Optimal Mobile Content Downloading in Device-to-Device Communication Underlaying Cellular Networks

Yong Li, *Member, IEEE*, Zhaocheng Wang, *Senior Member, IEEE*, Depeng Jin, *Member, IEEE*, and Sheng Chen, *Fellow, IEEE*

***Abstract*—With the emerging demands for local area services of popular content downloading, device-to-device (D2D) communication is conceived as a vital technological component for next-generation cellular communication networking to increase the spectral efficiency and to enhance the system capacity. Targeting the application of mobile content downloading, we investigate the fundamental problems of how D2D communication improves the system performance of cellular networks and what is the potential effect of D2D communication, with the aid of the optimal solutions for the system resource allocation and mode selection obtained under the realistic user and mobility conditions. Specifically, by formulating a max-flow optimization problem that maximizes the content downloading flows from all the cellular base stations to the content downloaders through any possible ways of transmission, we obtain the theoretical upper bound to system content-downloading performance. Using realistic mobility model and trace driven simulations, we evaluate the effects of the different system settings on the performance of the mobile content-downloading system, and reveal the fundamental influence of D2D communication.**

***Index Terms*—Device-to-device communication, mobile data downloading, cellular networks.**

I. Introduction

MOBILE Internet access is getting ever-increasingly popular and it today provides various services and applications, including audio, images and video. Cisco estimates that mobile data traffic grew at an annual rate of 170% in 2012, and will reach over 10 exabytes per month in 2017 [1]. Moreover, two-thirds of the world's mobile data traffic will be video by 2017, according to Cisco's forecast [1]. Huge portion of the total mobile data traffics delivered by mobile service providers, such as weather forecasts, multimedia newspapers, stock information, movie trailers, and etc, are typically delivered or broadcasted to large number of mobile users. Currently, these mobile content downloadings are supported by cellular networks, which is the most popular method of mobile access today [2]. With the explosive increase in mobile services and user demands, cellular networks will, very likely, be overloaded and congested in the near future. Pessimists already foresee near nightmare scenarios that, during peak time and in urban area, users face extreme performance hits in terms of low or even no network bandwidth, missed voice calls, and unreliable coverage. Indeed, the limited spectrum and over-the-air bandwidth constrain multimedia mobile content downloading applications, such as mobile TV and voice, streaming music or video downloads.

As one of the next generation wireless communication systems, Third generation partnership project Long Term Evolution (LTE) is committed to provide technologies for high data rate transmission system, and LTE-Advanced (LTE-A) is defined to support new components for LTE to enable higher-rate communication and mobile content downloading demands [3]. Combined with the emerging demands for local area services of popular content downloading, Device-to-Device (D2D) communication is proposed as a key component for LTE-A, which enables devices to communicate directly, and it is an underlay to the cellular network for increasing the spectral efficiency [4]–[6]. In D2D communications, under the control of Base Stations (BSs), User Equipments (UEs) transmit data to each other over direct links using the cellular resources, instead of through BSs. Thus, most of context-aware applications that involve discovering and communicating with nearby devices, including the popular content downloading, can benefit from the D2D communication by reducing the communication cost since it enables physical-proximity communication, which saves communication power while improving the spectral efficiency. Therefore, the potential of D2D communications to improve spectral utilization is being promoted in the recent years [5]. It is expected that D2D communication will be a key feature supported by the next-generation cellular networks [6].

Although D2D communication may enhance the spectral efficiency and increase the system capacity, it also causes interference to the cellular network as the results of spectrum sharing. Current works mainly focus on power control [7]–[9],

Manuscript received March 22, 2013; revised July 14, 2013, September 28, 2013, November 16, 2013, and February 6, 2014; accepted March 29, 2014. Date of publication April 4, 2014; date of current version July 8, 2014. This work was supported by the National Key Basic Research Program of China (973 Program Grant Nos. 2013CB329001 and 2013CB329203), the National Nature Science Foundation of China (Grant Nos. 61301080, 61171065, and 61273214), the National High Technology Research and Development Program (Grant Nos. 2013AA013501 and 2013AA013505), and the Program for Changjiang Scholars and Innovative Research Team in University. The associate editor coordinating the review of this paper and approving it for publication was Y.-C. Ko.

Y. Li, Z. Wang, and D. Jin are with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: liyong07@tsinghua.edu.cn; zcwang@tsinghua.edu.cn; jindp@tsinghua.edu.cn).

S. Chen is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K., and also with King Abdulaziz University, Jeddah 21589, Saudi Arabia (e-mail: sqc@ecs.soton.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2014.2315807

resource allocation [10], [11] and interference management [12]–[14]. Thus, significant research has been carried out on the resource allocation and power control algorithms that consider mutual interference between D2D and cellular communications where the D2D communications consume part of the cellular resources delicately. To further increase the spectrum reuse ratio, our work considers the even more challenging senario where all the D2D communications occur on the same part of the frequency resources. However, the main difference between our work and these existing works is the network scale. The existing works [7]–[14] consider the problem under a very restricted cellular system setting consisting of only four nodes—a pair of D2D UEs, a cellular UE, and a BS. Within such a small and simplified network scenario, these existing studies can only deal with some individual aspects of the D2D underlaying cellular network. None of these existing works, however, has tackled the underlying problem as a whole and, therefore, they are unable to quantify the actual potential of the D2D enabling cellular network under a realistic network scenario with tens of hundreds nodes and multiple D2D pairs sharing one cellular user's spectrum. In this paper, we identify the mobile content downloading performance bound achievable through the D2D underlaying large scale cellular network with multiple cellular users and D2D pairs, and answer the challenging questions of how D2D can improve the cellular network system performance and what are the potential effects of D2D.

Intuitively, D2D communication consumes some spectrum resources of the cellular network, while it may increase the resource utilization by the reuse of the spectrum for the communicating devices that are physically in close proximity to enable very high bit-rate, low delay and low power-consumption communication. Thus, D2D communication depends on how often the devices are in physical-proximity communication with each other and how long the communication can last. In other words, underlaying D2D communication opportunities and therefore the performance of the D2D underlaying cellular system depends on the node mobility patterns. On the other hand, the D2D underlaying cellular network's performance also depends on the adopted system resource allocation and mode selection schemes. Since D2D communications share resources with normal cellular communications and therefore impose new interferences, we need to optimize the resource allocation to effectively manage interference while maximizing the gains of D2D communications. Mode selection decides whether D2D candidate pairs should communicate in D2D mode or in cellular mode, and an appropriate mode should be selected according to the available bandwidth to achieve an optimal system performance. It is clear that to obtain the fundamental performance bound achievable by D2D communication, it is essential to first find the optimal solutions for the system resource allocation and mode selection under the practical network setting of a large number of users with realistic mobility patterns.

From a system engineering viewpoint, we make some realistic assumptions, namely, the availability of the preemptive knowledge of BS deployment and user mobility trajectories as well as the optimal scheduling of mode selection and resource sharing for the content downloading. With the aid of the available system information, we can cast the content downloading process in the D2D communication underlaying cellular network as a max-flow optimization problem that maximizes the content downloading flows from all the cellular BSs to the content downloaders through any possible communicating means that include directly cellular transmission, D2D enabled connected transmission and opportunistic transmission. The solution of this max-flow problem yields the optimal resource sharing and mode selection over the whole cellular network with a given set of users. Thus, it represents the theoretical performance upper bound to the content downloading system. To the best of our knowledge, this is the first study on the performance bound of the D2D communication underlaying cellular system under realistic networking scenarios with large-scale node mobility and all possible transmission modes. Our novel contributions are summarized as follows.

- We introduce the dynamic graph to model the D2D communication underlaying cellular network with multiple D2D pairs and node mobility patterns. This model is used to set up a realistic D2D underlaying cellular network for studying the performance limits of the content downloading services.
- We formulate the optimal mobile content downloading problem in D2D communication underlaying cellular networks as a flow maximization problem based on the dynamic graph model, which takes into the consideration of the resource allocation and content transmission mode selection. The solution of this max-flow problem reveals the theoretical performance bound of the system.
- To highlight the effects of the D2D communication on cellular networks, we evaluate the influence of the different networking environments on the performance of the mobile content downloading system with extensive simulation results.

The rest of the paper is organized as follows. The mobile content downloading system in the D2D communication underlaying network is given in Section II, while the dynamic graph is described in Section III to model the content downloading process which takes into account the user mobility and all possible transmission modes. In Section IV, we formulate the flow-max optimization problem with the consideration of the resource allocation and mode selection for the mobile content downloading system, and obtain its optimal solution that represents the achievable performance limit of the system. Section V introduces the experimental simulation environment and presents the simulation results for the performance evaluation of the proposed mobile content downloading system in the D2D communication underlaying network. In the light of our results, we conclude the paper in Section VI.

II. Mobile Content Downloading System

The envisaged mobile content downloading system in D2D underlaying cellular networks is illustrated in Fig. 1, where the cellular network provides the coverage over a certain region through BSs and the UEs here refer to the users with mobile phones or other devices that travel around. The UEs are naturally mobile nodes, and their positions change with

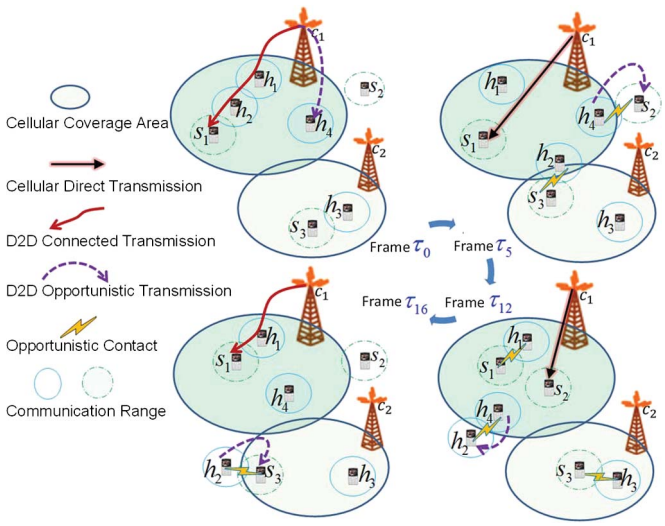


Fig. 1. Illustration of the mobile content downloading in D2D underlaying cellular networks, where there are 2 cellular BSs, c_1 and c_2, and 7 mobile UEs, which include 3 mobile content downloaders, s_1, s_2, and s_3, whose communication ranges are denoted by dotted and green circles, and 4 mobile content downloading helpers, h_1, h_2, h_3, and h_4, whose communication ranges are denoted by solid and green circles. A downloader can receive the data from the cellular network directly (cellular transmission mode) or from the downloading helpers/relays (D2D transmission mode). Furthermore, a D2D transmission can either be multi-hop connected transmission or delay-tolerant opportunistic transmission.

the time. Therefore, at different time frames,[1] their access and physical (location) relationships may be different. For example, Fig. 1 displays the access and physical relationships at the different system time frames of τ_0, τ_5, τ_{12}, and τ_{16}, respectively. The BSs are connected to the content servers in the Internet through wired-line links. The users requesting mobile contents send their data requests to the relevant content servers via the cellular network. Then, the requested data are delivered from the corresponding content servers to the related users either via directly cellular transmissions, if the users are in the cellular coverage, with certain transmission rates, or via D2D based transmissions when they physically encounter some other users that have their requested mobile contents.

In this mobile content downloading system, UEs that are requesting data are referred to as mobile content *downloaders*. Other UEs that are currently not retrieving mobile contents for themselves may participate in the content downloading to receive the data from BSs and then transmit them to the relevant downloaders via D2D communication. These UEs are referred to as content transmission *helpers*. In our proposed system, we assume that all UEs are rational and, therefore, they will participate in the content downloading as helpers only when they are not currently retrieving the contents for themselves. Incentives for the helpers can be given by using some micro-payment scheme, or the operator can offer the helpers a reduced cost for their service or better quality of service [15]–[17]. A full analysis of such incentives is beyond the scope of this paper. In the example depicted in Fig. 1, there are 2 cellular BSs marked as c_1 and c_2, three mobile content downloaders known as s_1, s_2, and s_3 whose communication ranges are denoted by the respective dotted and green circles, and four mobile content downloading helpers called h_1, h_2, h_3, and h_4 whose communication ranges are denoted by the corresponding solid and green circles. In this D2D underlaying cellular network, there exist the following three different modes for mobile content downloading.

1) *Directly cellular transmission*. The downloaders inside the coverage of the cellular network receive their mobile contents directly from BSs. This mode is the original way of communication in cellular networks. However, due to coverage leak and signal attenuation, this mode may fail to provide the required content transmissions when the downloaders are out of coverage.

   For example, in Fig. 1, cellular BS c_1 transmits directly to downloader s_1 during the time frame τ_5 and transmits directly to downloader s_2 during the time frame τ_{12}.
2) *D2D connected transmission*. A connected path from a BS via some helpers to a downloader is established by taking the advantage of the physical proximity of communicating devices. The mobile data are send via this D2D enabled connected path to the downloader.

   The example of Fig. 1 shows that during the time frame τ_0, BS c_1 transmits the content to downloader s_1 via the D2D connected link with the aid of helpers h_1 and h_2, while during the time frame τ_{16}, c_1 transmits the content to s_1 using the connected $c_1 \rightarrow h_1 \rightarrow s_1$ path.
3) *D2D opportunistic transmission*. Owing to the mobility of downloaders and helpers, a D2D connected path is prone to be broken. However, a helper can store the received content in its buffer, and transmits the data to the relevant downloader or other helpers when communication contact arises. This is known as store-carry-and-forwarding. Since UEs are inherently mobile, this opportunistic communication mode is capable of further enhancing the system performance.

   In Fig. 1, over the time frames τ_0 to τ_5, downloader s_2 is outside the coverages of both BSs. During the time frame τ_0, helper h_4 receives the data from BS c_1 and stores the data in its buffer. When the communication opportunity occurs between h_4 and s_2 during the time frame τ_5, h_4 transmits the content to s_2. Another example is that h_4 opportunistically transmits the data to h_2 during the time frame τ_{12} and then h_2 transmits the data to s_3 when they meet during the time frame τ_{16}.

The D2D connected and opportunistic transmission modes are inherently multi-hops. Clearly, either modes must have at least two hops with the first hop from a BS to a helper and the second hop from this helper to a downloader. In theory, there is no limit imposed on the maximum number of relays that can be used to deliver the content to a downloader. However, in practical systems, the hop count is usually limited because multi-hop communication may yield link cost since it requires multiple time slot resources in the half-duplex mode. In our problem formulation, the effect of hop number to the aggregated system throughput is implicitly considered.

[1]The time frames are used to mark the system time, which will be formally defined in the next section.

Since the traditional D2D technologies based on Bluetooth and WiFi work at 2.4 GHz unlicensed band [6], they are inadequate for the envisaged mobile content downloading system. We therefore focus on the operator controlled D2D communication where the devices communicate directly with each other under the control of the cellular network, which include access authentication, connection control and resource allocation [6]. In the operator controlled D2D system, D2D communications between devices use the same licensed band of cellular communication, and they assume the same air interface of the underlaying cellular network. Thus, D2D communications consume part of the resources allocated to the cellular network. This resource sharing with the directly cellular communication is orthogonal, since the D2D communication and the cellular transmission occur on the different frequency channels, and there exists no interference between the cellular and D2D communications. However, to reuse the network resource, all the D2D communications occur on the same frequency channel, and this shared channel bandwidth is allocated for D2D communication using the IEEE 802.11 based MAC protocol. Clearly, there exists the interference between two D2D communications, and this interference will influence the achievable communication rate. As a D2D communication only occurs when two UEs are close-proximity neighbors, by carefully defining the communication range and the transmit power, the interference between two different D2D communications can be limited/managed.

To achieve our objective of investigating the optimal content downloading system that maximizes the system throughput by appropriately utilizing the available content transmission opportunities in the D2D underlaying cellular network, we must have the solutions of the following two problems underpinning the optimal content downloading system.

(i) *Transmission Mode Selection*. Since the D2D communication uses the same air interface of cellular communication, a UE can only operate either in the D2D mode or in the directly cellular transmission mode. Furthermore, in the D2D mode, a decision must be made whether to use a connected or an opportunistic transmission. Given all the possible transmission modes involving all the UEs, the task is how to utilize them to maximize the content transmission throughput from all the cellular BSs to all the downloaders.

(ii) *Optimal Resource Allocation*. Given the available network resource of frequency channels and the content transmission opportunities between BSs and UEs as well as between UEs, the task is to allocate the resource between D2D communications and cellular communications to attain the maximum content transmission throughput.

Given the knowledge of the deployment and coverage of cellular BSs, UEs' mobility trajectories, and the prefect scheduling of data transmission in terms of resource allocation and mode selection, we first build a graph of the transmission evolution in the dynamic D2D underlaying cellular communication by processing the cellular network layout and the associated UEs' mobility traces. Based on this graph model, we then formulate the aggregated system throughput problem as a flow maximization problem. The solution of this max-flow problem provides the optimal network scheduling of mode selection and resource allocation, which allows us to derive the upper bound of the network throughput performance.

III. Dynamic Graph for D2D Underlaying Cellular Communication

Now, we utilize the model of dynamical graph [18], [19] to model the D2D D2D communication underlaying cellular networks, which includes all the possible transmission opportunities of different modes evolving in time. The graph considers three different types of nodes in the system. There are C BSs labelled as $\mathcal{C} = \{c_1, c_2, \ldots, c_C\}$, each of which provides a stationary area of cellular coverage. We further assume that there are no overlapped coverage area but there exist the areas of no cellular coverage. Among the U UEs, there are S content downloaders retrieving the contents from the Internet labelled as $\mathcal{S} = \{s_1, s_2, \ldots, s_S\}$, while the other $H = U - S$ UEs are the content transmission helpers labelled as $\mathcal{H} = \{h_1, h_2, \ldots, h_H\}$, who are involved in the D2D communication to help the content downloading for downloaders.

To describe the UEs' dynamic accessing relationships with the BSs as well as the time-varying communication opportunities and paths among the helpers and downloaders, we adopt the time-expanded graph, and generate the graph from the mobility trace. To model all the possible content transmission opportunities of different modes, we identify the communication contact events between a pair of nodes, which may be BS and UEs, helper and downloader, or helper and helper. These contact events include the following five types.

1) Cellular accessing starts: the time when a UE moves into the coverage of a cellular BS, and it establishes a link with the BS.
2) Cellular accessing ends: the time when a UE moves out of the coverage of a cellular BS, and the established communication link is disconnected.
3) D2D contact starts: the time when a pair of two UEs, either helper and helper or helper and downloader, physically move into the communication range of each other and they can transmit the mobile content from one to another by a D2D communication link.
4) D2D contact ends: the time when a pair of two UEs, which have been in the D2D communication state, moves outside the communication range of each other, and the established link is broken.
5) Link quality level changes: the time that the quality of communication link changes. Several metrics can be considered, and in this work we take the achievable data rate of the link as the link quality metric.

The above communication contact events, which are assumed to be sequentially occurring at the different time points[2] of $t_0, t_1, \ldots, t_N$, divide the continuous time into time periods. We define the period between two successive events as time

[2]This assumption is made purely for the purpose of simplifying the description of a dynamic graph model, while in practice more than one contact events can occur at the same time point in the graph.

frame. Specifically, the time period between t_{l-1} and t_l is labelled as frame τ_l, with the initial time frame τ_0 defining the time before t_0. Thus, all the time frames are labelled as $\tau_0, \tau_1, \ldots, \tau_N$. Within each time frame, we further assume that no contact event occurs and no link quality changes. In other words, within each time frame, the states of the nodes and the commutation contacts do not change. Otherwise, the system evolves into the next time frame. For notational simplification, we will also use l to denote time frame τ_l in the sequel.

In the dynamic D2D graph, each network node at each frame is represented by a vertex. For a BS $c_i \in \mathcal{C}$ where $1 \leq i \leq C$, we denote the corresponding vertex at frame l as c_i^l. Similarly, we denote the vertices of a downloader $s_i \in \mathcal{S}$ for $1 \leq i \leq S$ and a helper $h_i \in \mathcal{H}$ for $1 \leq i \leq H$ at frame l as s_i^l, and h_i^l, respectively. We further denote all the vertices generated at frame l by all the BSs, helpers and downloaders as $\mathcal{C}^l$, $\mathcal{H}^l$ and $\mathcal{S}^l$, respectively. For convenience, we define the set of the vertices of the helpers and downloaders, i.e., all the UEs, by $\mathcal{U}^l = \mathcal{H}^l \cup \mathcal{S}^l$. Within time frame l, a directional edge (c_i^l, u_j^l) exists from vertex c_i^l to vertex u_j^l if the node $u_j \in \mathcal{U}$ are in the coverage of $c_i \in \mathcal{C}$. We denote all these edges at l as $\mathcal{E}_c^l$, which represent all the directly cellular transmission links for the downloader vertices $s_j^l \in \mathcal{S}^l$ and all the first-hop links of D2D communications for the helper vertices $h_j^l \in \mathcal{H}^l$. Similarly, if a pair of helper and downloader, denoted by h_i and s_j, are in the contact at frame l, a directional edge (h_i^l, s_j^l) exists from vertex h_i^l to vertex s_j^l. All these edges are denoted by $\mathcal{E}_s^l$, and they represent the content transmission opportunities from the helpers to the downloaders. For a pair of helpers h_i and h_j, which are in the contact at frame l, they can transmit content to each other and, therefore, a bidirectional edge exists between h_i^l and h_j^l, and all these bidirectional edges are denoted by $\mathcal{E}_h^l$. For notational convenience, all the edges (u_i^l, u_j^l) where $u_i^l, u_j^l \in \mathcal{U}^l$ will be denoted as $\mathcal{E}_u^l$, i.e., $\mathcal{E}_u^l = \mathcal{E}_h^l \cup \mathcal{E}_s^l$. For any edge at time frame l, namely $(a_1, a_2) \in \mathcal{E}^l = \mathcal{E}_c^l \cup \mathcal{E}_u^l$, we assign a weight $w(a_1, a_2)$ to (a_1, a_2), which represents the achievable content transmission rate of (a_1, a_2).

All the edges in $\mathcal{E}^l$ represent the available content transmission opportunities in frame l, and $\{\mathcal{E}^l\}_{l=0}^N$ represent the available content transmission opportunities evolving in time. A directly cellular transmission from BS c_i to downloader s_k at frame l occurs at the edge $(c_i^l, s_k^l) \in \mathcal{E}_c^l$. A two-hop D2D connected transmission from c_i via helper h_j to s_k at frame l occurs at the link that consists of the two edges $(c_i^l, h_j^l) \cup (h_j^l, s_k^l)$, where $(c_i^l, h_j^l) \in \mathcal{E}_c^l$ and $(h_j^l, s_k^l) \in \mathcal{E}_s^l$. A D2D opportunistic transmission from h_i to s_j at frame l happens on the edge $(h_i^l, s_j^l) \in \mathcal{E}_s^l$. Note that h_i in this case must have received the content and stored the content in its buffer in some previous frame $l' < l$, either via a directly cellular link $(c_k^{l'}, h_i^{l'})$ from a BS vertex $c_k^{l'}$ or via an opportunistic link $(h_k^{l'}, h_i^{l'})$ from another helper vertex $h_k^{l'}$, before its opportunistic contact with s_j at time point t_{l-1}. Therefore, we need to model the time evolution of content buffering for helpers and similarly for BSs in the graph. For this aim, a directional edge (b_i^l, b_i^{l+1}) of infinite weight is drawn for the same BS or helper $b_i \in \mathcal{B} = \mathcal{C} \cup \mathcal{H}$ between two successive frames. All the these edges are denoted by $\mathcal{E}^b$.

TABLE I
NETWORK EVENTS FOR THE D2D UNDERLAYING CELLULAR NETWORK INVOLVING TWO BASE STATIONS (c_1, c_2), THREE DOWNLOADERS (s_1, s_2, s_3) AND FOUR HELPERS (h_1, h_2, h_3, h_4)

Time Point	Event	Meanings
t_0	$h_1 \nleftrightarrow h_2$	Contact between h_1 and h_2 ends
t_1	$h_2 \nleftrightarrow s_1$	Contact between h_2 and s_1 ends
t_2	$h_4 \leftrightarrow s_2$	Contact between h_4 and s_2 starts
t_3	$h_3 \nleftrightarrow s_3$	Contact between h_3 and s_3 ends
t_4	$h_2 \leftrightarrow s_3$	Contact between h_2 and s_3 starts
t_5	$h_1 \leftrightarrow s_1$	Contact between h_1 and s_1 starts
t_6	$h_4 \nleftrightarrow s_2$	Contact between h_4 and s_2 ends
t_7	$s_2 \rightarrow c_1$	s_2 moves into coverage of c_1
t_8	$h_2 \nleftrightarrow s_3$	Contact between h_2 and s_3 ends
t_9	$h_2 \nrightarrow c_1$	h_2 moves out coverage of c_1
t_{10}	$h_2 \leftrightarrow h_4$	Contact between h_2 and h_4 starts
t_{11}	$h_3 \leftrightarrow s_3$	Contact between h_3 and s_3 starts
t_{12}	$s_2 \nrightarrow c_1$	s_2 moves out coverage of c_1
t_{13}	$h_2 \nleftrightarrow h_4$	Contact between h_2 and h_4 ends
t_{14}	$h_3 \nleftrightarrow s_3$	Contact between h_3 and s_3 ends
t_{15}	$h_2 \leftrightarrow s_3$	Contact between h_2 and s_3 starts

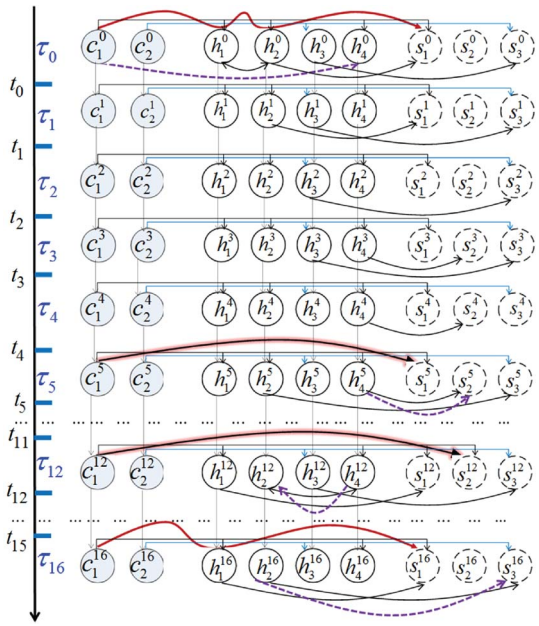


Fig. 2. The dynamic D2D underlaying cellular communication graph generated by the events of opportunistic contact and access relationship that are given in Fig. 1 and Table I.

Given all the vertices and edges, we can build the dynamic graph for the D2D underlaying cellular communication, which is a weighted directional graph representing the spatial-temporal evolution of the network topology involving all the BSs and UEs. Consider the D2D underlaying cellular network illustrated in Fig. 1, which involves two BSs (c_1, c_2), three downloaders (s_1, s_2, s_3) and four helpers (h_1, h_2, h_3, h_4). Assume that there are no link quality changing events, and we have counted all the network events of cellular access start and end as well as D2D contact start and end, for the time points t_0 to t_{15}, which are summarized in Table I. Fig. 2 depicts the dynamic graph for this D2D underlaying cellular network, which explicitly shows the spatial-temporal evolution of the network topology of all the BSs and UEs. For graphical clarification, we omit the weights of edges. In this dynamic graph, the contact events are highlighted by the time points at which

links are established and lost, and time frames correspond to rows of the vertices, while nodes correspond to columns of the vertices. Note that the dynamic graph allows us to capture all the possible content downloading modes, including the directly cellular transmission, and D2D connected as well as opportunistic transmission modes.

IV. Maximum Throughput Content Downloading

Based on the spatial-temporal dynamic graph for D2D communication underlaying cellular networks, we formulate the content downloading throughput optimization by expressing the maximization objective and analyzing the constraints.

A. Problem Formulation

The content downloading throughput optimization must jointly consider transmission mode selection and resource allocation. Since our dynamic graph represents all the possible content transmission opportunities and modes, we need to associate the resource allocation with this dynamic graph. In terms of the resource allocation, we assign the bandwidth between D2D communications and cellular communications, and the allocated resource directly influences the content transmission rates for the cellular communication and the D2D communication. Therefore, the weight of each edge, namely, the "flow" rate of each directional edge, is directly associated with the allocated resource. The resource allocation policies are assumed to be scheduled at the same frequency as the network events, while within a time frame, the allocated resources remain unchanged and, therefore, the content transmission rates are also kept constant within a time frame.

Consider the cellular coverage area by BS $c_i \in \mathcal{C}$, $1 \le i \le C$, at frame l. Denote the allocated spectrum resource for the directly cellular communication between this BS and UEs as x_i^l. Under the Rayleigh fading channel model, we can express the maximum achievable average data rate for the link between c_i and UE u_j, denoted by R_{c_i}, as follows:

$$R_{c_i} = x_i^l \log_2\left(1 + \frac{P_{c_i}\zeta_{c_i,u_j}^{-\ell}\left|h_{c_i,u_j}\right|^2}{N_0}\right) \tag{1}$$

where P_{c_i} is the transmitted signal power of BS c_i, ζ_{c_i,u_j} is the distance between c_i and u_j, ℓ is the path loss exponent, N_0 is the power of the receiver noise which is assumed to be the additive white Gaussian noise (AWGN), and $|h_{c_i,u_j}|^2$ denotes the average power or second-order statistic of the Rayleigh fading channel linking c_i and u_j. As expected, the content transmission rate to UE u_j is proportional to the allocated resource x_i^l. Thus, the weight for the edge (c_i^l, u_j^l) can be expressed as $w(c_i^l, u_j^l) = \alpha_i^{lu_j} x_i^l$ with $\alpha_i^{lu_j} = \log_2(1 + (P_{c_i}\zeta_{c_i,u_j}^{-\ell}|h_{c_i,u_j}|^2/N_0 0)$.

Denote the set of all the D2D communication pairs in the coverage area of c_i as $\mathcal{G}$, and assume that the allocated bandwidth resource at time frame l is y_i^l for D2D communication. Since the D2D communication pairs in the coverage area of c_i share the same spectrum, we must consider the interference between all the different D2D pairs, and the average transmission rate of the D2D pair $g \in \mathcal{G}$, denoted by R_g, can be expressed as

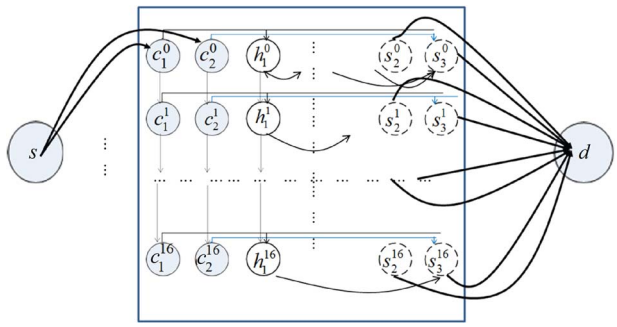


Fig. 3. Illustration of the flow maximization formulation for the optimal throughput content downloading problem.

$$R_g = y_i^l \log_2\left(1 + \frac{P_g\zeta_g^{-\ell}|h_g|^2}{\sum\limits_{g' \in \mathcal{G}\setminus g} P_{g'}\zeta_{g',g}^{-\ell}|h_{g',g}|^2 + N_0}\right) \tag{2}$$

where P_g and $P_{g'}$ are the transmission powers of the two D2D pairs g and g', respectively, and ζ_g denote the distance between the transmitter and receiver of the D2D pair g with $|h_g|^2$ being the average power of the corresponding Rayleigh fading channel, while $\zeta_{g',g}$ is the distance from the transmitter of the D2D pair g' to the receiver of the D2D pair g with $|h_{g',g}|^2$ being the average power of the corresponding Rayleigh fading channel. Clearly, the content transmission rate of the D2D pair $g = (u_j^l, u_k^l)$ can be expressed as $w(u_j^l, u_k^l) = \beta_{jk}^l y_i^l$ with $\beta_{jk}^l = \beta_g^l = \log_2(1 + (P_g\zeta_g^{-\ell}|h_g|^2/\sum_{g' \in \mathcal{G}\setminus g} P_{g'}\zeta_{g',g}^{-\ell}|h_{g',g}|^2 + N_0))$. The expression (2) is generic, including the situation with inactive D2D pairs. In fact, if a D2D pair g' is inactive, we simply set the corresponding transmission power $P_{g'}$ to zero in (2).

To formulate the max-flow problem for the dynamic graph of the D2D underlaying cellular communication, we introduce two virtual vertices, s and d, which represent the source and destination of the total flow over the graph. To model the flow of the content downloading, we add the edges (s, c_i^0) of infinite weight from s to all the vertices $c_i^0 \in \mathcal{C}^0$ for $1 \le i \le C$, and introduce the edges (s_i^l, d) of infinite weight from all the vertices $s_i^l \in \mathcal{S}^l$, for $1 \le i \le S$ and $0 \le l \le N$, to d. In this way, we obtain the directional connected graph, on which the content downloading is modelled as the flow from s to d, which represents the total amount of the downloaded contents by the D2D communication underlaying cellular network. An illustration of this content downloading model is shown in Fig. 3. Given this directional connected graph, we can formulate the optimization problem whose goal is to maximize the flow from s to d. For any edge, $(a_1, a_2) \in (\bigcup\limits_{0 \le l \le N} \mathcal{E}^l) \cup \mathcal{E}^b$, we further denote the traffic over the edge connecting two vertices of a_1 and a_2 as $f(a_1, a_2)$. Then, the objective of maximizing the total downloaded content of all the downloaders can be expressed as follows:

$$\max \sum_{l=0}^{N}\sum_{i=1}^{S} f\left(s_i^l, d\right). \tag{3}$$

The solution of this max-flow problem depends on the transmission mode selection as well as the resource allocation of x_i^l and y_i^l, for $0 \le l \le N$, $1 \le i \le C$. This max-flow problem is subject to several constraints, including the constraints of flow conservation and system that have fundamental influence on the achievable theoretical performance bound of the D2D communication underlaying cellular system.

B. Constraints of Transmission Flow

We now consider the system constraints in terms of flow conservation and channel transmissions. The flow on every existing edge much be non-negative. Thus, for any edge (a_1, a_2), we have $f(a_1, a_2) \ge 0$. Moreover, for any vertex in the graph, the flow conservation means that the amount of incoming flow must equal to the amount of outgoing flow. In the system, there are three different types of vertices, namely, BSs, helpers and downloaders, which have different content transmission behaviors that influence the flow. For each one, we constrain that the incoming flow amount equals to outgoing flow amount. In terms of the channel transmission, for the directly cellular communication with the UEs, we assume the unicast transmission. Since the available transmission resources are limited, the total transmitted flow to all the UEs during a time frame must satisfy the system constraints on the transmission rates and the given time frame duration. Specifically, for BS vertex $c_i^l \in \mathcal{C}^l$ at time frame l whose transmission duration is τ_l, we have

$$\sum_{u_j^l \in \mathcal{U}^l : \left(c_i^l, u_j^l\right) \in \mathcal{E}_c^l} \frac{f\left(c_i^l, u_j^l\right)}{\alpha_i^{l u_j} x_i^l} \le \tau_l. \tag{4}$$

As for the D2D communication, the shared resources are accessed based on the IEEE 802.11 based MAC protocol, and the nodes in the same access domain are not allowed to transmit simultaneously to avoid collision or interference. According to the allocated D2D communication resource, we also need to limit the transmitted content flows among the "connected" UEs at each time frame to meet the system constraints on the transmission rates and the given time frame duration. Specifically, for receiving UE vertex $u_k^l \in \mathcal{U}^l$ that is in the coverage of BS c_i at time frame l and whose receiving time duration is τ_l, we have

$$\sum_{u_m^l, u_n^l \in \mathcal{U}^l : (u_m^l, u_n^l) \in \mathcal{E}_u^l} \frac{\xi f\left(u_m^l, u_n^l\right)}{\beta_{mn}^l y_i^l} + \sum_{u_j^l \in \mathcal{U}^l : \left(u_j^l, u_k^l\right) \in \mathcal{E}_u^l} \frac{f\left(u_j^l, u_k^l\right)}{\beta_{jk}^l y_i^l} \le \tau_l \tag{5}$$

where $\xi = I_{[(u_m^l, u_k^l) \| (u_n^l, u_k^l)]}$ and $[(u_m^l, u_k^l) \| (u_n^l, u_k^l)]$ indicates the condition that either edge (u_m^l, u_k^l) or (u_n^l, u_k^l) exists, while the first term in the left hand side of equation obtains the time duration for other nodes "connected" with u_k^l to transmit their flows, and the second term is the time duration for u_k^l to transmit the flow.

C. Constraints of Resource Allocation and Mode Selection

We now discuss how the transmission mode selection and resource allocation problems are inherently modelled in the above max-flow optimization formulation.

A directly cellular transmission at time frame l corresponds to the flow from a BS vertex $c_i^l \in \mathcal{C}^l$ to a content downloader vertex $s_j^l \in \mathcal{S}^l$, which is inherently a one-hop flow occurring on the edge (c_i^l, s_j^l). A D2D connected transmission at time frame l represents the data flow originated from a BS vertex $c_i^l \in \mathcal{C}^l$, via one or more helper vertices $h_{k_1}^l, \ldots, h_{k_m}^l \in \mathcal{H}^l$ at the same time frame l, to a content downloader vertex $s_j^l \in \mathcal{S}^l$, which is a multi-hop flow occurring on the "connected" multi-edges $(c_i^l, h_{k_1}^l) \cup \cdots \cup (h_{k_m}^l, s_j^l)$. This kind of transmission is related to the scenario that at a given time frame, a multi-hop or multi-edge connected path exists from a transmitting BS with the aid of some helper relays to a receiving downloader. A D2D opportunistic transmission represents the multi-hop flow occurring on different edges related to different time frames. More specifically, a D2D opportunistic flow is originated at some time frame l' on an edge $(c_i^{l'}, h_k^{l'})$. The flow may be carried by the "vertical edges" of helper h_k through the subsequent time frames until at a future time frame $l > l'$, the flow continues on the edge (h_k^l, u_j^l). If the receiving vertex is a downloader $u_j^l = s_j^l$, this D2D opportunistic flow reaches its destination at time frame l. Otherwise, the helper $u_j^l = h_j^l$ will carry the flow into future time frames. From the above discussion, it is clearly that all the transmission modes are naturally modelled by the flows $f(a_1, a_2)$ occurring on all the edges $(a_1, a_2) \in (\bigcup_{0 \le l \le N} \mathcal{E}^l) \cup \mathcal{E}^b$ existing in the graph, and the transmission mode selection problem is solved by optimizing the flow allocation on these edges.

In general, the hop count or the number of helpers used to deliver the content to a downloader can be unlimited in the D2D connected and opportunistic transmissions. In practice, the number of relay helpers is likely to be limited, and in our model we set the maximum number of hops or helpers used to deliver the content. A special case is the two-hop forwarding, which only allows one helper relay for assisting the content downloading from a BS to a downloader. In the extreme case of limiting the hop count to one, the only transmissions allowed are one-hop transmissions directly from BSs to downloaders, and the system degenerates into a pure cellular communication without the assistance of D2D communication. In our performance evaluation, we will consider these three different cases of hop count in the mode selection.

As stated before, the system transmission resources are allocated to directly cellular communication and D2D communication as x_i^l and y_i^l, respectively, for each time frame l and each cellular BS c_i. We further assume that the total resource for the coverage area of c_i is limited to B_i. Thus, we have the following constraint on the resource allocation:

$$x_i^l + y_i^l \le B_i,\ 1 \le i \le C,\ 0 \le l \le N. \tag{6}$$

It is clear that the resource allocation problem is naturally modelled in the process of determining the resources of x_i^l and y_i^l for each cellular BS c_i.

D. Solution

Combing the objective of (3) and the system constraints of (4)–(6), we have the flow maximization problem for the D2D underlaying cellular system with the decision variables of $f(a_1, a_2)$ for $(a_1, a_2) \in (\underset{0\le l\le N}{\cup} \mathcal{E}^l) \cup \mathcal{E}^b$ as well as x_i^l and y_i^l for $1 \le i \le C$ and $0 \le l \le N$, which are all real-valued. Note that the objective (3) is a linear composition of $f(s_i^l, d)$, while the flow constraints and constraint (6) are linear constraints. But the constraints (4) and (5) are not expressed in linear forms. If we can transform these two constraints related to the resource allocation and channel access into the linear expressions of the decision variables, the flow maximization problem can be solved by linear programming techniques.

To transform the constraint (4) into a linear form, we introduce the "communicating duration" vector for BS vertex c_i^l at time frame l, whose elements are denoted by $p_{u_j^l}$ for $u_j^l \in \mathcal{U}^l$. Specifically, $p_{u_j^l}$ indicates how long c_i^l communicates with UE vertex u_j^l at time frame l. Consequently, (4) can be transformed into the following form:

$$\sum_{u_j^l \in \mathcal{U}^l:\left(c_i^l, u_j^l\right)\in\mathcal{E}_c^l} \alpha_i^{lu_j} x_i^l p_{u_j^l} \ge \sum_{u_j^l \in \mathcal{U}^l:\left(c_i^l, u_j^l\right)\in\mathcal{E}_c^l} f\left(c_i^l, u_j^l\right) \times \sum_{u_j^l \in \mathcal{U}^l:\left(c_i^l, u_j^l\right)\in\mathcal{E}_c^l} p_{u_j^l} \le \tau_l. \quad (7)$$

Similarly, we introduce the two communicating duration vectors for receiving UE vertex u_k^l at frame l, whose elements are denoted by $q_{u_j}^l$ and $r_{u_m u_n}^l$, respectively. In particular, $q_{u_j}^l$ indicates how long UE vertex u_j^l communicates with u_k^l at frame l, while $r_{u_m u_n}^l$ measures how long UE vertex u_m^l communicates with UE vertex u_n^l during frame l. It can then readily be seen that the constraint (5) can be transformed into the form of

$$\begin{aligned}
&\sum_{u_j^l \in \mathcal{U}^l:\left(u_j^l, u_k^l\right)\in\mathcal{E}_u^l} q_{u_j}^l \\
&\quad + \sum_{u_m^l, u_n^l \in \mathcal{U}^l:(u_m^l, u_n^l)\in\mathcal{E}_u^l} I_{\left[\left(u_m^l, u_k^l\right)\|\left(u_n^l, u_k^l\right)\right]} r_{u_m u_n}^l \\
&\le \tau_l, \sum_{u_j^l \in \mathcal{U}^l:\left(u_j^l, u_k^l\right)\in\mathcal{E}_u^l} q_{u_j}^l \beta_{jk}^l y_i^l \\
&\quad + \sum_{u_m^l, u_n^l \in \mathcal{U}^l:(u_m^l, u_n^l)\in\mathcal{E}_u^l} I_{\left[\left(u_m^l, u_k^l\right)\|\left(u_n^l, u_k^l\right)\right]} r_{u_m u_n}^l \beta_{mn}^l y_i^l \\
&\quad \ge \sum_{u_j^l \in \mathcal{U}^l:\left(u_j^l, u_k^l\right)\in\mathcal{E}_u^l} f\left(u_j^l, u_k^l\right) \\
&\quad + \sum_{u_m^l, u_n^l \in \mathcal{U}^l:(u_m^l, u_n^l)\in\mathcal{E}_u^l} I_{\left[\left(u_m^l, u_k^l\right)\|\left(u_n^l, u_k^l\right)\right]} f\left(u_m^l, u_n^l\right). \quad (8)
\end{aligned}$$

We have now expressed all the constraints of our flow maximization problem in linear forms. Thus, this flow maximization problem falls into the category of linear programming problems, which can be solved using the existing optimization tool kits, such as CPLEX [20] and YALMIP [21]. In the following section, we will simulate the system with the realistic mobility model and traces to evaluate and reveal the achievable performance upper bound of the mobile content downloading in the D2D underlaying cellular system.

V. Performance Evaluation

We considered realistic BS deployment with both synthetic and real-world human mobility traces. The synthetic trace employed was the recently proposed human mobility model of Self-Similar Least Action Walk) (SLAW) [22], where the number of system nodes can be changed in the simulation. The two real-world human mobility traces considered were the contact trace *Infocom05* and the GPS trace *KAIST*.

A. Evaluation Environment and Experimental Settings

Infocom05 was gathered by the Haggle Project [23]. It recorded the contacts among users carrying Bluetooth devices, which periodically discovered their contacts with peers in the communication range and recorded them. For this trace, we assumed that the devices having contacts could communicate with each other in the D2D mode and could also choose to communicate in the cellular mode. Since this trace was collected in conference site, we assumed that all the users were covered by one cellular BS. To simulate more generic and realistic user mobility environments, we used a GPS trace *KAIST*. *KAIST* was collected in the university campus at Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in which the GPS traces were recorded by Garmin GPS 60CSx handheld receivers, which are GPS devices with a position accuracy of better than 3 m 95% of the time [24]. Specifically, the GPS receivers took the readings of their current positions every 10 s and recorded them into a daily track log, and the readings were taken by the students that were randomly selected from those enrolled in a course at the computer science department. As for the SLAW, we used its typical settings given in [22], where the speed of every user was set to 1 m/s, and a truncated Pareto distribution was used to generate the pause time of which the minimum and maximum values were 30 s and 700 s, respectively, while the network area was set to $1000 \times 1000\,\mathrm{m}^2$. These three human mobility traces are summarised in Table II.

Except for the trace of *Infocom05* which had a single BS deployment in the simulation, multiple BSs, each with the coverage radius of 400 m, were deployed for the other two cases, and the number of BSs deployed depended on the area covered. Specifically, we randomly varied the deployed BSs in the SLAW model based simulation from 5 to 11, while the deployed BSs in *KAIST* were varied from 3 to 9. For the D2D communication, we limited the maximum transmission range of a node to 50 m. The achievable network-layer rate between any two UE nodes was adjusted according to the distance between them. In our network simulation, the wireless propagation was modeled by WINNER II channel models. In particular, the D2D communication channel was based on the scenario that two communicating UEs were physically in close proximity, while the cellular communication channel was simulated according to the urban microcell scenario.

TABLE II
TRACE SUMMARY

	Mobility Contact Trace: *Infocom05*	Mobility GPS Trace: *KAIST*	Mobility Model: SLAW
Device Type	iMote	Garmin GPS	N/A
Number of devices	41	92	Varying
Number of contacts	22,459	N/A	5,540,667
Contact frequency (per day)	4.6	N/A	N/A
Maximal number of "time frames"	20,231	1,803,632	6,230,588

Under the above simulation settings, we generated the dynamic graph that represented the mobile content downloading process and solved the corresponding flow maximization problem on the generated dynamic graph. Specifically, we obtained the optimal resource allocation and transmission mode selection as well as the flows related to the amount of the mobile content that the downloaders received. According to the solution of the max-flow problem, we evaluated the following three performance metrics:

1) The average amount of the mobile content downloaded, which is defined as the total amount of the downloaded mobile content divided by the number of downloaders.
2) The Jain's fairness measure [25], which determines whether downloaders are receiving fair share of the system resources. Specifically, since the amount of content received by downloader s_i, $1 \leq i \leq S$, is given by

$$\bar{f}_i = \sum_{l=0}^{N} f\left(s_i^l, d\right) \tag{9}$$

the Jain's fairness measure of the content downloading process is computed as

$$J(\bar{f}_1, \bar{f}_2, \cdots, \bar{f}_S) = \frac{\left(\sum_{i=1}^{S} \bar{f}_i\right)^2}{S \cdot \sum_{i=1}^{S} \bar{f}_i^2}. \tag{10}$$

3) The amounts of downloaded content through directly cellular transmission, D2D connected transmission and D2D opportunistic transmission, respectively.

Since these performance metrics were evaluated based on the output of the flow maximization problem which yields the optimal mode selection and resource allocation, these results represented the best performance that could be attained under the simulated D2D communication underlaying cellular network system. To assess the impact of different system settings, we first evaluated the system with the above mentioned metrics using the SLAW mobility model by changing the number of the UEs with the fixed helper-downloader ratio of 4 : 6. Next, we used the GPS trace of *KAIST* to investigate the influence of the helpers to the achievable performance bound. Finally, we used the contact trace of *infocom05* to study the influence of the content transmission modes and hop limits.

B. Results of SLAW Mobility Model

In the SLAW mobility model based simulation which lasted 600 s, we varied the number of UEs from 10 to 170 and we changed the number of BSs from 5 to 11. Note that the BS deployment with 11 BSs provided a completed coverage of the 1000×1000 m^2 network area. No limit was imposed on the number of hops in D2D transmission. The results obtained by the flow maximization problem are shown in Fig. 4, in terms of the three evaluation metrics, the average amount of the content downloaded, the content downloading fairness, and the amounts of the content downloaded by different modes.

Fig. 4(a) indicates that the average received content flow increases with the number of UEs. This is because as the number of UEs increases, the D2D transmission opportunities also increase due to the increase in the helper density. The steepest growth in content downloading seems to occur when the number of UEs in this simulated network lies between 50 to 90. In terms of the impact of the number of deployed BSs, we note from Fig. 4(a) that increasing the number of BSs from 5 to 7 as well as increasing the number of BSs from 9 to 11 bring large system performance enhancement. The former represents the scenarios of sparse BS deployment, while the latter corresponds to the situations of dense BS deployment. Therefore, in the D2D enabling cellular system, we should pay special attention to the cellular coverages achieved under sparse and dense BS deployments. We also note that a more dense cellular coverage helps the both cases of low and high numbers of UEs. Indeed, a pervasive 11-BS deployment results in a performance gain of about 2.4 times over a sparse 5-BS deployment when the number of UEs is 170, while this gain is about 2.8 times when the number of UEs is 10.

To obtain some insight on how the content downloading task is actually shared among the downloaders, we depict the Jain's fairness index of (10) in Fig. 4(b). We observe that changing the number of UEs has a non-obvious influence on the fairness of the content downloading process since the numbers of helpers and downloaders are changed with the fixed ratio of 4 : 6 in this simulated system. But the increase in the number of deployed BSs has a major impact on the fairness of the content downloading process. More specifically, we observe from Fig. 4(b) that some unfairness arises under a sparse BS deployment, and this is because in the system of a low-number BS deployment, the downloaders traveling in the areas with little cellular coverage will have far fewer chances to benefit from the content downloading than the downloaders traveling in the cellular coverage areas. The system however becomes increasingly fair for the pervasive BS deployment where all the areas are enjoying cellular BS coverage.

Next we examined the amounts of content downloading by the three different transmission modes, separately. The results obtained by solving the flow maximization problem of the content downloading system with the deployment of 7 BSs are shown in Fig. 4(c). From Fig. 4(c), we observe that the largest amount of content downloading is achieved by the directly cellular mode, which accounts for about 65.3% of the total content downloaded when the number of UEs is 10 and for about 53.4% of the total content flow when the number of UEs is 170,

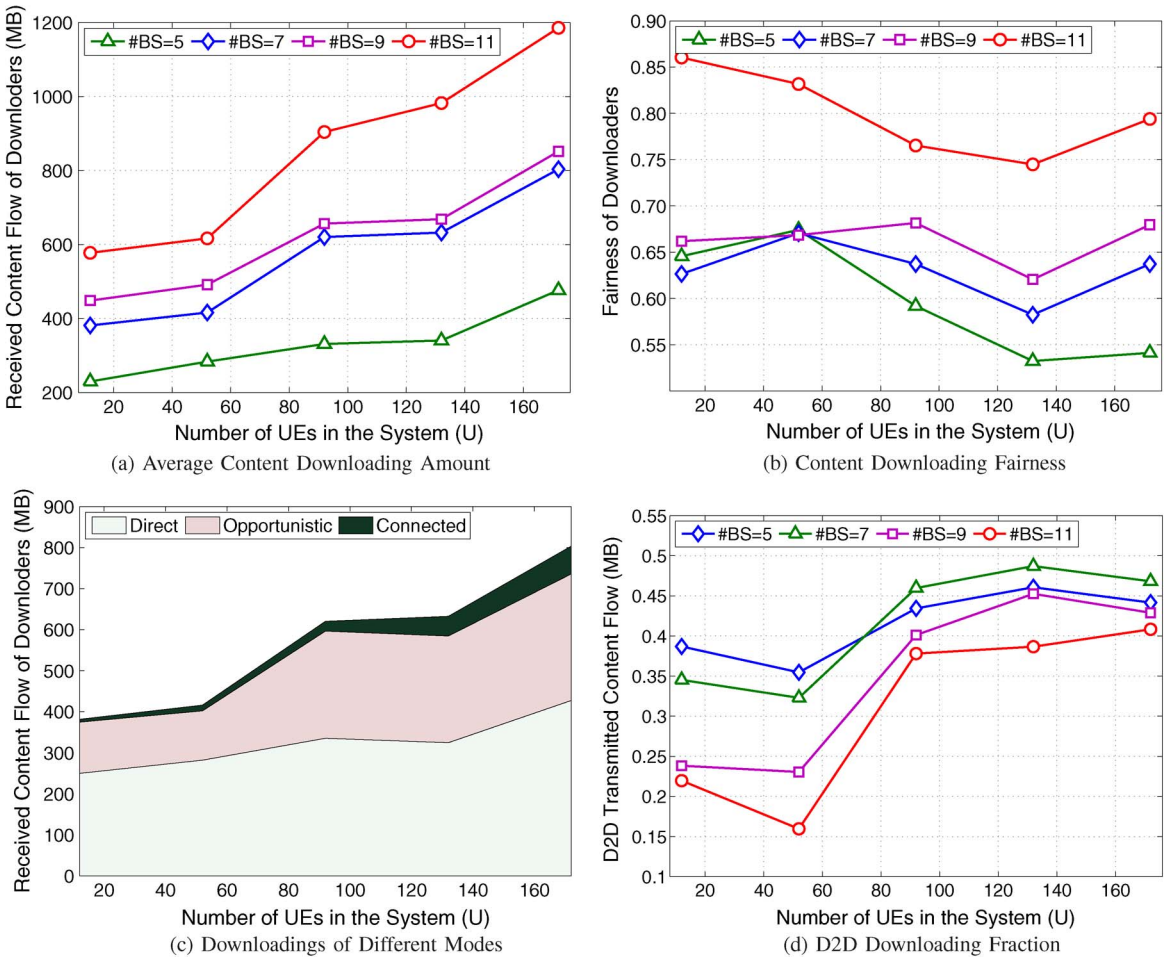


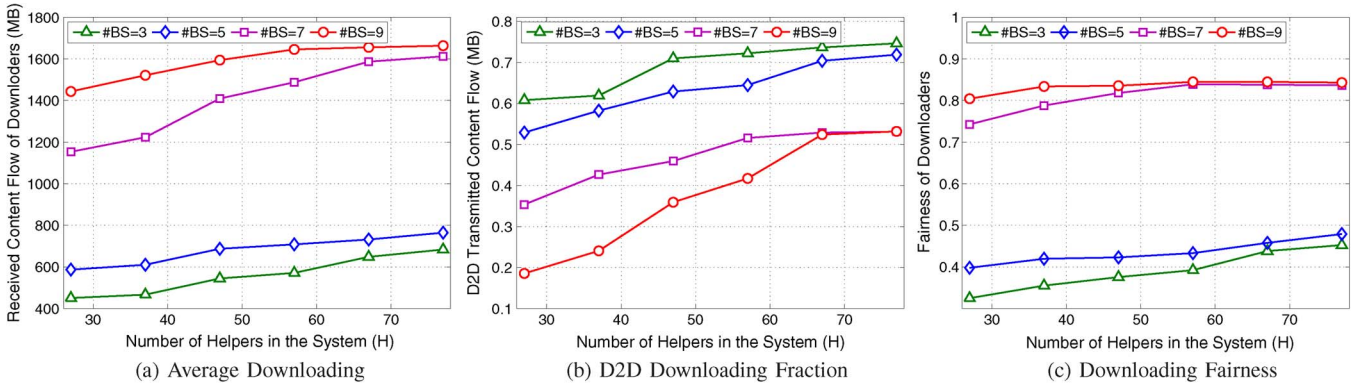
Fig. 4. Performance of mobile content downloading as a function of the numbers of UEs and BSs in the D2D underlaying cellular network with the SLAW mobility model. (a) Average amount of content downloading. (b) Fairness of content downloading process. (c) Content downloading amounts by directly cellular mode, D2D connected mode, and D2D opportunistic mode, given the deployment of 7 BSs. (d) Fraction of content downloading by D2D.

respectively. It is also clear that most of the data downloaded by D2D communications are via the D2D opportunistic mode, and the importance of opportunistic transmission tends to increase with the number of UEs. On the other hand, the D2D connected mode is only responsible for a very small amount of the content downloaded, which indicates that the connected path from a BS with the assistance of some helper relays to a downloader may not be a most efficient way for mobile content downloading.

To further observe the important role played by the D2D communication in the mobile content downloading, we plot the ratio of the content downloaded by the two D2D transmission modes over the total downloaded content in Fig. 4(d). We note from Fig. 4(d) that the percentage of the D2D downloading decreases when the number of UEs changes from 10 to 50, while the D2D downloading percentage increases with the number of UEs when the number of UEs in the system is larger than 50. The reason we believe is due to the fact that for the system with very few UEs, the direct cellular transmission can meet the needs of most UEs and D2D transmission opportunities are also very few. By contrast, when the density of UEs is high, the directly cellular mode along becomes insufficient for the needs of all the UEs, and the D2D modes, particularly the opportunistic one, play an increasingly important role. In terms of the impact of the number of deployed BSs, we observe that the D2D downloading fraction has the smallest value when the number of deployed BSs is the largest, i.e., 11 BSs. The reason is that when the system employs a pervasive BS coverage, the directly cellular mode is available for most of the UEs and, consequently, D2D transmissions are less needed. Overall, the D2D mobile content downloading accounts for about 31% to 48% of the total content flow when the number of UEs is larger than 80. This clearly demonstrates the significant role of D2D transmissions in the mobile content downloading system.

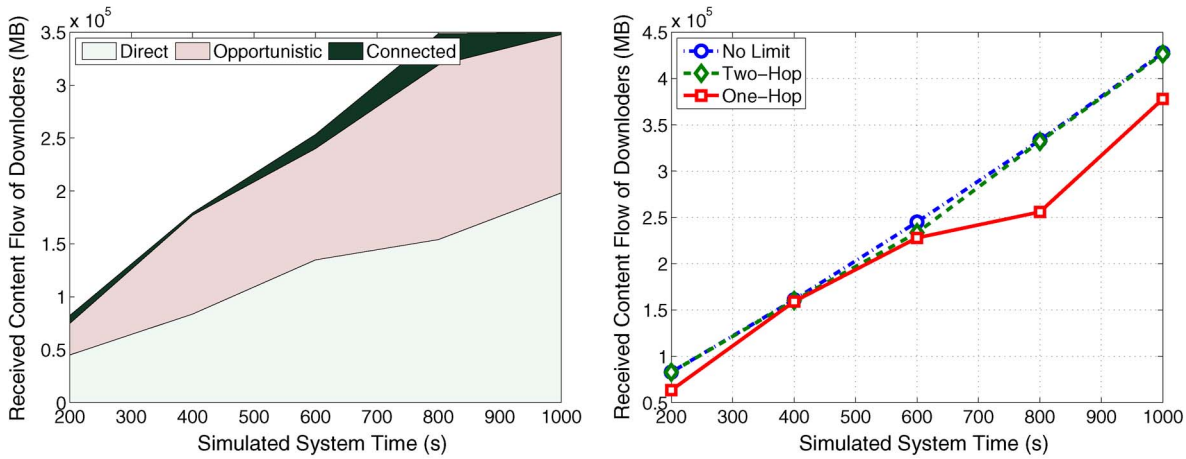
C. *Results of GPS Mobility Trace KAIST*

We ran the the simulated D2D underlaying cellular system with the GPS human mobility trace of *KAIST* for 600 s. To observe the role played by the D2D transmission in the system, we varied the number of helpers in the system. The number of deployed BSs for the system also varied from 3 to 9, with the last BS deployment of 9 BSs providing a complete cellular coverage of the whole UEs' mobility area. Again, we did not set the limit on the number of hops in D2D transmission. The system performance metrics were then evaluated, in terms of the average amount of content downloading, D2D downloading fraction and content downloading fairness, and the results obtained are depicted in Fig. 5.



(a) Average Downloading (b) D2D Downloading Fraction (c) Downloading Fairness

Fig. 5. Performance of mobile content downloading as a function of the numbers of helpers and BSs in the D2D communication underlaying cellular network with the human mobility trace of *KAIST*. (a) Average amount of content downloading. (b) Fraction of content downloading by D2D. (c) Fairness of content downloading process. The number of UEs in *KAIST* trace is 92.



(a) Content Downloading of Different Modes (b) Content Downloading of Different Hop Limits

Fig. 6. Performance of mobile content downloading as a function of the downloading time and the hop limit for the D2D communication underlaying cellular network with the human mobility trace of *Infocom05*. (a) Content downloading amounts by directly cellular mode, D2D connected mode, and D2D opportunistic mode, respectively, with no limit imposed on the number of hops. (b) Amounts of content downloading given different hop limits.

We observe that the average received content flow increases with the increase of the number of helpers in the system, as clearly seen in Fig. 5(a). This is because the increase of helpers creates more opportunities of both the D2D opportunistic and connected transmissions. Fig. 5(a) also shows that the impact of varying the number of deployed BSs is similar to the case of the SLAW based system. From Fig. 5(b), we can see that the percentage of the D2D downloading increases with the number of helpers in the system. Specifically, with the dense coverage of 9 BSs the fraction of D2D downloading is about 18.6% to 53.2%, while with the sparse coverage of 3 BSs the percentage of D2D downloading is about 60.8% to 74.6%, as the number of helpers increases from 25 to 80. From Fig. 5(c), we observe that the increase of the helpers only has a small influence on the fairness of the content downloading process. By contrast, the fairness of the content downloading process is mainly determined by the number of deployed BSs. In particular, when the system is under the pervasive coverage of 9 BSs, the system's fairness is above 0.8, which indicates that the content downloading is well shared among the downloaders. The results for the GPS mobility trace of *KAIST* are consistent with those observed in the SLAW based simulation, which indicates that these properties reflect the true system characteristics of the optimal content downloading in the D2D underlaying cellular system.

D. Results of Contact Mobility Trace Infocom05

We used the contact mobility trace of *Infocom05* to further investigate the impact of different content transmission modes and different hop limits on the achievable system performance. In this mobility trace, since we only have the nodes' contact information but without their trajectories, we varied the time of the content downloading to measure the average received content flows corresponding to the different transmission modes as well as corresponding to different hop limits. Also it is worth mentioning again that only one BS was deployed. The results obtained are shown in Fig. 6.

As shown in Fig. 6(a), both the directly cellular mode and the D2D opportunistic mode are responsible for large percentages of the content downloaded, while the D2D connected mode only provides a very small fraction of the content downloaded. This observation is also consistent with the results obtained by the SLAW based simulation. From Fig. 6(a), we can compare

the behaviors of the directly cellular transmission and the D2D opportunistic transmission under small and large content downloading times. Specifically, for the system downloading time smaller than 400 s, the amount of the content downloaded by the directly cellular mode is larger than that of the D2D opportunistic mode. But for the system downloading time higher than 400 s, the amount of the content downloaded via the D2D opportunistic transmission exceeds that of the directly cellular transmission. This is not surprising since D2D opportunities occur with inherently long delay. When the content downloading latency is short, there are fewer D2D opportunistic transmission opportunities and the system must rely mainly on the directly cellular mode. As the system's allowed downloading latency increases, the opportunities of D2D opportunistic transmission increase too.

To observe the influence of the number of hops on the achievable performance, we set the hop count to one, two and no limit, respectively, and simulated the corresponding content downloading systems. From the results of Fig. 6(b), we can see that for the system with the downloading time restricted to smaller than 400 s, increasing the number of hops has little impact on the achievable system performance. This is not surprising. Since the two D2D modes inherently require long delay, the system must rely on the directly cellular (one-hop) transmission to meet the time deadline. When the downloading time is allowed to exceed 400 s, the addition of two-hop transmission leads to a considerable performance enhancement. For example, the average amount of the downloaded content is increased by about 13.5% by allowing the two-hop transmission when the downloading time is 1000 s, compared with the cellular only system. However, performance enhancement achievable by increasing the number of hops from two to no limit is negligible for this one-BS system.

E. Summary

Based on the above system simulation results involving the synthetic human mobility model and real-world human mobility traces, we have observed the critical influence of the numbers of UEs and deployed BSs in the system to the achievable system performance metrics of the optimal D2D underlaying cellular system. Specifically, under the optimal system resource allocation and mode selection for the mobile content downloading, our general observations are as follows.

1) Increasing the number of UEs, especially the number of helpers, in the system significantly enhances the average amount of mobile content downloading as well as the amount of the mobile content downloaded by D2D transmissions, but it has little influence on the fairness of the mobile content downloading process.
2) The cellular coverage has important influence on the achievable system performance. In particular, increasing the number of deployed BSs increases the amount of content downloading as well as improves the fairness of the mobile content downloading process, but decreases the percentage of the D2D content downloading.
3) The D2D underlaying cellular system with only two-hop D2D transmissions is capable of attaining almost all the achievable performance of the optimal mobile content downloading system. The D2D opportunistic mode plays a much more significant role than the D2D connected mode in the mobile content downloading system.

As a further note, our results are obtained under the condition that the mobile data is delay tolerant and the content dissemination latency is not critical. If the contents are delay sensitive, some of the outcomes will need to be changed.

VI. Conclusion and Future Work

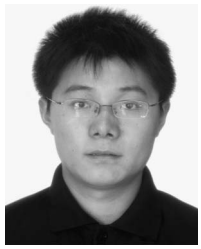
We has studied the performance bound of the D2D communication underlaying cellular system under realistic networking scenarios with large-scale node mobility and all possible transmission modes. By formulating the optimal mobile content downloading as an optimization problem that maximizes the content downloading flows from all the cellular BSs to the content downloaders through the three possible modes of directly cellular transmission, D2D enabled connected transmission and D2D opportunistic transmission, we have obtained the solution representing the theoretical upper bound to the achievable performance of the content downloading process. Using realistic human mobility model and real-world human mobility traces driven simulations, we have evaluated and validated the effects of the different system settings on the performance bound of the mobile content-downloading system. In particular, in our extensive simulation study, by varying the number of UEs, the number of helpers, the number of deployed BSs, and the content downloading time, we have revealed some fundamental influences of the D2D communication to the D2D underlaying cellular network.

In this paper, we assume the knowledge of the deployment and coverage of BSs and the UEs' mobility trajectories as well as the scheduling of data transmission in terms of resource allocation and mode selection. Therefore, our solution represents the idealised system throughput upper bound. Considerable future works are required to investigate how to implement such a D2D communication underlaying cellular network in practice by designing optimal scheduling schemes with the consideration of QoS guarantee, such as content delivery delay. In particular, further research is warranted to study the acquisition of the required system information, including the UEs' mobility statistics. In our current simulation investigation, we mainly use human mobility. Our future work will evaluate the proposed mobile content downloading system in realistic D2D underlaying cellular networks under vehicular mobility environments.

References

[1] Global mobile data traffic forecast update, 2012–2017, Feb. 6, 2013. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html
[2] B. Han *et al.*, "Cellular traffic offloading through opportunistic communications: A case study," in *Proc. 5th ACM Workshop Challenged Netw.*, Chicago, IL, USA, Sep. 24, 2010, pp. 31–38.

[3] S. Sesia, I. Toufik, and M. Baker, Eds., *LTE—The UMTS Long Term Evolution: From Theory to Practice*. Chichester, U.K.: Wiley, 2009.
[4] K. Doppler, M. Rinne, C. Wijting, C. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 42–49, Dec. 2009.
[5] G. Fodor *et al.*, "Design aspects of network assisted device-to-device communications," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 170–177, Mar. 2012.
[6] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-advanced networks," *IEEE Wireless Commun.*, vol. 19, no. 3, pp. 96–104, Jun. 2012.
[7] C.-H. Yu, O. Tirkkonen, K. Doppler, and C. Ribeiro, "Power optimization of device-to-device communication underlaying cellular communication," in *Proc. ICC*, Dresden, Germany, Jun. 14–18, 2009, pp. 1–5.
[8] C. Yu, O. Tirkkonen, and K. Doppler, "On the performance of device-to-device underlay communication with simple power control," in *Proc. VTC-Spring*, Barcelona, Spain, Apr. 26–29, 2009, pp. 1–5.
[9] H. Xing and S. Hakola, "The investigation of power control schemes for a device-to-device communication integrated into OFDMA cellular system," in *Proc. 21th PIMRC*, Istanbul, Turkey, Sep. 26–30, 2010, pp. 1775–1780.
[10] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2752–2763, Aug. 2011.
[11] C. Xu *et al.*, "Efficiency resource allocation for device-to-device underlay communication systems: A reverse iterative combinatorial auction based approach," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 348–358, Sep. 2013.
[12] P. Jänis *et al.*, "Interference-aware resource allocation for device-to-device radio underlaying cellular networks," in *Proc. VTC-Spring*, Barcelona, Spain, Apr. 26–29, 2009, pp. 1–5.
[13] S. Xu, H. Wang, T. Chen, Q. Huang, and T. Peng, "Effective interference cancellation scheme for device-to-device communication underlaying cellular networks," in *Proc. VTC-Fall*, Ottawa, ON, Canada, Sep. 6–9, 2010, pp. 1–5.
[14] H. Min, J. Lee, S. Park, and D. Hong, "Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 3995–4000, Dec. 2011.
[15] H. Luo, X. Meng, R. Ramjee, P. Sinha, and L. Li, "The design and evaluation of unified cellular and *ad hoc* networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 9, pp. 1060–1074, Sep. 2007.
[16] B. Han, P. Hui, and A. Srinivasan, "Mobile data offloading in metropolitan area networks," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 14, no. 4, pp. 28–30, Oct. 2010.
[17] X. Zhuo, W. Gao, G. Cao, and Y. Dai, "Win-coupon: An incentive framework for 3G traffic offloading," in *Proc. 19th IEEE Int. Conf. Netw. Protocols*, Vancouver, BC, Canada, Oct. 17–20, 2011, pp. 206–215.
[18] F. Harary and G. Gupta, "Dynamic graph models," *Math. Comput. Modelling*, vol. 25, no. 7, pp. 79–87, Apr. 1997.
[19] F. Malandrino, C. Casetti, and C. Chiasserini, "Content downloading in vehicular networks: What really matters," in *Proc. IEEE INFOCOM*, Shanghai, China, Apr. 10–15, 2011, pp. 426–430.
[20] CPLEX: Linear Programming Solver. [Online]. Available: http://www.ilog.com/
[21] I. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *Proc. IEEE Int. Symp. Comput. Aided Control Syst. Design*, Taipei, Taiwan, Sep. 2–4, 2004, pp. 284–289.
[22] K. R. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "Slaw: A new mobility model for human walks," in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 19–25, 2009, pp. 855–863.
[23] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: Social-based forwarding in delay tolerant networks," in *Proc. 9th ACM MobiHoc*, Hong Kong, China, May 26–30, 2008, pp. 241–250.
[24] I. Rhee *et al.*, "On the levy-walk nature of human mobility," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 630–643, Jun. 2011.
[25] R. K. Jain, D. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Eastern Research Lab, Digital Equipment Corp., Hudson, MA, USA, Sep. 1984.



Yong Li (M'09) received the B.S. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012.

He is now a Postdoctoral Researcher at Tsinghua University, Beijing. He serves as a paper Reviewer for international conferences including IEEE ICC, VTC, ICOIN, PIMRC, APCC and many others. His research fields include mobile delay tolerant networks, topics including forwarding polices design, buffer management design and performance evaluation; mobility modeling; and mobility management in next generation wireless IP networks, topics including Mobile IP, SIP, Proxy mobile IP, cross-layer design, multicast mobility, modeling for mobility performance evaluation, enhancing handoff performance and proposing mobility management architecture.

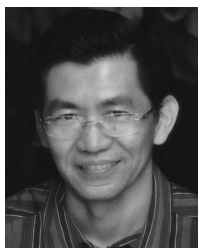
Zhaocheng Wang (M'09-SM'11) received the B.S., M.S., and the Ph.D. degrees from Tsinghua University, Beijing, China, in 1991, 1993, and 1996, respectively.

From 1996 to 1997, he was a Post Doctoral Fellow at Nanyang Technological University, in Singapore. From 1997 to 1999, he was a Research Engineer and then a Senior Engineer at OKI Techno Centre (Singapore) Pte. Ltd. From 1999 to 2009, he was a Senior Engineer and then a Principal Engineer at SONY Deutschland GmbH. He is currently a Professor in the Department of Electronic Engineering, Tsinghua University. His research areas include wireless communications, digital broadcasting and millimeter wave communications. He holds 32 granted US/EU patents and has published over 100 technical papers. He has served as technical program committee co-chair/member of many international conferences. Dr. Wang is a Fellow of IET.

Depeng Jin (M'09) received the B.S. and the Ph.D. degrees in electronics engineering from Tsinghua University, Beijing, China, in 1995 and 1999, respectively.

He is an Associate Professor and Vice Chair of the Department of Electronic Engineering, Tsinghua University, Beijng.

Dr. Jin was awarded the National Scientific and Technological Innovation Prize (Second Class) in 2002. His research fields include telecommunications, high-speed networks, ASIC design and future Internet architecture.



Sheng Chen (M'90–SM'97–F'08) received the B.Eng. degree from the East China Petroleum Institute, Dongying, China, in January 1982, the Ph.D. degree from the City University, London, U.K., in September 1986, both in control engineering, and the D.Sc. degree from the University of Southampton, Southampton, U.K., in 2005.

From 1986 to 1999, he held research and academic appointments at the Universities of Sheffield, Edinburgh and Portsmouth, all in the U.K. Since 1999, he has been with the School of Electronics and Computer Science, University of Southampton, U.K., where he currently holds the post of Professor in Intelligent Systems and Signal Processing. Dr. Chen is also a Distinguished Adjunct Professor at King Abdulaziz University, Jeddah, Saudi Arabia. He is a Chartered Engineer (CEng) and a Fellow of IET. His recent research interests include adaptive signal processing, wireless communications, modelling and identification of nonlinear systems, neural network and machine learning, intelligent control system design, evolutionary computation methods and optimisation. He has published over 470 research papers. Dr. Chen is an ISI highly cited Researcher in the engineering category (March 2004).