An Efficient Content Delivery System for 5G CRAN Employing Realistic Human Mobility

Chun Pong Lau, Student Member, IEEE, Abdulrahman Alabbasi, Member, IEEE, and Basem Shihada, Senior Member, IEEE

Abstract—Today’s modern communication technologies such as cloud radio access and software defined networks are key candidate technologies for enabling 5G networks as they incorporate intelligence for data-driven networks. Traditional content caching in the last mile access point has shown a reduction in the core network traffic. However, the radio access network still does not fully leverage such solution. Transmitting duplicate copies of contents to mobile users consumes valuable radio spectrum resources and unnecessary base station energy. To overcome these challenges, we propose huMan mObility-based cOntent Distribution (MOOD) system. MOOD exploits urban scale users’ mobility to allocate radio resources spatially and temporally for content delivery. Our approach uses the broadcast nature of wireless communication to reduce the number of duplicated transmissions of contents in the radio access network for conserving radio resources and energy. Furthermore, a human activity model is presented and statistically analyzed for simulating people daily routines. The proposed approach is evaluated via simulations and compared with a generic broadcast strategy in an actual existing deployment of base stations as well as a smaller cells environment, which is a trending deployment strategy in future 5G networks. MOOD achieves 15.2% and 25.4% of performance improvement in the actual and small-cell deployment, respectively.

Index Terms—content delivery, energy efficient, 5G, human mobility, delay tolerant

1 INTRODUCTION

Mobile cellular networks are considered as fundamental infrastructures for content delivery to mobile users. The high-speed data rate, low latency, and comprehensive coverage enable the emerging Internet of Things, bringing more devices, such as smartphones, driverless cars, wearable, and machine automated devices into the Internet. Cloud technology, one of the key enablers of the fifth-generation (5G) network, has received the highest attention for deploying in the core network (CN) and radio access network (RAN). Cloud radio access network (CRAN), software defined network (SDN), and network function virtualization (NFV) functionalities provide a high degree of flexibility and scalability for the network adapting to the needs of individual services [1]. Efficiently broadcasting television and multicasting video through software defined mobile networks are examples of content-driven traffic management [2], [3], [4]. Deploying SDN onto mobile networks for video-on-demand delivery solves the problem of numerous redundant independent flows by adjusting delivery strategies through monitoring the popularity growth of contents to increase the capacity of the network [5].

Proactive content caching at the network edge, which is caching in the last-mile access points or base stations, reduces the traffic load as well as the latency in the network backhaul CN. For instance, OpenCache proposed by Georgopoulos et al. utilizes SDN to cache popular on-demand videos as close to the end-user as possible [5]. However, it reduces traffic load in the CN but could not contribute to the reduction of the traffic in the RAN due to the pull-on-demand characteristic of content requests. Mobile users download the common content from the same base station (BS) by unicast transmission. From the network operator point of view, it is inefficient to repeatedly transmit the same content from a BS. Although a discussion on various caching locations including the CN and RAN is presented in [6], the physical nature of broadcasting over the air is underutilized. In fact, the more users in a cell at a given time, the more efficient it becomes to deliver popular contents in a single broadcast wireless transmission [7]. Therefore, caching in the BS is insufficient for improving spectral efficiency in the RAN.

In the convention, random walk and random waypoint models are used in mobility modeling [8]. These models follow the stochastic approach in moving direction, velocity, and independent with the previous status. These mobility models may be suitable for small-scale within a base station coverage. However, these random and small-scale mobility models are insufficient for mobile operators to massively allocate radio resources in an urban scale for broadcasting or multicasting. Moreover, human mobility is far from being considered random. In contrast, it exhibits a high degree of spatiotemporal regularity, following a reproducible pattern such as traveling between home and work locations [9]. Furthermore, the mobility and the mobile traffic have periodical patterns and can be linked with social ecology [10].

1.1 Motivation

Considering the following scenario, a group of students enrolls in an online course. The course announces and distributes subscribed lecture videos to students twice a
week at a scheduled time through a mobile network. If all students download the video by unicast, it will generate many redundant duplicated traffic. This problem is also found in other on-demand multimedia services, such as software update, TV serial, movie, music, and game file subscription. Note that these large size contents are produced before the official announcement time, thus, caching or multicasting to the users before the actual consumption time is a solution for this problem. However, it is foreseeable that students would gather at the school during school hours. If the system leverages the timing of content broadcasting when the subscribers are congregated, it will reduce the traffic congestion, radio resource usage as well as the energy consumption in the mobile network.

We conclude that mobile networks can utilize human mobility patterns to schedule transmissions to deliver large size delay-tolerant contents in the crowded area for the best usage of radio resources and energy. Furthermore, it is straightforward to be implemented by the network operators without the need of requesting mobile users for contributing their resources in device-to-device communication and reducing the uncertainty in delivering through opportunistic networks. Therefore, it motivates us to develop a broadcasting system in 5G networks.

1.2 Contribution

Our contributions in this work are folded as follows. We first formulate the content delivery problem as searching for optimal timing, where mobile users are the most congregated in a region. Second, we analyze the synthetic human activity-based mobility model statistically by deriving related probabilities of users’ activities and staying location. Then, an efficient content delivery approach, huMan mObility-based cOntent Distribution (MOOD), is proposed. The mobility is transformed into a figure, namely broadcasting threshold, which is linked to the number of subscribers of content and cells. The proposed approach utilizes temporal and spatial resources in allocating radio resources for delivering popular subscribed delay-tolerant contents. Unlike the conventional unicast methods, we utilized the broadcasting nature of the physical medium of wireless communications to reduce the number of transmissions in the RAN to save radio resource usage and energy consumption. Furthermore, MOOD reduces significant wireless transmissions compared to the generic broadcasting method which broadcasts a content immediately to the entire region on its arrival. Moreover, MOOD mitigates the difficulties of compensating the mobile users for device-to-device communication and the uncertainty in opportunistic networks. Finally, the proposed approach is evaluated by conducting simulations incorporated with a real geographical location and realistic schedules of people daily activities to demonstrate the human movement and the proper timing for content distribution.

The rest of this paper is organized as follows. First, we review the related work in Section 2. Section 3 presents the reference system, system model, human mobility model, and problem formulation. A statistical analysis is conducted in Section 4. The proposed approach is then presented in Section 5. Section 6 and 7 describe the details of the simulation setup and evaluation, respectively. Section 8 discusses the use cases and limitations of the approach as well as future research directions. Finally, the paper is concluded in Section 9.

2 RELATED WORK

In [11], The authors propose algorithms to minimize the total cost of Ternary Content Addressable Memories occupation in rule caching and remote processing in the software-defined switches. Finamore et al. suggested to push the caching paradigm further onto the mobile terminals in [12]. Their solution utilizes the physical nature of broadcast in the mobile network to reduce the number of wireless transmissions and to cache popular content into mobile devices. However, this approach faces a problem that the data proactively pushed to the users by popularity prediction may be wasted if the prediction goes wrong or if the content is not eventually consumed by users. This problem could be solved by understanding the human mobility for a smarter multicasting solution.

Abou-zeid et al. proposed a resource allocation framework by predicting user movement for optimizing long-term radio resource allocation in multiple cells [13]. Xu et al. empirically studied human mobility from the big data of a mobile network in a metropolitan city with over 9600 cellular towers. The results reveal that the human mobility has strong correlations with the mobile traffic. Do et al. conducted a study on human daily life visiting patterns through data collected by mobile devices [14]. From the research, authors found that people usually follow simple routines involving a few frequent places. Similar mobility is observed in [15] for analyzing in Saudi Arabia. Nousias et al. found human mobility patterns are following a universal law, which is the probability of people traveling from one place to another is related to the relative rank of places [16]. Zhang et al. constructed a mobility prediction model from a telecom cloud that can be run at the infrastructure level for online mobility prediction [17]. Qiao et al. proposed a rating framework employing big data to measure human mobility and applications usage behavior [18]. In order to reproduce synthetic realistic mobility patterns close to reality, daily activities of the human schedule are considered in the mobility models. Ekman et al. proposed the working day movement model in [19] by presenting the everyday life of average people, such as sleeping at home, working in the office, and evening activities. Issacman et al. proposed WHERE, which models large populations movement in different metropolitan areas from real-world probability distributions [20]. This model primarily generates synthetic traces for the people moving between two places. It is scalable to more locations but with introducing extra complexity. Zheng et al. proposed the agenda driven mobility model in [21] emphasizing the social role of humans in making movement decisions.

Human mobility has been considered for caching and content delivery fields. Lee et al. analyzed human mobility from traces of location-based social networks to develop a method to deliver video data by moving people to static kiosks [22]. Authors in [23] exploited the human mobility patterns and social tie for caching contents in the mobile devices for distribution through device-to-device communication. Chen et al. proposed a machine-learning algorithm
on both content request and human mobility for proactive caching at BBU’s and RRHs [24]. The approach demonstrated the effectiveness on determining which contents to cache at which location. The suggested work shows the significance of reducing the traffic in backhaul and fronthaul network between the RRHs and BBU’s. However, the radio access between the UEs and the RRHs are remained in the unicast fashion. Furthermore, there is a lack of discussion on radio resources and energy consumption in these fields.

3 System Models and Problem Formulation

In this section, the reference system is first presented, followed by the system model. Then the human mobility model and the energy consumption model are introduced. Finally, the content delivery problem is formulated.

3.1 Reference System

In order to provide efficient broadcast and multicast services in the traditional mobile network, Evolved Multimedia Broadcast Multicast Service (eMBMS) has been introduced in the Long-Term Evolution Advanced (LTE-A) standard [25]. There are maximum 60% of downlink radio frames could be allocated for the eMBMS service for content multicasting or broadcasting. The eMBMS system architecture consists of the following major components in the CN, Broadcast/Multicast Service Center (BM-SC), MBMS Gateway (MBMS-GW), Mobility Management Entity (MME), and modulation and coding scheme (MCS).

BM-SC is the entry point for content providers transmitting multimedia from external networks to the users in the mobile network. It also accomplishes service announcement. MBMS-GW is a logical node for forwarding user plane data to Evolved NodeB (eNB). It also sends MBMS session control signals to the MME, which is responsible for authentication, security, and mobility management procedures. Multicell/multicast Coordinating Entity (MCE) receives control signal from the MME and performs admission control and radio resources allocation for the eNBs. The RAN of LTE is composed of eNBs. An eNB is the base station that controls one or more cells in the RAN and distributes multimedia data to the mobile devices through a wireless radio connection.

CRAN is a strong candidate for the future 5G network [26]. The BS in the CRAN is decomposed into a signal processing unit and a radio unit. The signal processing unit is called baseband unit (BBU) or data unit. It is responsible for receiving data and control signal from the core network through the backhaul links, processing the signal into the baseband signal and send to the radio unit through a line-card (LC) by the fronthaul optical fiber link. The radio unit in the remote site is called remote radio head (RRH). It consists of a optical network unit (ONU) to convert the optical signal from the BBU to the radio signal and transmit it through a power amplifier to the antenna. A BBU can be located up to 40 km away from RRH and served one or more RRH. The BBU’s could be centralized into an entity called BBU Pool. It is a centralized cluster providing network function virtualization for baseband processing. The centralized approach helps to reduce the cost of site rental, maintenance, air-conditioning, as well as power consumption by sharing most of the hardware facilities from the traditional BS. Figure 1 illustrates the 5G CRAN system architecture considered in this framework. It is an extension of the typical eMBMS standard in 4G LTE-A integrated with the CRAN.

Besides the components above, a cloud-based software defined controller (SDC) is included. In literature, the SDC is responsible for making control decision and sends the control signal to forwarding devices such as routers, gateways, and switches [27], [28]. In the proposed system, the controller receives high-level service policies from the content providers in the cloud. The SDC implements control signal onto the MCE and MBMS-GW for radio resources allocation, content distribution scheduling, and cooperated broadcasting and multicasting. In this approach, when a content is ready to be distributed, the content provider informs the mobile network by sending a control signal to the SDC. The SDC then instructs the MCE to collect the subscribers’ interests as well as the mobility history of UEs. The analysis of mobility history of the subscribers will be performed in the SDC. Meanwhile, the SDC keeps monitoring the subscriber’s real-time mobility through the aggregated reports from MCE. Once the broadcasting conditions are fulfilled, the SDC will instruct the MCE and MBMS-GW for distributing the content in the mobile network. Each content is individually considered in the SDC. The content distribution policies could be different among contents.

3.2 System Model

We consider a mobile network that has a set of cell \( \mathcal{B} = \{b_1, ..., b_B\} \) be deployed in a certain region. Let \( t \) be the time segment. A set of contents \( \mathcal{C} = \{c_1, ..., c_C\} \) is ready to be delivered to subscribers. In the region, there is a set of mobile users \( \mathcal{U} = \{u_1, ..., u_U\} \). A subscriber is a mobile user who subscribes to content which has not yet been delivered. Let \( q_{u,c} \) be a binary variable where \( q_{u,c} = 1 \) if a mobile user \( u \) is a subscriber of a content \( c \). Let \( q_{u,c,b,t} \) be a binary variable where \( q_{u,c,b,t} = 1 \) if a subscriber of a content \( c \) is associated with cell \( b \) at time segment \( t \). The total number of subscribers of content \( c \) in cell \( b \) at time \( t \) is equal to,

\[
N_{c,b,t} = \sum_{u \in \mathcal{U}} q_{u,c,b,t}.
\]
An active cell for a content is defined as a cell that has at least \( m_u \) active subscribers. The number of active cells of content \( c \) at time \( t \) is denoted as \( a^c_t \) as,

\[
a^c_t = \sum_{b \in B} \left[ N_{c,b,t} \geq m_u \right],
\]

where \( m_u \) is the threshold of the minimum number of user to declare that a cell \( b \) is active. The notation \( [ \cdot ] \) is the Iverson bracket. If the condition in the square bracket is fulfilled, the number is 1, while 0 otherwise.

The number of active cells varies according to the movement of the mobile users. Therefore, studying a human mobility model is essential for understanding the daily mobility patterns of users in the region.

3.3 Human Mobility Model

The human mobility model is described in this subsection. It makes use of the daily life routines of people. The mobile users are categorized into different sets of user groups \( G = \{g_1, ..., g_G\} \) according to their occupation, life habit, and behavior for generating individual mobility traces with a degree of randomness, while representing the realistic environment. The human states mobility model makes use of the human daily life routine for each user group which is composed of a set of states \( \mathcal{V} = \{v_1, ..., v_N\} \). The users within the same user group have similar daily routines from the same set of states with same activity stages and state starting range, but different staying locations and duration.

Each state \( v_i \) consists of the following components for each user \( u \), an activity stage \( A_{vi} \), a staying location \( L_{vi} \), a staying duration \( D_{vi} \), a state starting range and a set of transition probabilities.

3.3.1 Activity Stage

An activity stage \( A_{vi} \) is the name of a daily activity such as ‘Sleeping’ and ‘Working’. A series of activity stages forms a life routine for a user group. Each user group has different sets of activity stages in their corresponded mobility model. For example, a group of office staff has a ‘Working in office’ stage while a group of students has a stage ‘At school’.

3.3.2 Staying Location

A staying location \( L_{vi} \) is randomly chosen from a set of places, depends on the activity and the user group. There is static and dynamic spatial information for a mobile user in different states. For instance, home and work locations are static. These places remained unchanged for a mobile user in this model. On the other hand, the dining and recreation locations are dynamic. The mobile user may visit different places for dining and leisure on various days. These locations are randomly chosen from a set \( L_{vi} \) of related locations within a reasonable distance. The selection process of a location is independent of the other state parameters.

3.3.3 Staying Duration

The staying duration \( D_{vi} \) of each state \( v_i \) is a random variable which follows a truncated distribution with a lower bound \( t_{D_{vi}} \) and an upper bound \( T_{D_{vi}} \). Each state has individual mean \( \mu_{D_{vi}} \), variance \( \sigma_{D_{vi}} \), and truncation boundary for the staying duration according to the activity stage and user group. For example, the staying duration of a ‘Sleeping’ state of an adult may have a mean of 7 hours with a larger variance whereas 10 hours for a child with a smaller variance.

3.3.4 State Starting Range

The state starting range consists of a pair of lower bound \( t_{S_{vi}} \) and upper bound \( T_{S_{vi}} \) for controlling the start of a state. It is independent of the staying duration. A state starts if and only if the starting time \( S_{vi} \) of the user \( u \) of a state \( i \) is within this range.

In addition to the components mentioned above, the starting time and the finishing time of a state for a user are introduced. The finishing time of a state is defined as the sum of the starting time and the staying duration. Let \( E_{vi} \) be the finishing time of user \( u \) staying at the state \( v_i \), and it is formulated as,

\[
E_{vi} = S_{vi} + D_{vi},
\]

where \( S_{vi} \) is the starting time of state \( v_i \) of user \( u \). We assume, the starting time of initial state \( S_{vi} \) is a constant at \( t = t_0 \).

3.3.5 Transition Probability

The transition probability \( m_{vi,v_j}^u(t) \) is the probability of transiting from state \( v_i \) to \( v_j \) at time \( t \). It is a function of time depends on the finishing time of the current state \( i \), the starting range of state \( j \) and a state selection probability \( \rho_{vi,j} \). The transition from a state \( i \) to a future state \( j \) can be triggered when the finishing time of the current state \( i \) is within the range of the starting time in state \( j \).

An example graph of a human mobility model for the routine of an office staff is shown in Figure 2. It consists of six activity stages. These stages are repeated daily and each day will have new random variables for the duration and transition probabilities. The locations of some states such as ‘Lunch’ and ‘Dinner’ may change on different days.
3.4 Energy Consumption Model

It has been proven in [29] that discontinuous transmission (DTX) achieves significant energy reductions in Long-Term Evolution (LTE) networks. The energy consumption of the baseband and the power amplifier could be reduced by decreasing the number of radio transmissions, assuming the mobile subscribers of content have a foreseeable individual mobility pattern. In the network system point of view, the subscribers’ distribution in a region varies from sparse to dense in a daily cycle. When multicasting a content in a region, the system delivers it to active cells as per (2). The cell which has no active subscribers utilizes the DTX technology to suspend the radio transmission and turn the related components into sleep mode. Therefore, the overall energy consumption in the network system is reduced.

The total power consumption $P_t$ in the RAN for delivering a content is divided into two parts, the consumption in the central site $P_{pool}$, where the BBU Pool located, and the remote site, where the RRHs placed [30]. It is formulated as follows,

$$P_t = P_{pool} + P_{fh} + \sum_{b \in B} P_{RRH_b}$$  \hspace{1cm} (4)

where $P_{pool}$, $P_{fh}$, and $\sum_{b \in B} P_{RRH_b}$ are the total power consumption for the BBU pool, fronthaul network, and all active radio units. $P_{pool}$ is expressed as,

$$P_{pool} = \sum_{\phi \in \Phi} P_{BBU,\phi} + P_{CS},$$  \hspace{1cm} (5)

where the first term is the total power consumption for all active BBU $\phi$, and $P_{CS}$ is the infrastructure power of the central site such as cooling and monitoring. The power of the fronthaul is formulated as,

$$P_{fh} = \sum_{\gamma \in \Gamma} P_{LC,\gamma} + \sum_{\sigma \in O} P_{ONU,\sigma},$$  \hspace{1cm} (6)

where the first term is the total power consumption for all active wavelength $\gamma$ in LC and $\sum_{\sigma \in O} P_{ONU,\sigma}$ are the total power consumption for all active ONUs in the region.

In the RRH, the linear approximation of power consumption in [31] is considered in this work which is formulated as follows,

$$P_{RRH_b} = \begin{cases} N_{TRX} P_0 + \Delta_p P_{out}, & 0 < P_{out} \leq P_{max} \\ N_{TRX} P_{sleep}, & P_{out} = 0 \end{cases}$$  \hspace{1cm} (7)

where $N_{TRX}$ is the number of transceiver chains, $P_0$ is the power consumption calculated at the minimum possible output power, $\Delta_p$ is the slope of the load dependent power consumption, $P_{out}$ is the radio frequency output power. $P_{max}$ is the maximum radio frequency output power at maximum load, and $P_{sleep}$ is the power consumption in the sleep mode. For simplicity, we denoted $P_{RRH_b}^{a} = N_{TRX} P_0 + \Delta_p P_{out}$ for power consumption of RRH when it is active, and $P_{RRH_b}^{s} = N_{TRX} P_{sleep}$ when it is in the sleep mode.

The energy consumption $E_c$ for delivering a content in the radio access, the fronthaul, and the backhaul networks is formulated as,

$$E_c = t_{BBU} P_{pool} + t_{opt} P_{fh} + t_{rx} (a_c^{e} P_{RRH_b}^{a} + (|\mathbb{B}|-a_c^{e}) P_{RRH_b}^{s}),$$  \hspace{1cm} (8)

where $t_{BBU}$ is the BBU processing duration, $t_{opt}$ is the optical transmission duration, $t_{rx}$ is the radio transmission duration, $a_c^{e}$ is the number of active cells which multicasting the content, and $(|\mathbb{B}|-a_c^{e})$ is the number of non-active cells which are in sleep mode. By minimizing the number of active cells, the overall energy consumption of the system can be reduced.

3.5 Problem Formulation

As mentioned earlier, an active cell $a_c^{e}$ implies that there is at least one active subscriber of a content $c$ within the cell coverage area at time $t$. From the network operator’s point of view, if the radio reception levels of the mobile users are good, broadcasting content in a cell to all subscribers yields the most efficiency. It requires only a single radio transmission instead of multiple duplicate transmission compared to the unicast transmission. From the network point of view, perceiving a time with the minimum number of broadcasting transmissions will be the most efficient way to deliver content, i.e., using the minimum amount of radio resource and the least amount of system energy for transmitting to all of the subscribers. Therefore, the problem could be described as searching a time segment $t$ for each content $c$ that has the minimum number of active cells $a_c^{e}$ for content delivery, which is formulated as follow,

$$\min_t \quad a_c^{e}$$  \hspace{1cm} (9a)

subject to \hspace{1cm} $t_a^{e} < t < t_{c}^{e}$, \hspace{1cm} (9b)

where $t$ is the search time segment, as an optimization variable and it is an integer. The variables $t_a^{e}$ and $t_{c}^{e}$ are the arrival time and the latest time for delivering the content $c$.

4 Statistical Analysis

The primary objective of the problem above is to broadcast the content to all users with the minimum number of transmissions, i.e., the minimum number of active cells. For this purpose, we need to find a time and optimal locations that have the maximum number of subscribers for a content. Therefore, we calculate the expected number of users at a time $t$ in location $L$.

The building blocks of these calculations start with finding the probability of several events. First, a user starting the state $j$ at time $t$, which can be calculated based on the staying duration and starting time of that user at previous states. Second, we find the probability of finishing a state $j$ at time $t$. Third, we incorporate the selection of location $l$ in the starting and finishing time probabilities. Fourth, we calculate these probabilities for all users, at all possible time segments. Finally, we find the optimal time which guarantees a minimum number of active cells.

In order to calculate the probability of a user $u$ starting a state $j$ at time $t$, we recognize five events which contribute to the starting point as follows,

- $R$: The finishing time of previous state $i$ is in between time $t$ and $t-1$.
- $Z$: The finishing time $E_{v_i}^u$ of previous state $i$ must be larger than the lower bound of the starting time
where $t^j_{v_i}$ is the current state $j$ and lower than the upper bound of starting time $t^j_{v_i}$. 
- $X$: For every previous state $i$, the staying duration $D^u_i$ is lower and upper bounded by $t^D_i$ and $t^U_i$, respectively.
- $Y$: The current searching time $t$, must be lower then the maximum staying duration in the previous state $j$.
- $W$ is a transition probability from state $i$ to state $j$.

Event $R$ limits the finishing time of previous state $i$ within the time segment $t$ as follows,

$$ R : t - 1 \leq E^u_{v_i} \leq t. \quad (10) $$

Event $Z$ ensures that the finishing time of the previous state $i$ is within the range of starting time of the current state $j$. It is formulated as,

$$ Z : t^S_{v_j} \leq E^u_{v_i} \leq t^S_{v_j}. \quad (11) $$

Event $X$ describes that the staying duration $D^u_i$ is a truncated random variable which is upper bounded by $t^D_i$ and lower bounded by $t^D_i$, which is,

$$ X : t^D_i \leq D^u_i \leq t^D_i. \quad (12) $$

Event $Y$ indicates the current time $t$, should be less than the maximum staying duration of previous state $i$ in order to have a valid transition from state $i$ to current state $j$, which is formulated as,

$$ Y : t \leq t^S_{v_i} + t^D_{v_i}. \quad (13) $$

Finally, we define $W$ as a positive transition probability from state $i$ to state $j$. This probability excludes all the other possibilities for transiting from previous state $i$ to states $k$ other than the current state $j$,

$$ W : m_{v_i,v_j}(t) = 1 - \left( \sum_{k \neq j, k \neq l} m_{v_i,v_k}(t) \right) \geq 0. \quad (14) $$

The probability of a user $u$ starting a state $j$ at time $t$ is a combination of the aforementioned events, and it is formulated as follows,

$$ \Pr\{S^u_{v_j} = t\} = \Pr\{R, Z, X, Y, W\}. \quad (15) $$

The probability $\Pr\{S^u_{v_j} = t\}$ is the probability of intersection between all events, $X, Y, Z, W$. Hence, we can reformulate it using conditional probability to the following expression,

$$ \Pr\{S^u_{v_j} = t\} = \sum_{j \in \{1,...,V\}} \Pr\{R, Z|X, Y\}\Pr\{X\}\Pr\{Y\} m_{v_i,v_j}(t). \quad (16) $$

Note that event $Y$ does not contain any random variable, hence, we express the intersection with event $Y$ as an indicator function, $\mathbb{I}\{Y\} = 1$ when $Y$ is true and $\mathbb{I}\{Y\} = 0$ when $Y$ is false. Event $Z$ is converted from its original definition in (11) to $Z : t^S_{v_i} \leq t \leq E^u_{v_i}$, because we already bound $E^u_{v_i}$ by $t - 1$ and $t$ at (10). It follows that using (11)-(14) and the new definition of event $Z$, expression (16) is expanded as in (17).

$$ \Pr\{S^u_{v_j} = t\} = \sum_{j \in \{1,...,V\}} \Pr\{t - 1 \leq E^u_{v_i} \leq t|X, Y\} \mathbb{I}\{t^S_{v_i} \leq t \leq E^u_{v_i}\} \left[ F_{D^u_i}(t^D_{v_j}) - F_{D^u_i}(t^D_{v_i}) \right] I(Y) m_{v_i,v_j}(t), \quad \forall j \in [1, V]. \quad (17) $$

where $F(.)$ is the cumulative distribution function.

We calculate the probability of the finishing state at time $v_j$ since it is necessary to obtain the next state probabilities. We let $n_{v_j} = t - d^u_{v_j}$ and $\delta_{v_j} = \Pr\{D^v_{v_j} = d^v_{v_j}\}$. The probability of finishing a state $v_j$ is formulated as follows,

$$ \Pr\{E^u_{v_j} = t\} = \Pr\{S^u_{v_j} + D^v_{v_j} = t\} = \sum_{d^v_{v_j} \in \mathbb{R}} \Pr\{S^u_{v_j} + d^v_{v_j} = t|D^v_{v_j} = d^v_{v_j}\} \Pr\{D^v_{v_j} = d^v_{v_j}\} = \sum_{d^v_{v_j} \in \mathbb{R}} \Pr\{S^u_{v_j} = t - d^v_{v_j}|D^v_{v_j} = d^v_{v_j}\} \Pr\{D^v_{v_j} = d^v_{v_j}\}. \quad (18) $$

Recall that the staying duration $D^u_{v_j}$ at $v_j$ is independent from the starting time $S^u_{v_j}$ of $v_j$. It follows that $\Pr\{S^u_{v_j} = n_{v_j}|D^v_{v_j} = d^v_{v_j}\}$ is in similar form to that of (17), hence we can solve it by replacing $t$ with $n_{v_j}$.

The joint probability of finishing a state $v_j$ with the possibility of being at different locations is expressed as,

$$ \Pr\{E^u_{v_j} = t, L^u_{v_j}\} = \sum_{l \in L^u_{v_j}} \Pr\{E^u_{v_j} = t|L^u_{v_j} = l\} \Pr\{L^u_{v_j} = l\}, $$

where $L^u_{v_j}$ is the random variable that spans all possible locations for the same state, i.e., $L^u_{v_j}$. Recall our assumption that the user selects the location of the activity in state $v_j$ independent from the starting and finishing time of that activity. Hence, the probability in (19) is calculated as follows,

$$ \Pr\{E^u_{v_j} = t, L^u_{v_j}\} = \sum_{l \in L^u_{v_j}} \Pr\{E^u_{v_j} = t|L^u_{v_j} = l\} \Pr\{L^u_{v_j} = l\} = \sum_{l \in L^u_{v_j}} \sum_{d^v_{v_j} \in \mathbb{R}} \Pr\{S^u_{v_j} = n_{v_j}|D^v_{v_j} = d^v_{v_j}\} \Pr\{L^u_{v_j} = l\} \delta_{v_j}. \quad (20) $$

In similar lines, we link the probability of starting a state $v_j$ at a time $t$ with the set of locations $L^u_{v_j}$, from (17), as in (21).

$$ \Pr\{L^u_{v_j} = l\} \left[ F_{D^u_i}(t^D_{v_j}) - F_{D^u_i}(t^D_{v_i}) \right] I(Y) m_{v_i,v_j}(t), \quad \forall j \in [1, V]. \quad (21) $$

Then, we calculate the expected value of a number of users at specific time and location. We begin by finding the
No terministic values of the random variables \(S_t\) using the probabilities found in (20)-(21), as follows,

\[
Pr \left\{ \{u_{vj}|t, L_{vj} = I\} \right\} = \sum_{s^{u} \in (0, \ldots, t)} \sum_{e^{u} \in (0, \ldots, t)} Pr \left\{ S_{vj}^{u} = s^{u}, E_{vj}^{u} = e^{u}, L_{vj} = I \right\} \tag{22}
\]

where \(s^{u}\) and \(e^{u}\) are the starting and finishing time deterministic values of the random variables \(S_{vj}^{u}\) and \(E_{vj}^{u}\), respectively, and in notation they are replaced by \(t\) at (17) and (18).

Utilizing the probability of each user being in location \(l\) at time \(t\), expressed in (22), the expected numbers of users is obtained as,

\[
\mathbb{E} \{N_U(t, l)\} = \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{V} \cup \{l\} \in \mathcal{I}} Pr \left\{ \{u_{vj}|t, L_{vj} = I\} \right\}. \tag{23}
\]

Finally, the time segment \(t^*\) at which the minimum number of active cells that is enough to serve all subscribed users for a content can be found by solving the following optimization problem,

\[
t^* = \arg \min_t \sum_{l \in \mathcal{B}} \mathbb{E} \{N_U(t, l)\} \geq m_u, \tag{24}
\]

where \(m_u\) is the threshold of the minimum number of user to declare that a cell \(l\) is active.

Since the human movement in the future is uncertain. The network system makes the delivery decision based on incomplete information. Although we comprehensively lay out the theoretical model in this section, it is infeasible to guarantee an optimal solution without knowing the future. Therefore, we attempt at solving the problem heuristically and propose an online approach for efficient content delivery in the next section.

5 **Human Mobility-based Content Distribution System**

In this section, our proposed approach MOOD is introduced. The approach uses the historical data for online decision making without the future information. Figure 3 shows the flow of the content delivery process of our proposed approach and Algorithm 1 describes the procedure.

First, when new content is ready to be delivered, the content provider sends a notification to the SDC in the mobile network. Then, the SDC sends an announcement control packet through the MCE, BBU, and RRH to collect the subscription information from the UEs. In the meanwhile, the content can be transmitted from the content provider to the mobile network and be cached at the MBMS-GW at any time before the content arrival time \(t^*_a\). After the subscription requests are collected, the SDC analyzes the mobility history of subscribers to generate an initial broadcasting threshold. The Broadcasting threshold \(\beta^*_c\) is a threshold for determining the broadcast decision. The MCE keeps monitoring the users’ locations and reports to the SDC for evaluating the number of active cells and update the broadcasting threshold in every time segment. Furthermore, at every period of time \(t^#\), the broadcasting threshold is updated with the moving minimum active cells \(a^c_1\#\) in the last \(t^#\) periods. The process is repeated until the number of active cells is lower than the broadcasting threshold. Then the content cached in the MBMS-GW will be forwarded to the BBU and be broadcast in all the active RRHs to the subscribers.

**Algorithm 1 Human Mobility-based Content Distribution (MOOD)**

**Input:** fading constant \(d\), number of cells \(|\mathcal{B}|\)

- Content \(c\) is announced
- Collect subscription requests \(i = 1\)

1. Analyze mobility history of subscribers, obtain \(a^c_1\#\), \(\omega^c_1\)
2. Calculate the initial \(\beta^c_1\)
3. while \(t^*_a \leq t \leq t^*_c\) do
4. if \(t \mod t^# = 0\) then
5. 
6. if \(a^c_1 \leq \beta^c_1\) then
7. Broadcast content \(c\) into all active RRHs
8. break
9. end if
10. end if
11. end while

In the proposed algorithm, calculating the broadcasting threshold is the most critical phase. The broadcasting threshold \(\beta^c_1\) of content \(c\) at time \(t\) is subjected to content remaining lifetime and the historical minimum number of active cells as follows,

\[
\beta^c_t = e^{-dp^c_1} (|\mathcal{B}| - \omega^c_t) + \omega^c_t, \tag{25}
\]

where \(e\) is an exponential function for raising the quantity from \(\beta^c_1\). A time-related fading constant \(d\) is multiplied by \(p^c_1\) for controlling the
increasing rate. The $|B|$ is the total number of cells in the region, and $\omega^c_i$ will be explained in later paragraphs. The broadcasting threshold increases faster with a larger $d$. The percentage of remaining lifetime $p_i^c$ of content $c$ at time $t$ is defined as follows,

$$p_i^c = \frac{t \in_c - t}{t \in_c - t_a^c},$$

(26)

where $t$ is the current time, $t_a^c$ and $t \in_c$ are the arrival time and latest time for delivering content $c$. The latest time $t \in_c$ constrains the system should deliver the content within the recommended delivery period. Therefore, the broadcasting threshold is designed to relax the broadcasting constraint exponentially along the lifetime of the content. The broadcasting threshold of each content is subjected to different user preferences and subscriber mobility. A variable $\omega^c_i$ in (25) is applied to engage these features to the distribution algorithm. It is a weighted arithmetic mean of the periodic minimum number of active cells and formulated as follows,

$$\omega^c_i = \frac{\sum (i^2 \times a_{i^2}^c)}{\sum i^2},$$

(27)

where $i$ is the number of previously moving periods from the mobility history and $i^2$ is a weight coefficient. By the quadratic weight, the data from recent periods contribute more than the former data. The weights are multiplied by the minimum number of active cells $a_{i^2}^c$ of period $i$ which is,

$$a_{i^2}^c = \min_{j} \{a_{i,j}^c\}, \forall j$$

(28)

where $j$ is the time segment index within period $i$. The subscriber mobility history of each content is logged in the record for obtaining a daily minimum number of active cells $a_{i^2}^c$. As shown in (25), the variable $\omega^c_i$ is used as a starting point of the threshold once the content arrived.

6 Evaluation Setup

In order to evaluate MOOD, simulations are conducted in the weekday scenario. This section presents the simulation setup and the associated assumptions.

6.1 Geographical Location

A town located in Thuwal, Makkah Province, Saudi Arabia, is considered in the simulation. It is a moderate density living compound that facilitates both working and living environment. In the simulation, there are about 2000 townhouses and 80 two-story apartment buildings. Each townhouse populates a family or 3-8 people and an apartment building populates about 20-40 people. In the compound, the university campus is the major working area for the residents and three schools for primary and secondary school students. Furthermore, there are six buildings for recreation, dining, and shopping. Six BSs with their actual deployment locations in reality are applied in the simulation named as a 4-digit number from 5912-5919, where each BS consists of three 120 degrees sectors. Each sector has been considered as a cell with a 5-digit number name. The sector name is constructed by extended one more digit from the BS name to the rightmost digit, such as 59121(northeast), 59122(southeast), 59123(west) for three sectors of BS 5912. In total, 18 cells are deployed in the 9.57 km² simulation area. A map of the simulation with the BS deployment and cell boundaries is shown in Figure 4.

6.2 Mobility

In the simulation, a daily life of a user is modeled with random locations and duration. The model first randomly selects a home and a work location for certain user. These locations are static for a user throughout the simulation period. Then, the duration of staying and traveling are randomly generated following various truncated normal distributions. Furthermore, a user has a certain probability of visiting the recreational and dining area with a random duration. The following are four groups of mobile users with different daily mobility patterns employed in the simulation:

6.2.1 Staff

A model for the movements of faculties, researchers, and other workforce is adopted for 2200 users. There are 75% of users live in townhouses and 25% of users live in apartment buildings. Their office locations are randomly chosen in university campus buildings. The daily mobility patterns start by staying at home at the midnight until morning. Then, users start moving to offices and stay until lunch hours. After an average one-hour lunch break, users go back to offices until evening. A percentage of people will go to the
recreational and dining areas after work. Finally, most of the
users return home in the evening.

6.2.2 School Students
This group includes 1400 primary and secondary school
students who live in townhouses. Starting at midnight,
the mobility of this user group is similar to the others, staying
at home until morning. At 7 am-7:30 am, all of the school
students go either to the primary or secondary school areas
and stay until 3-3:30 pm. After school, school students start
to travel around the community actively and return
home in the evening. This group of users has a significantly
synchronized mobility pattern to travel and stay during the
school period.

6.2.3 University Students
There are 900 university students live in apartment build-
ings. The mobility patterns start from the midnight, while
most of the university students stay in apartments until
morning. In the morning and afternoon, university students
move between the campus buildings and stay for classes
and activities. In the evening, students may go to recrea-
tional and dining areas or go home. The major difference
between adult staff and university students is that univer-
sity students have higher mobility when moving inside the
campus and a shorter average staying period.

6.2.4 Dependents
There are 1700 dependents in the simulation. They have
fixed home locations but without working locations. Their
staying locations and duration are random and unpre-
dictable. In general, the mobility patterns start from mid-
night while users stay at home, until morning. Then users

travel and stay randomly in the simulation area. The aver-
age staying duration is three hours with one-hour variance.

6.2.5 Distinct Patterns
These four user groups have distinct mobility patterns.
From the mobile network operator perspective, these move-
ments generate various daily periodic patterns in terms of
the number of users in each cell. For instance, the cells
covering the university campus area have more users in
working hours. The cells covering primary and secondary
schools have a significant decrease in users after the school
hours. Figure 5 shows the number of users in four selected
cells for each user group over a five-day simulation period.
As shown in Figure 4, the Cell 59191 covers half of the
university campus. Figure 5a indicates that it has more staff
in this cell in working hours. The number of staff increases
steadily in Cell 59191 starting in the morning and reaches
a local maximum in the morning and a global maximum in
the afternoon on a daily basis. In Figure 5b, the Cell 59123,
where the secondary school located, shows a significant
sharp increase of school students during the school hours.
In Figure 5c, the Cell 59143, which covers most of the univer-
sity student apartment, the peak numbers of university
students appear in the night when university students stay
in apartment buildings. In Figure 5d, the dependents have
no static locations to travel or stay. Therefore, the number
of users in each cell is chaotic. Figure 6 shows the number of
users in these four selected cells. It shows that each cell has
different peak hours according to its coverage area.

6.3 Contents
In the simulation, thirteen types of contents are evaluated.
Each type has different combinations of subscribers from the
previously mentioned user groups. There are ten contents
in each type, and each content has a random number of sub-
scribers. Table 1 shows the subscriber combinations for these
contents. The arrival time $t_a$ of contents follows a uniform
distribution between the second and the third day of the
simulation. The lifetime of contents also follows a uniform
distribution between one to three days. Each content has a
random size range from 250 MB to 350 MB.

6.4 Parameters and Assumptions
The time segment $t$ is set to be 10 minutes and the fading
constant $d$ is configured to be 8, respectively. The threshold
of the minimum number of subscribers to declare an active cell \( m_a \) is set as 1. The period for updating the broadcasting threshold \( t^b \) is set to be 24 hours for accompanying with the daily human mobility. There are total 6200 mobile users in 18 cells, and 130 contents to be distributed. For the network configuration, we assume the bandwidth is 10 MHz with 2 \( \times \) 2 MIMO. Without the loss of generality, it is assumed that the modulation and coding scheme, the transmission power, and the operation power for broadcasting are the same for all transmissions in each cell. The cells are synchronized for multicasting contents as configured in MBMS single-frequency network (MBSFN), with the transport block index of 9 for all cells. According to [30], [31], [32], [33], we assume a BBU in the BBU pool processes the baseband signal for one cell. Each BBU has one LC, each LC has an optical channel connected to the ONU at each RRH. The BBU processing time for each radio frame is assumed as 500 \( \mu \)s. The optical line bitrate is 2457.6 Mbit/s, the size of each optical frame is 38 880 B, and the optical frame duration is 125 \( \mu \)s. The \( P_{LC} \) is 20 W, \( P_{BBU} \) is 100 W, \( P_{CS} \) is 500 W, the number of transceiver chain \( N_{TRX} \) for each cell is 2, \( P_{ONU} \) is 20 W, \( P_0 \) is 84 W, \( P_{max} \) is 40 W, \( \Delta_p \) is 2.8, and \( P_{sleep} \) is 56 W.

### 7 Evaluation Results

In this section, a simulation with actual existing BS deployment is first evaluated. Secondly, a simulation in a small-cell environment is conducted to show the advantage of the proposed algorithm. A benchmark method is implemented for comparison. It is a generic multicasting method which delivers the content immediately in all active cells to subscribers once the content arrives. The cells without subscriber for that content do not transmit. The total number of active cells for the benchmark method will not be more than the total number of cells in the area. The proposed algorithm aims at lowering the number of active cells, the radio resource consumption, and the energy consumption below the benchmark method. Furthermore, the optimal solution found offline by exhaustive search is provided in the results for performance comparison.

#### 7.1 Actual Existing BS Deployment

In this deployment, there are 18 cells in the simulation area. In Figure 7, the number of active cells for multicasting each content by MOOD, the benchmark, and the optimal solution are shown as the red circle (○), blue cross (×) and black plus (+) curves, respectively. The red dash, blue dotted and black solid straight lines are the averages of performances for MOOD, benchmark, and the optimal solution, respectively.

Since contents arrived the network randomly, benchmark method delivered at any time in a day without considering the number of active cells. Therefore, the numbers of active cells were varying among contents. The average number of active cells for the benchmark is 14.13. In MOOD, since most of the contents were delivered at the time of the daily minimum number of active cells, the curve is steadier. The average number of active cells for the proposed algorithm is 11.98, having 15.2% of improvement. The optimal solution has an average of 11.69 number of active cells, which has only 2.5% of advantage comparing to MOOD.

In Figure 8, the total energy consumption of the radio access, the fronthaul, and the backhaul networks for transmitting each content is presented, based on the energy model in (8). It shows that MOOD had a closer performance to the optimal solution, and outperforms the benchmark method, especially in Type 2 contents. MOOD had 9.5% improvement in energy consumption compared to the benchmark method. Recall that the energy consumption of each content is increased with the radio transmission time, which depends on the content sizes. Therefore, the curves of the number of active cells are steadier than the curves of energy consumption.

It is notable that the first three types of contents have a larger improvement in the performance of the proposed algorithm. The history of numbers of active cells of content is discussed and is individually drawn in Figure 9. In the sub-figures, the x-axis is the time of the simulation period, and the y-axis is the number of active cells. The curve shows the number of active cells at a given time. The left red vertical dotted line is the content arrival time. The right green vertical dotted line is the content latest delivery time. The symbol circle (○), cross (×), and plus (+) indicate the delivery decisions made by MOOD, the benchmark, and the optimal solution, respectively.

Four contents of the first four types, which were only subscribed by one group users, are chosen to be investigated. These patterns of the number of active cells are highly correlated with the number of users’ pattern shown earlier in Figure 5. Figure 9a shows the history of Content 10 which was subscribed by only staff. The timing of reaching the minimum number of active cell in the afternoon of day 1, was the time of the peak number of users in Cell 59191. It was the only time segment that most subscribers were congregated in the cells which cover the university campus area. It is challenging to identify the optimal delivery timing in real time for the minimum number of transmissions. When the content arrived on day 2, it was delivered at the similar timing as discovered as minimal on day 1. Four more cells were put into the sleep mode comparing to the benchmark method which the content was immediately delivered. In Figure 9b, the pattern is the most obvious among the user groups. The school students congregated in the school at the same time every day. It is evident to deliver school student subscribed contents during school hours that having the minimum number of transmissions. Therefore, the proposed algorithm had the best performance on delivering Type 2 contents. The lowest number of active

### Table 1: Subscriber Group Distribution of Contents

<table>
<thead>
<tr>
<th>Type (Content ID)</th>
<th>Staff</th>
<th>S. Students</th>
<th>U. Students</th>
<th>Dependents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1 (1-10)</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Type 2 (11-20)</td>
<td>25%</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>Type 3 (21-30)</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Type 4 (31-40)</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Type 5 (41-50)</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Type 6 (51-60)</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Type 7 (61-70)</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Type 8 (71-80)</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>Type 9 (81-90)</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>Type 10 (91-100)</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>Type 11 (101-110)</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>Type 12 (111-120)</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>Type 13 (121-130)</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>
cells was four which covering all of the primary and secondary schools. The university students congregated in the apartment buildings after midnight; that is the best timing for content delivery to university students as shown in Figure 9c. The proposed algorithm successfully minimized the number of transmissions in both Type 2 and Type 3 contents. However, since the mobility patterns of dependents group are mostly unpredictable, the minimum numbers of active cells in Type 4 contents were large comparing to other types. As shown in Figure 9d, the content was delivered in 14 cells. Therefore, the proposed algorithm was slightly improved only.

For other types of content, Type 6 had an outstanding performance among the others. It is because half of the subscribers of Type 6 were school students that contributed to the congregation. The remaining types show similar performance that having only 1 to 2 active cells less than the benchmark method. In summary, MOOD outperformed the benchmark method and close to the optimal solution.

7.2 Small-cell environment

In this section, a simulation with a small-cell deployment is conducted, i.e., micro cells, to show the performance of the proposed MOOD algorithm. The simulation has the same configuration and assumptions as the previous simulation environment excepted the cell deployment and power model. According to [31], number of transceiver chain $N_{TRX}$ for each cell is 2, $P_0$ is 56 W, $P_{max}$ is 6.3 W, $\Delta P$ is 2.6, and $P_{sleep}$ is 39 W. The new deployment is shown in Figure 10. It is a typical hexagonal cell deployment with about 320 meters inter-BS distance. The number of cells increases rapidly from 18 to 115 in the entire region. The main characteristic of this scenario is that a cell is more focus to cover a single type of buildings. It gives higher potential to MOOD for making content delivery decisions according to the users’ mobility.

Figure 11 shows the number of users of each user group in the five selected cells. Each cell covers a specific type of buildings. Cell 1016 covers building apartments only, where mainly occupied by university students and a few staff. Cell 1024 covers staff housings only. Cell 1030 covers part of the university campus and the campus canteen. Cell 1044 mainly covers a school and some staff housing. Cell 1054 covers a recreational and dining building. In Figure 11a, Cell 1030 shows a significant peak daily during the lunch hour when most of the staff have their meals in the canteen. In Figure 11b, Cell 1044 shows a larger daily difference
than before, from almost no users to close to 500 users during school hours. In Figure 11c, Cell 1030 has similar pattern observed in staff, having a sharp peak in lunch hours. Moreover, Cell 1016 which is covering the student apartment, shows significantly a wider daily peak in the night when students are staying in the apartment buildings. Finally, in Figure 11d, the cells covering the dining area have distinct daily patterns which do not appear in the actual BS deployment.

The number of radio resource blocks consumed for multicasting each content is shown in Figure 12. The benchmark method had an average result of 2.416e+5 but only 1.802e+5 in the proposed algorithm. The proposed algorithm had reduced 25.4% of radio resource blocks in the small-cell environment comparing to only 15.2% in the previous cell deployment. In average, MOOD achieved 96.5% of the optimal solution. Similar to the previous results, the first three types of content had significant improvement by the proposed algorithm. Surprisingly, both Type 5, 6 and 11 contents outperformed to the benchmark method to have a low number of active cells, especially contents subscribed by staff that the nadir was very short in time. Second, a short lifetime of the content leads to a fast-growing broadcasting threshold. This problem is severe especially when the lifetime of content is close to a day. The algorithm delivers the content before the next nadir. In Figure 15a, Content A arrived at the time just after the nadir on Day 2. Benchmark approach caught a good timing to deliver the content in only 30 cells. Since the short lifetime of this content, the broadcasting threshold grew quickly, and the content was delivered just before it expired. MOOD delivered the content in 47 cells, 56% and 80% worse than the benchmark and the optimal solution, respectively. Content B is the same type of content as Content A. The lifetime of Content B was longer than one day, and it did not fail like Content A as shown in Figure 15b.

8 Discussions and Future Work

In this section, some use cases, which are considered suitable for MOOD, are discussed as well as some limitations are highlighted for future investigations.

The proposed algorithm is suitable for delivering contents which are usually large in file size and delay-tolerant, such as software updates and subscribed contents. Software updates for mobile applications and operating systems are frequently released for new features, bugs fixing, new contents publishing and so on [34]. First, Carbunar et al. found that some popular applications consume an enormous amount of bandwidth for pushing updates to the users in unicast connection [35]. Second, Möller et al. found that almost half of the users do not install an update a week after it published, which is a critical security issue [36]. Therefore, by designing an algorithm to push software updates efficiently within a short period is essential to reduce network congestion and security risks.

User-subscribed contents, such as games, video, and music, could be pre-loaded to user devices before the official
release date and time. The contents are typically encrypted during the transmission. A decryption key will be sent to devices at the release time for unlocking. Most of the computer and video gaming platforms are now supporting pre-loading service. The large size game installation files can be downloaded before the release date. The network traffic can be distributed within the pre-loading period to reduce the sudden congestion on the server. A game file has a large
size of up to tens of gigabytes and days for the pre-loading period. On the other hand, new television program episodes and new songs could also be delivered to the subscribers’ devices before the release time. A multimedia content could have file sizes up to gigabytes and short pre-loading periods within a day. For these massive on-demand contents with a pre-loading period, it is essential to be delivered efficiently.

Our proposed algorithm, in its current shape, addresses several important issues but has few limitations. First, it is not suitable for live broadcast. Live contents, such as television and radio broadcast require real-time transmission with shortest end-to-end delay. It should be delivered to the users immediately without considering users’ location or mobility. Second, it is challenging to multicast contents which are shared on the online social network. For instance, if a user uploads a picture to share onto his/her social network account, the friends of this user will view this picture once they access the social network website or application. However, it is hard to predict when will his/her friends access the social network to view this picture. Therefore, MOOD is not suitable for contents which have fast spreading rate and short lifetime. The 5G network should have the ability to promptly detect popular contents shared on the social media and broadcast them efficiently to the geographically closed social network. Nevertheless, combining MOOD with the concept of social network switches proposed by Su et al. in [37] may solve this problem. Finally, the human mobility model presented in this article cannot represent all of the mobility cases. It merely simulates typical daily routine of people. Modeling special events such as protests or carnivals will be difficult. The mobility of people in these events is tremendously different from a normal working day. Furthermore, some user groups that do not follow any routine in their daily lives are difficult to be captured in MOOD. Therefore, applying machine learning techniques on extracting mobile users’ moving patterns would be our future target [38].

9 Conclusion

In this paper, an efficient content delivery approach named MOOD is proposed to reduce radio resources and energy consumption for transmitting duplicated contents in the radio access network. MOOD transforms urban scale users’ mobility into a broadcasting threshold for determining the time and the location for broadcasting contents. Four types of users, with entirely different daily mobility patterns, were modeled in the simulation for evaluation. The results show that MOOD outperforms the benchmark method, which broadcasts a content immediately to the entire region on its arrival, in reducing the number of transmissions and energy consumption. Furthermore, the simulation results indicate that MOOD achieves near-optimal solutions. Our approach can be further improved by identifying popular contents from online social networks and predicting users’ mobility through machine learning methods.

References


