Stochastic Muting with Short-range Relay Analysis

Christian Vitale*, Vincenzo Sciancalepore†, Vincenzo Mancuso‡
*KIOS Research Center †NEC Laboratories Europe GmbH ‡IMDEA Networks Institute

***Abstract*—Resource *muting* in the time-frequency domain is a tool for controlling co-channel interference in cellular networks. However, coordinated resource of multiple cells implies non-trivial complexity to be effective. Moreover, resource muting does not guarantee fair resource allocation to users. However, we analytically show that *stochastic* resource muting jointly with *opportunistic* short-range relay operated by neighbor mobile devices permits to abate the complexity of muting coordination and boosts throughput fairness among users. We therefore formulate a convex proportional fairness optimization problem and leverage our analysis to design `VSM-PF`, a practical optimization scheme that significantly outperforms legacy systems.**

I. Introduction

Inter-Cell Interference Coordination (ICIC) and Device-to-Device (D2D) communications represent two key-elements along the innovation roadmap towards advanced cellular network deployments jointly with the advent of efficient mmWave communication standards suitable for direct network access as well as for relay [1].

Among the supported interference control mechanisms (i.e., ICIC, eICIC and FeICC), time-frequency domain traffic scheduling serves as a valid alternative to existing approaches, e.g., beamstearing, power control and massive MIMO. When designing a flexible radio for 5G communications [2] there have been several proposals to apply transmission coordination in the time-frequency domain and introduce energy efficient mechanisms to keep interference under control. The most appealing approach is resource muting [3], which includes alternating transmission activity from neighboring cells using the same radio resource. These efforts pave the road towards flexible and high-efficient cellular networks able to accommodate impelling service requirements. However, scalability may represent still a challenge due to the complexity of adopted coordination algorithms, which involve sampling of interference [4] or advanced base station coordination [5].

In this framework, we start by proving that existing interference coordination schemes are throughput-unfair. This follows from imposing (*almost*) interference-free transmissions at the expenses of drastically limiting the number of transmissions in a neighborhood. Nevertheless, we also show that resource muting may offer the possibility to schedule base station activities with very little overhead and low computational complexity. Therefore, playing with cell muting frequencies can be used to enforce fairness among cells.

Within a cell, the throughput of individual users can be boosted by exploiting channel diversity, and it is known that using opportunistic mobile relay over short-range communication channels can help to achieve the most of such diversity [6]. Thus, we analyze the performance of stochastic muting with opportunistic mobile relay within short inter-mobile distances. For relay, we use short-range mmwave cellular sidelinks based on IEEE 802.11ad standards and its evoluted versions 802.11ax and 802.11ay. Therefore, in the rest of the manuscript, we use the term mmD2D to refer to mmWave D2D sidelinks.

Based on the results of the analysis, we formulate a convex proportional fairness optimization problem and design `VSM-PF`, a practical optimization scheme that achieves fairness and high throughputs.

A. Background and related work

eICIC and FeICIC become popular due to the trade-off they offer between performance improvement and low implementation complexity [7]. Almost Blank Subframe (ABS) is the most common example of time-domain muting scheme adopted in 3GPP heterogeneous networks [8]. Specifically, ABS has been proposed for throttling macro base station power transmissions in presence of micro and pico cells so as to reduce the caused interference. However, much more interesting results have been shown when ABS has been adopted for all kinds of heterogeneous cells [9], whereas deterministic ABS approaches have shown how pre-computed time-patterns can lead to near-optimal working points [10]. Machine learning has been also proposed to learn interference patterns and build resource muting sequences accordingly [11]

Differently, [12], [13] propose a two-level beamforming coordination approach based on mmWave technology that partitions the network into clusters while performing an intra-cluster coordination similar to the user selection algorithms in a MIMO-based scenario. More advanced solutions for time-domain muting of resources focus on the pattern reuse, which directly guides the ABS activity pattern. In particular, [14] derives the best temporal pattern duration, given a set of chosen patterns to maximize the total user throughput. Near interference free transmissions are also possible with selective resource muting combined with highly directional antennas at base stations and mobile devices using mmWave technologies, although they have a high complexity in practice [15]. Other highly efficient approaches, at least in terms of boosting throughput, are based on using CoMP techniques [16], which

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under Grant 101003439 (C-AVOID), and in part by the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development. This work has been also supported by the Project AEON-CPS (TSI-063000-2021-38), funded by the Ministry of Economic Affairs and Digital Transformation and the European Union NextGeneration-EU in the framework of the Spanish Recovery, Transformation and Resilience Plan.

ISBN 978-3-903176-49-2 © 2022 IFIP

are however highly demanding in terms of distributed coordination. Moreover, as proved in our work, while some scenarios may adversely impact the system throughput a pure throughput maximization can lead to highly unfair throughput distributions.

There exist several studies on how to trade-off throughput for fairness using different approaches to the service architecture, which invoke the coordination of cellular and D2D transmissions. [17] and [18], for example, propose to apply user association and D2D multi-hop offloading, respectively. The authors of [6] and [19] implemented an opportunistic relay on WiFi Direct on a millisecond timescale. However, such solutions call for deep changes to the current cellular architecture. In contrast, our proposal is based on standard relay features of 3GPP sidelinks.

By adopting directional and electronically steerable antennas, mmWave sidelinks offer the advantage of having virtually unlimited bandwidth over the range of few meters, with respect to sub-6 GHz cellular communications, and being practically interference-free.

B. *Novelty and contributions*

To fully understand the potentials of cellular resource muting and mmD2D in combination, and to make the basis for advanced control architectures for 5G networks, we derive a theoretical framework. We pioneer a joint scheme to provide high spectral efficiency by leveraging cooperative D2D opportunistic communications using mmWave sidelinks, while at the same time adjusting user fairness by means of coordinated resource muting across neighboring cells.

Specifically, we (i) present a theoretical study on the limitations of resource muting and on the advantages of mmD2D sidelinks, (ii) derive stochastic conditions to show how resource muting can be orchestrated to steer user fairness, (iii) formulate a novel and convex optimization problem to set stochastic muting patterns by leveraging on the advantages of mmD2D relay and (iv) show that the joint operation of stochastic muting and mmD2D is practical and brings dramatic gains with respect to state-of-the-art solutions.

II. A Novel Sidelink-assisted ICIC Framework

We consider downlink transmissions in a cellular access network with a set $\mathcal{B}$ of interfering base stations (namely, "gNBs"), operated on the same frequency band by the same operator. Users are provided with multi-RAT connectivity, i.e., 3GPP cellular and IEEE 802.11ad/ax/ay physical interfaces which are considered for 3GPP 5G-NR extensions over unlicensed bands [20], for instance as sidelink technologies for V2X communications [21].

Base stations implement resource muting at slot level while users can leverage D2D sidelinks to form relay groups within few meters distance. The base station dynamically elects a relay within the group. The relay is the responsible for handling the traffic of the entire group. Intra-group relay transmissions adopt mmWave outband D2D, and the base stations select relay nodes opportunistically, according to who is experiencing the best channel condition, similarly to what proposed and

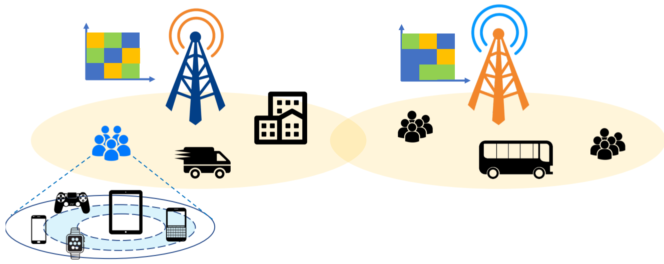


Fig. 1. Enhanced D2D assisted cellular framework

implemented with WiFi-Direct in [6]. Throughout the paper we use *mmWave outband D2D* (mmD2D) to refer to intra-group D2D relay. We assume that all groups always have packets to receive, i.e., the downlink queue of each users' group is saturated. The set of groups will be denoted by $\mathcal{C}$, and the size of group $c \in \mathcal{C}$ will be denoted as $U_c = |\mathcal{C}|$.

In such mmD2D-assisted cellular framework, depicted in Fig. 1, we propose a solution that retains the key strengths provided by mmD2D and resource muting. Differently from standard applications, we design a practical scheme to tune the use of cellular resources *stochastically* and not deterministically, ensuring user fairness additionally to inter-cell interference reduction, while at the same time counting on mmD2D to boost the system throughput by means of packet relay. The building blocks of the framework outlined above are *user groups*, *mmD2D sidelinks*, *opportunistic relay election*, and *stochastic muting pattern generation*.

We go beyond existing schemes, and apply the resource muting paradigm to mmD2D-enabled networks wherein the use of bad channels is limited and the number of users accounting for interference reduces to the number of (mmD2D) groups. With our proposal, we neither impose the burden of making fully tailored per-cell muting decisions in a centralized way, nor force any base station to generate myopic resource muting patterns completely unaware of others' cell activity. Instead, we let a *network controller* generate and distribute stochastically built muting patterns at low complexity. Interestingly, this can be adjusted over time to pursue fairness but also, indirectly, system throughput.

III. Analysis

We analyze the framework described in the previous section and characterize transmission efficiency and throughput. The analysis provides us with the tools needed to optimize the stochastic generation of gNB activity patterns.

A. *Transmission Efficiency*

Interfering cellular transmissions may cause a severe performance degradation. To analytically quantify the impact of interference we might use the Signal-to-Interference-plus-Noise-Ratio (SINR). Nevertheless, SINR does not account for the specific modulation and coding scheme (MCS) used, and therefore, it does not account for the real efficiency of a transmission. We formally define the transmission efficiency of a group (of one or more users) as follows:

Definition 1 (Transmission efficiency). *The efficiency $\zeta_c(X_c, Y_c)$ of base station transmissions operated towards a group c whose U_c members are located at $(X_c, Y_c) = [(x_1, y_1), ..., (x_{U_c}, y_{U_c})]$ is the average number of bits transmitted by the base station to the group for each transmission symbol, considering that the base station always uses the fastest MCS allowed by the best SINR experienced by the users in the group.*

Computing transmission efficiency does not depend solely on the position and number of users in the group, but also on the mapping between SINR and MCS. Depending on the distribution of the SINR, we may achieve largely different values for the transmission efficiency, even with the same SINR average. Specifically, given the location of the users of group c at positions (X_c, Y_c) and denoting by b_k the number of bits transmitted per symbol using MCS k, the transmission efficiency for the group is

$$\zeta_c(X_c, Y_c) = \sum_{k \in \mathcal{M}} b_k \left[F_c(T_k^{\max}) - F_c(T_k^{\min})\right], \qquad (1)$$

where $F_c(x)$ denotes the cumulative distribution of the r.v. $\gamma_c(X_c, Y_c) = \max_{i \in c}\{\gamma_{ci}(X_c, Y_c)\}$, being $\gamma_{ci}(X_c, Y_c)$ the SINR of user i in group c. The summation in (1) accounts for the number of bits per symbol transmitted by the users on a discrete set $\mathcal{M}$ of MCSs, as suggested by the standard [8], i.e., by casting the SINR function in a continuous subset of values comprised between $T_k^{\min}$ and $T_k^{\max}$, representing lower and upper SINR levels, respectively, for assigning the MCS k.

To compute the distribution $F_c(x)$, we consider that the SINRs experienced by users are independent random variables with averages imposed by the user positions. The independence comes from the fact that the fast fading process is very much affected by tiny position differences. For the computation of the average of such random variables, we will consider two cases and show that they yield similar results in practice. We either consider that the average SINR depends on the exact position of the user, or that the position of a user can be approximated with the center of gravity of its group, so that all users see the same average signal and interference.

Having assumed that the SINR values in a group are independent, and considering that $\max_{i \in c}\{\gamma_{ci}(X_c, Y_c)\} \leq x \iff \forall i \in c, \gamma_{ci}(X_c, Y_c) \leq x$, we can write

$$F_c(x) = \prod_{i \in c} F_{\gamma_{ci}}(x). \qquad (2)$$

Instead, if we approximate the positions of the users with the center of gravity of the group, denoting by $\gamma_c^\star$ the SINR computed at the center of gravity, the following expression holds and approximates (2):

$$F_c(x) = \left[F_{\gamma_c^\star}(x)\right]^{U_c}. \qquad (3)$$

This approximation makes sense when a group mobility model can be used to describe the dynamics of the user's topology.

Moreover, as pointed out in [22], in an urban environment, the power received by a user from a base station at any given location follows a negative exponential distribution (whereas the instantaneous signal follows a Rayleigh distribution) whose average value only depends on the pathloss effect. Therefore, the distributions $F_{\gamma_{ci}}$ and $F_{\gamma_c^\star}$ can be computed using the result reported in the following proposition.

Proposition 1. *The distribution of the SINR γ resulting from an exponential useful signal with average power $1/\lambda_S$, k independently exponentially distributed interfering signals I_j with average power $1/\lambda_j$, $j = 1, 2, \cdots k$, and additive Gaussian white noise with zero mean and power N is, $\forall x \geq 0$,*

$$F_\gamma(x) = 1 - \frac{1}{\sqrt{1 + 2\lambda_S N x}} \prod_{j=1}^{k} \frac{\lambda_j}{\lambda_j + x\lambda_S}. \qquad (4)$$

Sketch of the proof. The result is easily obtained by writing $F_\gamma(x)$ as $E_Y\left[\Pr\left\{\frac{S}{Y^2 + \sum_{j=1}^{k} I_j} \leq x\right\}\right]$, where S is the useful signal and the average is computed over an additive zero-mean Gaussian noise Y, which has variance N. The quantity within the average can be computed by conditioning over all possible values of I_j r.v.'s, which are all independent and exponentially distributed. Therefore, we obtain that $F_\gamma(x) = E_Y\left[1 - e^{-x\,\lambda_S Y^2} \prod_{j=1}^{k} \frac{\lambda_j}{\lambda_j + x\,\lambda_S}\right] = 1 - E_Y\left[e^{-x\,\lambda_S Y^2}\right] \prod_{j=1}^{k} \frac{\lambda_j}{\lambda_j + x\,\lambda_S}$. Computing the average $E_Y\left[e^{-x\,\lambda_S Y^2}\right]$ in closed form is straightforward for a Gaussian r.v., which leads to the result. □

Let us define each potential combination of base stations active on a time-frequency resource as one of the possible *states* of the system, and let us denote it by s. Let $\mathcal{B}_s$ be the set of base stations transmitting data in state s, whereas $|\mathcal{S}| = 2^{|\mathcal{B}|}$ be the set of all possible states. With each state s the interference changes and so the transmission efficiency does (hence the system throughput), as expressed in the following definition.

Definition 2 (Transmission efficiency in state s). *We denote by ζ_c^s the transmission efficiency of group c in state s, which can be computed with* (1)–(4) *by ignoring the contributions due to base stations not active in state s. The transmission efficiency of a group under the coverage of an inactive base station is set to* 0.

B. Instantaneous System Throughput

Each active base station implements a scheduler such that each group obtains transmission opportunities proportionally to the group size, i.e., for any group $c \in \mathcal{C}$ under the coverage of an active base station $b \in \mathcal{B}_s$ at time t, the resources allocated are expressed as follows:

$$D_c(t) = K_{\text{sym}} \frac{U_c}{\sum_{i | b \text{ covers groups } i \text{ and } c \text{ at time } t} U_i}, \qquad (5)$$

where K_{sym} is the total available number of symbols per second at the base station b serving group c at time t. $D_c(t)$ does not depend on state s if the system is in saturation; otherwise, in (5) we should use the actual number of backlogged groups at time t instead of the total number of groups.

With the foregoing definitions, the instantaneous per-group throughput $\Gamma_c^s(t)$ and the corresponding aggregate system throughput $\Gamma^s(t)$ in state s at time t are computed with the following expressions:

$$\Gamma_c^s(t) = D_c(t)\, \zeta_c^s(x_c, y_c); \tag{6}$$

$$\Gamma^s(t) = \sum_{c\in\mathcal{C}} \Gamma_c^s(t). \tag{7}$$

C. Effect of Mobility

Let us now consider the impact of group mobility and muting states, since they directly affect experienced SINR levels. The objective here is to compute the mean transmission efficiency and throughput of group c, averaged over time.

1) Asymptotic Transmission Efficiency: The asymptotic transmission efficiency is defined as a time average:

$$\bar{\zeta}_c = \lim_{T\to+\infty} \frac{\int_0^T \zeta_c^{s(t)}(x(t), y(t))dt}{T}, \tag{8}$$

where resource muting state s and group's location can change over time. Assuming that the system is ergodic, the above quantity is equal to the stochastic average of a random process ζ_c^V, in which V is the random process for the muting state:

$$\bar{\zeta}_c = E[\zeta_c^V] = \sum_{s\in\mathcal{S}} E[\zeta_c^s] P^s, \tag{9}$$

where we have used the total probability formula and defined $P^s = \Pr(V{=}s)$. The above relation unveils that the asymptotic transmission efficiency can be expressed in terms of per-state transmission efficiency. Most importantly, it conveys that the asymptotic transmission efficiency is affected by the probability of using a particular state, and the following fundamental result holds:

Proposition 2. *The asymptotic transmission efficiency $\bar{\zeta}_c$, for a given user mobility model and for a fixed topology of base stations is solely affected by two components that can be tuned independently: resource muting state probabilities and mmD2D group composition:*

$$E[\zeta_c^s] = \int_A L_c(x,y)\zeta_c^s(x,y)dA, \tag{10}$$

where $A = \cup_{b\in\mathcal{B}} A_b$ is the total covered area and $L_c(x,y)$ is denoting the spatial distribution of group c's center of gravity.

Sketch of the proof. The result is obtained by writing the conditional average transmission efficiency of a group in state s under the coverage of base station b. Then one can remove the condition by using the spatial distribution to obtain the probability to be covered by b, which is $p_c(b) = \int_{A_b} L_c(x,y)dA$, where A_b is the area covered of b. □

2) Asymptotic Throughput: We can derive the asymptotic average throughput (over time) achieved by group c with an approach similar to the one described above for the asymptotic transmission efficiency:

$$\bar{\Gamma}_c = E[\Gamma_c] = \sum_{s\in\mathcal{S}} E[\Gamma_c^s] P^s. \tag{11}$$

In the above formula, the conditional average throughput in state s can be computed as follows:

$$\begin{aligned} E[\Gamma_c^s] &= \sum_{b\in\mathcal{B}_s} p_c(b)\, E[\zeta_c^s|b]\, E[D_c|b] \\ &= \sum_{b\in\mathcal{B}_s} E[D_c|b] \int_{A_b} L_c(x,y)\zeta_c^s(x,y)dA, \end{aligned} \tag{12}$$

where $E[D_c|b]$ represents the average number of symbols allocated to group c under base station b. $E[D_c|b]$, is obtained by considering all possible combinations of groups that fall under the coverage of base stations b, as follows:

$$E[D_c|b] = \sum_{Z\in\mathcal{P}(\mathcal{C},c)} \prod_{i\in Z} p_i(b) \prod_{j\notin Z} [1-p_j(b)] \frac{U_c}{U_c + \sum_{i\in Z} U_i} K_{sym}, \tag{13}$$

where $\mathcal{P}(\mathcal{C}, c)$ is the power set of the groups in $\mathcal{C}$ when group c is taken out (Z is therefore a set too). The calculation of $E[D_c|b]$ assumes a proportional resource scheduling based upon group sizes, i.e., proportional to the number of users building up the mmD2D groups.

Since (13) does not depend on state probabilities, (12) behaves likewise, and the asymptotic average throughput (11) has a similar structure as the transmission efficiency. Therefore, the above derivation directly leads to the following result, similar to what presented for the transmission efficiency:

Proposition 3. *The asymptotic throughput $\bar{\Gamma}_c$, for a given user mobility model and for a fixed topology of base stations is solely affected by two components that can be tuned independently: resource muting state probabilities and mmD2D group composition.*

The distribution of resources expressed in (13)—and therefore the asymptotic system throughput in state s expressed in (12)—is strictly dependent on the set of active groups $\mathcal{C}$ and their movements.

The asymptotic system throughput achieved in state s is simply given by the sum of group's asymptotic throughputs, i.e., $E[\Gamma^s] = \sum_{c\in\mathcal{C}} E[\Gamma_c^s]$.

Remark 1. *Note that independent users not joining any group can be analyzed by regarding those users as groups of size one.*

D. Comments on the Model

Transmission efficiency and throughput achieved by each group—and hence the fairness level experienced in the network—depend on the fraction of time spent in each state and on the gain attained by means of mmD2D when relay groups form. The equations we have derived show that the two components operate orthogonally because the group composition does not depend on the state and vice versa. Besides, mobility (or the distribution of user positions) plays an important role, and we have shown how to analytically derive the achievable throughput as a function of group position distributions.

Most importantly, all derived results show that there is no dependency on how the resource muting state alternate over time, except the probability to observe a state has to be known, so as to be able to predict system performance. Thus, muting state probabilities can be tuned to optimize performance.

In Section IV, we will study the optimization of muting patterns in terms of state probabilities P^s, which can be computed with no need to coordinate between gNBs and only requires the enforcement of simple random assignments—thus achieving extremely low complexity—for fixed mmD2D configurations and mobility parameters.

IV. Proportional Fairness Optimization

Here, we design an easy-to-deploy stochastic muting mechanism that jointly achieves high transmission efficiency as well as maximal user fairness, closely following the variations of channel qualities and group locations in the system.

Implementing the resulting stochastic muting patterns has a twofold advantage: (i) random patterns do not incur systematic discretization issues that might arise with deterministic allocations of states, and (ii) they make it possible to generate distinct patterns for distinct base stations independently, thus reducing the complexity of *network controllers* issuing the patterns.

Note that the formation of groups could be optimized as well, as it emerges from the analysis. However, it is left out of this work since it would require a more comprehensive study on user behavior, trustworthiness, costs and incentives, and other aspects that deserve a stand-alone project.

A. Dynamic Proportional Fairness Optimization

To formulate our proportional fairness optimization problem, let us consider that, in short intervals of time of duration T, in which mobility effects are negligible, the throughput achievable in each state s by each group does not change and can be indicated as the group throughput computed at any point in time within that interval. Thus, considering a time-slotted optimization framework starting at time t_0, composed by intervals $I_n \triangleq [t_0 + nT; t_0 + (n+1)T)$, $n \geq 0$, we can denote the throughput as $E\left[\Gamma_c^s(I_n)\right] = E\left[\Gamma_c^s(t)\right]$ computed with the group positions evaluated at any t chosen in I_n. During intervals I_n, it is however possible to chose subsequently different muting states, so to achieve as performance the average of what is achieved over the selected states. We denote with $P^s(I_n)$ the fraction of time during which state s is enforced in interval I_n, the resulting throughput of group c in interval I_n is:

$$E\left[\Gamma_c(I_n)\right] = \sum_{s \in \mathcal{S}} P^s(I_n) E\left[\Gamma_c^s(I_n)\right]. \tag{14}$$

Clearly, the order in which states are visited is not important and the computation of state probabilities must be repeated at every interval I_n, due to network dynamics. The choice for the duration T of such interval is pivotal for system performance: a *short* interval allows to consider the network as static, while a *long* interval accounts for including several muting states, which in turn increases the accuracy resulting from implementing optimal probabilities with a finite-length muting pattern.

We formulate a Versatile Stochastic Muting for Proportional Fairness (`VSM-PF`) problem to select state probabilities based on a long-term proportional fairness metric. The throughput is observed over p past intervals and predicted for the next interval I_n. To achieve so, the optimization is repeated at the begin of each interval I_n, and we define a utility function $\widehat{\eta}$ based on the $\log$ of group throughput (to introduce proportional fairness):

Problem `VSM-PF` :

At time $t = t_0 + nT$, select $P^s(I_n), \forall s \in \mathcal{S}$, so to:

$$\begin{aligned} \text{maximize} \quad & \widehat{\eta} = \sum_{c \in \mathcal{C}} w_c \, \log\left(\sum_{k=n-p}^{n} \alpha_{n-k} E\left[\Gamma_c(I_k)\right]\right); \\ \text{subject to:} \quad & \sum_{s \in \mathcal{S}} P^s(I_n) = 1, \\ & P^s(I_n) \in [0,1], \quad \forall s \in \mathcal{S}; \end{aligned} \tag{15}$$

where weights w_c are used to tune the group fairness target and coefficients α_k define how past samples of throughput affect future decisions. Since $E\left[\Gamma_c(I_n)\right]$—which is the only unknown term in the sum inside the $\log$ argument because past values have been observed—is linear in the decision variables P^s, the optimization problem is convex, admits a global maximum and can be easily linearized and solved in polynomial time.

Every time probabilities $P^s(I_n)$ are computed, the network controller stochastically builds and distributes a new muting pattern by choosing, for each element of the pattern, a state at random according to new optimal probabilities.

B. Remarks on the Stochastic Optimization

Weights w_c can be selected based on the desired fairness target. E.g., for targeting equal throughput on a *per-user* basis, given that the group throughput is equally shared by group members, w_c can be set as the number of users forming group c, so that group throughputs will be as much as possible proportional to group sizes.

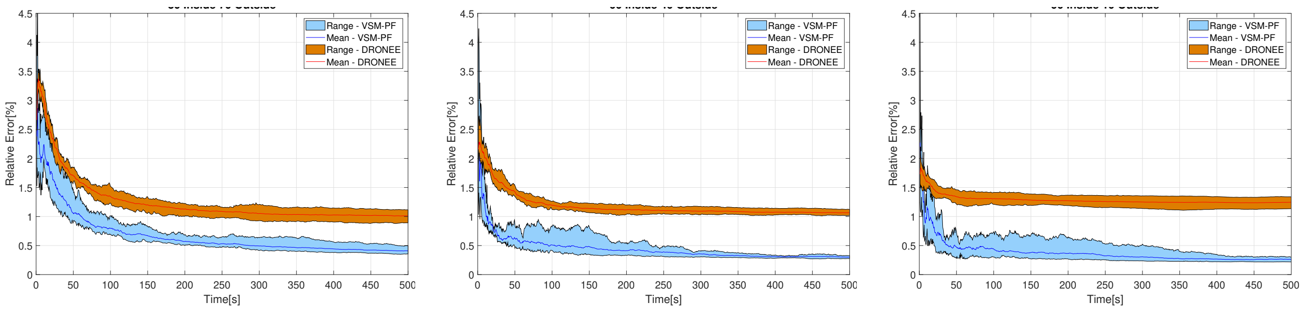
Coefficients α_k can be taken as a non-decreasing sequence of non-negative weights, so that past values of the throughput receive less or equal importance with respect to the prediction for next interval I_n. For example, exponentially decaying coefficients or constant coefficients represent simple and widely adopted solutions for this kind of digital filtering problems.

Interestingly, Problem `VSM-PF` is simple to adapt also to cover the case in which the traffic of groups is not saturated. In such a case, the time window $(p+1)T$ has to be smaller than the interval during which the set of receivers changes. In fact, during such interval all active receivers can be considered as saturated and the presented analysis holds.

V. Practical Details

Before proceeding with the numerical evaluation, here we comment on a few practical implementation details.

Groups. The formation and presence of groups of users leaning toward cooperation is key for the success of opportunistic relay approaches. Groups might form using services



(a) Peak hour (b) Business hour (c) Night hour

Fig. 2. Difference between simulation and analytically predicted throughput (minimum, maximum, and average values)

like Google Nearby, which is an Android feature to discover D2D peers and request connection [23]. Other groups might form by static user configuration, e.g., by pre-authorizing communication between devices belonging to the same owner (like it happens for wireless mice and hands-free speakers). In any case, the group has an effective role in the system only if relay opportunities last a sufficiently large period of time, so that the group setup overhead can be neglected.

Resource allocation with groups. Groups are served by the gNB like they were normal UEs. For the sake of fairness, the gNB should use a weighted round robin policy and assume the number of users within the group as the scheduling weight. Therefore, the average number of cellular resources allotted to each user would remain constant considering all possible group configurations.

Relay node selection. Selecting relay nodes opportunistically and switching them swiftly, as soon as channel conditions vary, is a key-enabler for mmD2D. We assume that the relay node is selected by the gNB on a per-packet basis, leveraging CSI reports, so that cellular transmissions always occur on the strongest cellular channel. A similar approach has been suggested and experimentally validated in [6] for WiFi Direct. Please note that continuous re-election of relay nodes has no practical drawbacks on end-to-end latency due to the huge available bandwidth.

Patterns for muting. 3GPP standard guidelines already allow to implement muting schemes like ABS [8] in a conventional cellular system without imposing any constraint on the specific set of resources to mute. Current implementations are limited to time-domain muting, although extensions have been proposed for time-frequency muting [3]. For instance, standard specifications for ABS describe an *application ratio* defined as the number of used subframes over the total number of subframes within a pattern. Once this fixed ratio is imposed by a *network controller*, base stations may make random choices to select the specific pattern of subframes to be muted, which can be updated at second time scale for ABS. The extension to arbitrary time-frequency resource blocks is straightforward, although it requires the definition of a common block structure across base stations, e.g., a common way to partition the set of available time-frequency resources.

VI. Performance Evaluation

In this section, we validate and evaluate our proposal for realistic cellular scenarios built on top of a heterogeneous dense-urban area of $400\text{m} \times 320\text{m}$ close to the Oxford Circus metro station in London city (UK). We consider only the base stations under the control of O2 mobile network operator, whose base stations' position and transmission power are publicly available.[1]

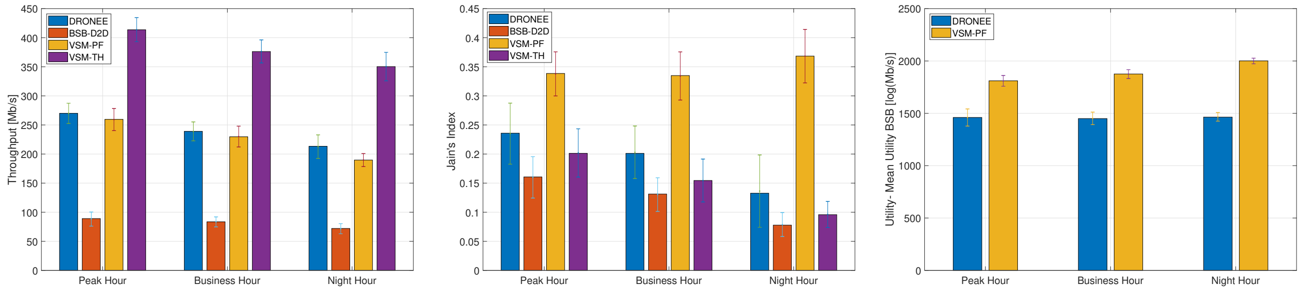
A. Experimental Setup

We have developed a Matlab event-driven packet simulator wherein only downlink transmissions are taken into account with a 20MHz bandwidth and with the pathloss in [24]. Network simulations last 500 s, whereas user channel conditions are evaluated on a 3GPP subframe basis, e.g., each 1 ms.

We consider a total of 1000 users, as the user density for such area is ~ 8000 users/km^2 and the area considered is 0.128 km^2 [25]. Users placed within the considered area can be divided into two groups: (i) users on the streets, which follow a random waypoint mobility model [26]; (ii) users in the building, which instead do not move from their location. Specifically, the outdoor user mobility works as follow: (i) users select randomly a speed and a destination location within a valid street of the map, (ii) they follow the shortest path and reach the new destination according to the streets of the map.

In our simulations, only users on the streets form groups, in contrast to indoor users. When relay groups are in place, their center of gravity follows the mobility model rules on the streets while the users of the group are randomly placed around the center of gravity. Users within the buildings are statically allocated uniformly at random. Along our simulations, we consider the performance in three operational timeframes. For each timeframe, we properly model the ratio between the distribution of users moving along the streets and the users staying within the buildings according to the report in [25]. Specifically, during *Peak Hours*, i.e., at lunch time, 70% of the users is on the street, at *Business Hours*, during morning and afternoon, 40% of the population stays outdoor, while at *Night Hours*, 90% of the users is at home. In the three timeframes analyzed, the percentage of outdoor users, hence of users that

[1]All information are retrieved from the OFCOM reports available at http://stakeholders.ofcom.org.uk/sitefinder/sitefinder-dataset/

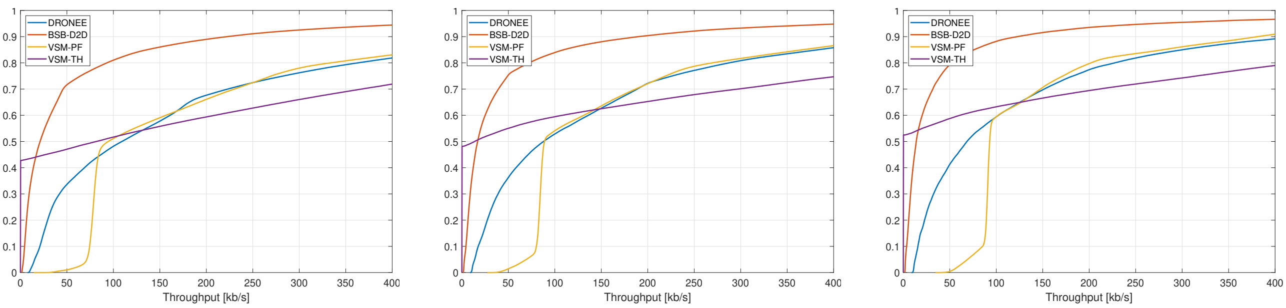


(a) Aggregate network throughput

(b) Per-user throughput fairness

(c) Network utility according to a proportional fairness metric (relative to `BSB-D2D`)

Fig. 3. Performance evaluation in the London Oxford Circus scenario



(a) Peak hour

(b) Business hour

(c) Night hour

Fig. 4. Distribution of per-user throughput in the London Oxford Circus scenario

can group and exploit opportunistic forwarding, decreases. When active, gNBs apply a weighted round robin policy to deliver the offered traffic to the relay groups, using the sizes of the groups as weights. Outdoor user speed uniformly ranges from 1 to 10 m/s.

Optimization. We consider two versions of the problem presented in Section IV-A. The first version, denoted as `VSM-PF`, optimizes proportional fairness and uses exactly the formulation described in (15). The second version, denoted as `VSM-TH`, maximises the throughput, and is obtained by omitting the log function in the objective of (15). To solve the problems, we run the optimization every $T = 500$ ms, with weights $w_c = U_c$ to achieve per-user fairness, $\alpha_k = 1$, and $p = 20$, i.e., we aim at the maximization of the average throughput fairness over a window of 10s.

Metrics and benchmarks. We evaluate `VSM-PF` and `VSM-TH` in terms of system throughput and fairness, the latter being measured through the proportional fairness $\widehat{\eta}$ and by means of the well-known Jain's Fairness Index (JFI). For the sake of comparison, we consider two practical state of the art heuristics: (i) `BSB`, proposed in [10], in which time-domain muting (ABS) is used to target a max-min utility function; and (ii) `DRONEE` [19], in which relay groups form dynamically to improve throughput, but no muting is considered. Although `DRONEE` was proposed for WiFi D2D, it is straightforward to use it with the bandwidth of mmWave D2D sidelinks.

B. Model Validation

In order to evaluate and validate the analytical model presented in Section III, in Fig. 2 we compare the system throughput Γ^s obtained according to the analysis and the one computed with the simulation in different traffic scenarios. We report average plus extreme deviations (the range between the largest and the smallest variation observed in our tests) for `VSM-PF` and for DRONEE. As shown in the figure, the relative error is always very small in all timeframes, below 5%, and tends to reduce as simulation time passes by, which is a proof that the simulations actually converge to the average values used in the analysis. The figure also shows that mobility has an impact on accuracy. In particular, the smaller the portion of pedestrian users considered, the faster the relative error drops, because the scenario becomes less and less dynamic. Interestingly, the relative error observed for `DRONEE` is always higher than the error for `VSM-PF`, which is due to the fact that in `DRONEE` all users are always active and can continuously switch to the gNB that offers the best signal, so that resource allocation is more dynamic than in `VSM-PF`. This also points out that `VSM-PF` is less complex and more scalable than regular opportunistic approaches.

C. Performance in a Heterogeneous Deployment

In Fig. 3, we compare the system performance obtained with different schemes. Fig. 3(a) illustrates how the aggregate system throughput increases during peak hours, as most of the people are moving outside and, thus, can exploit opportunistic relay over mmD2D sidelinks. `VSM-PF` performs quite well, showing similar throughput figures as `DRONEE`. `VSM-TH` achieves dramatic throughput gain, although at the expenses of fairness, as shown in Fig. 3(b) in terms of Jain's fairness index computed on the average per-user throughput. The `BSB-D2D` scheme does not perform well in terms of throughput, which was expected, and does not either achieve high levels of fairness, with respect to other schemes. Therefore, in Fig. 3(c), we use the fairness of `BSB-D2D` as a benchmark. In there, we observe that the proportional fairness level achieved by `VSM-PF` is much higher than with `DRONEE` (note the log scale used in the metric). Fig. 3(c) does not report values for `VSM-TH` because a pure throughput-based optimization might lead to observe intervals of time in which one or more users obtain zero throughput, with correspond to $-\infty$ in terms of fairness. Fig. 4 offers a detailed view of the fairness advantages offered by `VSM-PF` for the three considered timeframes. The figure compares the cumulative distribution functions of the per-user throughput across the adoption of different schemes. `VSM-PF` is the only scheme that avoids low throughput with high probability, while other schemes often starve users in as much as 50% of the cases.

The figures shown in this section unveil the great potentials of properly applying a dynamic muting scheme in conjunction with short-range opportunistic mmD2D sidelinks. The gain stemming from the use of `VSM-PF` is mainly expressed in terms of fairness, while throughput performance remains practically as high as the one of pure opportunistic mmD2D. Therefore, `VSM-PF` achieves significantly better fairness than traditional muting schemes while retaining most of the advantages of opportunistic mmD2D.

VII. Conclusions

We have modeled the performance of a cellular network with stochastic muting and short-range D2D relay with mmWave sidelinks, which are now becoming commodities for 3GPP and IEEE protocols. We have shown that interference coordination and transmission efficiency can be addressed simultaneously by stochastically coordinating the muting of resources at the gNBs and by leveraging opportunistic mmWave sidelinks. Indeed, our analytical model provides accurate results in realistic scenarios and under truly simulated user mobility models. Our practical optimization solution, `VSM-PF`, goes beyond state-of-the-art solutions and offers significant advantages in terms of throughput, per-user throughput fairness, and network-aggregate proportional fairness metrics.

References

[1] Z. Xiao, L. Zhu, J. Choi, P. Xia, and X.-G. Xia, "Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter wave communications," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 2961–2974, 2018.

[2] "System Architecture for the 5G System," 3GPP, 3GPP TS 23.501 Version 15.2.0 - Release 15, 2018. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/15.02.00_60/ts_123501v150200p.pdf

[3] B. Soret, A. D. Domenico, S. Bazzi, N. H. Mahmood, and K. I. Pedersen, "Interference coordination for 5G New Radio," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 131–137, JUNE 2018.

[4] N. Tenny, L. Bingzhao, and A. Sang, "Coordination of measurement gaps across sets of multiple frequencies," Apr. 25 2019, US Patent App. 15/789,751.

[5] W. Chen, I. Ahmad, and K. Chang, "Co-channel interference management using eICIC/FeICIC with coordinated scheduling for the coexistence of PS-LTE and LTE-R networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, p. 34, Feb 2017. [Online]. Available: https://doi.org/10.1186/s13638-017-0822-6

[6] A. Asadi, V. Mancuso, and R. Gupta, "An SDR-based Experimental Study of Outband D2D Communications," in *IEEE INFOCOM*, 2016.

[7] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. Thomas, J. Andrews, P. Xia, H. Jo, H. S. Dhillon, and T. D. Novlan, "Heterogeneous cellular networks: From theory to practice," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 54–64, 2012.

[8] Third Generation Partnership Project (3GPP), "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); X2 application protocol (X2AP)," 3GPP TS 36.423 v. 14.0.0, September 2016.

[9] M. Cierny, H. Wang, R. Wichman, Z. Ding, and C. Wijting, "On number of almost blank subframes in heterogeneous cellular networks," *IEEE Trans. on Wireless Comm.*, vol. 12, pp. 5061–5073, October 2013.

[10] V. Sciancalepore, V. Mancuso, A. Banchs, S. Zaks, and A. Capone, "Enhanced content update dissemination through D2D in 5G cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 7517–7530, Nov 2016.

[11] J. Guo, Z. Xu, and C. Yang, "Learning fairly with class-imbalanced data for interference coordination," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 7, pp. 7176–7181, 2021.

[12] J. Wang, J. Weitzen, O. Bayat, V. Sevindik, and M. Li, "Interference coordination for millimeter wave communications in 5G networks for performance optimization," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 46, Feb 2019.

[13] W. Yi, Y. Liu, and A. Nallanathan, "Modeling and analysis of D2D millimeter-wave networks with poisson cluster processes," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5574–5588, Dec 2017.

[14] K. Son, Y. Yi, and S. Chong, "Utility-optimal multi-pattern reuse in multi-cell networks," *IEEE Trans. on Wireless Communications*, vol. 10, no. 1, pp. 142–153, 2011.

[15] Z. Sha, S. Chen, and Z. Wang, "Near interference-free space-time user scheduling for mmWave cellular network," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2022.

[16] S. Gulia, A. Ahmad, S. Singh, and M. D. Gupta, "Interference management in backhaul constrained 5G hetnets through coordinated multipoint," *Computers and Electrical Engineering*, vol. 100, p. 107982, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790622002476

[17] S. Singh and J. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. on Wireless Communications*, vol. 13, no. 2, pp. 888–901, February 2014.

[18] Y. Niu, C. Gao, Y. Li, L. Su, D. Jin, and A. V. Vasilakos, "Exploiting device-to-device communications in joint scheduling of access and backhaul for mmWave small cells," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2052–2069, 2015.

[19] A. Asadi and V. Mancuso, "DRONEE: Dual-radio opportunistic networking for energy efficiency," *Computer Communications*, vol. 50, pp. 41–52, September 2014.

[20] K. Aldubaikhy, W. Wu, N. Zhang, N. Cheng, and X. Shen, "mmWave IEEE 802.11ay for 5G Fixed Wireless Access," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 88–95, 2020.

[21] K. Sehla, T. M. T. Nguyen, G. Pujolle, and P. B. Velloso, "Resource Allocation Modes in C-V2X: From LTE-V2X to 5G-V2X," *IEEE Internet of Things Journal*, pp. 1–1, 2022.

[22] D. P. Meyer and H. A. Mayer, "Radar target detection- handbook of theory and practice," *New York, Academic Press, Inc.*, 1973.

[23] J. Chioino, I. Contreras, A. Barrientos, and L. Vives, "Designing a decision tree for Cross-device communication technology aimed at iOS and Android developers," in *Proceedings of the 2nd International Conference on Information System and Data Mining*. ACM, 2018.

[24] Third Generation Partnership Project (3GPP), "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Radio Frequency (RF) system scenarios," 3GPP TS 36.942 v. 13.0.0, January 2016.

[25] P. Rode, C. Hoffman, J. Kandt, D. Smith, and A. Graff, "Towards new urban mobility the case of London and Berlin," *LSE Cities*, Sep 2015.

[26] C. Bettstetter, G. Resta, and P. Santi, "The node distribution of the random waypoint mobility model for wireless ad hoc networks," *IEEE Transactions on Mobile Computing*, vol. 2, no. 3, pp. 257–269, July 2003.