Spectrum Sharing and Scheduling in D2D-Enabled Dense Cellular Networks

Subhashini Krishnasamy and Sanjay Shakkottai
Department of Electrical & Computer Engineering
The University of Texas at Austin, USA
Email: subhashini.kb@utexas.edu, shakkott@mail.utexas.edu

***Abstract*—We study device-to-device (D2D) enabled hierarchical cellular networks consisting of a macro base station (BS), a dense network of access nodes (ANs) and mobile users, where spectrum is shared between cellular traffic and D2D traffic. Further, (the receivers of) mobile users dynamically time-share between the cellular and D2D networks. We develop algorithms for channel allocation and mobile-user receiver mode selection (choosing which network to participate in) with the objectives of minimizing delay for cellular traffic, and capacity maximization for D2D traffic. Our proposed solution takes advantage of the unique features offered by large and densified cellular networks such as multi-point connectivity, channel diversity, spatial reuse and load distribution.**

Given a BS-to-mobile delay requirement of $d + 1$ time-slots, we show that by appropriately scheduling channels and receiver modes, we can (with exponentially high probability) guarantee that cellular traffic reaches its intended destination within d time-slots. By leveraging spatial channel reuse, we show that this is achieved by utilizing a vanishingly small fraction of the available *spatial* capacity. Further, in the presence of delay-constrained cellular traffic, our scheduling algorithm guarantees D2D traffic can achieve rates within a $(1 - \frac{1}{d})$ factor of the corresponding achievable rates *without* cellular traffic.

***Index Terms*—D2D, resource allocation, channel scheduling, receiver mode, network densification**

I. Introduction

With the rise in demand for data services, there have been ongoing efforts to incorporate device-to-device (D2D) enabled devices in the cellular architecture. This combination of D2D communication and cellular infrastructure can potentially achieve high data rates through efficient spectrum utilization while availing the benefits offered by the cellular infrastructure for synchronization, peer discovery and other control purposes. D2D technology is especially anticipated to support the evolving demand for proximity based services which require peer-to-peer communication between nearby devices [1].

Although introducing D2D technology into the cellular network promises remarkable capacity gains and support for novel commercial applications, it brings with it new challenges in the design of network architecture and resource allocation policies. With D2D traffic sharing channel resources with conventional cellular traffic, a poorly designed architecture could potentially disrupt the long-standing performance established in cellular networks. One of the outstanding issues in this context is that of spectrum sharing for cellular and D2D traffic. While statically partitioning spectrum between cellular and D2D communication is a simple method to manage interference, using the entire spectrum on a shared basis through dynamic frequency allocation can substantially increase spectral efficiency. The second scheduling difficulty arises because each mobile node receives data either from the base station (cellular mode) or from a peer node (D2D mode). Thus, for each time-slot, one needs to decide if a mobile user is going to be part of the cellular network or part of the D2D network – we term this scheduling decision as mobile user *receiver mode* selection.

In this paper, we consider the above problems of resource allocation (spectrum sharing and receiver mode selection) in a densfied cellular network with a single base station (macro cell) and multiple access nodes (small cells). Specifically, we consider an OFDMA-based network with the base station (BS), access nodes (ANs) and user devices sharing the same spectrum resources. Users act as destination for both cellular traffic and D2D traffic with the cellular flows having latency constraints. In this setting, we propose allocation policies suitable for dense networks with large number of small cells and users and the total bandwidth scaling with number of users. Our framework also incorporates user mobility which could play an important role in a densified network.

A. Related Work

There is a rich history of research on spatial link scheduling for wireless networks. These include maximal matching algorithms [2], [3], [4], contention based scheduling algorithms [5], [6] for different interference models such as graph-based and SIR-based models. For a detailed discussion on various scheduling algorithms under different channel models, please refer to [7].

More recently, the problem of resource allocation has been studied from the perspective of D2D-enabled cellular networks. Various models and design approaches have been adopted. [8], [9] model the D2D network using stochastic geometry and consider the problem of resource optimization from the PHY layer perspective. In [10], the authors model the problem of spectrum allocation as an optimization problem with the objective of maximizing spatial reuse. [11] proposes a centralized graph theoretic approach for channel allocation and a distributed game-theoretic approach for power allocation. [12] recommends dividing the cell into disjoint spatial zones and dedicating a fraction exclusively for D2D communication.

A majority of these solutions prescribe opportunistic access of resources for D2D traffic through spatial reuse. However, prior studies on resource allocation for D2D traffic do not consider the densified network setting with multiple small cells. Moreover, none investigate the performance of allocation algorithms with increasing network size. However, there has been some recent research on resource allocation algorithms for conventional cellular traffic in densified networks [13], [14], [15].

Our model of the densified network closely follows that proposed in [15], where it is shown that features like multipoint connectivity (user simultaneously connected to multiple ANs) and AN-to-AN communication can be exploited to ensure throughput optimality and good delay performance for mobile users.

B. Contributions

In this work, we propose a resource allocation framework for a D2D-enabled cellular network in which users have two types of downlink traffic – conventional cellular traffic and single-hop D2D traffic. We consider dynamic sharing of the spectrum among the D2D links, ANs and the BS. Analytical results show some interesting performance benefits of the proposed solution in the dense setting.

Low Spectrum Usage at Access Nodes: We show that with some minimal coordination between the BS and the ANs for channel usage, a small fraction of the spectrum ($o(n)$ out of a total n channels) is sufficient for communication at AN level (both among the ANs and from AN to user) to ensure throughput optimality and good delay performance for cellular flows. Specifically, we show that if the cellular traffic has a delay constraint of $d+1$ time slots, then all packet arrivals for all users in any time slot reach their destination within $d+1$ time with exponentially high probability. In addition, we show a positive rate function for the buffer overflow event for queues at the BS and the ANs. This is possible due to the diversity of channel across transmitter-receiver pairs, high degree of spatial reuse of the spectrum and load distribution across the ANs in densified networks.

High Spectral Efficiency for D2D: We show that reuse of the spectrum by the BS and D2D links combined with load distribution enables usage of a large fraction of the spectrum ($n - o(n)$ out of n channels) to D2D links with high probability.

Receiver Access for D2D: We propose a scheduling policy which seeks to maximize the receiver time obtained by the D2D flows while giving sufficient priority to delay sensitive cellular flows. We show that the D2D flow for the user can be allocated all except $\frac{1}{d}$ fraction of the time slots. Let $\mathcal{C}_0(T, \tilde{n})$ denote the capacity region in T time slots for D2D flows in a network without cellular traffic and with $\tilde{n}$ channels. We show that the capacity region $\mathcal{C}_d(T, n)$ with n channels in presence of cellular traffic includes $\left(1 - \frac{1}{d}\right) \mathcal{C}_0(T, n - o(n)) - o(1)$.

II. System Model

We consider an OFDMA-based, D2D-enabled cellular network. As in [15], the cellular network has a hierarchical structure with base-station providing coverage at the macro level and a dense deployment of access-nodes (small cells) providing short-range coverage. The network has a single base-station, n users, and M access nodes with $M \sim poly(n)$. The system has n OFDM channels for communication.

Mobile users in the network act as destination for two types of downlink traffic – conventional cellular traffic and D2D traffic. Cellular traffic for all users arrive at the BS which then delivers them to the users either directly or through ANs acting as relays. On the other hand, D2D communication occurs with some transmitter device that is close to the user at all times. We assume that this communication is single hop and that the D2D link corresponding to each user moves along with the user. For each user, cellular and D2D arrivals are assumed to be independent of each other. Cellular traffic is delay sensitive – a packet that arrives at the BS at time t should reach the user at the beginning of time $t + d + 1$. For simplicity, we assume that the delay requirement is the same for all users. Each packet is to be delivered to the destination irrespective of whether the delay requirement is met or not. All users' receivers are subject to orthogonal access, i.e., at any time a user's device can act as a receiver for either cellular traffic or D2D traffic but not both.

Other details about the model are given below.

A. AN-AN Connectivity

An AN can communicate with other nearby ANs and forward users' packets to them. Let $\mathcal{V}_m$ be the set of *neighboring* ANs that AN m can reliably communicate with.

B. AN-AN Interference

Simultaneous transmission by multiple nearby ANs on the same channel could cause interference at the intended receivers. We adopt a simple graph based interference model to characterize interference relationship between various nodes. We assume that receivers are likely to be close to their corresponding transmitters, and thereby model interference using an undirected graph on a vertex set consisting of all ANs as transmitters. An edge between two ANs implies that simultaneous transmission on the same channel by the two ANs could lead to failure in delivery of the packets at the intended receivers. Analogously, absence of an edge indicates successful reception at both the receivers. Under this interference model, let $\mathcal{I}_m$ denote the set of ANs that interfere with AN m. We make the following assumption that this interference set can be at most polynomial in the network size. This bound on the interference set enables *spatial reuse* of frequency spectrum.

M.1 There exists a constant $\nu_1 \in (0, 1)$ such that for any $m \in [M]$, $|\mathcal{I}_m| \leq n^{\nu_1}$.

M.2 In addition, we also assume that the interference set contains the neighborhood set, i.e., $\mathcal{I}_m \supseteq \mathcal{V}_m \; \forall m \in [M]$.

C. User Mobility

We use a general model for user mobility which allows changes in users' association with ANs in consecutive time slots. Multipoint connectivity enables users to connect to multiple ANs simultaneously. For any user u, we denote by $\mathcal{M}_u(t)$ the set of ANs that the user is connected to in time slot t.

M.3 We assume that every user is connected to $\log n$ ANs, i.e., at any time t, $|\mathcal{M}_u(t)| \geq \log n$.

M.4 In addition, $(\mathcal{M}_u(t))_{u\in[n]}$ is an irreducible aperiodic positive recurrent DTMC.

We also make the following assumptions about users' mobility pattern. Let $\mathcal{U}_m(t)$ denote the set of users connected to AN m at time t. In a dense network with a large number of ANs, it is unlikely that a very large number of users are connected to any single AN. Specifically, we assume that, with exponentially high probability, the number of users that are connected to each AN in any time slot is no more than a polynomial in n. We refer to this feature of dense networks as *load distribution*.

M.5 There exist constants $\nu_2 \in (0, 1 - 2\nu_1)$ and $c_1 > 0$ such that for any $t \in \mathbb{N}$,

$$\mathbb{P}\left[\max_{m\in[M]} |\mathcal{U}_m(t)| > n^{\nu_2}\right] \leq e^{-c_1 n}.$$

Multipoint connectivity ensures that users are connected to overlapping sets of ANs in consecutive time slots. The following property characterizes this behavior:

M.6 For any user u, constant $\epsilon > 0$, and $t \in \mathbb{N}$ and for any set $\mathcal{M} \subseteq \mathcal{M}_u(t)$, $|\mathcal{M}| \geq \epsilon \log n$ implies $\mathcal{M} \cap \mathcal{M}_u(t+1) \neq \varnothing$.

In other words, we assume that for any set of $\Omega(\log n)$ ANs that a user is connected in a particular time slot, the user is connected to at least one of these ANs in the next time slot.

We also make the following assumption which relates user mobility with the physical proximity of an AN's neighbors.

M.7 For any user u and $t \in \mathbb{N}$ and for any $m \in \mathcal{M}_u(t)$, $\mathcal{V}_m \cap \mathcal{M}_u(t+1) \neq \varnothing$.

This assumption states that a user connected to an AN in a particular time slot stays connected to at least one of the neighbors of the AN in the next time slot.

D. User Receiver Modes

At any time, users can receive either cellular traffic or D2D traffic. The mode of users' receivers is denoted using the terms *Cell* mode and *D2D* mode to indicate reception of cellular traffic and D2D traffic respectively.

E. BS-AN-D2D Interference

A D2D link (Tx_2 to Rx_2) is much shorter than a direct link from the BS to its intended receiver (Rx_1 see Figure 1). Thus, as is well known, simultaneous transmissions between the BS to Rx_1 and Tx_2 to Rx_2 is feasible as long as the SIR requirements at both Rx_1 and Rx_2 are simultaneously satisfied.[1] In this paper, instead of specifying the physical interference model, we simply abstract this phenomenon via exclusion sets. We assume that if the BS is transmitting to an AN on a particular channel, then no D2D link which lies within the AN's footprint can use this channel. Let $\mathcal{S}_u(t)$ be the set of ANs with which user u's D2D link can interfere.

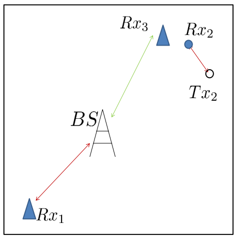

Fig. 1. Interference Model: D2D links can reuse the channels used by the BS if they do not interfere with BS transmission. Here, the D2D link $Tx_2 - Rx_2$ does not interfere with BS-Rx_1, but does interfere with BS-Rx_3.

M.8 We assume that for all $u \in [n]$ and $t \in \mathbb{N}$, $|\mathcal{S}_u(t)| \leq n^{\nu_1}$.[2] D2D flows for User u can be scheduled on all channels used by the BS other than those which are used to transmit to ANs in $\mathcal{S}_u(t)$.

Interference between different D2D transmissions depends on the spatial distribution of users in the network. For simplicity, we assume that this spatial distribution is captured by the user mobility process and that the interference relationship between different D2D flows is determined by $(\mathcal{M}_u(t))_{u\in[n]}$.

M.9 At any time t, interference relationship between different users is a function of $(\mathcal{M}_u(t))_{u\in[n]}$.

F. Channel Statistics

We make the following assumptions regarding the distribution of channel rates across different transmitter-receiver pairs.

M.10 Rates of all channels are independent of each other across time and transmitter-receiver pairs. They are also identically distributed across time slots.

M.11 The rates of all channels are upper bounded by a constant $\overline{R} \geq 1$.[3] Moreover, for any $m, m_1, m_2 \in [M]$, $j \in [n]$, $t \in \mathbb{N}$, $u \in \mathcal{U}_m(t)$, the channel rates $R^{BS-AN}_{0,m,j}(t)$, $R^{AN-AN}_{m_1,m_2,j}(t)$, $R^{AN-U}_{m,u,j}(t)$ have a probability mass on $\overline{R}$ of at least $q > 0$. The rates from the BS to the users could be lower due to hardware constraints and the location of the users with respect to the BS. Let $\overline{R}_0 \in (1, \overline{R})$ be the maximum rate possible for BS to user links. For all $u \in [n]$, $j \in [n]$, $t \in \mathbb{N}$, the channel rates $R^{BS-U}_{0,u,j}(t)$ have a probability mass on $\overline{R}_0$ of at least $q > 0$.

M.12 The channel rates for D2D links are independent and identically distributed across users, channels and time slots.

G. Arrival Statistics

We model arrivals for cellular traffic as a Markov process that satisfies the following assumptions.

[1] Indeed, recent architectures such as FlashLinQ [16], [17] focus on practical signaling mechanisms in a D2D setting to exploit this phenomenon for concurrently packing "long" and "short" links.

[2] We use the same constant ν_1 as in M.1 to indicate that the interference set for D2D transmissions is not likely to be greater than that for AN transmissions.

[3] All values of arrival and channel rates are specified in units of packets per time slot.

TABLE I
NOTATION

Symbol	Description
n	Number of users Number of OFDM Channels
M	Number of ANs
$d+1$	Delay requirement of cellular packets
$\mathcal{V}_m$	Neighborhood set of AN m
$\mathcal{I}_m$	Interference set of AN m
$\mathcal{M}_u(t)$	AN association set for user u at time t
$\mathcal{U}_m(t)$	User association set for AN m at time t
$R^{BS-U}_{0,u,j}(t)$	Channel j's rate from the BS to user u at time t
$R^{BS-AN}_{0,m,j}(t)$	Channel j's rate from the BS to AN m at time t
$R^{AN-AN}_{m_1,m_2,j}(t)$	Channel j's rate from AN m_1 to AN m_2 at time t
$R^{AN-U}_{m,u,j}(t)$	Channel j's rate from AN m to user u at time t
$\overline{R}$	Maximum service rate for any channel
q	Probability mass of channels on $\overline{R}$
$A_u(t)$	Arrival for user u at the BS at time t
ρ	Load of the arrival vector $\boldsymbol{A}(t)$
$\chi_u(t)$	Indicates mode of user u at time t
$Q_u(t)$	Queue of user u at the BS at time t
$A^d_u(t)$	Packets of user u that arrived at the BS at $t-d-1$ but not delivered by time t
$Q_{u,m}(t)$	Queue of user u at AN m at time t
$\hat{\mathcal{N}}$ & $\hat{n}$	Set & number of channels used by the ANs
$\mathcal{J}^{AN-AN}_{m_1,m_2}(t)$	Set of channels used by the BS at time t that cannot be used for transmission from AN m_1 to AN m_2
$\mathcal{J}^{AN-U}_{m,u}(t)$	Set of channels used by the BS at time t that cannot be used for transmission from AN m to user u
$\mathcal{J}^{BS}_u(t)$	Set of channels used by the BS at time t that cannot be used for D2D transmission by user u
$\mathcal{J}^{AN}_u(t)$	Set of channels used by ANs at time t that cannot be used for D2D transmission by user u

M.13 The arrival vector for cellular traffic $\boldsymbol{A}(t) = (A_u(t))_{u\in[n]}$ is a positive recurrent DTMC that satisfies the following condition: there exists a $\rho > 0$ and a positive function $f_2 : (0,\infty) \to (0,\infty)$ such that for any $\epsilon > 0$ and $t \in \mathbb{N}$,

$$\mathbb{P}\left[\frac{1}{n}\sum_{u=1}^{n}\frac{A_u(t)}{\overline{R}} > \rho + \epsilon\right] \leq e^{-f_2(\epsilon)n}$$

We refer to the value ρ as the *load* of the arrival vector $\boldsymbol{A}(t)$. It was shown in [15] that $\rho < 1$ is a necessary and sufficient condition for throughput stability of the cellular network queues (even without D2D traffic). Therefore, we assume that ρ satisfies the stability condition $\rho < 1$.

M.14 For all users, arrivals in any given time slot are upper bounded by n^{ν_3}, for some ν_3 such that $2\nu_1+\nu_2+\nu_3 < 1$.

M.15 Arrivals, channel rates and user mobility are independent of each other.

III. Resource Allocation Framework

We propose a scheduling policy that allocates two types of resources to cellular and D2D flows – (a) Channel Access, and (b) Receiver Access. D2D traffic is transmitted in an opportunistic fashion by accessing only those system resources unused by conventional cellular traffic. At the same time, the policy seeks to optimize the resources allotted to cellular traffic while satisfying the delay constraints so that a large fraction of the resources can be utilized for D2D transmissions.

Channel Access: We propose a hierarchical architecture for channel allocation in which priority is given to the BS, ANs and D2D transmitters in that order. Specifically, the BS is first allowed to schedule its transmissions on any of the n available channels. The ANs then schedule their transmissions on channels that do not interfere with BS or other AN transmissions. Finally, D2D transmissions are restricted to those channels which do not interfere with all BS and AN links using the same channels. We do not specify the exact scheduling policy to be used for channel allocation within ANs or within D2D transmissions. Instead we consider the class of all policies that allocate channels such that transmissions do not interfere with each other.

Receiver Access: For scheduling receiver access at the user devices, again, cellular traffic is prioritized over D2D traffic. A user's receiver switches to Cell mode whenever the BS or some AN has some cellular packet scheduled for delivery to the user.

The following sections describe the resource allocation algorithm in detail.

In every time slot, events occur in the following sequence 1) Arrival of packets and association of users with ANs, 2) Resource scheduling and signaling at the MAC level, 3) Packet transmission, 4) Queue update.

A. Signaling Scheme

We now describe the signaling mechanism at the MAC level that facilitates the implementation of the scheduling policy. The following signaling scheme is adopted at the beginning of any time slot t.

1) The BS allocates channels to transmit packets to ANs and users. This information is communicated to the ANs and users.
2) After receiving the BS channel allocation, according to the available set of channels, the ANs choose an initial set of users to transmit packets (Round I in Algorithm 2). This information is communicated to the users.
3) The users who are scheduled to receive packets from either the BS or the ANs switch to Cell mode and send this information to the BS. The BS broadcasts this information to all the ANs.
4) The ANs receive information about the user modes and augment the set of users to transmit packets by choosing more users from among those in the Cell mode through a second round of channel allocation (Round II in Algorithm 2).
5) The channel allocation decisions of the ANs is then communicated to the users in D2D mode so that they can make channel allocation decisions appropriately.

B. Resource Allocation at the BS

The number of packet arrivals for user u at time t is denoted by $A_u(t)$. The BS maintains the state of each user by a sliding window $W_u(t;t-d)$ and an "overflow" queue $Q_u(t)$ (see Figure III-B). The sliding window $W_u(t;t-d)$, which is of size $d+1$ time slots, stores the packets that arrived for the user over the last $d+1$ time slots. At time $t-1$ it contains packets arrived from $t-d-1$ to $t-1$. At the beginning of time t, all packets that arrived at $t-d-1$ and have not yet been delivered to user u are moved to the "overflow" queue Q_u. The sliding window is then shifted to the right (i.e., corresponds to the time periods $(t-d)$ to t). The new arrivals that occur at time t are now appended to the sliding window (see Figure III-B).

In summary, the sliding window $W_u(t;t-d)$ keeps track of the packet arrivals over the time-interval $[t-d,t]$ and Q_u contains those packets that arrived at the BS until (and including) time slot $t-d-1$ and have not reached their destination by time t.

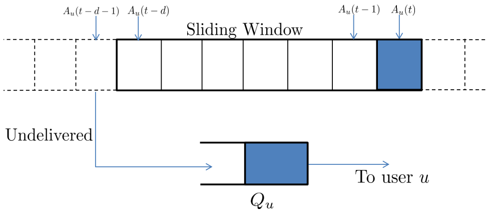

Fig. 2. State evolution at the BS: Arrivals in last $d+1$ time slots maintained by the sliding window. Current sliding window is shown by solid lines. The BS allocates channels to serve new arrivals and packets in the queue (shown in blue).

The scheduling algorithm used for channel allocation at the BS is given by Algorithm 1. At any time t, the BS sends packet arrivals at time t ($\{A_u(t)\}_{u\in[n]}$) to multiple ANs to which the user is connected. The BS transmits these packets only if they can be received by multiple connected ANs. For each channel, the user queue chosen for transmission on the channel and the set of ANs to which the user's packets are transmitted on that channel is determined by appropriate weights given in Algorithm 1. The weights are obtained through iterative update of the queue lengths similar to the SSG algorithm in [18]. If after forwarding all the current arrivals there are channels left unallocated, it allocates them to transmit packets from the queues $\{Q_u\}_{u\in[n]}$ directly to the users. Again, for each channel, the user queue that is chosen for transmission is determined by weights computed through iterative updates of queue lengths.

Let $A_u^d(t)$ denote the number of packets of user u that arrived at time $t-d-1$ but could not reach their destination by time t. Thus, the queue update equations at the BS are given by

$$Q_u(t) = \left(Q_u(t-1) + A_u^d(t) - R_{0,u,j}^{BS-U}(t)\Upsilon_{0,u,j}^{BS-U}(t)\right)^+,$$

Algorithm 1 Resource Allocation at the BS at time t

```
Initialize $A_u^I \leftarrow A_u(t)$, $Q_u^I \leftarrow Q_u(t)$ $\forall 1 \le u \le n$, $\alpha \leftarrow \frac{q}{2}$.
for $j = 1$ to $n$ do
    $$\mathcal{U} \leftarrow \left\{u \in [n] : \left|\underset{m:m\in\mathcal{M}_u(t)}{\arg\max}\ R_{0,m,j}^{BS-AN}(t)\right| \ge \alpha\,|\mathcal{M}_u(t)|\right\}$$
    if $\max_{u\in\mathcal{U},m\in\mathcal{M}_u(t)} A_u^I R_{0,m,j}^{BS-AN}(t) > 0$ then
        Forward New Arrivals to ANs
        Choose some
            $$u^* \in \underset{u\in\mathcal{U}}{\arg\max}\, A_u^I\left(\max_{m:m\in\mathcal{M}_u(t)} R_{0,m,j}^{BS-AN}(t)\right)$$
        breaking ties arbitrarily and forward new arrivals of User $u^*$ to every AN $m^* \in \underset{m:m\in\mathcal{M}_{u^*}(t)}{\arg\max}\ R_{0,m,j}^{BS-AN}(t)$ on Channel $j$.
        Update variables:
            $$A_{u^*}^I \leftarrow \left(A_{u^*}^I - R_{0,m^*,j}^{BS-AN}(t)\right)^+.$$
    else if $\max_u Q_u^I R_{0,u,j}^{BS-U}(t) > 0$ then
        Forward Old Packets to Users
        Choose some
            $$u^* \in \underset{u\in[n]}{\arg\max}\, Q_u^I R_{0,u,j}^{BS-U}(t)$$
        breaking ties in favor of the smaller user index and forward old packets directly to User $u^*$ on Channel $j$.
        Update variables:
            $$Q_{u^*}^I \leftarrow \left(Q_{u^*}^I - R_{0,u^*,j}^{BS-U}(t)\right)^+.$$
    end if
end for
```

where $\Upsilon_{0,u,j}^{BS-U}(t) = 1$ if the BS allocates channel j to directly transmit to user u and is equal to zero otherwise.

C. Resource Allocation at the ANs

Each AN maintains one queue for every user. The queue length for user u at time t at AN m is given by $Q_{u,m}(t)$. At any time, an AN stores packets of only those users who are currently connected. All packets of users who are not connected are discarded. Moreover, only packets that arrived in the last d time slots are stored i.e., at time t for any user u and AN m, $Q_{u,m}(t)$ consists of only those packets that arrived at the BS after $t-d$. Since these queues store only those packets that arrive in a period of d time slots and the arrivals in any time slot are bounded by n^{ν_3}, the queue lengths do not exceed dn^{ν_3}.

Scheduling at the ANs at any time t is given by Algorithm 2. Packets at the ANs are either transmitted directly to users or to neighboring ANs. Scheduling for transmission to users occurs in two rounds. At time t, each AN first determines the set of currently connected users who have packets that reach their delay deadline in the next time slot, i.e., packets that arrived at the BS at time slot $t-d$. Even though the selection of the users is dependent on arrivals at time $t-d$, all packets in the queue (even those that arrived after $t-d$) of the selected user are scheduled for transmission depending on channel availability. Channels are allocated to transmit packets of these users such that the they do not affect the transmissions of the BS or nearby ANs (determined by the interference sets described in Section II-B). Let $\mathcal{J}_{m,m_1}^{AN-AN}(t)$ and $\mathcal{J}_{m,u}^{AN-U}(t)$ be the set of channels used by the BS that cannot be used by AN m to transfer packets to a neighboring AN m_1 and a connected user u respectively. According to these restrictions, channels are allocated to ANs from a small set $\hat{\mathcal{N}}$ of $\hat{n}$ channels. Specifically, this set is given by $\hat{\mathcal{N}} = \{n, n-1, \ldots, n-\hat{n}+1\}$. $\hat{n}$ is chosen such that it is $\omega(n^{\nu_2+2\nu_1+\nu_3})$ but $o(n)$.

The users who are scheduled for reception from either the BS or the ANs in the first round switch their receiver to Cell Mode. For user u, $\chi_u(t)$ is equal to 1 if the user is in Cell mode and 0 otherwise. This information is communicated to all ANs. Based on this information, a second round of channel allocation is made for each AN to transmit all packets of connected users in the Cell mode. The second round of scheduling helps in transmitting as many stored packets as possible to users in the Cell mode (subject to availability of channels) so that the users can operate in the D2D mode in other time slots, thus optimizing the receiver access time for the D2D flows.

The packets of all other connected users are scheduled for transmission to all neighboring ANs. This is to ensure, in the event of user movement and change of AN associations, that the packets of the mobile user is always present at some connected AN.

Algorithm 2 Resource Allocation at the ANs at time t

```
Forward Packets to Users
Round I:
Initialize $\mathcal{U}'_m \leftarrow \{u \in \mathcal{U}_m(t) : A_u(t-d) > 0\}\ \forall m \in [M]$,
$\mathcal{J}_{u,m}^{U} \leftarrow \hat{\mathcal{N}} \setminus \mathcal{J}_{m,u}^{AN-U}(t)\ \forall u \in \mathcal{U}'_m,\ \forall m \in [M]$.
Allocate channels to $Q_{u,m}(t)$ from $\mathcal{J}_{u,m}^{U}\ \forall u \in \mathcal{U}'_m,\ \forall m \in [M]$ according to feasible set determined by the AN interference graph.
Round II:
Initialize $\mathcal{U}''_m \leftarrow \{u \in \mathcal{U}_m(t) : \chi_u(t) = 1\} \setminus \mathcal{U}'_m\ \forall m \in [M]$,
$\mathcal{J}_{u,m}^{U} \leftarrow \hat{\mathcal{N}} \setminus \mathcal{J}_{m,u}^{AN-U}(t)\ \forall u \in \mathcal{U}''_m,\ \forall m \in [M]$.
Allocate channels to $Q_{u,m}(t)$ from $\mathcal{J}_{u,m}^{U}\ \forall u \in \mathcal{U}''_m,\ \forall m \in [M]$ according to feasible set determined by the AN interference graph and the channels allocated in Round I.
Forward Packets to Neighboring ANs
Initialize $Q_{m,m_1}^{I,AN} \leftarrow \sum_{u \in \mathcal{U}_m(t) \setminus (\mathcal{U}'_m \cup \mathcal{U}''_m)} Q_{u,m}(t)$,
$\mathcal{J}_{m,m_1}^{AN} \leftarrow \hat{\mathcal{N}} \setminus \mathcal{J}_{m,m_1}^{AN-AN}(t)\ \forall m_1 \in \mathcal{V}_m$.
Allocate channels to $Q_{m,m_1}^{I,AN}$ from $\mathcal{J}_{m,m_1}^{AN}\ \forall m, m_1 \in [M]$ according to feasible set determined by the AN interference graph and the channels allocated for user transmissions in Round I and II.
```

Remark 1. *Although, in this paper, we present a centralized channel allocation algorithm for the ANs, we note that it is possible to design distributed channel allocation algorithms in the same framework and provide similar performance guarantees. One approach to designing such distributed channel allocation is through graph coloring algorithms. A comprehensive summary of distributed graph coloring algorithms can be found in [19].*

D. Resource Allocation for D2D Communication

We do not specify the exact scheduling policy to be used for D2D transmissions. Instead, we consider the class of scheduling policies that allocate channels for D2D transmissions such that they do not interfere with BS and AN transmissions. Let $\mathcal{J}_u^{BS}(t)$ and $\mathcal{J}_u^{AN}(t)$ denote the set of channels that cannot be used by a user u in D2D mode at time t due to their usage by the BS and the ANs respectively. Any channel allotted to user u must be from the set $[n] \setminus \left(\mathcal{J}_u^{BS}(t) \cup \mathcal{J}_u^{AN}(t)\right)$.

IV. Theoretical Guarantees

We now present analytical results that give asymptotic performance guarantees for the proposed resource allocation framework as the network size n grows large. Detailed proofs for the theorems presented in this section can be found in [20]. We give delay guarantees for cellular traffic when the arrivals are within the stability region $\rho < 1$. As mentioned in Section III-C, the size of the channel set used by the ANs is given by $\left|\hat{\mathcal{N}}\right| = \hat{n}$.

A. Delay Guarantee for Cellular Traffic

The following result gives a probabilistic bound on the total number of packets not delivered to their destination within $d+1$ time slots. Specifically, it shows that all packets that arrive in a particular time slot are delivered to the destination through the ANs within $d+1$ time slots with exponentially high probability.

Theorem 1. *There exists a constant $c_3 > 0$ such that, for n large enough and for any $t \in \mathbb{N}$, $\mathbb{P}\left[\sum_{u=1}^{n} A_u^d(t) > 0\right] \leq e^{-c_3 \hat{n}}$.*

Proof Outline. The proof mainly relies on concentration results with scaling network size. Consider arrivals to the BS at time $t-d-1$. We can show (see [20] for details) that all these arrivals can be forwarded by the BS to ANs connected to the users with high probability using a constant fraction of the available channels (depending on the arrival load ρ). Note that the BS forwards packets only if they can be received by multiple connected ANs. This ensures (by M.6) that even if the destination user moves in the next time slot, it is still connected to at least one of the ANs that received its packets.

In the subsequent $d-1$ time slots, the ANs connected to the user store the user's packets and relay them to neighboring ANs. In the event of change in user-AN association due to mobility, the relaying ensures that some AN connected to the user always has the users' packets (by M.7). This is possible only if the ANs are allocated sufficient number of channels to relay the packets to multiple ANs. We show ([20, Lemma 7]) that this can be achieved with high probability as the network size increases. Since the BS requires only a constant fraction of the channels, with increasing number of channels, all the $\hat{n} = o(n)$ channels can be used by the ANs for transmission. In addition, due to high spatial reuse among the ANs (M.1) and load distribution across the ANs (M.5) with increasing network size, it is possible to allocate channels to serve packets at all ANs without causing interference at nearby ANs.

Finally, at time $t-1$, the packets are forwarded to the destination users and reach the users at the beginning of time slot t. By the same reasoning as above, sufficient number of channels can be allocated to forward all scheduled packets to users. Thus, with high probability, all packets reach the user through the AN relays within $d+1$ time slots without the BS requiring to serve them directly to the user. □

B. Stability and Queue Length of BS Queues

As in [15], we give stability and maximum queue length guarantees for the the buffers at the BS. The process $\{Z(t)\}_{t\in\mathbb{N}}$ defined as

$$Z(t) := \big(A_u(t), Q_u(t-1), (Q_{u,m}(t-1))_{m\in[M]}, \mathcal{M}_u(t)\big)_{u\in[n]} \tag{1}$$

is an irreducible, aperiodic DTMC. This follows from assumptions M.4, M.10, M.13 and the fact that the scheduling policy at any time depends only on current queue lengths, arrivals, user-AN associations and channel rates. By stability, we mean positive recurrence of this DTMC. In addition to stability, we show a positive rate function for the maximum queue length at the BS. Since the queues at the ANs store only packets that arrived in d time slots, $Q_{u,m}(t) \le dn^{\nu_3}$ for all u, m, t.

Theorem 2. *The process $\{Z(t)\}_{t\in\mathbb{N}}$ is positive recurrent. Let $\mathbb{P}_\pi$ denote the probability measure under stationarity. Then for any integer $b \ge 0$,*

$$c_4 := \liminf_{n\to\infty} \frac{-1}{(b+1)\hat{n}} \log\left(\mathbb{P}_\pi\left[\max_{i\in[n]} Q_i(0) > b\right]\right) > 0.$$

The proof of this theorem is fairly straightforward given the result from Theorem 1 and omitted for the sake of brevity.

C. Achievable Rates for D2D Traffic

The scheduling policy allocates resources to the D2D traffic on an opportunistic basis making use of resources that are not used by the cellular traffic. Therefore, it is useful to understand the best rates that can be guaranteed for the D2D flows. To evaluate the performance of the proposed scheduling policy with respect to the D2D flows, we compare the *achievable rates* and *capacity region* for the D2D flows in two scenarios:

1) When the network has no cellular traffic,
2) When the network has cellular traffic with delay constraints $d+1$ which is scheduled according to the proposed resource allocation policy.

Achievable rates and capacity are defined in terms of average service rates offered to users under stationarity. We note that stationarity here is well defined. Since the D2D channel rate process is i.i.d. across time (M.12), the system evolution can be represented by the user mobility process $\{(\mathcal{M}_u(t))_{u\in[n]}\}$ in the absence of cellular traffic. In the presence of cellular traffic, it can be represented using the process $\{Z(t)\}$ defined by (1). Both the processes are positive recurrent DTMCs. We use the notation $\mathbb{E}_\pi[\cdot]$ and $\mathbb{E}_{\pi_0}[\cdot]$ to denote the mean under stationarity in the presence and absence of cellular traffic respectively.

Definition 3 (D2D Capacity Region without Cellular Traffic)**.** *In the absence of cellular traffic, the triplet $(\mathbf{R}, T, \tilde{n}) \in \mathbb{R}^n \times \mathbb{N} \times \mathbb{N}$ is said to be achievable if, with a total of $\tilde{n}$ channels in the system, there exists a D2D scheduling policy (possibly randomized) such that the mean D2D rate offered to the users by the scheduling policy averaged across T time slots is at least R, i.e., if $R_u^D(t)$ is the rate offered to user u by the scheduling policy and $\mathbf{R}^D(t) = (R_u^D(t))_{u\in[n]}$, then*

$$\mathbb{E}_{\pi_0}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{R}^D(t)\right] \ge \mathbf{R}.$$

The D2D capacity region without cellular traffic $\mathcal{C}_0(T, \tilde{n})$ is the closure of the set of all $\mathbf{R} \in \mathbb{R}^n$ such that triplet $(\mathbf{R}, T, \tilde{n})$ is achievable in the absence of cellular traffic.

Definition 4 (D2D Capacity Region with Cellular Traffic)**.** *In the presence of cellular traffic with delay constraints $d+1$ and scheduled according to Algorithms 1 and 2, the triplet $(\mathbf{R}, T, \tilde{n}) \in \mathbb{R}^n \times \mathbb{N} \times \mathbb{N}$ is said to be achievable if, with a total of $\tilde{n}$ channels in the system, there exists a D2D scheduling policy (possibly randomized) in the proposed resource allocation framework such that the mean D2D rate offered to the users by the scheduling policy averaged across T time slots is at least $\mathbf{R}$, i.e., if $R_u^D(t)$ is the rate offered to user u by the scheduling policy and $\mathbf{R}^D(t) = (R_u^D(t))_{u\in[n]}$, then*

$$\mathbb{E}_{\pi}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{R}^D(t)\right] \ge \mathbf{R}.$$

The D2D capacity region with cellular traffic $\mathcal{C}_d(T, \tilde{n})$ is the closure of the set of all $\mathbf{R} \in \mathbb{R}^n$ such that triplet $(\mathbf{R}, T, \tilde{n})$ is achievable in the presence of cellular traffic with delay constraints $d+1$.

Note that in the presence of cellular traffic, any scheduling policy in the proposed framework can allocate to D2D traffic only those resources not utilized by the BS and the ANs for cellular traffic. Cellular flows compete with D2D flows for resources in two respects – channel access and receiver access. To evaluate the D2D performance that can be achieved under the proposed framework from the standpoint of channel access alone, we present the following lemma. It shows that even with shared spectrum, $n - o(n)$ channels are accessible to all users in the D2D mode.

Lemma 5. *Suppose that the users had separate receivers for cellular and D2D traffic. Then, for any $\epsilon > 0$, for n large enough and $\forall\ d, T \in \mathbb{N}$,*

$$\mathcal{C}_d(T, n) \supseteq \mathcal{C}_0(T, n(1-\epsilon)) - \epsilon\mathbf{1}.$$

Proof Outline. We bound the number of channels that cannot be scheduled for D2D flow of any user $u \in [n]$. Since the ANs use only $o(n)$ channels, for large enough n, it is sufficient to consider the channels used by the BS that cannot be used for D2D transmission. We show that this number is not very large because of load distribution (M.5) and spatial reuse (M.8) . The result then follows since the D2D channel rates are i.i.d. across different channels (M.12). □

Thus, if the receivers were capable of simultaneously receiving both types of traffic, the D2D capacity region with cellular traffic is close to the capacity region with D2D traffic alone. The following theorem gives an evaluation of the proposed solution with respect to both spectrum and receiver time allocation. It shows that, for n large enough, the D2D capacity with cellular traffic is close to $1 - \frac{1}{d}$ factor of the capacity without cellular traffic.

Theorem 6. *Let $T = O(n^{\nu_4})$ for some constant $\nu_4 \in (0,1)$ and $\epsilon > 0$. For any $d \in \mathbb{N}$, for n large enough,*

$$\mathcal{C}_d(T,n) \supseteq \left(1 - \frac{1}{d}\right) \mathcal{C}_0(T, n(1-\epsilon)) - \epsilon \mathbf{1}.$$

Proof Outline. Given the result in Lemma 5, the crux of the proof lies in showing that the "loss" in D2D capacity due to reception of cellular traffic is at most $\frac{1}{d}\mathcal{C}_0(T, n(1-\epsilon)) + \epsilon\mathbf{1}$. We show that with high probability, a user switches to Cell mode at most once in d time slots. If the times at which these events occur are independent of the user mobility, the result easily follows by M.9 since the rates obtained by the users in the D2D mode depends only on the channel rates and the interference between users. For each user, we construct a random sequence of time slots which are determined by the arrival process of the cellular traffic for that user and show that, in time interval $[1, T]$ this sequence of time slots is equal to the times at which the user switches to Cell mode with high probability. By M.15, this sequence is independent of user mobility and D2D channel rates. Thus the rate obtained is, with high probability, equal to the rate obtained if the actual time slots were equal to these random times. Bounds on the maximum possible rate enables us to extend this result to the mean average rate. □

V. Conclusion

In this paper, we studied the joint problem of channel allocation and receiver time allocation to device-to-device (D2D) flows and conventional cellular flows in a dense network setting. Our analysis shows that it is possible to offer high rates to D2D flows even while prioritizing delay sensitive cellular flows. The key to achieving this performance comes from two factors. One is reuse of spectrum used by the BS for cellular flows. The other is appropriate scheduling of receiver access times. This can be achieved by intelligent packing of cellular transmissions to users across time slots so as to maximize their receiver time for D2D flows.

Acknowledgements

This work is partially supported by NSF Grants CNS-1161868 and CNS-1343383, and the US DoT supported D-STOP Tier 1 University Transportation Center.

References

[1] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3gpp device-to-device proximity services," *Communications Magazine, IEEE*, vol. 52, no. 4, pp. 40–48, 2014.

[2] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-layer rate control in wireless networks," in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, vol. 3. IEEE, 2005, pp. 1804–1814.

[3] P. Chaporkar, K. Kar, X. Luo, and S. Sarkar, "Throughput and fairness guarantees through maximal scheduling in wireless networks," *Information Theory, IEEE Transactions on*, vol. 54, no. 2, pp. 572–594, 2008.

[4] L. X. Bui, S. Sanghavi, and R. Srikant, "Distributed link scheduling with constant overhead," *Networking, IEEE/ACM Transactions on*, vol. 17, no. 5, pp. 1467–1480, 2009.

[5] Y. Yi, G. De Veciana, and S. Shakkottai, "On optimal mac scheduling with physical interference," in *INFOCOM 2007. 26th IEEE International Conference on Computer Communications*. IEEE, 2007, pp. 294–302.

[6] L. Jiang and J. Walrand, "A distributed csma algorithm for throughput and utility maximization in wireless networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 18, no. 3, pp. 960–972, 2010.

[7] C. Joo, X. Lin, J. Ryu, and N. B. Shroff, "Distributed greedy approximation to maximum weighted independent set for scheduling with fading channels," in *Proceedings of the fourteenth ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2013, pp. 89–98.

[8] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Resource optimization in device-to-device cellular systems using time-frequency hopping," *arXiv preprint arXiv:1309.4062*, 2013.

[9] X. Lin, J. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks."

[10] D. H. Lee, K. W. Choi, W. S. Jeon, and D. G. Jeong, "Resource allocation scheme for device-to-device communication for maximizing spatial reuse," in *Wireless Communications and Networking Conference (WCNC), 2013 IEEE*. IEEE, 2013, pp. 112–117.

[11] S. Maghsudi and S. Stanczak, "Hybrid centralized-distributed resource allocation for device-to-device communication underlaying cellular networks," *arXiv preprint arXiv:1502.04539*, 2015.

[12] M. Botsov, M. Klugel, W. Kellerer, and P. Fertl, "Location dependent resource allocation for mobile device-to-device communications," in *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*. IEEE, 2014, pp. 1679–1684.

[13] A. Abdelnasser, E. Hossain, and D. I. Kim, "Tier-aware resource allocation in ofdma macrocell-small cell networks," *arXiv preprint arXiv:1405.2000*, 2014.

[14] D. T. Ngo, S. Khakurel, and T. Le-Ngoc, "Joint subchannel assignment and power allocation for ofdma femtocell networks," *Wireless Communications, IEEE Transactions on*, vol. 13, no. 1, pp. 342–355, 2014.

[15] S. Moharir, S. Krishnasamy, and S. Shakkottai, "Scheduling in densified networks: Algorithms and performance," in *Proceedings of Annual Conference on Communication, Control and Computing (Allerton)*, 2014.

[16] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "Flashlinq: A synchronous distributed scheduler for peer-to-peer ad hoc networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 21, no. 4, pp. 1215–1228, 2013.

[17] F. Baccelli, J. Li, T. Richardson, S. Shakkottai, S. Subramanian, and X. Wu, "On optimizing csma for wide area ad hoc networks," *Queueing Syst. Theory Appl.*, vol. 72, no. 1-2, pp. 31–68, October 2012.

[18] S. Bodas, S. Shakkottai, L. Ying, and R. Srikant, "Scheduling for small delay in multi-rate multi-channel wireless networks," in *Proceedings of IEEE Infocom*, 2011.

[19] L. Barenboim and M. Elkin, *Distributed Graph Coloring: Fundamentals and Recent Developments*. Morgan & Claypool Publishers, 2013.

[20] S. Krishnasamy and S. Shakkottai, "Spectrum sharing and scheduling in d2d-enabled dense cellular networks," Tech. Rep., 2015.