This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

Resource Allocation for Heterogeneous Applications with Device-to-Device Communication Underlaying Cellular Networks

Xiaoqiang Ma, *Student Member, IEEE*, Jiangchuan Liu, *Senior Member, IEEE*, and Hongbo Jiang, *Senior Member, IEEE*

***Abstract*—Mobile data traffic has been experiencing a phenomenal rise in the past decade. This ever increasing data traffic puts significant pressure on the infrastructure of state-of-the-art cellular networks. Recently, Device-to-Device (D2D) communication that smartly explores local wireless resources has been suggested as a complement of great potential, particularly for the popular proximity-based applications with instant data exchange between nearby users.**

Significant studies have been conducted on coordinating the D2D and the cellular communication paradigms that share the same licensed spectrum, commonly with an objective of maximizing the aggregated data rate. The new generation of cellular networks however have long supported heterogeneous networked applications, which have highly diverse Quality of Service (QoS) specifications. In this paper, we jointly consider resource allocation and power control with heterogeneous QoS requirements from the applications. We closely analyze two representative classes of applications, namely *streaming-like* and *file-sharing-like*, and develop optimized solutions to coordinate the cellular and D2D communications with the best resource sharing mode. We further extend our solution to accommodate more general application scenarios and larger system scales. Extensive simulations under realistic configurations demonstrate that our solution enables better resource utilization for heterogeneous applications with less possibility of under- or over-provisioning.

***Index Terms*—Quality of Service (QoS), Resource Allocation, Device-to-Device Communication.**

I. Introduction

With the proliferation of such high performance mobile devices as smartphones and tablets, and the advances in cellular network technologies, data-intensive applications including Voice over IP (VoIP), video streaming, and instant file sharing become increasingly popular. As a result, mobile data traffic has been experiencing a phenomenal rise in the past decade, which is expected to reach 11.2 Exabytes per month by 2017, a 13-fold increase over 2012 [1].

Such ever increasing data traffic has put significant pressure on the infrastructure of state-of-the-art cellular networks. There have been great efforts on the development and deployment of next generation wireless communication systems, notably the 3GPP Long Term Evolution (LTE)[1]. The widespread penetration of WiFi networks have also successfully offloaded a certain portion of the traffic [2]. Yet the cellular Base Stations (BSes) and WiFi Access Points (APs) remain bottlenecks that limit the achievable data rate for individual mobile devices. Also the availability of WiFi APs can hardly be guaranteed, particularly in rural areas, not to mention that most of the APs are not readily shared.

On the other hand, it is known that proximity-based services have constituted a considerable portion of the mobile data traffic [3]. Such services enable geographically close users to directly exchange data, which is of particular interest in the new generation of social applications. As an example, the popular WhatsApp[2], can utilize a Near Field Communication (NFC) [4] module that is readily available on the latest smartphones and tablets to support peer-to-peer file sharing for nearby users, albeit with a slow speed of 424 kbps. The more powerful Bluetooth [5] has served for proximity-based data exchange for a long period, which however needs cumbersome manual device pairing and still has a rather limited communication range as well as data rate; new standards, e.g., Wi-Fi Direct [6], remain at a very early stage to be widely adopted. Moreover, Bluetooth and Wi-Fi Direct are both standalone standards that are independent of the cellular networks; they operate on unlicensed spectrums, which often incur severe and unpredictable interferences [7].

Recently, Device-to-Device (D2D) communication underlaying cellular networks has been suggested as a new paradigm of great potential toward supporting proximity-based applications [3], [8]. With this new paradigm, the cellular BS-based and the direct D2D communications are coordinated to operate on the same licensed spectrum. Different resource allocation strategies can be applied to allocate the spectrum and to adjust the transmit power to optimize the overall system performance [9].

Significant studies have been conducted with a common objective of maximizing the aggregated data rate [10], [11]. The new generation of cellular networks however have long supported heterogeneous applications, which can have highly diverse Quality of Service (QoS) specifications. For example, file sharing applications generally demand high data rate but can smoothly adapt to a wide range of data rates. On the

Manuscript received xx, xxxx ; revised xx, xxxx. This research is supported by an NSERC Discovery Grant and an NSERC Strategic Project Grant.

X. Ma and J. Liu are with the School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada (e-mail: xma10@sfu.ca; jcliu@sfu.ca). J. Liu is the corresponding author.

H. Jiang is with the School of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China (email: hongbojiang2004@gmail.com).

[1]http://www.3gpp.org/

[2]http://www.whatsapp.com/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

other hand, such streaming applications as VoIP and Internet Protocol Television (IPTV) generally have a lower limit for the minimum acceptable quality, and often encode the source into multiple versions with different encoding bitrates [12]. Even their bottlenecks, whether on the uplink or the downlink, can be different. Maximizing the overall data rate without differentiating the needs of these applications can often lead to under- or over-provisioning, as revealed by our experiments.

In this paper, we consider a modern D2D underlay to cellular networks serving diverse types of applications. We jointly consider resource allocation and power control with heterogeneous QoS requirements from the applications for selecting the best resource sharing mode. We closely analyze two representative classes of applications, namely *streaming-like* and *file-sharing-like*, and develop optimized solutions for coordinating the cellular and D2D communications. We further extend our solution to accommodate more general application scenarios and systems of larger scales. The effectiveness of our solution has been validated through extensive simulations with realistic configurations. The results demonstrate that, as compared with state-of-the-art allocation schemes that maximize the total data rate only, our solution enables with better resource utilization for different types of applications with less possibility of under- or over-provisioning.

The remainder of this paper is organized as follows. We introduce the background of D2D communication underlaying cellular networks and review the related works in Section II. We present our system model and analyze the QoS requirements of representative applications in Section III. We then investigate the resource allocation problem and its solution for the dedicated, cellular, and reuse modes in Section IV. We further extend the solution to more general application scenarios and larger system scales in Section V. Section VI presents the evaluation results of the proposed solutions and Section VII concludes this paper.

II. Background

The concept of D2D communication as an underlay to a cellular network is illustrated in Fig. 1, where BS represents a Base Station and UE represents a User Equipment. The UEs can be served by the BSes, as in traditional cellular networks; they can also communicate with each other directly through D2D links. A distinct feature here is that the two types of communications share the same spectrum, which apparently needs careful coordinations [3], [10]:

(1) *Dedicated mode*: The cellular network allocates an exclusive portion of resources dedicated for the direct communications between D2D device pairs. There is no interference between the cellular and D2D communications;

(2) *Cellular mode*: D2D devices work as traditional cellular devices, and D2D communications are relayed by the BS;

(3) *Reuse mode*: D2D communications reuse a portion of or the whole resources allocated to cellular network. This mode can be further divided into *downlink reuse* (DLre) and *uplink reuse* (ULre), where the downlink/uplink of the D2D communications reuse the shared resources and may cause interference to the downlink/uplink of cellular users.

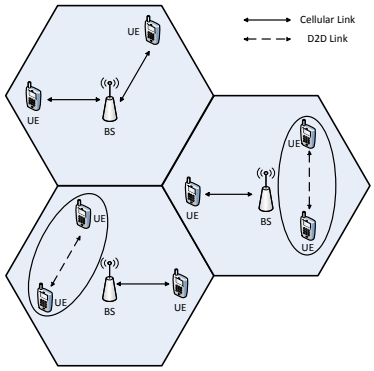

Fig. 1. D2D communication as an underlay to a cellular network.

Maximizing the total data rate for the cellular uplink and the D2D pairs has been widely used as the optimization objective in this context [10]. Doppler *et al.* [10] studied the optimal mode selection strategy for both single-cell and multi-cell scenarios, aiming at reliable D2D communications with limited interference to the cellular network. They showed that the mode selection highly depends on the locations of the devices. Liu *et al.* [13] studied the mode selection problem and showed that the introduction of relay nodes offers D2D pairs a higher probability to share the resources with cellular users. For each of the above modes, both power control and resource sharing need careful examination in order to achieve the maximum data rate.

A. Power Control with D2D Communications

Smart power control mitigates the interference among users sharing the same spectrum, which is critical for the coexistence of D2D and cellular users. Early efforts have been spent on exploiting the capacity gain of D2D connections without generating significant interference to cellular users [3], [11], which is closely related to the problem in the cognitive radio context that secondary users should not generate harmful interference to primary users [14]. Yet recent works have shown that the overall performance can be improved by giving slight priority to D2D links [15]. Yu *et al.* [16] further derived the optimal power allocation approach under both prioritized or non-prioritized cellular communications.

B. Resource Allocation with D2D Communications

Resource allocation is usually jointly considered with mode selection and power control to improve the total data rate or spectrum efficiency. Zulhasnine *et al.* [17] formulated this problem as a mixed integer nonlinear programming and proposed a greedy heuristic algorithm to reduce the interference to cellular users. Yu *et al.* [9] analyzed the optimal resource allocation and power control between cellular and D2D links that share the same resources for different sharing modes. Xu *et al.* [18] further proposed a sequential second price auction-based mechanism to allocated the resources to D2D pairs. Our work differs from the above works in that we pave an application-oriented avenue toward power control and resource allocation. We take the QoS specifications of heterogeneous

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

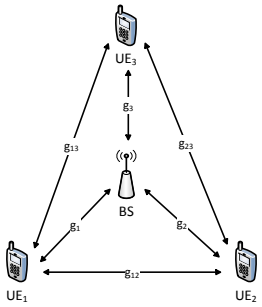

Fig. 2. Single cell scenario with a pair of D2D UE and a cellular UE.

applications into consideration, which calls for a revisit to the problem.

Most of the above studies assumed that the BSes have the CSI of all the involved links and adopt centralized schemes to allocate the resources for both cellular and D2D users. Recent works have shown that cell statistics can be used instead of the instantaneous CSI in resource allocation, although its accuracy remains to be examined [19]. A distributed game-theoretic allocation scheme was proposed in [20], but the solution is suboptimal due to the lack of accurate resource management and tight cooperation. We advocate a centralized control with readily available CSI in our work, which however can be extended in the future when smarter CSI data collection tools are available.

III. QoS-Aware Resource Allocation: Model and Problem

We start from a single cell scenario with one cellular user UE_1 and a D2D pair (UE_2 and UE_3), as illustrated in Fig. 2. We assume that the inter-cell interference is well managed by cooperative power control or resource allocation mechanisms across cells [26], which allow us to focus on the spectrum within individual cells. In line with existing studies [9], [10], we assume that the BS has all the CSI available and aim at designing a centralized resource allocation scheme. The cellular network adopts Frequency Division Duplexing (FDD) such that the uplink and the downlink each occupies half of the whole spectrum (denoted by W), as in the LTE standard [21]. We also assume symmetric channels, and use g_i to denote the channel gain between the BS and $\mathrm{UE_i}$, and g_{ij} the channel gain between $\mathrm{UE_i}$ and $\mathrm{UE_j}$. Typically, the channel gain includes the path loss, shadow fading and fast fading [22]. We denote the power of the Additive White Gaussian Noise (AWGN) at the receiver by N_0, and the allocated transmit power of $\mathrm{UE_i}$ by P_i. The maximum transmit power of the UEs, denoted by P^{max}, is up to 23 dBm in LTE standard. We also denote the allocated transmit power of the BSes by P_B. The maximum transmit power of the BSes, denoted by P_B^{max}, depends on their cover range, for example, up to 20 dBm for a Home BS, 24 dBm for a Local Area BS, and no upper limit for a Wide Area BS in LTE [21]. In most wireless communication systems, there is an upper limit on the spectrum efficiency such that a Signal to Interference plus Noise Ratio (SINR) higher than a maximum value, γ_h, does not further increase the data rate when the link spectrum efficiency is limited to r_h bps/Hz. A link adaptation technique will select a proper MCS from a limited number of options according to the current channel condition [23] and r_h is achieved when the current SINR is high enough to support the highest MCS, e.g., 64QAM Rate 11/12 for LTE [24]. On the other hand, the SINR should be no lower than a minimum value, γ_l, to support the lowest MCS with the spectrum efficiency of r_l bps/Hz.

Both the cellular and D2D communications can support heterogeneous networked applications. A user's experience largely depends on such network conditions as delay and data rate. In our system, the delay of the cellular communication is mainly determined by the backhaul and core networks, which are relatively independent of the operations in a cell; the delay of the D2D communication is very low given short distance between a D2D pair. Hence, in this work we focus on the data rate as the key factor that impacts user experience. We summarize the notations in Table I.

The relationship between user experience and data rate however is not homogeneous for different classes of applications. Assume that there are K classes of applications, each of them having a utility function $U_i, \forall i \in \{1, ..., K\}$ that quantifies the relationship between user experience and data rate. We then define *cell utility* and *D2D utility* as the total utility of the cellular applications and the D2D applications, respectively. We can further assign different weights to the cellular and D2D utilities, which give different priorities to each of them. Our target is then to identify the optimal strategy to allocate the resources and to adjust the transmit power of the BS and UEs to maximize the *weighted cell utility*. This QoS-aware resource allocation problem can be formulated as follows:

$$\begin{aligned} &\text{Maximize} && WCU = \lambda U_c(R_c) + \lambda' U_d(R_d) \\ &\text{Subject to} && P_i \le P^{max}, \forall i \in \{1,2,3\}, \\ &&& P_B \le P_B^{max}, \\ &&& \gamma_l \le \gamma_c, \gamma_d \le \gamma_h, \\ &&& \lambda' = 1 - \lambda, \\ &\text{Given} && U_c, U_d \in \{U_1, ..., U_K\}, \end{aligned} \tag{1}$$

where λ ($0 < \lambda < 1$) is the weight assigned to the cellular utility; R_c and R_d are the data rates of the cellular and D2D communications, respectively, and will be derived in the later section; U_c and U_d are the utility functions of the cellular and D2D communications and are determined by the corresponding applications, respectively.

A. Utility Functions of Applications

We first focus on two representative classes of applications, namely, *file sharing* for typical generic data exchange applications and *streaming* for typical multimedia communication applications.

1) File Sharing Applications: File sharing applications generally expect a short finish time or equivalent, high data rate; yet they are highly adaptive to a broad range of data rates with no stringent demand. Given the file size of a specific task, the utility function thus depends on the date rate. Let R^{max} be the maximum achievable data rate, we have the utility function $U_f(R) = \frac{R}{R^{max}}$ if the user's experience is linear with the data rate, or finish time. To ensure proportional fairness in resource allocation, however, logarithmic relation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

has also been widely used [27], leading to a utility function of $U_f(R) = \log_2(1 + \frac{R}{R^{max}})$.

2) Streaming Applications: Likewise, streaming applications also benefit from high data rate and adapt to a certain range, but generally has a lower bound for most of the audio/video multimedia data. On the other hand, if the data rate is higher than a certain encoding bitrate, the marginal utility quickly diminishes. In between, the operational rates of the encoder are discrete given the limited set of admissible quantizers [25]. Moreover, to meet the heterogeneous capacities or capabilities of users, a stored source video has often been encoded into multiple versions, each with a different encoding bitrate [28]. For example, the videos on YouTube can have 3-5 versions, of such resolutions as 240p, 360p, 480p, 720p and 1080p for different users [12].

Assume there are M admissible quantizers in source coding, or the source video is encoded into M versions. The encoding bitrate for version i is R_i, $i = 1, 2, ..., M$, where version 1 obviously has the lowest quality and version M the highest quality. The utility value of version i is given by u_i, $i = 1, 2, ..., M$, which denotes the perceived user experience.

The utility function of a typical streaming application can then be described as:

$$U_s(R) = \begin{cases} u_M & \text{if } R \geq R_M, \\ u_i & \text{if } R_i \leq R < R_{i+1}, \forall i \in \{1, ..., M-1\}, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where R represents the available data rate, and a user always chooses the version with the highest quality that is commensurate with the user's data rate.

It is worth noting that delay, particularly its jitter, is also a critical concern in streaming applications that demand continuous playback. In practice, if the data rate can be maintained above the source encoding rate, then the delay jitter can be effectively masked through buffering, which is available in all modern media streaming engines, e.g., Windows Media Player, Adobe Flash Player, and Apple QuickTime [29], [30]. Advertisements can also be inserted to mitigate the impact of the delay perceived by end users and to serve as a major income source, which are very common in today's commercial video sharing platforms, notably YouTube. Hence, in this paper, we use the data rate as the key parameter for utility calculation, and we consider both the linear relation $u_i = \frac{R_i}{R_M}, \forall i \in 1, ..., M$, and the logarithmic relation $u_i = \log_2(1 + \frac{R_i}{R_M}), \forall i \in 1, ..., M$. The latter not only addresses the inter-user fairness but also reflects the non-linear relation between the perceived video quality and encoding bitrate of state-of-the-art video encoders [31], [32].

IV. Optimal Sharing with Different Modes

Given the resource allocation problem and the utility functions of the applications, it is necessary to first derive the optimal allocation strategy for each of the sharing modes between the cellular and D2D communications.

A. Resource Allocation with Dedicated Mode

We first investigate the dedicated mode, in which the D2D communications take an exclusive portion of the spectrum resources from the cellular network and leave the remaining resources to the cellular users. Hence, the cellular and D2D communications do not cause interference to each other.

TABLE I
SUMMARY OF NOTATIONS

Notation	Description
W	System bandwidth
g_i	Channel gain between UE_i and the BS
g_{ij}	Channel gain between UE_i and UE_j
γ	SNR or SINR value
r	Spectrum efficiency
U_s	Utility function for streaming applications
U_f	Utility function for file sharing applications
P_i	Transmit power of UE_i (the maximum value is P^{max})
P_B	Transmit power of the BS (the maximum value is P_B^{max})
R_c	Data rate of cellular communications
R_d	Data rate of D2D communications
N_0	Noise power
WCU	Weighted cell utility
DM	Dedicated mode
CM	Cellular mode
ULre	Uplink reuse mode
DLre	Downlink reuse mode
$\alpha \in \Omega_\alpha$	The portion of resources allocated to D2D communications
M	Number of video versions

We use α to denote the portion of resources reserved for the cellular communications. Assuming that UE_2 is transmitting, with the Shannon capacity formula [22], we can obtain the data rate of the cellular and D2D communications, respectively, as follows:

$$R_{c\to BS}^{DM} = \frac{\alpha W}{2}\log_2(1 + \gamma_{c\to BS}^{DM}) = \frac{\alpha W}{2}\log_2(1 + \frac{g_1 P_1}{N_0 \alpha W/2}), \quad (3)$$

$$R_{BS\to c}^{DM} = \frac{\alpha W}{2}\log_2(1 + \gamma_{BS\to c}^{DM}) = \frac{\alpha W}{2}\log_2(1 + \frac{g_1 P_B}{N_0 \alpha W/2}), \quad (4)$$

$$R_d^{DM} = \frac{\alpha' W}{2}\log_2(1 + \gamma_d^{DM}) = \frac{\alpha' W}{2}\log_2(1 + \frac{g_{23} P_2}{N_0 \alpha' W/2}), \quad (5)$$

where $0 \leq \alpha \leq 1, \alpha' = 1 - \alpha$, W is the total frequency bandwidth that is equally occupied by the uplink and downlink, as previously described, and γ reflects the channel condition of the corresponding link.

It is worth noting that here we distinguish between the uplink and downlink of a cellular user, which extends the existing works that take the uplink data rate as the cellular data rate when maximizing the sum rate [9], [10]. The reason is twofold. First, the transmit power of the BS is much higher than that of the UEs and thus the downlink peak data rate is also higher in most cellular systems. Second, we deal with heterogeneous applications, which can be throttled by either the uplink, e.g. file sharing when a cellular user is transmitting data, or the downlink, e.g. video streaming; certain applications can be even throttled by both, e.g. 2-way video calling [33]. As such, only considering the resource allocation for the uplink of may lead to over/under-provisioning of resources for applications with different demands, as will be validated in Section VI.

First we assume that the cellular communications are serving a video streaming application, and the D2D communications are serving a file sharing application. We will extend to other application scenarios and larger system scales in Section V. Then the weighted cell utility becomes:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

$$\begin{aligned}&\lambda U_s(R_{BS\to c}^{DM})+\lambda' U_f(R_d^{DM})\\&=\lambda U_s(\frac{\alpha W}{2}\log_2(1+\gamma_{BS\to c}^{DM}))+\lambda' U_f(\frac{\alpha' W}{2}\log_2(1+\gamma_d^{DM}))\\&=\lambda U_s(\frac{\alpha W}{2}\log_2(1+\frac{g_1P_B}{N_0\alpha W/2}))+\lambda' U_f(\frac{\alpha' W}{2}\log_2(1+\frac{g_{23}P_2}{N_0\alpha' W/2})).\end{aligned} \tag{6}$$

The domain of α can be either continuous between 0 and 1 if the spectrum can be partitioned arbitrarily, which is an ideal situation; or a set of values if the spectrum is allocated at a granularity of subcarrier, which is adopted in practical cellular networks [34]. In this paper, we consider the latter case and use Ω_α to denote the set of all the possible values of α.

Algorithm 1 Resource Allocation for Dedicated Mode

1: $P_B = P_B^{max}, P_i = P^{max}, \forall i \in \{1,2,3\}$;
2: $WCU^{max} = 0$;
3: **for** $\alpha \in \Omega_\alpha$ **do**
4: Calculate $R_{BS\to c}^{DM}$, R_d^{DM}, and WCU according to Eqs. (4), (5) and (6), respectively;
5: **if** $WCU > WCU^{max}$ **then**
6: $WCU^{max} = WCU$; $\alpha^* = \alpha$; $R^* = \max_{R_i \in R_1,\ldots,R_M} R_i \le R_{BS\to c}^{DM}$;
7: **end if**
8: **end for**
9: $\alpha = \alpha^*$; $P_B = (2^{\frac{2R^*}{\alpha W}} - 1)N_0\alpha W/2)/g_1$;
10: Return WCU^{max}, P_B, α;

Since in the dedicated mode, there is no interference between the cellular and D2D communications, the BS and UEs can use the maximum power to transmit if necessary. Hence, we can simply set the transmit power of the BS and all UEs to the maximum values P_B^{max} and P^{max}, respectively. By calculating the weighted cell utility with each of the possible values of α, we can obtain the value of α giving the maximum weighted cell utility. Then we can reduce the transmit power of the BS to the highest value that does not degrade the cellular utility for the purpose of power efficiency. The pseudo-code is shown in Algorithm 1. After running the algorithm, the obtained α and P_B determine the optimal resource allocation strategy for the dedicated mode that offers the highest weighted cell utility (WCU^{max}). The computational complexity is $O(|\Omega_\alpha|\log_2 M)$, where $|\Omega_\alpha|$ denotes the number of values of α, which is in the order of the number of subcarriers, and we use binary search in step 10.

B. Resource Allocation with Cellular Mode

The operations in the cellular mode are quite similar to those of the dedicated mode except that the BS works as a relay node for the communications between D2D pairs. Hence, we can easily extend the system model and problem formulation of the dedicated mode to the cellular mode. Similar to the dedicated mode, a portion of the cellular resources are exclusively allocated to D2D communications. A D2D device first needs to transmit the data to the BS, and then the BS relays the data to the paired D2D device. Assuming that UE_2 is transmitting, the data rates in the cellular mode are as follows:

$$R_{BS\to c}^{CM} = \frac{\alpha W}{2}\log_2(1+\gamma_{BS\to c}^{CM}) = \frac{\alpha W}{2}\log_2(1+\frac{g_1P_B}{N_0\alpha W/2}), \tag{7}$$

$$\begin{aligned}R_d^{CM} &= \frac{\alpha' W}{2}\cdot\frac{1}{2}\log_2(1+\gamma_d^{CM})\\&= \frac{\alpha' W}{4}\log_2(1+\min(\frac{g_2P_2}{N_0\alpha' W/2},\frac{g_3P_B}{N_0\alpha' W/2})).\end{aligned} \tag{8}$$

The weighted cell utility can be calculated as follows:

$$\begin{aligned}&\lambda U_s(R_{BS\to c}^{CM})+\lambda' U_t(R_d^{CM})\\&=\lambda U_s(\frac{\alpha W}{2}\log_2(1+\frac{g_1P_B}{N_0\alpha W/2}))\\&\quad+\lambda' U_t(\frac{\alpha' W}{4}\log_2(1+\min(\frac{g_2P_2}{N_0\alpha' W/2},\frac{g_3P_B}{N_0\alpha' W/2}))).\end{aligned} \tag{9}$$

Similar to the dedicated mode, we need to find the optimal partitioning of the spectrum resources. We can reuse Algorithm 1 with slight modifications to obtain the optimal resource allocation strategy for the cellular mode as follows. First we need to find the link (from the transmitter to the BS or from the BS to the receiver) having lower SNR, which determines the achievable data rate of the D2D communications. We then calculate the WCU according to Eq. (9) for different values of α, and find the optimal one offering the highest WCU. The computational complexity is also $O(|\Omega_\alpha|\log_2 M)$.

C. Resource Allocation with Reuse Mode

In the reuse mode, the D2D communications can use either the uplink or downlink spectrum resources of the cellular users. We do not need to consider the partitioning of the spectrum resources in the reuse mode since the D2D communications will reuse the whole uplink/downlink spectrum. On the other hand, we need to carefully set the transmit power of the BS and UEs to control the interference, which is more challenging. The transmit power of the BS and UEs cannot be simply set to the respective maximum values as in the dedicated and cellular modes, because the interference will also be maximized and significantly impact the data rate of the interfered links. The interference may come from any D2D user depending on which one is transmitting at the moment. Similar to the previous section, we assume that UE_2 is the transmitter in the file sharing application. The derivation follows the same steps when UE_3 is the transmitter.

Since the reuse mode can be further categorized into uplink reuse and downlink reuse modes, we need different resource allocation strategies for each of them.

1) Uplink Reuse: We start from the uplink reuse mode, which is relatively easier to analyze since in our application scenario the cellular communication is throttled by the downlink that does not interfere with the D2D communication.

The data rates of the cellular and D2D communications in the uplink reuse mode are as follows:

$$R_{BS\to c}^{ULre} = \frac{W}{2}\log_2(1+\gamma_{BS\to c}^{ULre}) = \frac{W}{2}\log_2(1+\frac{g_1P_B}{N_0W/2}), \tag{10}$$

$$R_d^{ULre} = \frac{W}{2}\cdot\frac{1}{2}\log_2(1+\gamma_d^{ULre}) = \frac{W}{4}\log_2(1+\frac{g_{23}P_2}{g_{13}P_1+N_0W/2}). \tag{11}$$

And the weighted cell utility can be calculated as follows:

$$\begin{aligned}&\lambda U_s(R_{BS\to c}^{ULre})+\lambda' U_f(R_d^{ULre})\\&=\lambda U_s(\frac{W}{2}\log_2(1+\frac{g_1P_B}{N_0W/2}))+\lambda' U_f(\frac{W}{4}\log_2(1+\frac{g_{23}P_2}{g_{13}P_1+N_0W/2})).\end{aligned} \tag{12}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

Since the downlink of the cellular and D2D communications do not generate interference to each other, we can optimize them separately. We set the transmit power the BS to the maximum value, P_B^{max}, to maximize the cell utility. We also set the transmit power of UE_2 and UE_3 to the maximum value P^{max}, and set the transmit power of UE_1 to the value that can support the lowest MCS to minimize its interference to the D2D communications. This strategy will offer the highest weighted cell utility for the uplink reuse mode.

2) Downlink Reuse: Similarly, we can derive data rates in the downlink reuse mode as follows:

$$\begin{aligned} R_{BS\to c}^{DLre} &= \frac{W}{2}\log_2(1+\gamma_{BS\to c}^{DLre}) \\ &= \frac{W}{2}\log_2(1+\frac{g_1P_B}{\max(g_{12}P_2, g_{13}P_3)+N_0W/2}), \end{aligned} \tag{13}$$

$$R_d^{DLre} = \frac{W}{2}\cdot\frac{1}{2}\log_2(1+\gamma_d^{DLre}) = \frac{W}{4}\log_2(1+\frac{g_{23}P_2}{g_3P_B+N_0W/2}). \tag{14}$$

In this case, the downlink of the cellular communication experiences the interference caused by the D2D communication and vice versa. Therefore, we need to jointly adjust the transmit power of the BS and UEs. The weighted cell utility can be derived as:

$$\begin{aligned} &\lambda U_s(R_{BS\to c}^{DLre}) + \lambda' U_f(R_d^{DLre}) \\ &= \lambda U_s(\frac{W}{2}\log_2(1+\frac{g_1P_B}{\max(g_{12}P_2, g_{13}P_3)+N_0W/2})) \\ &\quad + \lambda' U_f(\frac{W}{4}\log_2(1+\frac{g_{23}P_2}{g_3P_B+N_0W/2})). \end{aligned} \tag{15}$$

Since the utility function of streaming applications (U_s) is not continuous and thus is not differentiable, we cannot obtain a closed form of the optimal values of P_B, P_2 and P_3. Fortunately the optimal solution can be obtained by exploiting the discreteness of the utility function U_s. The main idea is as follows. First we compute the highest feasible SINR γ', where $\gamma' = \frac{g_1P_B^{max}}{N_0W/2}$. Further we use $\gamma^{(i)}$ to denote the required SINR for version i. Then for each $\gamma^{(i)} \le \gamma'$ we solve the following optimization problem to obtain the highest weighted cell utility $WCU^{(i)}$ in this case:

$$\begin{aligned} \text{Maximize} \quad & \frac{g_{23}P_2}{g_3P_B+N_0W/2} \\ \text{Subject to} \quad & P_B \le P_B^{max}, \\ & P_2 \le P^{max}, \\ & \frac{g_1P_B}{g_{12}P_2+N_0W/2} \ge \gamma^{(i)}, \\ & P_B, P_2 \ge 0. \end{aligned} \tag{16}$$

To maximize the objective function, P_2 should be as high as possible and P_B should be as low as possible. The optimal value is reached when the SINR constraint $\frac{g_1P_B}{g_{12}P_2+N_0W/2} = \gamma^{(i)}$ is satisfied. Substituting this equality into the objective function, we have:

$$\begin{aligned} \frac{g_{23}P_2}{g_3P_B+N_0W/2} &= \frac{g_{23}P_2}{\frac{g_3}{g_1}\gamma^{(i)}(g_{12}P_2+N_0W/2)+N_0W/2} \\ &= \frac{g_{23}}{\frac{g_3}{g_1}\gamma^{(i)}(g_{12}+\frac{N_0W/2}{P_2})+N_0W/2}. \end{aligned} \tag{17}$$

We can see that the objective function increases monotonically with P_2. Hence, the maximum of the objective function is obtained when P_2 takes the maximum value subject to the SINR constraint as follows:

$$P_2 = \min(\frac{1}{g_{12}}(\frac{g_1P_B^{max}}{\gamma^{(i)}} - N_0W/2), P^{max}). \tag{18}$$

Substituting this into the SINR constraint, we have

$$P_B = \frac{\gamma^{(i)}(g_{12}P_2+N_0W/2)}{g_1}. \tag{19}$$

We also consider the case where the cellular user has no enough data rate to watch the video of the lowest version. Then highest weighted cell utility $WCU^{(0)}$ is obtained by setting $P_B = 0$ and $P_2 = P^{max}$. At last the value of P_B and P_2 resulting in the highest WCU is selected as the optimal strategy for the downlink reuse mode (we can simply set $P_3 = \frac{g_{12}P_2}{g_{12}}$ such that the SINR constraint is not violated). The pseudo-code is shown in Algorithm 2. The computational complexity is $O(M)$.

Algorithm 2 Resource Allocation for Downlink Reuse Mode

$\gamma' = \frac{g_1P_B^{max}}{N_0}$;
for $i = 1 : M$ **do**
  if $\gamma^{(i)} \le \gamma'$ **then**
    Calculate P_2, P_B and $WCU^{(i)}$ according to Eqs. (18), (19) and (15), respectively;
    if $WCU^{(i)} > WCU^{max}$ **then**
      $WCU^{max} = WCU^{(i)}$; $P_2^* = P_2$; $P_B^* = P_B$;
    end if
  else
    break
  end if
end for
$WCU^{(0)} = \lambda' U_f(\frac{W}{4}\log_2(1+\frac{g_{23}P^{max}}{N_0}))$;
if $WCU^{(0)} > WCU^{max}$ **then**
  $WCU^{max} = WCU^{(0)}$; $P_2^* = P^{max}$; $P_B^* = 0$;
end if
$P_2 = P_2^*$; $P_B = P_B^*$; $P_3 = (g_{12}P_2)/g_{12}$;
Return WCU^{max}, P_B, P_2, P_3;

The strategies for all the above resource sharing modes refer to the transmit power of the BS and each UE, plus the value of α that determines the allocation of bandwidth resources for the dedicated and cellular modes. After obtaining the resource allocation strategies for all the resource sharing modes, we can select the one with the highest weighted cell utility as well as the corresponding mode. The overall computational complexity is $O(\max(|\Omega_\alpha|\log_2 M, M))$.

V. Extension and Further Discussion

We now discuss how to extend our solutions to other general application scenarios and larger systems with multiple cellular users and D2D pairs.

A. General Application Scenarios

If the cellular communications serve a file sharing application, the bottleneck is the uplink, and the utility function changes to $U_f(R_{c\to BS})$. If the cellular communications serve a 2-way video calling application, both the uplink and downlink can be the bottleneck. Assuming the utility function of video calling applications as U_{vc}, the utility function of the cellular communications changes to $U_{vc}(\min(R_{c\to BS}, R_{BS\to c}))$. If the D2D communications serve different applications other

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

than the file sharing applications, we can also change the utility function accordingly.

Since there is no interference in both the dedicated and cellular modes, the optimization is almost the same. We can set the transmit power of the BS and UEs to the respective maximum values, and search for the optimal value of α offering the highest WCU.

The case for the reuse mode is more complex due to the interference. If the cellular communications serve file sharing applications and the D2D communications serve streaming applications, for uplink reuse, we can set P_2 and P_3 to P^{max}. We calculate $\gamma^{(i)}$ according to Eq. (18). The strategy of adjusting transmit power offers the highest WCU under different values of $\gamma^{(i)}$ is selected, which is similar to Algorithm 2. For downlink reuse, we can set the transmit power of all the UEs to the maximum value and the transmit power of the BS to the value that can support the lowest MCS for all the UEs.

If both the cellular and D2D communications serve streaming applications, the solution for uplink reuse is the same as in our original scenario. For downlink reuse, the approach to finding the optimal strategy is similar to Algorithm 2, and the worst case complexity is also $O(M)$. For each $\gamma^{(i)}$ received at the cellular user, we also set $P_2 = \min\{\frac{1}{g_{12}}(\frac{g_1 P_B^{max}}{\gamma^{(i)}} - N_0 W/2), P^{max}\}$ to maximize the D2D utility.

If both the cellular and D2D communications serve file sharing applications, the solution for the uplink reuse mode is the same as the original scenario. The solution for the downlink reuse case is different. The weighted cell utility now is given by:

$$\begin{aligned} &\lambda U_f(R_{BS\to c}^{DLre}) + \lambda' U_f(R_d^{DLre}) \\ &= \lambda U_f(\frac{W}{2}\log_2(1+\frac{g_1 P_B}{\max(g_{12}P_2, g_{13}P_3)+N_0 W/2})) \\ &\quad + \lambda' U_f(\frac{W}{4}\log_2(1+\frac{g_{23}P_2}{g_3 P_B + N_0 W/2})). \end{aligned} \tag{20}$$

Since both utility functions are continuous and differentiable, we can obtain a closed form of the optimal solution by letting the partial derivative of the expression on the right side of Eq. (29) with respect to P_B and P_2 to be zero, respectively, and then solving the system of equations to get the transmit power of the BS and UEs. The method can be generalized to other application scenarios with given continuous or discrete utilities functions.

B. Larger Systems with Multiple Users

For larger systems with multiple cellular users and D2D pairs, we can assume that the spectrum resources are equally shared among the cellular users [10], or are allocated based on the link qualities of different users [35], [36]. We further assume that the base station adopts some admission control mechanisms such that the number of D2D pairs allowed is no more than the number of cellular users and each reuse group consists of one cellular user and at most one D2D pair. This matching can be obtained by randomly picking a cellular user and a D2D pair, or picking a cellular user and a D2D pair who are far away enough such that the maximum interference is below a given threshold.

After the matching, the spectrum allocation and transmit power adjustment problem of the whole system now transforms to independent subproblems for each group that consists of one cellular user and at most one D2D group, which is exactly the scenario we were discussing in the previous section. Assuming that there are N cellular users and N D2D pairs, then the worst case complexity of the proposed centralized algorithm is $O(N * \max(|\Omega_\alpha| \log_2 M, M))$. The centralized algorithm can be distributed as follows such that the computational burden on BSes can be effectively mitigated. We assume that all the UEs will report their location information to the BSes. Hence, the base station can deliver the location information of the matching D2D pair to each cellular UE (thus the channel gain can be calculated). Then each cellular UE will find the optimal strategy for its own group, with the worst case complexity of $O(\max(|\Omega_\alpha| \log_2 M, M))$, and send back to the BSes, which then deliver the strategy to the corresponding D2D pair.

C. Implementation Requirements of D2D Communications

The infrastructure of existing cellular systems needs several modifications to effectively implement D2D communications. For example, UEs need to be able to directly communication with each other using the spectrum resources of cellular systems. Further, the channel gain information between UEs is required for resource allocation. The dedicated and cellular modes are easy to implement since the cellular and D2D communications operate on different spectrum and thus all the UEs can transmit at the maximum power to achieve the highest data rate, without generating interference to each other. While for the reuse mode, sophisticated power control mechanisms are needed to limit the interference and more channel gain information is required. Further, the movement of users would change the extent of interference significantly, and thus demanding more frequent updating channel gain and tuning the spectrum allocation and power control strategy. On the other hand, the reuse mode can provide higher spectrum utilization in many occasions, as will be validated in Section VI.

VI. Performance Evaluation

We have performed extensive simulations to evaluate the performance of the proposed QoS-aware resource allocation scheme. We developed a customized simulator using the Python programming language (version 2.7.3) to capture the essence of state-of-the-art LTE systems. The simulator was run on a PC with an Intel Core i7-3770 CPU at 3.40 GHz, 8 GB of RAM, and the 64bit Linux Ubuntu 12.04 operating system. Table II summarizes the simulation parameters and their default values, mostly adapted from [17], [24]. We allocate the spectrum resources at a granularity of Resource Blocks (RBs), each composed of 12 adjacent subcarriers of 15 KHz and thus the RB bandwidth is 180 KHz, as in the LTE system [24]. The carrier frequency is 2 GHz, and the path loss is composed of the distance attenuation 35.3 + 37.6$\times \log(d)$, where d is the distance in terms of meters, and shadow fading. We first simulated a single cellular network with one cellular user (UE_1) coexisting with a pair of D2D users (UE_2 and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Area size	200 m × 200 m
Carrier frequency	2.0 GHz
System bandwidth	5 MHz, 10 MHz, 20 MHz
Number of subcarriers per RB	12
RB bandwidth	15 kHz
RB bandwidth	180 kHz
Number of RB	24, 50, 100
Max BS Tx power	20 W (43 dBm)
BS antenna gain	14 dBi
BS noise figure	5 dB
Max UE Tx power	100 mW (20 dBm)
UE antenna gain	0 dBi
UE noise figure	9 dB
Distance between D2D UEs	1 to 50 m
Antenna pattern	Omni
MCS	QPSK: 1/12, 1/9, 1/6, 1/3, 1/2, 3/5 16QAM: 1/3, 1/2, 3/5 64QAM: 1/2, 3/4, 3/5, 5/6, 11/12
Distance attenuation	35.3 + 37.6× $\log(d)$
Log-normal shadowing std	8 dB
Noise density	-174 dBm/Hz
Bandwidth efficiency	0.83
User distribution	Uniform
Video encoding bitrate	500, 1000, 2500, 5000, 8000 kbps

UE_3). We further conducted a simulation with larger system scale. In both simulations, the BS is located at the center of a rectangular area of 200 m × 200 m. The location of the UEs are uniformly distributed in the area while the distance between the D2D users ranges from 1 to 50 m. The mean and standard deviation of the shadow fading variables are 0 dB and 8 dB, respectively. The CSI is calculated at the UEs and then fed back to the BS. We adopt an advanced link adaptation technique in [24], where a proper MCS is selected from the available MCSes (e.g., QPSK, 16QAM and 64QAM) with different coding rates ranging from 1/12 to 11/12 according to the estimated SINR value. Each MCS has a SINR threshold value that corresponds to 10% BLER (see [24] for details).

A. One Cellular User and One D2D Pair: A Case Study

For this scenario, we have experimented with a total bandwidth of both 5 MHz and 10 MHz, which is equally occupied by the uplink and downlink. The number of RBs is 24 with 5 MHz system bandwidth and 50 with 10 MHz system bandwidth. The maximum data rate R^{max} is obtained by allocating all the RBs, coded using the MCS with the highest coding rate, to the D2D communications. The source video is encoded into 5 versions, namely 240p, 360p, 480p, 720p and 1080p, with the corresponding bitrates ranging from 500 to 8000 kbps, which are the recommended bitrates for standard quality uploads of YouTube[3].

1) Performance of Different Modes: We first evaluate the performance of the resource allocation of different resource sharing modes. For each mode, we also investigate the impact of the two different types of utility functions. Here we set the value of λ to 0.5 such that the cellular and D2D communications are given equal weight. We will investigate the impact of different values of λ later. We perform 500 times of simulations with different locations of UEs to mitigate randomness.

We find all of the sharing modes can offer the cellular user the highest quality video, yet different data rates of the D2D users (referred to as D2D data rate in the following). We plot the average over 500 simulations (5 MHz and 10 MHz) in Fig. 3, and also report the detail statistics in Table III. W represents the overall system bandwidth, and the same in the following tables. When the system bandwidth is 5 MHz, both the uplink reuse and downlink reuse modes have higher average D2D data rates than those of the remaining two modes. The reason is that in the two reuse modes, half of the system bandwidth is available for the D2D users, and the cellular user exclusively occupies the other half. Yet, in both dedicated and cellular modes, the D2D users need to compete for the bandwidth resources with the cellular user. The D2D data rates of both dedicated and cellular modes are rather consistent, which are largely determined by the distance between the D2D users. The D2D data rates of both two reuse modes however incur very high variation, likely caused by the interference from the cellular communications. For downlink reuse, the D2D data rate will be higher if the receiving UE (UE_3) is far away from the BS and could be zero if it is too close. Similarly, the D2D data rate with uplink reuse depends on the distance between UE_1 and UE_3. The uplink reuse mode has a higher D2D data rate since the transmit power of UE_1 is generally lower than that of the BS and thus the interference caused by UE_1 is smaller. The cellular mode is only feasible when the D2D users are far apart from each other, as compared with their respective distance to the BS. Recall that we have limited the maximum distance between the D2D users to 50 m, and so the cellular mode is rarely selected in our simulation setting.

On the other hand, when the system bandwidth is 10 MHz, the dedicated mode offers a significantly higher D2D data rate, as compared with other modes. The reason is that, after allocating bandwidth resources enough for the video of the highest quality to the cellular communications, all the remaining bandwidth resources are allocated to the D2D communications. While in the uplink reuse mode, half of the resources are always allocated to the downlink of the cellular communications, which is far beyond the encoding bitrate of the highest quality video. This over-provisioning leaves less resources to the D2D communications, as compared with the dedicated mode. When the system bandwidth keeps growing or Multiple Input Multiple Output (MIMO) that supports higher spectrum efficiency is adopted, the gap between the dedicated and reuse modes will be further expanded.

We present the number of each mode selected as the best using the proposed scheme in Table IV. The results verifies the above discussion on mode selection. The cellular mode is selected in very few cases since in this mode, D2D communications need two steps. Whether to select the dedicated mode or the reuse mode mainly depends on the system bandwidth. When the system bandwidth is limited, say 5 MHz in our simulation, the reuse mode is preferred. Specifically, the uplink reuse mode is more preferred than the downlink reuse mode since the bottleneck links of the two applications are decoupled. According to Eq. (19), we can set the transmit

[3]According to the advanced encoding settings of YouTube: http://support.google.com/youtube/bin/answer.py?hl=en&answer=1722171.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

TABLE III
STATISTICS OF D2D DATA RATE (MBPS).

Mode	W	Max	Min	Mean	Median	Std
DM	5 MHz	1.658	1.351	1.657	1.658	0.020
	10 MHz	12.438	3.317	12.258	12.438	1.045
CM	5 MHz	0.829	0.176	0.804	0.829	0.103
	10 MHz	6.219	0.448	5.269	6.219	1.614
ULre	5 MHz	4.975	0	3.653	4.975	1.884
	10MHz	10.365	0	7.536	10.365	3.931
DLre	5 MHz	4.975	0	2.071	1.058	2.144
	10MHz	10.365	0	4.953	2.764	4.527

TABLE IV
OF EACH MODE SELECTED IN SIMULATIONS

W	DM	CM	ULre	DLre
5 MHz	116	0	363	21
10 MHz	497	3	0	0

power of the BS and UE_2 to the maximum without causing interference to each other. While for the downlink reuse mode, the BS and UE_2 will cause interference to each other, leading to higher SINR. Further examination shows that the downlink reuse mode is superior only when UE_1 is far from the BS but close to UE_3 such that even UE_1 even using the lowest MCS (and thus the lowest transmit power) would cause significant interference at UE_3. On the other hand, when the system bandwidth becomes larger, say 10 MHz, the dedicated mode dominates other three modes since it only allocates the exact bandwidth resources needed to support the highest quality video, which do not increase with the system bandwidth. Hence, the increased bandwidth resources are all exclusively allocated to the D2D communications.

TABLE V
OF VIDEOS IN EACH VERSION FOR LINEAR UTILITY FUNCTION

Version	W	λ				
		0.1	0.2	0.3	0.4	0.5
0	5 MHz	4	4	0	0	0
	10 MHz	9	3	0	0	0
1	5 MHz	495	492	486	128	0
	10 MHz	482	472	0	0	0
2	5 MHz	1	3	2	2	0
	10 MHz	6	0	0	0	0
3	5 MHz	0	1	0	0	0
	10 MHz	0	4	0	0	0
4	5 MHz	0	0	5	4	0
	10 MHz	1	1	0	0	0
5	5 MHz	0	0	7	366	500
	10 MHz	2	20	500	500	500

TABLE VI
OF VIDEOS IN EACH VERSION FOR LOG UTILITY FUNCTION

Version	W	λ				
		0.1	0.2	0.3	0.4	0.5
0	5 MHz	0	0	0	0	0
	10 MHz	0	0	0	0	0
1	5 MHz	497	485	180	0	0
	10 MHz	484	493	0	0	0
2	5 MHz	1	4	2	0	0
	10 MHz	6	3	0	0	0
3	5 MHz	2	5	2	139	0
	10 MHz	0	1	0	0	0
4	5 MHz	0	3	5	5	0
	10 MHz	3	3	0	0	0
5	5 MHz	0	3	311	356	500
	10 MHz	7	0	500	500	500

2) Impact of Weight: We next investigate the impact of the weight value λ on the system performance. We vary the value of λ from 0.1 to 0.9 with a step of 0.1, and for each λ we select the mode with the highest weighted cell utility. We report the number of videos in each version with different values of λ for the two utility functions in Table V and Table VI, respectively. Version 0 refers to that the cellular data rate is lower than the bitrate of version 1 and thus even the lowest quality video can not be played smoothly. We omit the results when the value of λ is higher than 0.5 since with $\lambda = 0.5$, the cellular user can already watch the highest quality video and the results will remain the same. When the system bandwidth is 5 MHz, the video quality quickly shifts from the lowest to the highest with increasing λ for both utility functions. When the system is 10 MHz, the two functions offer almost the same video quality with different values of λ. Further, we can see that the benefit of increasing system bandwidth for the cellular user is insignificant when too little weight is assigned to the cellular communications.

TABLE VII
OF VIDEOS IN EACH VERSION

Version	W	Proposed	Baseline1	Baseline1 (0.7)	Baseline1 (0.9)	Baseline2	Baseline2 (0.7)	Baseline2 (0.9)
0	5 MHz	0	1	4	7	0	0	0
	10 MHz	0	1	3	4	0	0	0
1	5 MHz	0	56	34	1	0	0	0
	10 MHz	0	82	32	3	0	0	0
2	5 MHz	0	278	22	13	41	0	0
	10 MHz	0	2	19	8	0	0	0
3	5 MHz	0	10	29	33	2	0	0
	10 MHz	0	254	23	5	39	0	0
4	5 MHz	0	5	33	24	3	0	0
	10 MHz	0	9	25	9	0	0	0
5	5 MHz	500	150	378	422	454	500	500
	10 MHz	500	152	398	471	461	500	500

We also plot the average D2D data rate with different λ for the two utility functions in Fig. 4. When the system bandwidth is 5 MHz, the log utility function offers almost identical D2D data rate, as compared with the linear utility function with λ = 0.1, 0.2, and 0.5. Yet the log utility function offers slightly lower D2D data rate with λ = 0.3 and 0.4, since more resources are allocated to cellular communications, which is consistent with the observation that the average video quality is better. When the system bandwidth is 10 MHz, the two utility functions have almost the same average D2D data rate since they also offer almost the same video quality which quickly shifts from the lowest to the highest when λ reaches 0.3.

3) Comparison with Baseline Schemes: We further compare our solution with a state-of-the-art scheme that maximizes the total data rate with no QoS differentiation [9]. The original scheme, referred to as *baseline1* (Base1), defines the total data rate as the sum of the uplink data rate of the cellular user and the D2D data rate. This baseline scheme ignores the fact different applications can be throttled by either the uplink or the downlink, e.g., the data traffic of both streaming and file sharing applications can be highly asymmetric. On the other hand, our scheme considers the data rate of the communication link which carries the major traffic. To ensure a fair comparison, we modify baseline1 to maximize the total data rate of the communication link carrying the major traffic, referred to as *baseline2* (Base2).

Since baseline1 does not give priority to either cellular or D2D communications, we also set the weight parameter λ to 0.5 in our scheme, and use the linear utility function, which, as shown before, performs identically to the log utility in this case. We report the number of videos in each version of all the schemes in Table VII and plot the average D2D data rates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

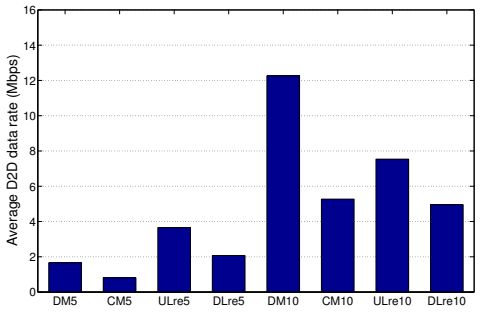

Fig. 3. Average D2D data rate for different resource sharing modes.

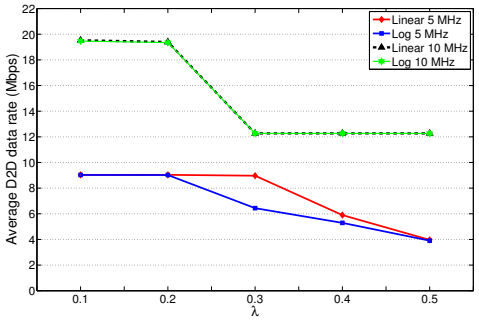

Fig. 4. Average D2D data rate with different λ.

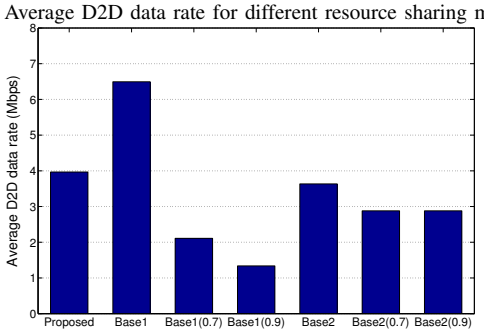

Fig. 5. Average D2D data rate with 5 MHz system bandwidth.

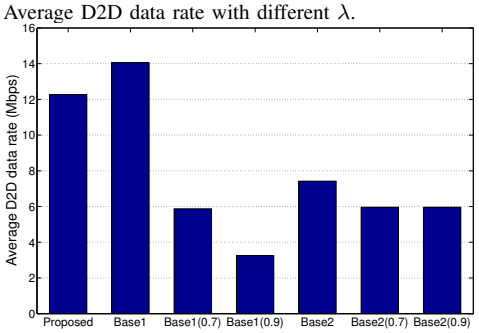

Fig. 6. Average D2D data rate with 10 MHz system bandwidth.

TABLE VIII
RUNNING TIME OF 500 SIMULATIONS (SECOND)

W	Proposed	Baseline1	Baseline2
5 MHz	0.09	0.07	0.07
10 MHz	0.13	0.12	0.12

TABLE IX
OF VIDEOS IN EACH VERSION WITH DIFFERENT SYSTEM SCALES

System scale	Version	Proposed	Baseline1	Baseline2
5-5	0	0	12	0
	1	0	6	0
	2	0	24	0
	3	0	25	0
	4	0	75	0
	5	500	358	500
10-10	0	0	22	0
	1	0	17	0
	2	0	37	0
	3	500	424	500
	4	0	0	0
	5	0	0	0
25-25	0	0	56	0
	1	0	26	0
	2	500	418	500
	3	0	0	0
	4	0	0	0
	5	0	0	0
50-50	0	0	81	0
	1	500	419	500
	2	0	0	0
	3	0	0	0
	4	0	0	0
	5	0	0	0

of all the schemes in Fig. 5 and Fig. 6 with different system bandwidth, respectively. The numbers in the bracket are the values of weights assigned to the cellular communications.

Compared with baseline1, our solution offers much better video quality for the cellular user. Although the average D2D data rate of the our solution is lower, the gap quickly decreases with increasing system bandwidth, and with more system bandwidth, our solution would eventually have a higher D2D data rate. Further, if we slightly reduce the value of λ without impacting the video quality, say to 0.3, the D2D data rate with our solution will be higher than that with baseline1. When a higher weight is assigned to the cellular user in baseline1, the video quality can be improved, but is still worse than ours, and meanwhile its D2D data rate will become much lower than ours.

On the other hand, baseline2 offers similar video quality as compared with our scheme since it optimizes the bottleneck communication links of applications. Its D2D data rate however is lower than our scheme and the gap keeps growing with more bandwidth. The reason is that baseline2 assigned more bandwidth resources which is far beyond the requirement of the highest quality video, leading to unnecessary over-provisioning.

The running time of simulations is shown in Table VIII. We can see that the efficiency of our scheme is comparable to that of the two baseline schemes.

In summary, baseline1 does not consider the different bottleneck links of diverse applications and thus the QoS specifications of applications may not be satisfied; baseline2 is only feasible when all the applications are of the file sharing type. When streaming applications are involved, baseline2 may lead to over-provisioning for the streaming applications and the precious spectrum resources will not be fully utilized to better serve the file sharing applications. Although we focus only on the two classes of applications, they are quite representative in real world, and our solution and discussions can be easily extended to other applications once given their specific QoS utility functions.

B. System Performance with Larger User Population

In this scenario, we set the total bandwidth to 20 MHz, which corresponds to 100 RBs. Our simulation is conducted on four system scales, namely 5 cellular users/5 D2D pairs, 10

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2015.2452586, IEEE Journal on Selected Areas in Communications

TABLE X
AVERAGE D2D DATA RATE WITH DIFFERENT SYSTEM SCALES (MBPS)

System scale	Proposed	Baseline1	Baseline2
5-5	3.104	1.003	2.409
10-10	1.703	0.450	1.205
25-25	0.632	0.169	0.482
50-50	0.316	0.082	0.241

TABLE XI
OF EACH MODE SELECTED IN SIMULATIONS WITH DIFFERENT SYSTEM SCALES

System scale	DM	CM	ULre	DLre
5-5	0	0	477	23
10-10	122	0	355	23
25-25	0	0	472	28
50-50	0	0	472	28

cellular users/10 D2D pairs, 25 cellular users/25 D2D pairs, and 50 cellular users/50 D2D pairs,respectively. We run the simulator 100, 50, 20 and 10 times for the four system scales, respectively, such that the number of total data points is 500 for all of them. In each simulation, each cellular user is randomly matched with exactly one D2D pair to form a reuse group. The total bandwidth is equally distributed to all reuse groups. We use linear utility function in our scheme and compare the performance of our scheme with the two baseline schemes. Given that the spectrum resources per reuse group becomes less as the system scale increases, we set $\lambda = 0.9$ to respect the priority of cellular users.

We report the number of videos in each version in Table IX and the average D2D data rate in Table X, respectively. We can see that the proposed scheme significantly outperforms *baseline1* in terms of both the video quality and D2D data rate. Compared with *baseline2*, the proposed scheme provides identical video quality to cellular users, and remarkably improves the average D2D rate at least 28.9% and up to 41.3%. The results again validate that the proposed scheme can better utilize the spectrum resources by considering the QoS specifications of applications.

We also report the number of each mode selected in simulations with different system scales in Table XI. We can see there is no clue that one mode dominates the others as the system scale increases. Yet we still have several interesting observations. Similar to the simulation with small scale, cellular mode is rarely selected; the uplink reuse mode is more preferred than the downlink reuse mode since the bottleneck links are decoupled. Both the dedicated and reuse modes have their own advantages depending on the system scale and topology. Different from the small scale system with plenty of spectrum resources (e.g. 10 MHz), in the system with larger scale where each reuse group is allocated with limited spectrum resources, the reuse mode is more preferred than the dedicated mode since it has the potential to achieve higher spectrum efficiency via sharing the spectrum resources.

VII. CONCLUSION

In this paper, we addressed the resource allocation problem for device-to-device (D2D) communications in cellular networks serving applications of heterogeneous QoS requirements. We systematically investigated the problem under different resource sharing modes, including dedicated, cellular and reuse modes. We developed optimized solutions for the cellular and D2D communications to coordinated using the same licensed spectrum, so as to maximize the users' utility. Our solution was evaluated under diverse configurations and we also compared it with state-of-the-art schemes tuned for homogeneous applications. The results demonstrated that the superiority of our solution in terms of better resource utilization that effectively differentiates applications and users, and less possibility of under- or over-provisioning.

There are many possible directions toward extending our solution. We have presented preliminary discussion on accommodating more general applications and large system scales, which is worth of further investigations. We are also interested in extending our solution to a multi-cell scenario to better allocate the resources across cells. Energy consumption for the devices is another important aspect that can be taken into spectrum allocation.

REFERENCES

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017," *available at* http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf.

[2] A. Balasubramanian, R. Mahajan and A. Venkataramani, "Augmenting Mobile 3G Using WiFi," *in Proc. ACM MobiSys '10*, San Francisco, USA, 2010.

[3] P. Janis, C. Yu, K. Doppler, C.B. Ribeiro, C. Wijting, K. Hugl, O. Tirkkonen and V. Koivunen, "Device-to-Device Communication Underlaying Cellular Communications Systems," *Intl J. of Commun. Netw. Syst. Sci.*, vol. 2, no. 3, pp. 169-178, 2009, doi: 10.4236/ijcns.2009.23019.

[4] R. Want, "Near Field Communication", *IEEE Pervas. Comput.*, vol. 10, no. 3, pp. 4-7, Jul. 2011.

[5] C. Bisdikian, "An Overview of the Bluetooth Wireless Technology," *IEEE Commun. Mag.*, vol. 39, no. 12, pp. 86-94, Dec. 2001.

[6] Wi-Fi Alliance, P2P Technical Group, "Wi-Fi Peer-to-Peer (P2P) Technical Specification v1.0," December 2009.

[7] S. Gollakota, F. Adib, D. Katabi, and S. Seshan, "Clearing the RF Smog: Making 802.11n Robust to Cross-Technology Interference," *in Proc. ACM SIGCOMM '11*, Toronto, Canada, 2011.

[8] K. Doppler, M. Rinne, C. Wijting, C.B. 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.

[9] C.-H. Yu, K. Doppler, C.B. Ribeiro, and O. Tirkkonen, "Resource Sharing Optimization for Device-to-Device Communication Underlaying Cellular Networks," *IEEE Trans. Commun.*, vol. 10, no. 8, pp. 2752-2763, Aug. 2011.

[10] K. Doppler, C.-H. Yu, C.B. Ribeiro, and P. Janis, "Mode Selection for Device-to-Device Communication underlaying an LTE-Advanced Network," *in Proc. IEEE WCNC '10*, Sydney, Australia, Apr. 2010.

[11] P. Janis, V. Koivunen, C.B. Ribeiro, J. Korhonen, K. Doppler, and K. Hugl, "Interference-Aware Resource Allocation for Device-to-Device Radio Underlaying Cellular Networks," *IEEE VTC '09 Spring*, Barcelona, Spain, Apr. 2009.

[12] A. Finamore, M. Mellia, M.M. Munafò, R. Torres, and S.G. Rao, "YouTube Everywhere: Impact of Device and Infrastructure Synergies on User Experience," *in Proc. ACM IMC '11*, Berlin, Germany, Nov. 2011

[13] Z. Liu, T. Peng, S. Xiang, and W. Wang, "Mode selection for Device-to-Device (D2D) communication under LTE-Advanced networks," *in Proc. IEEE ICC '12*, Ottawa, Canada, Jun. 2012.

[14] I.F. Akyldiz, W.-Y. Lee, M. Vuran, and S. Mohanty, "A Survey on Spectrum Management in Cognitive Radio Networks," *IEEE Commun. Mag.*, vol. 46, no. 4, pp. 40-48, Apr. 2008.

[15] H. Min, W. Seo, J. Lee, S. Park, and D. Hong, "Reliability Improvement Using Receive Mode Selection in the Device-to-Device Uplink Period Underlaying Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 2, pp. 413-418, Feb. 2011.

[16] C.-H. Yu, O. Tirkkonen, K. Doppler, and C.B. Ribeiro, "Power Optimization of Device-to-Device Communication Underlaying Cellular Communication", *in Proc. IEEE ICC '09*, Dresden, Germany, Jun. 2009.

[17] M. Zulhasnine, C. Huang, and A. Srinivasan, "Efficient Resource Allocation for Device-to-Device Communication Underlaying LTE Network," *in Proc. IEEE WiMob '10*, Niagara Falls, Canada, Oct. 2010.

[18] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, and B. Jiao, "Interference-Aware Resource Allocation for Device-to-Device Communications as an Underlay Using Sequential Second Price Auction," *in Proc. IEEE ICC '12*, Ottawa, Canada, Jun, 2012.

[19] C.-H. Yu, O. Tirkkonen, K. Doppler, and C.B. Ribeiro, "On the Performance of Device-to-Device Underlay Communication with Simple Power Control," *in Proc. IEEE VTC '09 Spring*, Barcelona, Spain, Apr. 2009.

[20] M. Belleschi, G. Fodor, and A. Abrardo, "Performance Analysis of A Distributed Resource Allocation Scheme for D2D Communications," *in Proc. IEEE Workshop on Machine-to-Machine Communications '12*, Anaheim, USA, Dec. 2012.

[21] 3GPP TS 36.104. Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) Radio Transmission and Reception (Release 10), Apr. 2013.

[22] A. Goldsmith, "Wireless Communications," *Cambridge University Press*, 2005.

[23] K.L. Baum, T.A. Kostas, P.J. Sartori, and B.K. Classon, "Performance Characteristics of Cellular Systems with Different Link Adaptation Strategies," *IEEE Trans. Veh. Technol.*, vol. 52, no. 6, pp. 1497-1507, Nov. 2003.

[24] D. Lopez-Peres, A. Ladanyi, A. Juttner, H. Rivano, and J. Zhang, "Optimization Method for the Joint Allocation of Modulation Schemes, Coding Rates, Resource Blocks and Power in Self-Organizing LTE Networks," *in Proc. IEEE INFOCOM '11*, Shanghai, China, Apr. 2011.

[25] Y. Wang, J. Ostermann, Y.Q. Zhang, "Video Processing and Communications," *Prentice Hall*, 2001.

[26] M.C. Necker, "Interference Coordination in Cellular OFDMA Networks," *IEEE Netw.*, vol. 22, no. 6, pp. 12-19, Dec. 2008.

[27] G. Song, and Y. Li, "Cross-Layer Optimization for OFDM Wireless Networks-Part II: Algorithm Development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625-634, Mar. 2005.

[28] T. Stockhammer, "Dynamic Adaptive Streaming over HTTP –: Standards and Design Principles," *in Proc. ACM MMsys '11*, San Jose, USA, Feb. 2011.

[29] A. Zambelli, "IIS Smooth Streaming Technical Overview," Microsoft Corporation, 2009.

[30] L.D. Cicco, S. Mascolo, and V. Palmisano, "Feedback Control for Adaptive Live Video Streaming," *in Proc. ACM MMsys '11*, San Jose, USA, Feb. 2011.

[31] H. Koumaras, C.-H. Lin, C.-K. Shieh, A. Kourtis, "A framework for end-to-end video quality prediction of MPEG video," *J. Vis Commun. Image R.*, vol. 21, no. 2, pp. 139-154, Feb. 2010.

[32] K. ur Rehman, O. Issa, F. Speranza, and T.H. Falk, "Quality-of-Experience Perception for Video Streaming Services: Preliminary Subjective and Objective Results," *in Proc. Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC '12)*, Hollywood, USA, Dec. 2012.

[33] X. Zhang, Y. Xu, H. Hu, Y. Liu, Z. Guo, and Y. Wang, "Profiling Skype Video Calls: Rate Control and Video Quality," *in Proc. IEEE INFOCOM '12*, Orlando, USA, Mar. 2012.

[34] S.H. Ali, and V.C.M. Leung, "Dynamic Frequency Allocation in Fractional Frequency Reused OFDMA Networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4286-4295, Aug. 2009.

[35] D. Kivanc, G. Li, and H. Liu, "Computationally Efficient Bandwidth Allocation and Power Control for OFDMA," *IEEE Trans. Wireless Commun.*, vol .2, no. 6, pp. 1150-1158, Nov. 2003.

[36] Z. Han, Z. Ji, and K.J.R. Liu, "Fair Multiuser Channel Allocation for OFDMA Networks Using Nash Bargaining Solutions and Coalitions," *IEEE Trans. Commun.*, vol. 53, no. 8, pp. 1366-1376, Aug. 2005.

Xiaoqiang Ma (S'11) received the BEng degree from Huazhong University of Science and Technology, Wuhan, China, in 2010, and the M.Sc. degree from Simon Fraser University, Canada, in 2012. He is now a Ph.D. student from School of Computing Science, Simon Fraser University, British Columbia, Canada. His areas of interest are wireless networks, social networks, and cloud computing.

Jiangchuan Liu (S'01-M'03-SM'08) is a Full Professor in the School of Computing Science, Simon Fraser University, British Columbia, Canada, and an NSERC E.W.R. Steacie Memorial Fellow. He is an EMC- Endowed Visiting Chair Professor of Tsinghua University, Beijing, China (2013-2016). From 2003 to 2004, he was an Assistant Professor at The Chinese University of Hong Kong.

He received the BEng degree (cum laude) from Tsinghua University, Beijing, China, in 1999, and the PhD degree from The Hong Kong University of Science and Technology in 2003, both in computer science. He is a co-recipient of the inaugural Test of Time Paper Award of IEEE INFOCOM (2015), ACM TOMCCAP Nicolas D. Georganas Best Paper Award (2013), ACM Multimedia Best Paper Award (2012), IEEE Globecom Best Paper Award (2011), and IEEE Communications Society Best Paper Award on Multimedia Communications (2009). His students received the Best Student Paper Award of IEEE/ACM IWQoS twice (2008 and 2012).

His research interests include multimedia systems and networks, cloud computing, social networking, online gaming, big data computing, wireless sensor networks, and peer-to-peer and overlay networks. He has served on the editorial boards of IEEE Transactions on Big Data, Transactions on Multimedia, IEEE Communications Surveys and Tutorials, IEEE Access, IEEE Internet of Things Journal, Elsevier Computer Communications, and Wiley Wireless Communications and Mobile Computing. He is the Steering Committee Chair of IEEE/ACM IWQoS from 2015 to 2017.

Hongbo Jiang (M'08-SM'15) received the B.S. and M.S. degrees from Huazhong University of Science and Technology, China. He received his Ph.D. from Case Western Reserve University in 2008. After that he joined the faculty of Huazhong University of Science and Technology, where he is now a full professor. His research concerns computer networking, especially algorithms and protocols for wireless and mobile networks. He is a senior member of the IEEE.