in several European Union research projects for several years. In 2003, he joined the University of Essex, U.K., where he is currently a Full Professor with the School of Computer Science and Electronic Engineering. He manages research projects funded by various sources, such as UK EPSRC, EU FP7, and industries. He has authored over 50 journal papers. His main research interests include heterogeneous wireless networks, fixed mobile convergence, pervasive service engineering, future Internet technology and network virtualization, and cloud computing. He is a fellow of IET. He serves on the editorial boards of IEEE and non-IEEE journals.