Journal of Computer Engineering
ISSN 10003428

Original Research Article
Vol 12, Issue 1, 2023 | ISSN: 1000-3428

Energy-Efficient Spectral Allocation and Adaptive Clustering in Next-Generation Cellular Networks

Andreas Vasilakis and Konstantinos Papadopoulos

Andreas Vasilakis, Department of Electrical Engineering, University of Thessaly, Volos, Greece; Konstantinos Papadopoulos, Institute for Telecommunications and Information Networking, University of the Aegean, Lesvos, Greece

Abstract—The current and envisaged increase of cellular traffic poses new challenges to Mobile Network Operators (MNO), who must densify their Radio Access Networks (RAN) while maintaining low Capital Expenditure and Operational Expenditure to ensure long-term sustainability. In this context, this paper analyses optimal clustering solutions based on Device-to-Device (D2D) communications to mitigate partially or completely the need for MNOs to carry out extremely dense RAN deployments. Specifically, a low complexity algorithm that enables the creation of spectral efficient clusters among users from different cells, denoted as enhanced Clustering Optimization for Resources' Efficiency (eCORE) is presented. Due to the imbalance between uplink and downlink traffic, a complementary algorithm, known as Clustering algorithm for Load Balancing (CaLB), is also proposed to create non-spectral efficient clusters when they result in a capacity increase. Finally, in order to alleviate the energy overconsumption suffered by cluster heads, the Clustering Energy Efficient algorithm (CEEa) is also designed to manage the trade-off between the capacity enhancement and the early battery drain of some users. Results show that the proposed algorithms increase the network capacity and outperform existing solutions, while, at the same time, CEEa is able to handle the cluster heads energy overconsumption.

I. Introduction

The envisaged increase of the cellular traffic, which according to [1] is expected to reach 30.6 exabytes per month by 2020 at a compound annual growth rate (CAGR) of 53%, imposes new capacity challenges to the fifth generation (5G) cellular networks. Specifically, this - increasing trend in data traffic demand will force 5G networks to meet a 1000× capacity increase, mainly based upon three pillars: the improvement of the spectral efficiency, the allocation of new spectrum bands, and the densification of the Radio Access Network (RAN) [2]. Focusing on the densification of the RAN, the research community has proposed the dense deployment of Small Cells (SC) as an enabler for the capacity increase required to meet the expected traffic demand. However, such densification of the RAN has posed technological challenges, such as interference management [3][4], and economic considerations [5].

As mobile devices are the main contributors to the traffic growth, high capacity demand is intrinsically linked to the boost in the number of mobile devices. For instance, and based on [1], the number of mobile devices and connections will globally reach 11.6 billion by 2020. Therefore, the need for denser RAN deployments run in parallel with the actual and envisaged growth of the density of mobile devices. In this context, where the densification of the network is jeopardized by the high deployment costs, we propose the exploitation of the cooperation among mobile devices (through Device-toDevice communications, D2D [6]) as a cost-efficient solution to expand the RAN when and where needed. The inclusion of mobile devices as an expansion of the RAN can provide high spatial diversity and improve the spectral efficiency of the whole network. Although cooperation among Base Stations (BS) has already been proposed as a mean to increase the spectral efficiency (e.g. [7]), cooperation among devices proposed in the sequel opens up new opportunities and challenges to get the network dynamically adapted to traffic needs.

The rest of the paper is organized as follows: The State of the Art and contributions are detailed in Section II. In Section III the system is modelled as an optimization problem, and two clustering algorithms, namely enhanced Clustering Optimization for Resources Efficiency (eCORE) and Clustering algorithm for Load Balancing (CaLB), are presented. Section IV analyses the energy consumption and proposes a Clustering Energy Efficient algorithm (CEEa) to prevent cluster heads from early battery drain. Numerical results are presented in Section V and conclusions in Section VI.

II. State of the Art and Contributions

The need to improve spectrum utilization, overall throughput and energy consumption in cellular networks has stimulated the research on the D2D field over the last years. In short, D2D communications are expected to become the basis to provide direct connectivity between users (with or without the support

How to cite: Andreas Vasilakis and Konstantinos Papadopoulos (2023). Energy-Efficient Spectral Allocation and Adaptive Clustering in Next-Generation Cellular Networks. Journal of Computer Engineering, 12(1), 63–86. DOI: https://doi.org/10.5281/zenodo.19347569

of network infrastructure), enable devices to play the role of relay in two-hops communications, and allow the multicast of common content to a multicast group [8].

Regarding the direct connectivity between users, *Feng et al.* proposed in [9] a resources' allocation framework to optimize the spectral efficiency of the network when a set of D2D pairs operate over the same frequency as the cellular users. In this study, however, D2D pairs are never connected to the cellular network and therefore the D2D pairs have only two options: transmit in D2D mode or remain silent. Similarly, [10] analyses the joint power control and frequency reuse of D2D pairs in the same scenario presented in [9]. Also in line with [9] and [10], *J. Huang et al.* proposed in [11] a significant step towards more efficient D2D communications by expanding these communications from intra-cell environments to inter-cell environments. The proposal, which is based on game theory, shows clearly the potential of this inter-cell cooperation. Yet, the scenario is restricted to a use case with disjoint sets of D2D and cellular users.

The works in [12]-[17] study the performance of D2D communications in multicast groups, where all users download a common content from the BS via a cluster head user. It is shown that a better efficiency in the resources' usage can be achieved in these scenarios, although the gain is bounded by the lowest quality link between the cluster head and the rest of users of the multicast group. In detail, the authors in [12] derive expressions to select the optimal number of D2D retransmitters in a multicast group, and [13] proposes a Conventional Multicast Scheme (CMS) to decide whether a user should be served by the BS or by the cluster head.

Similarly, *Meshgi et al.* [14] maximize the throughput in a single cell scenario with multicast D2D groups by proposing a heuristic resource allocation algorithm that achieves near optimal performance. In [15] the authors address the multicast clustering by setting up a Primary Cluster Head (PCH) and a Secondary Cluster Head (SCH). The PCH and the SCH are selected based on their residual energy and the received Signal to Interference Noise Ratio (SINR). Similarly, in [16] the authors analyse a set of different strategies for the establishment of multicast clusters. The work shows that D2D-based multicast clustering can increase the system capacity, although it is sensitive to parameters such as clusters' dimension. Finally, key features required to support network controlled D2D-based multicasting are analysed in [17].

Although the works described so far address the problem of D2D clustering in cellular networks, they are constrained by two assumptions: i) only downlink traffic is considered ; ii) the same content is delivered to all users in the cluster/group. Cooperative D2D moves a step forward in [18], where authors formulate the clustering problem as the maximization of the throughput constrained by energy efficiency. The proposed algorithm outperforms the results obtained without clustering but it neglects two important aspects: i) the mobility, that impacts on the quality of the links and on the role played by each user; ii) the energy consumption of the relay/cluster head could be higher in idle state than in transmission state.

In contrast with the State of the Art, we propose clustering algorithms aimed to improving the resources' utilization efficiency in scenarios where uplink (UL) and downlink (DL) traffic are considered in a LTE-A FDD system. Our algorithms are based on a previous work [19], where the clustering algorithm CORE was proposed. CORE restricted the creation of spectral efficient clusters to users within the same cell thus limiting the achieved gains in dense Heterogeneous Networks (HetNets). In order to go beyond this constraint, we propose a new algorithm, namely enhanced Clustering Optimization for Resources' Efficiency (eCORE), that extends clustering to multi-cell deployments. Specifically, eCORE is based on the cooperation among devices by leveraging the D2D communication concept, initially introduced in the framework of LTE-A to support *Proximity-based Services (ProSe)* for public safety [6]. In our solution, the mobile devices create spectral efficient clusters with a single cluster head (CH) characterized by good quality links with the serving BS and with the rest of cluster members. In eCORE clusters can be created among users from different cells as long as they result in a decrease of the required resources. The cluster head is responsible for receiving and forwarding packets from/to the BS and the cluster members. As traffic is more intense in the DL and D2D communications are usually carried out over UL bands to limit the interference caused to neighbouring users [9], [11], the proposal benefits from the imbalance between UL and DL traffic intensity and the high channel gain of D2D communications to increase the capacity of the network. Although the dynamic adaptation to the imbalance between UL and DL traffic has been addressed in [20], [21] for TDD HetNets, the problem is more challenging in FDD systems, where transferring traffic from DL to UL is more complex.

Following this rationale, it is shown that the capacity of the network can be further increased by establishing nonspectral efficient clusters that balance UL and DL traffic. This is the objective of the Clustering algorithm for Load Balancing (CaLB), the second proposed algorithm. CaLB shows that in some cases clustering can be beneficial despite increasing the number of required

spectrum resources. Yet, the proposed solutions present challenges in terms of energy consumption of the cluster head that are studied and addressed along the paper by complementing eCORE with the Clustering Energy Efficient Algorithm (CEEa). CEEa limits the cluster head energy overconsumption, thus minimizing the disincentive in the creation of clusters. Both CaLB and CEEa are designed to be executed after eCORE to improve its performance, but not to be implemented in a standalone manner.

In a nutshell, the three clustering proposals are a costefficient RAN densification solution based on D2D for FDDLTE networks, and the work's contributions are the following:

- A RAN densification solution based on D2D clustering in the framework of FDD LTE-A is presented to improve the spectral efficiency. The algorithm, which is an extension of CORE [19] and is denoted by eCORE, exploits the spatial diversity provided by the high density of users and the imbalance between UL and DL traffic. Contrary to CORE, eCORE enables the creation of inter-cell clusters.
- A load balancing clustering algorithm, namely CaLB, is proposed to increase the capacity of the network. In contrast with eCORE that creates spectral-efficient clusters, CaLB complements eCORE by establishing nonspectral efficient clusters. The capacity gain results from the UL and DL load balancing.
- We propose a complementary algorithm to eCORE, known as CEEa, that compensates the energy overconsumption suffered by cluster heads in eCORE. CEEa benefits from mobility and forces reclustering by limiting the time during which users play the role of cluster head to reduce the energy overconsumption.

III. Clustering Proposal

The proposed clustering solutions described in the sequel (eCORE, CaLB and CEEa) are all based on a set of premises:

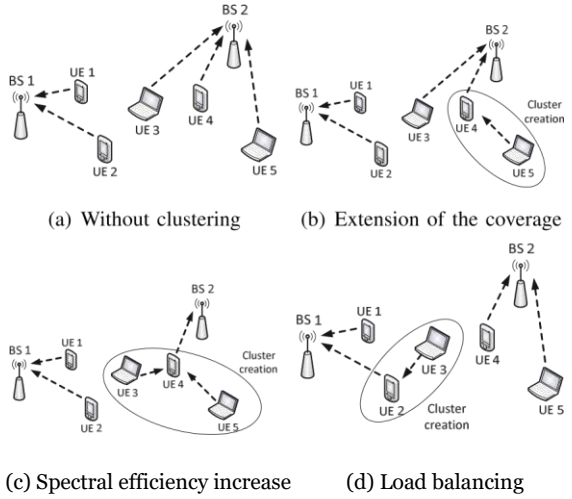
i) each cluster has a single cluster head; ii) each user/device can be direcly served by a BS, play the role of cluster head, or join a cluster to be served by a BS through the corresponding cluster head, but no more than a single role can be played simultaneously; iii) intra-cluster communications are D2D transmissions carried out in the UL band to limit the interference [9], [11]. In FDD, the creation of a cluster is translated into a transfer of resources' utilization from the DL band to the UL band, which is usually underutilized. For instance, the DL traffic of a clustered user is first served with DL resources (from the BS to the cluster head) and subsequently with UL resources in the D2D communication from the cluster head to the cluster member. If we assume that the channel gain from the BS to the cluster head is higher than the channel gain from the BS to the rest of clustered users, the required DL resources are reduced. Although the three algorithms share a set of premises, they differ in their objectives. Thus, in eCORE clustering is aimed to reduce the number of required resources (Section III-E). In CaLB, the creation of a cluster must decrease the load of the DL (Section III-F). Finally, in CEEa the energy overconsumption of cluster heads must be compensated (Section IV-C).

This Section is focused on the algorithms that improve the capacity of the network, i.e. eCORE and CaLB. The Section first describes a set of use cases where clustering can be applied (Section III-A). Then, the system model is stated in Section III-B and the general expressions of the required resources in UL and DL are developed in Section III-C. Based on these expressions, the optimal clustering problem aimed to minimize the total number of resources is formalized in Section III-D. Finally, eCORE is proposed in Section III-E as a low complexity algorithm and CaLB is introduced in Section III-F to further enhance the capacity.

A. Use cases

The clustering proposal addresses three use cases: the service of user equipments (UE) in coverage gaps, the enhancement of spectral efficiency and the load balancing. Fig. 1 sketches the initial scenario with 6 UEs served by one of the BSs (Fig. 1(a)) and the following cases:

- Extension of the coverage (Fig.1(b)): Assume that UE5 is in a coverage gap. If the quality of the links UE5-UE4 and UE4-BS2 is good enough, the clustering of UE4 (cluster head) and UE 5 can guarantee the service of the latter.
- Spectral efficiency enhancement (Fig.1(c)): Clustering UE5 and UE3 with UE4 (the cluster head) increases spectral efficiency if: i) the quality of links UE3-UE4, UE5-UE4 and UE4-BS2 is significantly better than the quality of links UE5-BS2 and UE3-BS2; ii) downlink is highly loaded while uplink is less loaded.
- Load balancing (Fig.1(d)): If BS2 is highly loaded and BS1 is less loaded, the clustering of UE3 with UE2 (cluster head) can balance the load of BS2 to BS1.



(a) Without clustering (b) Extension of the coverage

(c) Spectral efficiency increase (d) Load balancing

Fig. 1. Example of possible clustering use cases

B. System Model

The network is composed of a set of FDD-LTE BSs (macro eNBs and/or SCs), namely B, covering the scenario and serving a set of users, denoted by U. Each user $i \in U$ is connected to a BS $k \in B$ according to any of the existing cell association algorithms (e.g. algorithms based on Reference Signal Received Power (RSRP) with or without Cell Range Expansion). The set of users connected to BS k is referred to as U_k. As users are only served by a single BS simultaneously,

$U = \bigcup_{k\in B} U_k$ and $\bigcap_{k\in B} U_k = \emptyset$. Each user $i \in U$ is characterized by its traffic profile $\pi_i = (R_i^d, R_i^u)$, composed of the average transmission rate in the DL R_i^d and in the UL R_i^u. As in general UL and DL traffic are unbalanced, $R_i^u = \alpha_i R_i^d$, with $0 \le \alpha_i \le 1$. In LTE-A the transmission rate between two nodes depends on the selected Modulation and Coding Scheme (MCS), which is determined by the maximum allowed bit error rate (BER) and the SINR. Accordingly, the number of bits transmitted by user i during a subframe time $T^s = 1$ms, defined as Transport Block Size (TBS), can be approximated by an attenuated and truncated form of Shannon bound. Thus, the TBS of a transmission from i to j in the band v ($v = u$ if the transmission is in the UL band and $v = d$ if it is in the DL band) is approximatted as

$$\eta_{i,jv} = T^s r W \log_2(1 + \gamma_{i,jv}) \quad (1)$$

where r is the attenuating factor, W is the bandwidth of a Physical Resource Block (PRB) and $\gamma^v_{i,j}$ is the SINR received at j when data is transmitted by i. If the transmitter is a UE and the receiver is a BS, $i \in U$ and $j \in B$; if the transmitter is a BS and the receiver is a UE, $i \in B$ and $j \in U$; finally, if both transmitter and receiver are users in D2D mode, $i,j \in U$.

C. Resources required with and without clustering

The spectral efficiency is measured in bps/Hz. Therefore, the enhancement of the spectral efficiency is equivalent to the minimization of the PRBs required to serve a given traffic. Based on the definitions stated above, the expected number of PRBs required in the scenario to serve all the users in the UL (N^u) and in the DL (N^d) can be expressed as

$$N^d = \sum_{k\in\mathcal{B}} N_k^d = \sum_{k\in\mathcal{B}} \sum_{i\in\mathcal{U}_k} \frac{R_i^d T^s}{\eta_{k,i}^d} = \sum_{k\in\mathcal{B}} \sum_{i\in\mathcal{U}_k} R_i^d \phi_{k,i}^d \quad (2)$$

$$N^u = \sum_{k\in\mathcal{B}} N_k^u = \sum_{k\in\mathcal{B}} \sum_{i\in\mathcal{U}_k} \frac{R_i^u T^s}{\eta_{i,k}^u} = \sum_{k\in\mathcal{B}} \sum_{i\in\mathcal{U}_k} \alpha_i R_i^d \phi_{i,k}^u \quad (3)$$

where N_k^d and N_k^u are the expected number of PRBs per subframe required by base station k (eNB or SC) in DL and UL. For simplicity, we define $\phi_{k,i}^d = \frac{T^s}{\eta_{k,i}^d}$ and $\phi_{i,k}^u = \frac{T^s}{\eta_{i,k}^u}$.

Let us consider that groups of users can create clusters. Each cluster u is composed of cluster member users, among which a single user plays the role of cluster head. Hereafter, the set of users in cluster u will be denoted by C_u, and the cluster head by $h_u \in C_u$. The cluster head h_u is responsible for receiving the DL traffic of all cluster members from the BS and forward it to the corresponding cluster member. Likewise, for the UL traffic, the cluster head receives the traffic from the rest of the cluster members and forwards it to the BS. We will denote the set of all the clusters in the scenario by $C = \bigcup_u C_u$. Note that the communication within the cluster is carried out over the UL band to minimize the interference caused to the users outside the cluster. Therefore, intra-cluster communications are always carried out in the UL band. In real FDD networks, BSs are always full-duplex; conversely, user devices can be half-duplex (Half-Duplex FDD devices) or full-duplex (Full-Duplex FDD devices)[1]. We define the set of cluster heads as $H = \{h_u\}_{\forall u}$, and the set of cluster heads connected to BS k as $H_k = H \cap U_k$. Accordingly, the expected number of PRBs required in the DL band (N^d) and in the UL band (N^u) with clusters are written as

1

$$\tilde{N}^d = \sum_{k\in\mathcal{B}} \tilde{N}_k^d = \underbrace{\sum_{k\in\mathcal{B}} \sum_{i\in\mathcal{U}_k\setminus\mathcal{C}} \varphi_{dk,i} R_i^d}_{\text{non-clustered users}} + \underbrace{\sum_{k\in\mathcal{B}} \sum_{h_u\in\mathcal{H}_k} \varphi_{dk,h_u} R_{h_u}^d}_{\text{clustered users}} \quad (4)$$

$$\tilde{N}^u = \sum_{k\in\mathcal{B}} \tilde{N}_k^u = \underbrace{\sum_{\mathcal{C}_u\subseteq\mathcal{C}} \sum_{i\in\mathcal{C}_u\setminus\{h_u\}} \left(\phi_{i,h_u}^u R_i^u + \phi_{h_u,i}^u R_i^d\right)}_{\text{transmissions within the cluster}} \quad (5)$$

$$+ \sum_{k\in\mathcal{B}} \sum_{i\in\mathcal{U}_k\setminus\mathcal{C}} \varphi_{ui,k} R_{iu} + \sum_{k\in\mathcal{B}} \sum_{h_u\in\mathcal{H}_k} \sum_{i\in\mathcal{C}_u} \varphi_{uh_u,k} R_{iu}$$

where $\tilde{N}_k^d$ and $\tilde{N}_k^u$ are the expected number of PRBs required by base station k in the DL and UL. As observed in (5), intra-cluster communications do not interfere with UL communications from the cluster head to the BS (they are not simultaneous). Moreover, the number of PRBs required in the scenario is a function of the SINR, which in turn depends on the cell association algorithm. Yet, (2)-(5) are valid for a given SINR level and regardless of the cell association algorithm.

D. Optimal clustering for spectral efficiency

The aim of the clustering technique presented herein is the minimization of the spectral resources utilization, i.e. $\tilde{N} = \tilde{N}^u + \tilde{N}^d$. As it can be observed, the minimization of the required resources is an association problem, where a user must be associated to a BS directly or through a cluster head. Let us define the association matrix $\mathbf{X} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{B}|}$, where $|\cdot|$ is the cardinality operator of a set, and the elements of $\mathbf{X}$ are $x_{i,k} = 1$ if user i is directly served by BS k and $x_{i,k} = 0$ otherwise. We define $\mathbf{Y} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{U}|}$ as the intra-cluster association matrix, with the elements of $\mathbf{Y}$ such that $y_{j,i} = 1$ if user j is connected to a BS through user i (i is the cluster head) and $y_{j,i} = 0$ otherwise. Using matrices $\mathbf{X}$ and $\mathbf{Y}$, the total number of required resources can be expressed as,

$$\tilde{N}(\mathbf{X},\mathbf{Y}) = \sum_{i\in\mathcal{U}} \sum_{k\in\mathcal{B}} \Big[x_{i,k} R_i^d \left(\phi_{k,i}^d + \alpha_i \phi_{i,k}^u\right) + \sum_{j\in\mathcal{U}\setminus\{i\}} y_{j,i} R_j^d \left(\phi_{k,i}^d + \alpha_j \phi_{i,k}^u + \phi_{i,j}^u + \alpha_j \phi_{j,i}^u\right) \Big] \quad (6)$$

Therefore, the optimization problem is formulated as

$$\min_{\mathbf{X},\mathbf{Y}} \quad \tilde{N}(\mathbf{X},\mathbf{Y}) \quad (7)$$

$$\text{s.t.} \quad x_{i,k}, y_{i,j} \in \{0,1\}, \qquad \forall i,j \in \mathcal{U}, \forall k \in \mathcal{B} \quad (7a)$$

$$\sum_{k\in\mathcal{B}} x_{i,k} + \sum_{j\in\mathcal{U}} y_{i,j} = 1, \qquad \forall i \in \mathcal{U} \quad (7b)$$

$$\sum_{k\in\mathcal{B}} \sum_{i\in\mathcal{U}} x_{i,k} \geq 1, \quad (7c)$$

$$x_{i,k} - y_{j,i} \geq 0, \qquad \forall i,j \in \mathcal{U},\ k\in\mathcal{B} \quad (7d)$$

$$y_{i,j} + y_{j,i} \leq 1, \qquad \forall i,j \in \mathcal{U} \quad (7e)$$

$$y_{i,i} = 0, \qquad \forall i \in \mathcal{U} \quad (7f)$$

The problem in (7) is an integer (binary) linear programming problem (ILP) (7a), where UEs can be served by a BS or a cluster head (7b) and at least one UE is connected to a BS (7c). A cluster can only be created if the cluster head is directly connected to a BS (7d), since multi-hops are not allowed within the cluster. A clustered user can either be a cluster head or be associated to a cluster head (7e).

E. Enhanced Clustering Optimization for Resources Efficiency (eCORE)

As all 0-1 ILP problems are NP-hard [22], (7) is NPhard. A low complexity algorithm ($O(n^3)$), namely enhanced Clustering Optimization for Resources' Efficiency (eCORE), is presented. Based on the expressions derived in Section III-C, some results can be enunciated.

Lemma 1. *The number of resources required to serve a user $i \in \mathcal{U}_k$ is reduced when it joins a cluster with cluster head $j \in \mathcal{U}_q$ iff* $(\varphi_{dk,i} - \varphi_{dq,j} - \varphi_{uj,i}) + \alpha_i(\varphi_{i,ku} - \varphi_{uj,q} - \varphi_{ui,j}) > 0$.

Proof. Lemma 1 is calculated from the difference between PRBs required in (2)-(3) and PRBs required in (4)-(5). □

Lemma 2. *Given two users $i \in \mathcal{U}_k$ and $j \in \mathcal{U}_q$, the clustering gain $G_{i,j}$ when j is the cluster head is defined as,*

$$G_{i,j} = R_i^d(\phi_{k,i}^d - \phi_{q,j}^d - \phi_{j,i}^u) + \alpha_i R_i^d(\phi_{i,k}^u - \phi_{j,q}^u - \phi_{i,j}^u) \quad (8)$$

The set of possible cluster heads of user i is defined as $\mathcal{Y}_i = \{j : G_{i,j} > 0\}$. For two users $i \in \mathcal{U}_k$ and $j \in \mathcal{U}_q$, if $\mathcal{Y}_i = \{j\}$ and $\mathcal{Y}_j = \emptyset$, then i and j will create a cluster in which j is the cluster head. Conversely, if $\mathcal{Y}_j \neq \emptyset$, $j \in \mathcal{Y}_i$ and $|\mathcal{Y}_i| > 1$, i and j will create a cluster where j plays the role of cluster head if $G_{i,j} > G_{j,n} + G_{i,t}$ for $\forall n \in \mathcal{Y}_j$ and $\forall t \in \mathcal{Y}_i$.

Proof. Using Lemma 1, the clustering gain achieved by a cluster equals the aggregation of clustering gains of all cluster members. Thus, Lemma 2 can be derived from (2)-(5). □

According to Lemmas 1 and 2, clustering is not limited to users within the same cell. A cluster may be created by users served by different cells (eNBs and/or SCs). The proposed eCORE, described in Algorithm 1, is based on Lemmas 1 and 2 and it is aimed to create clusters that improve the total spectral efficiency. The key parameter of the algorithm is the clustering gain ($G_{i,j}$) defined in Lemma 1. eCORE starts with the computation of clustering gains for the different UEs, and initializing for each user i the set Y_i of users j that would result in a positive clustering gain, i.e. $G_{i,j} > 0$ (line 1). As mentioned, eCORE only considers single-hop intracluster communications to limit complexity and signalling. Accordingly, the term *conflict* is used in the sequel to describe situations where a user i has a positive clustering gain with a user j ($G_{i,j} > 0$) that, in turn, has a positive clustering gain with a third user n ($G_{j,n} > 0$). In these conflicting situations, either user j becomes the cluster head of user i or user n becomes the cluster head of user j, but not both of them. Both situations are enunciated in Lemma 2 and implemented in Algorithm 1. Initially, eCORE clusters users without *conflicts* (lines 3-17). In the second part, eCORE resolves the unsolved *conflicts*, stored in the set A (see Alg. 1), by selecting the option that provides the highest clustering gain (lines 19-32).

The computational complexity is reduced by dividing the problem into two steps: the first step (lines 1-17) discards unfeasible clustering solutions, whereas the second step (lines 18-32) resolves conflicting cases. The first step identifies potential cluster heads by figuring out if any of the associations would result in a reduction of the required resources. If not, that association is discarded (it is unfeasible for a spectral efficient cluster). In practice, the identification of potential cluster heads does not require a comparison of all users, since users farther than the D2D range can be discarded at the beginning. In a nutshell, eCORE is an algorithm that checks which clusters can reduce the overall required PRBs. With this, not only the overall number of PRBs is reduced but traffic imbalance is decreased by transferring load from the DL to the UL.

F. Clustering algorithm for Load Balancing (CaLB)

eCORE takes advantage of UL and DL traffic imbalance to decrease the DL usage at the expense of an increase of the UL usage (only if the DL usage decrease is higher than the UL

Algorithm 1: Enhanced Clustering Optimization for Resources Efficiency (eCORE) $(O(n^3))$

Data: $\mathcal{U}$, $\phi^d_{k,i}$, $\phi^u_{i,k}$, $\phi^u_{i,j}$, $\phi^u_{j,i}$

Result: Set of Clusters C = $\bigcup_u C_u$

1 Initialize the set of possible CHs (Y_i), $\forall i \in U$
2 A = ∅: A is a set of UEs with $Y_i \neq \emptyset$
3 for $i \in U$ do
4 if $Y = \emptyset$ then
5 j = argmax ($G_{i,j}$), $\forall j \in Y_i$ and $G_i^{max} = G_{i,j^*}$
6 if $Y_{j^*} = \emptyset$ then
7 if *UE j^* is CH of cluster u* then
8 $C_u \leftarrow C_u \cup \{i\}$
9 else
10 j^* is CH of a new cluster u: $C_u = \{j^*, i\}$
11 end
12 $Y_i = \emptyset$
13
14 $A \leftarrow A \cup \{i\}$
15
16
17 end
18 UEs in A sorted in G_i^{max} descending order
19 while $A \neq \emptyset$ do
20 $i \leftarrow$ First UE in A
21 $j^* = \arg\max_j (G_{i,j} - G_{j,n} - G_{i,t})$, $\forall j,t \in Y_i$ $\forall n \in Y_j$ with { $Y_t = \emptyset$ or $t \in A$} and { $Y_n = \emptyset$ or $n \in A$}
22 if $G_{i,j^*} \leq 0$ then
23 $Y_i = \emptyset$ and $A \leftarrow A \backslash \{i\}$
24 else
25 if *UE j^* is CH of cluster u* then
26 $C_u \leftarrow C_u \cup \{i\}$ and $A \leftarrow A \backslash \{i\}$
27 else
28 UE j^* is CH of a new cluster u ($h_u = j^*$)
29 $Y_{j^*} = \emptyset$, $C_u = \{j^*, i\}$ and $A \leftarrow A \backslash \{i, j^*\}$
30 end
31 end
32 end

j = argmax (G

Algorithm 1, only if the DL usage decrease is higher than the UL usage increase). This fact limits the maximum capacity. Let us define the maximum number of PRBs allocated in the DL and in the UL to BS^k as $N_k^{d,max}$ and $N_k^{u,max}$. The saturation point is of the cell (when the cell capacity reaches its limit) is defined as the situation when either the DL or the UL cannot serve more traffic. Mathematically, the saturation is met when $\min(N_k^{u,max} - N_k^u, N_k^{d,max} - N_k^d) \approx 0$, where N_k^u and N_k^d are

the PRBs used in each band in BS k without clustering. As DL is generally more loaded, when $N_k^d \gg N_k^u$ and $N_k^d \approx N_k^{d,max}$ it may be convenient to create clusters to increase the capacity even at the expense of a spectral efficiency decrease.

Lemma 3. *Given a BS $k \in B$ with an average number of required PRBs without clustering in the downlink and in the uplink equal to $N_k{}^d$ and $N_k{}^u$, respectively, the cell capacity is increased after creating the cluster u (with $C_u \subseteq U_k$) if*

$$\Delta N_k^u \leq \Delta N_k^d + (N_k^{u,max} - N_k^u) - (N_k^{d,max} - N_k^d) \quad (9)$$

even if the clustering gain is negative or null, i.e. $G_{Cu} = P_{i \in Cu \setminus \{hu\}} G_{i,hu} = -(\Delta N_{kd} + \Delta N_{ku}) \leq 0$[2], where

$$\Delta N_k^d = \sum_{i \in \mathcal{C}_u \setminus \{h_u\}} R_i^d \left(\phi_{k,h_u}^d - \phi_{k,i}^d\right) \quad (10)$$

$$\Delta N_k^u = \sum_{i \in \mathcal{C}_u \setminus \{h_u\}} R_i^d \left[\phi_{h_u,i}^u + \alpha_i \left(\phi_{h_u,k}^u + \phi_{i,h_u}^u - \phi_{i,k}^u\right)\right] \quad (11)$$

Proof. We define the number of available PRBs in the limiting band (the most loaded band) as $A = \min(N_k^{u,max} - N_k{}^u, N_k{}^{d,max} - N_k{}^d)$. If uplink is the limiting band, then $A = (N_k{}^{u,max} - N_k{}^u)$. Knowing that, by definition, $\Delta N_k{}^u > 0$ and $\Delta N_k{}^d < 0$, it can be found that (9) is not true. Therefore, $A = (N_k^{d,max} - N_k^d)$ must be true (the downlink is more loaded). Rearranging (9) we obtain that $(N_k^{d,max} - N_k^d - \Delta N_k^d \leq N_k^{u,max} - N_k^u - \Delta N_k^u)$, and therefore the number of available resources in the limiting band after clustering is $A' = N_k^{d,max} - N_k^d - \Delta N_k^d$. As $\Delta N_k^d < 0$, then $A' > A$. □

Lemma 4. *Given two users $i,j \in U_k$, where user i is not clustered and user j is a cluster head, the number of PRBs required in the DL decreases when i joins the cluster headed by j if $\Delta N_k{}^d(i,j) < 0$, with $\Delta N_k{}^d(i,j) = R_i{}^d(\varphi^d{}_{k,j} - \varphi^d{}_{k,i})$. If j is not clustered, and given two additional users m and n that minimize $x_{i,j} = \Delta N_k{}^d(i,j) - \Delta N_k{}^d(i,m) - \Delta N_k{}^d(j,n)$, user i must join the cluster headed by j to maximize the reduction in the required PRBs if $x_{i,j} \leq 0$. Conversely, if $x_{i,j} > 0$, users i and m should create a cluster and users j and n should create a second cluster.*

Proof. The first case is trivial, since $\Delta N_k^d(i,j)$ is, by definition, the increase in the downlink PRBs. If it is negative, the number of required PRBs decreases. If user j is not a cluster head (second case), user j can become the cluster head of user i or the cluster member of an alternative cluster. In that case, if $n = \text{argmin}_q\{\Delta N_k^d(j,q)\}$ and $m = \text{argmin}_q\{\Delta N_k^d(i,q)\}$, the maximum overall reduction of PRBs would be $\Delta N_k{}^d(i,m) + \Delta N_k{}^d(j,n)$. Therefore, the maximum reduction of the PRBs in the downlink would result from clustering i and j if $\Delta N_k{}^d(i,j) < \Delta N_k{}^d(i,m) + \Delta N_k{}^d(j,n)$ (i.e. if $x_{i,j} < 0$). □

In order to further extend the capacity provided by eCORE, CaLB is proposed, mainly based on Lemmas 3 and 4. It is aimed to improve the capacity when no additional spectral efficient clusters can be created, the DL reaches the capacity limit and the UL is still unloaded (see Alg. 2). Therefore, CaLB is always run after the execution of eCORE. The inputs of CaLB are the set of users and clusters created by eCORE and two load thresholds, $n^d{}_{min}$ and $n^u{}_{min}$ for the DL and UL, respectively. These thresholds are used to determine whether a BS DL and UL are loaded or not: if the number of available PRBs in the DL, denoted in Alg. 2 by n^d (line 2), is below $n^d{}_{min}$, the DL of the BS is loaded; if the number of available PRBs in the UL, denoted by n^u (line 2), is higher than $n^u{}_{min}$, the UL of the BS is considered unloaded. Only in this case, each BS executes CaLB and triggers the clustering procedure (line 6). The algorithm establishes the clusters that reduce the load in the DL, by joining users to existing clusters or by establishing new clusters. To do that, all possible pairs of users (defined as Q_k in Alg. 2) are ordered according to the reduction that would cause in the number of required DL PRBs if clustered (i.e. $\Delta N_k{}^d(i,j)$). There are constraints in this clustering process to prevent spectral efficient clusters (established by eCORE) from being destroyed. First, the cluster head of an existing cluster can serve new users by enlarging the cluster; that is, unclustered users can join existing clusters. A cluster head will not leave an existing cluster to become the cluster member of a new cluster. Finally, the clustering of a user must always result in a decrease of the DL resources; therefore, the channel gain to the BS is higher for the cluster head than for the rest of cluster members ($\phi_{k,h_u}^d < \phi_{k,i}^d$ when user i joins a cluster head $h_u \in U_k$). Based on these constraints and on Lemma 4, CaLB favours the clustering until the number of available PRBs in the DL is larger than n_{min}^d or the number of available PRBs in the UL reaches the minimum, n_{min}^u.

[2]

Algorithm 2: Clustering alg for Load Balancing (CaLB)

Data: $n_{dmin}, n_{umin}, \{U_k, H_k, N_{kd,max}, N_{ku,max}, \tilde{N}_{kd}, \tilde{N}_{ku}\}_{\forall k \in B}$

Result: Set of Clusters $C = \bigcup_u C_u$

for $k \in B$ do
  $n^d = N_k^{d,max} - \tilde{N}_k^d$ and $n^u = N_k^{u,max} - \tilde{N}_k^u$
  if $n^d < n_{min}^d$ **then**
    Define $\mathcal{Q}_k = \{(i,j) : \phi_{k,j}^d < \phi_{k,i}^d, \forall i \in \mathcal{U}_k \setminus \mathcal{C}, \forall j \in (\mathcal{U}_k \setminus \mathcal{C}) \cup \mathcal{H}_k\}$
    $(i,j) \in \mathcal{Q}_k$ are sorted in ascending order in $\mathcal{Q}_k$ based on $\Delta N_k^d(i,j) = R_i^d(\phi_{k,j}^d - \phi_{k,i}^d)$
    while $\mathcal{Q}_k \neq \emptyset$ **and** $n^u \geq n_{min}^u$ **and** $n^d < n_{min}^d$ **do**
      $(i,j) \leftarrow$ First pair of nodes in $\mathcal{Q}_k$
      $\Delta N_k^u(i,j) = R_i^d \left(\phi_{j,i}^u + \alpha_i(\phi_{j,k}^u + \phi_{i,j}^u - \phi_{i,k}^u)\right)$
      if $n^u + \Delta N_k^u(i,j) \geq \epsilon^u$ **then**
        if $\exists u : j = h_u$ **then**
          $\mathcal{C}_u \leftarrow \mathcal{C}_u \cup \{i\}$
          $\mathcal{Q}_k \leftarrow \mathcal{Q}_k \setminus \{(i,m) : \forall m \neq i\}$
          $n^v \leftarrow n^v + \Delta N_k^v(i,j)$ for $v = \{u, d\}$
        else
          Association according to Lemma 4 and update of $\mathcal{Q}_k$, $\mathcal{C}$, n^d and n^u
        end
      else
        $\mathcal{Q}_k \leftarrow \mathcal{Q}_k \setminus \{(i,j)\}$
      end
    end
  end
end

To sum up, CaLB resumes the clustering process initiated by eCORE. The created clusters are not spectral efficient, but reduce the UL and DL imbalance. CaLB is appropriate when the DL is highly loaded.

IV. Impact on Energy Consumption

eCORE and CaLB rely on the set-up of cluster heads under the conditions stated in Section III. However, the role of cluster head entails energy consuming tasks, e.g. receiving and retransmitting the data of the rest of cluster members. Therefore, the role of cluster head can cause early battery drain. In this section, the expression of the energy consumption of each stakeholder is derived, and the mitigation of possible energy overconsumption of the clustering approach is studied. In the following, the energy consumption expressions are derived in Section IV-A. In Section IV-B these expressions are used to modify the optimal clustering problem defined in Section III-D and to include energy overconsumption limits. Section IV-C proposes a low complexity Clustering Energy Efficient algorithm (CEEa).

A. Energy Consumption Analysis

The energy consumption of a UE depends on two main factors: the Radio Resource Control (RRC) state of the device, that can be RRC CONNECTED or RRC IDLE, and the transmitted power [23]. Let us define the RRC state space as $S = \{I, C_{tx}, C_{rx}\}$, where I stands for the RRC IDLE state and the RRC CONNECTED state has been decoupled into two states, the transmitting state C_{tx} and the receiving state C_{rx}. We define $S_C = \{C_{rx}, C_{tx}\}$. Based on this, the energy consumed by user i during a subframe time T^s is given by $E_i = T^s(P_{si} + P_{txi})$, where P_{si} is the power consumed when user i is in state $s_i \in S$ and P_{txi} is the transmitted power. The transmitted power differs in D2D mode (the intra-cluster communications) and in the communication with the BS, and for a user i is described in LTE [24] by,

$$P_{tx_i} = \begin{cases} M_i P_0 h_{i,k}^{-\xi} & \text{if connected to BS } k \\ M_i P_{d2d} & \text{if connected in D2D mode} \end{cases} \quad (12)$$

where M_i is the number of PRBs scheduled for user i, P_o is the target received power at BS k, $h_{i,k}$ is the channel gain between user i and BS k, $\xi \in [0,1]$ is the compensating factor and P_{d2d} is the transmitted power per PRB in D2D mode. In the following the role played by user i is denoted by $\rho_i = \{H, M, N\}$, with $\rho_i = H$ for a cluster head, $\rho_i = M$ for the rest of the cluster members and $\rho_i = N$ for the nonclustered users. Note that a user i is directly connected to a BS if $\rho_i = \{H, N\}$, and it is in D2D mode if $\rho_i = M$.

Each user is characterized by its profile π_i, the role ρ_i and the location (channel gains with the rest of UEs and BSs), and the expected energy consumed during a subframe is expressed as $E[E_i|\rho_i] = T^s E[P_i|\rho_i] = T^s E[P_{si}|\rho_i] + T^s E[P_{txi}|\rho_i]$ (13) where, by definition,

$$E[P_{si}|\rho_i] = P\{s_i = I|\rho_i\} P_I + P\{s_i \in S_C|\rho_i\} P_C \quad (14)$$

where P_I is the power consumed in state $s_i = I$ and P_C is the power consumed in state $s_i \in S_C$. Note that the probability of being in state s_i depends on the role of the user. For instance, $P\{s_i = I|\rho_i = H\} \leq P\{s_i = I|\rho_i = N\}$. Taking into account that the cluster head forwards both the UL traffic of all cluster members to the BS, and the DL traffic to the cluster members (intra-cluster communications in D2D mode), the expected transmitted power of a user i connected either to BS k or to cluster head h_u can be easily found using (12).

□ $-\xi$ $d\varphi_{ui,k}$ E if $\rho_i = N$ if $\rho_i = M$

□□□□□□□□□ if $\rho_i = H$

$PP_{0d02}h_{di,k-i,k}\alpha_{\xi i}\varphi\alpha R_{i,kuiid}R\varphi_i X_{i,hu\ u}\alpha_j R_{jd}+$

$P\ h$

$[P_{txi}|\rho_i] =$

□□□□□□□□□ $+P_{d2d}$
$j\in C X_{uj}\backslash\{\in Ci\}_u R_{jd}\varphi_{ui,j}$

(15)

B. Optimal clustering with energy consumption constraints

In order to limit the energy consumed by the cluster head, the problem defined in (7) must be modified to include the energy consumption constraint. If we define $w > 0$ as the maximum allowed increase of the expected power/energy of a cluster head, the expected power consumed by a cluster head should not exceed the power consumed if it was not clustered:

$$\mathrm{E}[P_i|\rho_i = H] \leq (1 + w)\mathrm{E}[P_i|\rho_i = N] \quad (16)$$

As shown in (13)-(15), the total power depends on the probability $\mathrm{P}\{s_i \in \mathrm{S}_C|\rho_i\}$ and on the transmitted power. Regarding the former, when the user i is the cluster head, the probability can be divided into two components: the probability of $s_i \in \mathrm{S}_C$ due to the time required to transmit/receive its own traffic from/to the BS^k ($\theta^N_{i,k}$) and due to the time required to forward the traffic of the rest of the cluster members ($\theta^H_{i,j,k}$, for all users j in the cluster).

$$\mathrm{P}\{s_i \in \mathrm{S}_C|\rho_i = H\} = \mathrm{X}(x_{i,k}\theta_{i,kN} + \mathrm{X}y_{j,i}\theta_{i,j,kH}\ ) \quad (17)$$
$$k\in\mathrm{B} \qquad j\in\mathrm{U}$$

where $x_{i,k} = 1$ if user i is served by BS k and $x_{i,k} = 0$ otherwise; and $y_{j,i} = 1$ when user i is the cluster head of user j and $y_{j,i} = 0$ otherwise (expressions for $\theta_{i,k}{}^N$ and $\theta_{i,j,k}{}^H$ are derived in Appendix A of [25]). By using (13)-(15) and (17), the components of (16) can be written as

$$\mathbb{E}[P_i|\rho_i = H] = P_I + \sum_{k\in\mathcal{B}} x_{i,k}\left(\Delta P_{CI}\theta^N_{i,k} + \mathbb{E}[P_{tx_i}|\rho_i = N]\right) + \sum_{j\in\mathcal{U}} y_{j,i}(\theta^H_{i,j,k}\Delta P_{CI} + \mathbb{E}[P_{tx_i}|\rho_i = H, j]) \quad (19)$$

$$\mathbb{E}[P_i|\rho_i = N] = \theta^N_{i,k}\Delta P_{CI} + P_I + \mathbb{E}[P_{tx_i}|\rho_i = N] \quad (18)$$

where $\Delta P_{CI} = P_C - P_I$ and $\mathrm{E}[P_{txi}|\rho_i = H,j]$ is the power consumed by the cluster head attributable to the traffic of cluster member j, and it is defined as

$$\mathbb{E}[P_{tx_i}|\rho_i = H, j] = R^d_j\left(P_0 h^{-\xi}_{i,k}\phi^u_{i,k}\alpha_j + P_{d2d}\phi^u_{i,j}\right) \quad (20)$$

Parameter w must be selected to limit the energy overconsumption of cluster heads while allowing the creation of clusters. For instance, if only a 5% power increase is allowed (w = 0.05), cluster heads will not suffer from rapid battery drain but, in many cases, the establishment of some clusters will be compromised. Therefore, the optimization problem constrained by the energy consumption of the cluster heads results from including (16) as a constraint into (7).

C. Clustering Energy Efficient algorithm (CEEa)

Due to the complexity of the optimization problem, in this Section we present a low complexity algorithm, namely CEEa, to manage the different energy consumption of each user. As the energy consumed by a cluster head is higher than the energy consumed by a non-clustered user, it is clearly a disincentive for users to become cluster heads, even when w is small. In a scenario without mobility, this disincentive can hardly be addressed (they can only be limited, as proposed in Section IV-B), but the changing environment offered by mobility opens up new possibilities. In order to analyse these possibilities, in the sequel the analysis is carried out as a function of time. Let us define the observation period T_ε as the time during which the energy consumption is analysed to prevent users from energy overconsumption. For each user i, T_ε can be divided into subperiods $T_{i,n} = [t^0_{i,n}, t^1_{i,n}) \in \mathrm{R}^2$ during which the role of user i remains constant, i.e. $\rho_i(t^0_{i,n}) = \rho_i(t^0_{i,n} - \delta t)$ for $\delta t \to 0$, and $t^1_{i,n} = \max\{t : \rho_i(t) = \rho_i(t^0_{i,n}), t > t_{i,n}{}^0\}$. Based on the definitions, the time during which each user plays a specific role is the aggregation of periods with the same $\rho_i(t)$. Thus, three sets of periods $\mathcal{T}^H_i$, $\mathcal{T}^M_i$ and $\mathcal{T}^N_i$ are defined as $\mathcal{T}^m_i = \{T_{i,n} : \rho_i(t^0_{i,n}) = m\}$ for $m = \{H,M,N\}$. If we denote the power consumed by user i at time t with role $\rho_i(t) = m$ as $P^m_i(t)$, and the power that would have been consumed by user i at time t in case of not being clustered as $\tilde{P}^N_i(t)$, the energy consumed over a subperiod $T_{i,n} \in \mathrm{T}^m$ with $m = \{H,M\}$ and the energy that

would have been consumed if $\rho_i(t) = N$ are given by $E_{im}(T_{i,n}) = R_{T_{i,n}} P_{im}(t)dt$ and $\tilde{E}_{iN}(T_{i,n}) = R_{T_{i,n}} \tilde{P}_{iN}(t)dt$ (the estimate of $\tilde{E}_i^N(t)$ can be found in Appendix B of [25]). If the definition of energy overconsumption, $w(T_\varepsilon)$, is given by $E_i^m(T_\varepsilon) = (1 + w(T_\varepsilon))\tilde{E}_i^N(T_\varepsilon)$, it can be rewritten as

$$w(T_\varepsilon) = \frac{\sum_{T_{i,n} \in (\mathcal{T}_i^H \cup \mathcal{T}_i^M)} E_i^{\rho_i(t_{i,n}^0)}(T_{i,n})}{\sum_{T_{i,n} \in (\mathcal{T}_i^H \cup \mathcal{T}_i^M)} \tilde{E}_i^N(T_{i,n})} - 1 \quad (21)$$

As $P_i^H(t) > \tilde{P}_i^N(t) > P_i^M(t)$, user i experiences energy overconsumption due to clustering if $w(T_\varepsilon) > 0$. Although the objective is to keep the overconsumption around 0 in the longterm, $\lim_{T_\varepsilon \to \infty} w(T_\varepsilon) \approx 0$, in practice overconsumption must be limited over finite periods of time.

CEEa (see Alg. 3) limits the overconsumption of users involved in the cluster by setting a maximum overconsumption threshold, referred to as w_{max}, that cannot be exceeded along the observation period T_ε. This observation period is divided into a set of n_ε subperiods of duration t_ε, such that $T_\varepsilon = n_\varepsilon t_\varepsilon$. Specifically, for a given set of users, CEEa creates a list of users that cannot become cluster heads due to excessive energy consumption in the past, denoted by Z, which is included as a constraint in eCORE. The maximum overconsumption condition, $E_i^m(t) > (1 + w_{max})\tilde{E}_i^N(t)$, is checked at the end of each subperiod of duration t_ε in two ways: first, the energy consumption condition is checked for the total time since the beginning of the observation period (line 4); secondly, the condition is checked for the subperiod (line 5). Despite experiencing total overconsumption, the user is not banned from remaining as cluster head if overconsumption is not experienced in the current subperiod (overconsumption is being compensated). If the time during which the user has had the role $\rho_i = M$ until time t, $\tau_i^M(t)$, is smaller than the time during which it has had $\rho_i = H$ until time t, $\tau_i^H(t)$, the user cannot be cluster head. This condition works proactively to cope with situations where the cluster head suffers from slight but constant overconsumption. As CEEa aims to compensate the overconsumption within T_ε, the threshold w_{max} is reduced at every observation subperiod with a factor ($\frac{n_\varepsilon - 1}{n_\varepsilon}$), since the higher n_i is, the more difficult to compensate the energy consumption in the remaining $n_\varepsilon - n_i$ subperiods is.

Although there is not apparent incentive for a user to become cluster head in the short-term, this is not actually true. In loaded scenarios, not only cell-edge users can benefit from the proposed clustering, but also most of the users (even the cluster heads themselves, since the depletion of resources can impact on the resources allocated to them). In this context, CEEa eliminates the disincentive to become cluster head. The detection of selfish users is out of the scope of CEEa, but the proposed clustering algorithm does not preclude the design and implementation of additional algorithms running on top of CEEa to prevent selfish behaviours.

Algorithm 3: Clustering Energy Efficient alg. (CEEa)

Data: U, $n_i \in [1...n_\varepsilon]$ for $\forall i \in U$

Result: Set of users banned as cluster heads: Z

1 Initialization: if $n_i = 1, \forall i \in U \Rightarrow w_i = w_{max}$; $E_i = 0$; $\tilde{E}_i = 0$

2 for $i \in U : \rho_i(n_i t_\varepsilon) = \{H, M\}$ do

3 | $E_i > (w_i + 1)\tilde{E}_i$ then

4 | if $E_i^m(t_i) > (w_i + 1)\tilde{E}_i^N(t_i)$ then

5 | | Z ← Z ∪ { i }

6 | else if $\tau_i^H(n_i t_\varepsilon) > \tau_i^M(n_i t_\varepsilon)$ then

7 | | Z ← Z ∪ { i }

8 | else

9 | | Z ← Z \ { i }

10 | end

11 | else

12 | | Z ← Z \ { i }

13 | end

14 | $w_i \leftarrow w_i \frac{n_\varepsilon - 1}{n_\varepsilon}$ and $n_i \leftarrow n_i + 1$

15 | $E_i \leftarrow E_i + E_i^m(t_i)$ and $\tilde{E}_i \leftarrow \tilde{E}_i + \tilde{E}_i^N(t_i)$

16 end

$t = [(n-1)t_\varepsilon, n t_\varepsilon]$ and $m = \rho(n t_\varepsilon)$

V. Numerical Results

A. Scenario

In this section the proposed algorithms are validated and compared with existing algorithms found in the literature and with the results when no clustering algorithms are implemented (labelled in figures as *Without Clustering* or *w/o Clust.*). A custom-made simulator implemented in C++ has been used to simulate a network, which consists of a central eNB (macro BS) and the first interfering ring of 6 eNBs, with and inter-site distance of 500m. Under the coverage area of each eNB, 4 small cells are randomly deployed. The minimum distance between the eNB and a SC is 125m and the minimum inter-SC distance is 25m [26]. All eNBs are equally loaded and simulated, but only results from the central eNB and the corresponding 4 small cells are collected. Results are averaged over 1000 iterations. Users move at a constant

speed of 3 km/h (pedestrian). The hit and bounce technique is used when users move out of the scenario under analysis [27]. 50% of the deployed users are characterized by symmetric VoIP traffic (64 kbps in DL and UL) while the rest of users demand
FTP or streaming traffic (700 kbps in the DL). The system is FDD and spectrum resource partition is considered between eNBs and SCs: eNBs and SCs operate in different bands
[28]. No interference coordination techniques are considered in the simulations, and the PRBs are allocated randomly among users. Although interference coordination could lead to higher SINR levels, it has been omitted to better characterize the performance of the proposed algorithms. Users and the BS have a single antenna (SISO), and the spectral efficiency look-up table has been obtained from [29]. The rest of the parameters can be found in Table I [30].

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Bandwidth	Macro: 10 MHz & Small Cell: 5 MHz
Macro cell Path-Loss	$128.1+37.6log_{10}$(distance km)
Small cell Path-Loss	$140.7+36.7log_{10}$(distance km)
D2D Path-Loss	$148+40log_{10}$(distance km)
Max. BS Transmission power	Macro: 46 dBm & Small Cell: 27 dBm
Max. UE Transmission power	Cellular: 20 dBm & D2D: 18 dBm

B. Results

The objective of the optimal clustering for spectral efficiency stated in Section III-D (problem (7)) and labelled in figures as *Optimal Clustering*, is the minimization of the total number of PRBs required to serve the traffic (i.e. the maximization of the spectral efficiency). Similarly, the optimal clustering with energy consumption constraints, detailed in Section IV-B and labelled hereafter as *Energy Constrained*, is aimed to minimize the required PRBs while imposing energy overconsumption constraints for cluster heads. The spectral efficiency (bps/Hz) of these two solutions can be observed in Fig. 2 for 60 users, along with the results for our previous work CORE [19], and the proposed eCORE, CaLB (with $n_{dmin} = 0.2N_{kd,max}$, $n_{umin} = 0.1N_{ku,max}$) and CEEa (with $w = 0.2$). It can be seen that the *Optimal Clustering* increases the spectral efficiency in the DL band by clustering users and exploiting the good quality of the link between BS and cluster head. For instance, spectral efficiency in the DL band rises a 54% (from 1.26 bps/Hz to 1.95 bps/Hz) when *Optimal Clustering* is applied with 60 users. Although clustering solutions incur in additional PRBs utilization in the UL band due to intra-cluster communications, it can be observed that the total spectral efficiency (UL and DL) increases. Thus, the higher UL band utilization is overcompensated by the DL improvement. As it will be seen in Fig. 3, when no clustering solution is applied, cell-edge users are not served due to low spectral efficiency. Fig. 2 also shows the spectral efficiency of *Energy Constrained* when maximum energy overconsumption of the optimal clustering is limited to 10% and to 50% ($w = 0.1$ and $w = 0.5$). As expected, the overconsumption constraint prevents clusters from being set up if they result in excessive energy overconsumption. Thus, only clusters that are simultaneously spectral efficient and keep cluster heads consumption below a threshold (i.e. w) are set up. This is the reason why the spectral efficiency is lower as the energy constraint becomes more restrictive (lower w). For instance, the DL spectral efficiency is 1.27 bps/Hz when $w = 0.1$ and 1.36 bps/Hz when $w = 0.5$. Some insights can be found in Table II, where the average number of clusters and the average size of each cluster are shown for 30 and 60 users. In the *Energy Constrained* solution, the reduction of w (lower overconsumption is allowed) has a higher impact on the number of clusters created than in the size of the cluster. That is, whereas the size of the cluster remains stable, overconsumption constraints cause a significant reduction in the average number of clusters.

Fig. 2 also includes the results for CORE, eCORE, CaLB and CEEa. eCORE achieves results very close to the optima, with a performance less than 5% lower than *Opt. Clust.* Moreover, eCORE increases the DL spectral efficiency with respect to CORE, since it enables the establishment of clusters among users from different cells.

Table II shows that the intensification in the creation of clusters promoted by eCORE results in the setup of more clusters, although with a similar size. For instance, for 60 users eCORE doubles the number of clusters with respect to CORE while the average size of each cluster is approximately the same. Something similar occurs with CaLB: the number of clusters grows more than the average size of the clusters. That is, CaLB creates new clusters rather than enlarge the clusters established by eCORE. However, CaLB enables the creation of non-spectral efficient clusters if the imbalance between UL and DL is reduced. This is the reason why although the DL spectral efficiency in CaLB is higher than in eCORE, the opposite occurs with the total spectral efficiency (UL and DL bands). Finally, as CEEa limits the energy consumption by deterring some users from being cluster heads, the spectral efficiency is reduced with respect to eCORE and CaLB. Table II shows that the energy consumption constraints reduces the number of clusters.

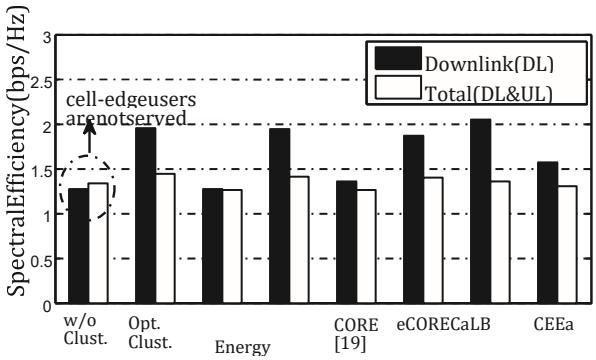


Fig. 3 shows the DL throughput for each algorithm and includes as baseline the algorithm proposed in [18], which is labelled as CS. As CS is a scheme based on the received SNR to allow or ban cooperation, results for two minimum SNR thresholds have been simulated: 4.73 dB and 2.84 dB. Fig. 3 shows how CaLB outperforms the rest of algorithms, reaching a 59.5% gain in the downlink throughput with respect to the case *Without Clustering* for 140 users. As expected, it can be also observed that eCORE outperforms CORE and, in turn, CaLB outperforms eCORE. In particular, CORE achieves a throughput 36.6% higher than *Without Clustering*, whereas eCORE reaches a 47.2% improvement and CaLB a 59.5%. As for CEEa, the additional constraints reduce the DL throughput, but still presents slightly better results than CORE.

$w = 0.1$ $w = 0.5$
(III-D) Constrained (IV-B)

Fig. 2. Downlink and total spectral efficiency for 60 users.

TABLE II AVERAGE NUMBER AND SIZE OF CLUSTERS

	Avg. Num. Clusters		Avg. Cluster Size	
Num users	30	60	30	60
Optimal Clustering	5.38	11.45	2.37	2.67
Energy Constrained (w=0.1)	1.44	3.27	2.29	2.88
Energy Constrained (w=0.5)	2.68	5.21	2.35	2.85
CORE	5.45	11.39	2.42	2.69
eCORE	5.71	11.64	2.43	2.75
CaLB	5.71	15.27	2.43	2.75
CEEa	3.38	7.62	2.50	2.98

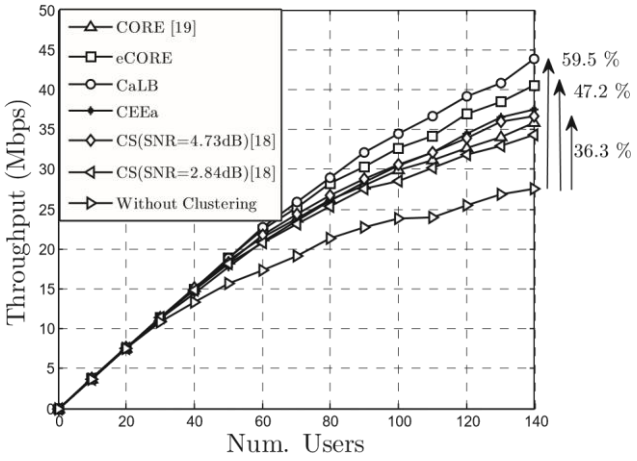


Fig. 3. Downlink throughput for the set of algorithms.

Focusing on how CEEa limits the energy overconsumption of cluster heads, Fig. 4 plots the Cumulative Distribution Function (CDF) of the energy overconsumption, w, for eCORE, CaLB and CEEa. The overconsumption is always expressed with respect to the case where no clustering algorithms are implemented. Therefore, without any clustering, the energy overconsumption would be w = 0%. As it can be observed in Fig. 4, the energy underconsumption from which cluster members (except for the cluster head) benefit is similar in eCORE, CaLB and CEEe. However, CEEa limits the overconsumption of cluster heads. For instance 99% of the users have an overconsumption w < 20% with CEEa; in turn, for eCORE the 99% of users experience an overconsumption w < 240% and with CaLB the same percentage of users experience w < 260%. Therefore, CEEa is able to limit the overconsumption of cluster heads.

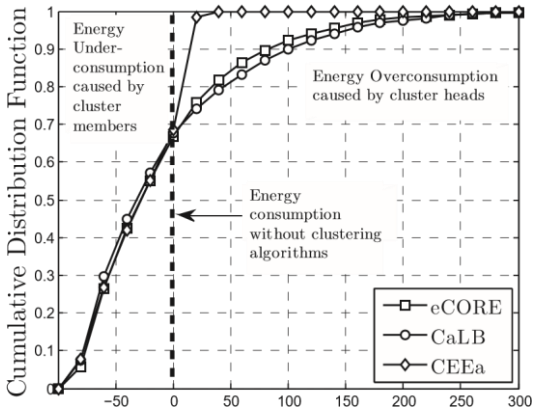


Fig. 4. CDF of the energy overconsumption with 60 users.

Given the trade-off between the maximum capacity gain

(CaLB) and the minimum impact on energy consumption (CEEa), Fig. 5 sheds light on the energy efficiency of eCORE, CaLB and CEEa for 60 users. Cluster heads present low energy efficiency because they forward traffic to/from cluster members. Therefore, the percentage of users with low energy efficiency grows with the number of cluster heads. This can be significant in CaLB and eCORE. Conversely, CEEa alleviates partially the high energy consumption of cluster heads but decreases the throughput. In none of the cases the low energy efficiency of cluster heads is compensated by the increased energy efficiency of the rest of cluster members. Accordingly clustering algorithms can improve the capacity of the network at the expense of lower energy efficiency.

In order to see how sensitive CaLB and CEE are to their key parameters (n^d_{min} and n^u_{min} for CaLB and w for CEEa), simulations have been run with different values. As for CaLB, differences in terms of throughput are not significant and below 2% for a wide range of values n^d_{min} and n^u_{min}. Although the creation/enlargement of clusters will start before as the values of n^d_{min} increase, it is also true that it will not be translated into a significant increase of the throughput. Therefore, CaLB is slightly sensitive to n^d_{min} variations in terms of throughput as long as $n^d_{min} > 0$, but should be selected small enough to avoid the creation of additional clusters when it is not actually needed (in terms of throughput)[3].

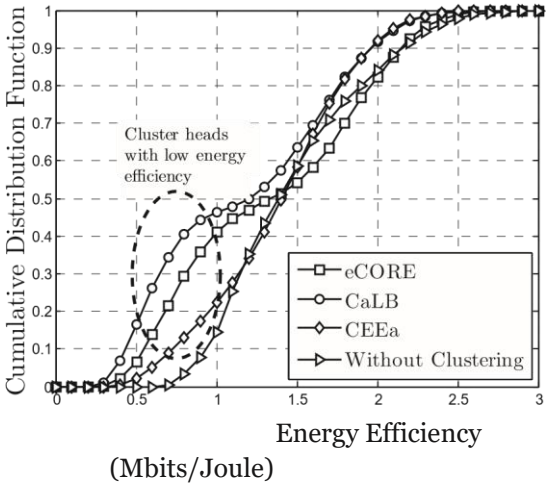


Fig. 5. CDF of the energy efficiency with 60 users.

Regarding CEEa, the key parameter is the maximum allowed energy overconsumption w. This parameter has a single objective that is attained in a two-fold manner: firstly, by preventing some users from becoming cluster heads (due to previous energy overconsumption), and secondly by forcing the release of the role of cluster head (if the energy overconsumption is too high). In a nutshell, the larger w is, the more aggressive the clustering is, thus achieving similar results to the ones obtained with eCORE (where no energy consumption constraints are imposed). Conversely, small w values impose additional constraints in the creation of clusters. This effect can be observed in Fig. 6, where the CDF of the energy efficiency is plotted for 60 users and w = {0.2,0.6,1.5}. Results for eCORE have been also included for the sake of comparison. It is observed that eCORE has cluster heads with low energy efficiency and in turn cluster members with high energy efficiency. The higher w is, the more closed results are to the ones of eCORE, since less constraints on energy consumption are imposed.

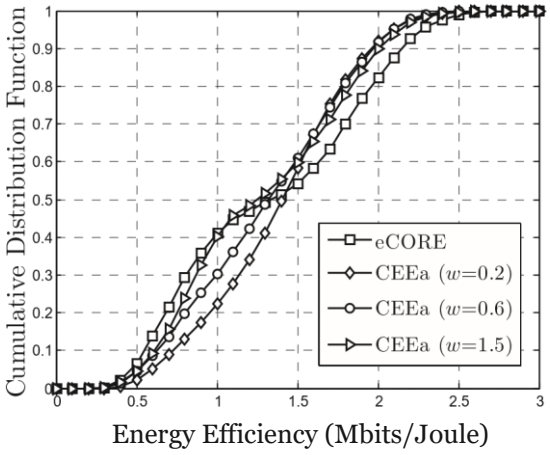


Fig. 6. CDF of the energy efficiency of CEEa with 60 users.

C. Discussion on signalling

Signalling is an important aspect of D2D communications. 3GPP establishes control and data plane paths for D2D communications (termed as Proximity Services -ProSe) in [6], and covers these aspects in more detail in [31]. The proposed algorithms are framed within the group of *UE-to-Network Relay* functions [31], since the cluster head acts as a relay from each of the cluster members to the network. In this context two important interfaces are defined: PC3, defined as the interface from the relay (i.e. the cluster head) to the network; and PC5, defined as the one-to-one or one-to-many interface between users (the so-called D2D communication). The proposed mechanisms implement the network-assisted D2D mode with the loosely-controlled scheme, in which the network allocates resources for the D2D communications, and the cluster head reallocates the resources within the

[3] No figure for throughput is included due to the slight observed differences

cluster. Network-assisted loosely-controlled D2D communications require additional signalling, particularly over PC5 interface. However, as shown in Table II, the proposed algorithms improve the throughput by creating a significant number of small size clusters rather than large size clusters, thus alleviating/reducing the increase of signalling over the PC5 interface. Therefore, although eCORE, CaLB and CEEa require additional signalling, the small size of the clusters limits the additional signalling burden over PC5.

Nevertheless, frequent cluster head (re-)selection could incur excessive signalling burden. There exists a trade-off between signalling and system performance. Algorithms eCORE and CaLB do not include neither parameters to control the number of clusters nor parameters to limit the duration of the clusters. Conversely, CEEa controls indirectly the number and size of the clusters, as well as how long they remain active or with the same cluster head, with parameters w_{max} and T_{ε}.

VI. Conclusions

This work presents a complement/alternative to the costly densification of cellular RANs based on the creation of clusters of users, where intra-cluster communications are carried out in a D2D mode. Three clustering algorithms are presented: eCORE, CaLB and CEEa. eCORE optimizes the usage of spectral resources by establishing spectral efficient clusters. Due to the significant imbalance between uplink and downlink traffic, CaLB creates non-spectral efficient clusters that improve the capacity of the network by reducing the aforementioned imbalance. Finally, CEEa is proposed to keep track of the overconsumption of users and ban some users from becoming cluster heads. Results show that the proposed clustering solutions increase the capacity of the network. In particular, the most aggressive clustering algorithm (CaLB) outperforms the rest of algorithms. Yet, any capacity improvement is translated into an increase of the consumed energy. In that sense, CEEa achieves a good energy consumption performance but it leads to the smallest capacity gain.

References

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014-2019*, Cisco Systems Inc, Feb. 2015.

[2] A. Gupta, R.K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies", *IEEE Access*, vol. 3, pp. 1206-123, July 2015.

[3] X. Zhang, N. Zhao, F. R. Yu and V. C. M. Leung, "Resource Allocation in Topology Management of Asymmetric Wireless Interference Networks," 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), Nanjing, 2016, pp. 1-5.

[4] N. Zhao; X. Zhang; F. R. Yu; V. Leung, "To Align or Not to Align: Topology Management in Asymmetric Interference Networks," in IEEE Transactions on Vehicular Technology , vol.PP, no.99, pp.1-1, Jan 2017.

[5] S.F.Yunas, J. Niemela, M. Valkama, T. Isotalo, "Techno-Economical Analysis and Comparison of Legacy and Ultra-dense Small Cell Networks", *in Proc. IEEE LCN (Workshops)*, pp. 768-776, 2014.

[6] 3GPP Technical Report 22.803, "Feasibility study for Proximity Services (ProSe)," V12.2.0, Jun. 2014. Available at www.3gpp.org

[7] A. Papadogiannis, D. Gesbert and E. Hardouin, "A Dynamic Clustering Approach in Wireless Networks with Multi-Cell Cooperative Processing," 2008 IEEE International Conference on Communications, Beijing, 2008, pp. 4033-4037.

[8] A. Asadi, Q. Wang, V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks", *IEEE Surveys & Tutorials*, vol.16, no.4, pp.1801-1819, 4th quarter 2014.

[9] R. Yin, C. Zhong, G. Yu, Z. Zhang, K. Wong, X. Chen, "Joint Spectrum and Power Allocation for D2D Communications Underlaying Cellular Networks", *IEEE Transactions on Vehicular Technology*, vol.65, no.4, pp. 2182-2195, Apr 2016.

[10] W. Zhibo, T. Hui, C. Nannan, "Clustering and power control for reliability improvement in Device-to-Device networks", *in Proc. IEEE Globecom (Workshops)*, pp.573-578, 9-13 Dec. 2013.

[11] J. Huang, Y. Yin, Y. Zhao, Q. Duan, W. Wang and S. Yu, "A Game-Theoretic Resource Allocation Approach for Intercell Device-to-Device Communications in Cellular Networks," in IEEE Transactions on Emerging Topics in Computing, vol. 4, no. 4, pp. 475-486, Oct.-Dec. 2016.

[12] B. Zhou, H. Hu, S.Q. Huang, H. Chen, "Intracluster Device-to-Device Relay Algorithm With Optimal Resource Utilization", *IEEE Transactions on Vehicular Technology*, vol.62, no.5, pp.2315-2326, Jun 2013.

[13] M. Condoluci, L. Militano, G. Araniti, A. Molinaro, A. Iera, "Multicasting in LTE-A networks enhanced by device-to-device communications", *in Proc. IEEE Globecom (Workshops)*, pp.567-572, 9-13 Dec. 2013.

[14] H. Meshgi, D. Zhao and R. Zheng, "Joint channel and power allocation in underlay multicast device-to-device communications", *in Proc. IEEE ICC*, London, 2015, pp. 2937-2942.

[15] S. Hassan, M.I. Ashraf, M.D. Katz, "Mobile Cloud based architecture for Device-to-Device (D2D) communication underlying cellular network", *in Proc. Wireless Days (WD) IFIP*, pp.1-3, 13-15 Nov. 2013.

[16] T. Koskela, S. Hakola, T. Chen and J. Lehtomaki, "Clustering Concept Using Device-To-Device Communication in Cellular System," 2010 IEEE Wireless Communication and Networking Conference, Sydney, Australia, 2010, pp. 1-6.

[17] J. Seppala, T. Koskela, T. Chen, S. Hakola, "Network controlled Deviceto-Device (D2D) and cluster multicast concept for LTE and LTE-A networks", *in Proc. IEEE WCNC*, pp.986-991, 28-31 March 2011.

[18] Q. Sun, L. Tian, Y. Zhou, J. Shi and X. Wang, "Energy efficient incentive resource allocation in D2D cooperative communications", *in Proc. IEEE ICC*, pp. 2632-2637, June 2015.

[19] G. Kollias, F. Adelantado, K. Ramantas, C. Verikoukis, "CORE: A Clustering Optimization Algorithm for Resource Efficiency in LTE-A Networks", *in Proc. IEEE Globecom*, pp. 1-6, 6-10 Dec. 2015.

[20] M. Ding, D. Lopez-Perez, R. Xue, A. V. Vasilakos and W. Chen, "On Dynamic Time-Division-Duplex Transmissions for Small-Cell Networks," in IEEE Transactions on Vehicular Technology, vol. 65, no. 11, pp. 8933-8951, Nov. 2016.

[21] M. Ding, D. Lopez-Perez, R. Xue, A. V. Vasilakos and W. Chen, ”Small cell dynamic TDD transmissions in heterogeneous networks,” 2014 IEEE International Conference on Communications (ICC), Sydney, NSW, 2014, pp. 4881-4887.

[22] A. Schrijver, ”Theory of Linear and Integer Programming”, Wiley 1986

[23] A. R. Jensen, M. Lauridsen, P. Mogensen, T. B. Srensen and P. Jensen, “LTE UE Power Consumption Model: For System Level Energy and Performance Optimization”, *in Proc. IEEE VTC Fall*, pp. 1-5, Sept.
2012.

[24] 3GPP Technical Report 36.213, ”Physical layer procedures,” V12.0.1, Mar. 2014. Available at: www.3gpp.org

[25] G. Kollias, F. Adelantado, and C. Verikoukis, Spectral Efficient and Energy Aware Clustering in Cellular Networks (Extended Version), 2017. [Online]. Available at https://arxiv.org/abs/1706.02146

[26] 3GPP Technical Report 36.842, ”Study on Small Cell enhancements for E-UTRA and E-UTRAN; Higher layer aspects ” V12.0.0, Dec. 2013, Available at www.3gpp.org

[27] 3GPP Technical Report 36.839, ”Mobility enhancements in heterogeneous networks” V11.0.0. Sep. 2012, Available at www.3gpp.org

[28] 3GPP Technical Report 36.872, ”Small cell enhancements for E-UTRA and E-UTRAN - Physical layer aspects ” V12.1.0, Dec. 2013, Available at www.3gpp.org

[29] A. ElNashar, M. A. El-saidny, M. Sherif, ”Design, Deployment and Performance of 4G-LTE Networks: A Practical Approach”, 2014, Wiley

[30] 3GPP Technical Report 36.814, ”Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects” V9.0.0, Mar. 2010. . Available at: www.3gpp.org

[31] 3GPP Technical Report 23.303, ”Proximity-based services (ProSe); Stage 2” V14.1.0 (2016-12), Available at www.3gpp.org

[32] S. Goyal, P. Liu and S. S. Panwar, ”User Selection and Power Allocation in Full-Duplex Multicell Networks,” in IEEE Transactions on Vehicular