
**SCUOLA DI DOTTORATO**
UNIVERSITA' *MEDITERRANEA* DI REGGIO CALABRIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE,
DELLE INFRASTRUTTURE E DELL'ENERGIA SOSTENIBILE
(DIIES)

DOTTORATO DI RICERCA IN
INGEGNERIA DELL'INFORMAZIONE

S.S.D. ING-INF/03
XXIX CICLO

# TOWARDS NATIVE DEVICE-TO-DEVICE INTERGRATION INTO EMERGING 5G CELLULAR SYSTEMS

CANDIDATO
Antonino ORSINO

TUTOR
PROF. ING. GIUSEPPE ARANITI

COORDINATORE
PROF. ING. CLAUDIO DE CAPUA

REGGIO CALABRIA, APRILE 2017

Finito di stampare nel mese di **Aprile 2017**

Edizione CSdA Centro Stampa d'Ateneo

Collana *Quaderni del Dottorato di Ricerca in Ingegneria dell'Informazione*
Curatore *Prof. Claudio De Capua*

ISBN 978-88-99352-21-9

Università degli Studi *Mediterranea* di Reggio Calabria
Salita Melissari, Feo di Vito, Reggio Calabria

ANTONINO ORSINO

# TOWARDS NATIVE DEVICE-TO-DEVICE INTEGRATION INTO EMERGING 5G CELLULAR SYSTEMS

Il Collegio dei Docenti del *Dottorato di Ricerca in Ingegneria dell'Informazione* è composto da:

Claudio DE CAPUA (coordinatore)
Francesco BUCCAFURRI
Francesco DELLA CORTE
Antonio IERA
Tommaso ISERNIA
Riccardo CAROTENUTO
Antonella MOLINARO
Domenico URSINO
Rosario CARBONE
Salvatore COCO
Giovanna IDONE
Giacomo MESSINA
Domenico ROSACI
Giuseppe ARANITI
Andrea MORABITO
Aime' LAY EKUAKILLE
Giovanni ANGIULLI
Mariantonia COTRONEI
Pasquale FILIANOTI
Giuliana FAGGIO
Sofia GIUFFRE'
Gianluca LAX
Fortunato PEZZIMENTI
Francesco RICCIARDELLI
Giuseppe RUGGERI
Valerio SCORDAMAGLIA
Claudia CAMPOLO
Rosario MORELLO
Leonardo MILITANO
Sandro RAO
Lorenzo CROCCO
Roberta NIPOTI
Ivo RENDINA
Lubomir DOBOS

# Acknowledgment

Firstly, a special thanks to my family. Words cannot express how grateful I am to my mother, sister, and father for all of the sacrifices that youve made on my behalf. Your prayer for me was what sustained me thus far. Nothing would be realized so far without your constant support. I will never thank you enough for what you have been made for me. I would also like to thank all of my friends who supported me in writing, and incentived me to strive towards my goal.

I would like to express my sincere gratitude to my advisor, but also a friend, Prof. Giuseppe Araniti for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

A sincere thank also to my fellow labmates for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years. In particular, a special thank to my colleague, but first of all a friend, Dr. Leonardo Militano. He has been of inspiration and support in a number of ways. Without their enthusiasm, encouragement, support and continuous optimism this thesis would hardly have been completed.

Besides my advisor, I would like to thank the Dr. Sergey Andreev and Dr. Yevgeny Koucheryavy for allowing me to spend a fantastic visiting period at Tamper University of Technology, Finland. Without their precious support it would not be possible to conduct this research and to continue my carrier in this field.

# Contents

# List of Figures

# List of Tables

# 1

# Introduction

## 1.1 Motivations

The exponential growth of multimedia applications driven by enhanced devices (i.e., smartphones, tablets, sensors, and wearable equipment) has triggered the investigation of the future fifth-generation (5G) cellular networks. As shown in Fig. 1.1, it is expected that around 2020, new 5G system will take shape and start to be a fundamental part for the connectivity in our daily life [1] [2]. To cope with this scenario, support for demanding multimedia applications with a wide variety of requirements, including higher peak and user data rates, reduced latency, enhanced indoor coverage, and improved energy efficiency, has to be addressed [3].

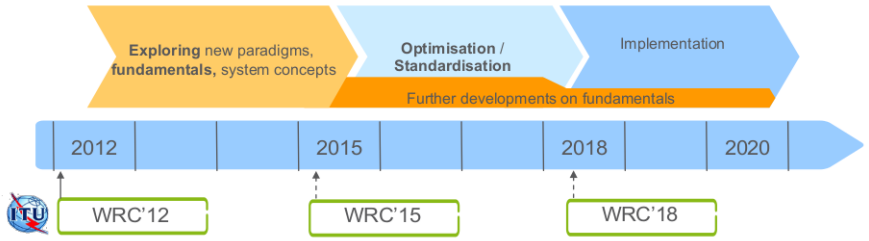



**Fig. 1.1.** The 5G timeline.

In addition, future wireless systems will be infinitely richer and more complex than those of today. The expectation is that the network infrastructure will be capable of connecting everything that benefits from being connected: actors, applications, technologies, people, things, processes, goods. Future networks will encompass connected sensors, connected vehicles, smart meters and smart home gadgets way beyond our current experience of tablet and smartphone connectivity [4].

As a consequence, the increasing trend of network densification will lead to a mulit-tier heterogeneous network consisting of a large number of low power infrastructure nodes, e.g. small cells, relays, remote radio heads (RRHs) deployed within the macro cellular coverage. In particular, the deployments of heterogeneous nodes in 5G systems

will increasingly see much higher density than today's conventional single-tier (e.g., macrocell) networks [5] as clearly stated also in Fig. 1.2.

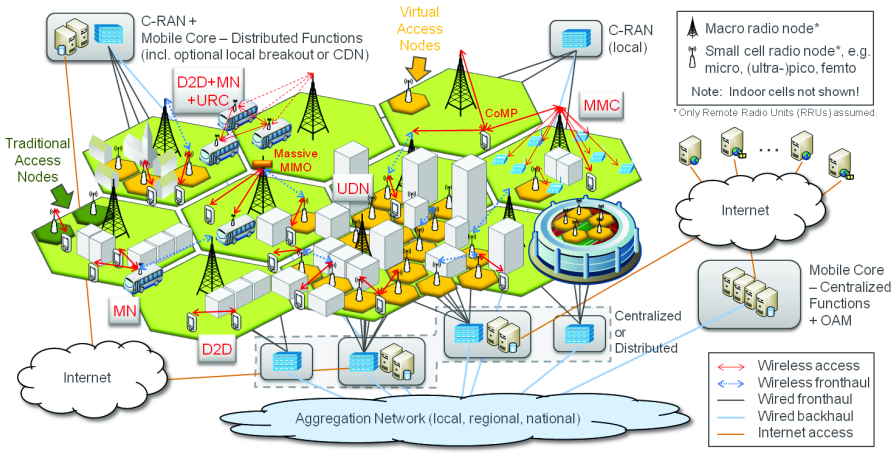



**Fig. 1.2.** High level view of 5G reference model.

In such a scenario, new paradigms represented by Device-to-Device (D2D) and Machine-to-Machine (M2M) gained momentum. In particular, D2D refers to the direct communication between two nearby devices without the need of the network infrastructure, whereas M2M refers to a communication paradigm that enables machines to communicate with each other with little or no human interaction. The latter can be achieved using the peer-to-peer model (i.e., D2D), or over a centralized model. Therefore, short-range transmissions, also supported with *Proximity Services* (ProSe) [6] by the $3^{rd}$ Generation Partnership Project (3GPP), will play an important role in increasing the spectral efficiency, managing effectively the radio spectrum, and providing lower latencies. As example of the D2D architecture is illustrated in Fig. 1.3. In addition, the network-controlled D2D communications in 5G systems will allow other nodes (such as users, relays, M2M gateways, sensors), rather than the macrocell base stations, to arbitrate the communications among D2D nodes [7].

However, the provisioning for short-range transmissions paradigm goes far beyond the current 3GPP ProSe concepts, which are limited to a single-hop communication and typically rely on a network-assisted infrastructure or multi-hop routing with limited network performance characteristic of the classical ad-hoc networks [8]. Indeed, whereas currently deployed wireless technologies are helpful to cope with those challenges [9], it is predicted that they will be insufficient to meet the exponential multimedia service growth aggravated by the rapid proliferation in types and numbers of wireless devices. All these technological challenges push the telco operators to investigate innovative solutions in order to transform the user experience in a revolutionary manner across both network infrastructure and device architecture. With

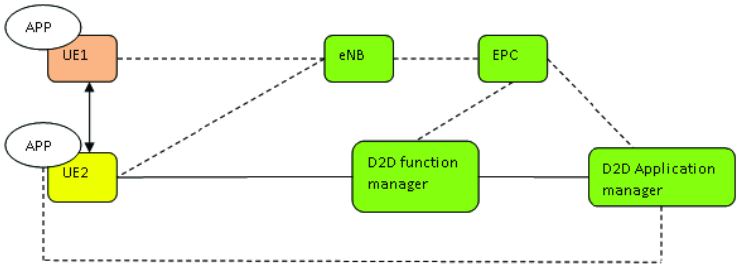



**Fig. 1.3.** 3GPP supported D2D architecture. The eNB represents the base stations, EPC is evolved packet core, APP is application functionality. Dotted lines show the control plane and thick lines show the data plan.

cellular assistance, the considered D2D technology has the potential to automate user/service discovery and connection establishment procedures, as well as enable secure D2D connectivity between proximate users that are currently outside each others social spheres [10]. This is expected to further broaden the use of assisted proximate communication, as well as enable novel ways of interaction between users, particularly those not known to each other previously (e.g., communication between unfamiliar users).

As will be surveyed in **Chapter 2**, the research conducted by the scientific community on these aspects is getting momentum in a number of publications and projects [11] [12] [13]. The new trend for proximate transmissions is in that it will be built on top of a variety of network elements at the wireless edge, such as base stations owned by different operators, WiFi access points and a diversity of users that could act as relays (i.e., See Fig. 1.4). Such personalized networks can adapt their characteristics to the user profile, location and application context. They also allow for providing enhanced content-aware connectivity to the end-users, by considering different network layers including, for example, D2D cachers, small cells and cellular macrocell. By doing so, the users or their devices are no longer concerned with managing different access networks, access-specific authentication mechanisms, etc.

In summary, although proximity services push the data transmission (i.e., both in downlink and uplink) to achieve high-data rates and low delays, licensed spectrum typically used by the network operator continues to be scarce and expensive. In such a situation, it is obvious that conventional methods to manage and improve the available radio spectrum utilization are not sufficient to handle the uncontrolled growth of multimedia applications and bandwidth hungry services (i.e, video-based traffic). Therefore, we expect that the majority of gains will come from innovative architectures and protocols that would employ a combination of licensed and unlicensed spectrum, by taking advantage of the intricate interactions between the device

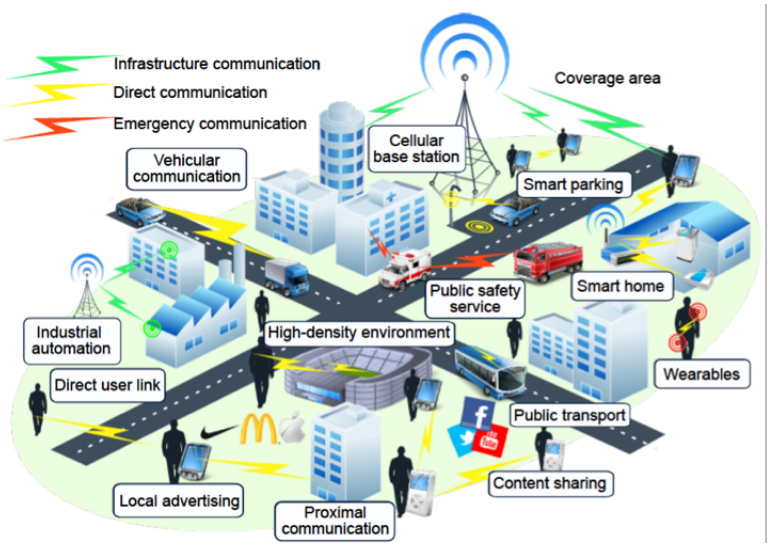



**Fig. 1.4.** Envisioned uses cases, architecture, and support of D2D into emerging 5G systems

and the network, as well as between the devices themselves, across the converged heterogeneous deployments [14].

The aim of this thesis is to design a rich set of D2D-based innovations in order to (i) improve the wireless network connectivity, (ii) increase the perceived satisfaction and Quality of Experience (QoE) of the users, and (iii) deliver new 5G-grade broadcast and multimedia services with high-data rate and low delays. In order to cope with these goals, we believe that emerging concept of D2D communications has to be comprehensively explored. Therefore, the contributions of this thesis (discussed in the remainder of this Chapter) cover (i) the design and implementation of new uploading D2D-aware models, (ii) the understanding of mobility effects in D2D-based scenarios, (iii) the accounting for the interplay between the social sphere and the communication properties in the D2D communications systems, and (iv) the integration of all the results of the aforementioned research directions for the native support of D2D communications into the future 5G Internet of Things (IoT) systems.

## 1.2 Outline of the Thesis and Contributions

The contributions of this thesis mainly deal with the native support of D2D communications into the emerging 5G mobile systems. In particular, four main research lines have been conducted during the Ph.D period regarding the integration of D2D into 5G systems: (i) *uplink transmission*, (ii) *mobility*, (iii) *security and social aspects*, and (iv) *exploitation of D2D into future 5G Internet of Things scenarios*. In this Section, is provided a brief description, with the relative scientific publications obtained, for all the research lines with particular focus on motivations and possible solutions. It

is worth noticing that the order with which the research topics are presented respect also the outline of the structure for the following Ph.D Thesis.

### 1.2.1 Uplink Transmissions

The first aspect that has been taken into consideration in **Chapter 3** of this thesis is related to the enhanced uploading cellular transmission that can be provided by exploiting D2D links. In fact, the traditional uploading technique used in cellular systems, i.e., with separate links from each User Equipment (UE) to the eNodeB, may be enhanced (or in some cases substituted) with innovative relay-based schemes that exploit D2D communications between two (or more) UEs in proximity to each other. Differences in the channel quality experienced by the UEs offer an opportunity to develop proximity-based solutions, where *(i)* the UE with a poor direct link to the eNodeB will forward data to a nearby UE over a high-quality D2D link; and *(ii)* the receiving UE then uploads its own generated data and the relayed data to the eNodeB over a good uplink channel. Indeed, following these opportunities, a straightforward gain in the data uploading time can be obtained for the first UE. In addition, extending the benefits also to the relaying UE, enhanced D2D-based solutions may be proposed in order to decrease the uploading time of the UE based on cooperative sharing of the radio resources allocated by the eNodeB to cooperating devices.

Taking into account the above issues and possible solution, the fist contribution of this thesis has been to exploit proximity-based communications (i.e., D2D) for enhancing uplink transmission in LTE/LTE-A network and beyond (e.g., 5G) and to discuss in detail the related challenges and benefits. The obtained results of this research line are published in:

- L. Militano, A. Orsino, G. Araniti, A. Molinaro, A. Iera, A Constrained Coalition Formation Game for Multihop D2D Content Uploading, in *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 2012-2024, March 2016 (reference [15] of this thesis).
- A. Orsino, G. Araniti, L. Militano, J. Alonso-Zarate, A. Molinaro, A. Iera, Energy Efficient IoT Data Collection in Smart Cities Exploiting D2D Communications, *Sensors 2016*, 16, 836 (reference [16] of this thesis).
- L. Militano, A. Orsino, G. Araniti, A. Molinaro and A. Iera, Overlapping Coalitions for D2D-supported Data Uploading in LTE-A Systems, *Personal, Indoor, and Mobile Radio Communications (PIMRC), 2015 IEEE 26th Annual International Symposium on*, Hong Kong, 2015, pp. 1526-1530 (reference [17] of this thesis).
- M. Condoluci, L. Militano, A. Orsino, J. Alonso-Zarate and G. Araniti, LTE-Direct vs. WiFi-direct for Machine-Type Communications over LTE-A Systems, *Per-*

*sonal, Indoor, and Mobile Radio Communications (PIMRC), 2015 IEEE 26th Annual International Symposium on*, Hong Kong, 2015, pp. 2298-2302 (reference [18] of this thesis).

- A. Orsino, L. Militano, G. Araniti, A. Molinaro and A. Iera, Efficient Data Uploading Supported by D2D Communications in LTE-A Systems, *European Wireless 2015; 21th European Wireless Conference; Proceedings of*, Budapest, Hungary, 2015, pp. 1-6 (reference [19] of this thesis).
- L. Militano, A. Orsino, G. Araniti, A. Molinaro, A. Iera and L. Wang, Efficient Spectrum Management Exploiting D2D Communication in 5G Systems, *2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, Ghent, 2015, pp. 1-5 (reference [20] of this thesis).

### 1.2.2 Mobility

Next-generation D2D communications technology is rapidly taking shape today, where a cellular network assists proximal users at all stages of their interaction. To this end, the respective D2D performance aspects have been thoroughly characterized by past research, from discovery to connection establishment, security, and service continuity. However, prospective D2D-enabled applications and services envision highly opportunistic device contacts as a consequence of unpredictable human user mobility. Therefore, as described in **Chapter 4**, the impact of mobility on D2D communication needs a careful investigation to understand the practical operational efficiency of future cellular-assisted D2D systems. Along these lines, this research topic offers a first-hand tutorial, solution, and methods on various implications of D2D mobility, across different user movement patterns and mobility-related parameters and proposes an assessment methodology for D2D-enabled systems. The rigorous system-level evaluation conducted by this study also delivers important conclusions on the effects of user mobility in emerging D2D-centric systems. The results of this thesis in the field of the effect of mobility over D2D-based 5G systems are published in:

- A. Orsino, G. Weisi, G. Araniti, Multi-Scale Mobility Models in the Forthcoming 5G Era: A General Overview, Submitted to *IEEE Vehicular Technology Magazine,* July 2016 (under review, reference [21] of this thesis).
- A. Orsino, D. Moltchanov, M. Gapeyenko, A. Samuylov, S. Andreev, L. Militano, G. Araniti, Y. Koucheryavy, Direct Connection on the Move: Characterization of User Mobility in Cellular-Assisted D2D Systems, in *IEEE Vehicular Technology Magazine*, vol. 11, no. 3, pp. 38-48, Sept. 2016 (reference [22] of this thesis).
- A. Orsino, M. Gapeyenko, L. Militano, D. Moltchanov, S. Andreev, Y. Koucheryavy, G. Araniti, Assisted Handover Based on Device-to-Device Communica-

tions in 3GPP LTE Systems, 2015 *IEEE Globecom Workshops (GC Wkshps)*, San Diego, CA, 2015, pp. 1-6 (reference [23] of this thesis).

- A. Orsino, A. Samuylov, D. Moltchanov, S. Andreev, L. Militano, G. Araniti, Y. Koucheryavy, Time-Dependent Energy and Resource Management in Mobility-Aware D2D-Empowered 5G Systems, *Submitted to IEEE Wireless Communications*, Oct. 2016 (under review, reference [24] of this thesis).

### 1.2.3 Security and Social Relationships

Driven by the unprecedented increase of mobile data traffic, D2D communications technology is rapidly moving into the mainstream of 5G networking landscape. While D2D connectivity has originally emerged as a technology enabler for public safety services, it is likely to remain in the heart of the 5G ecosystem by spawning a wide diversity of proximate applications and services. In this research topic, we argue that the widespread adoption of the direct communications paradigm is unlikely without embracing the concepts of trust and social-aware cooperation between end users and network operators. However, such adoption remains conditional on identifying adequate incentives that engage humans and their connected devices into a plethora of collective activities. To this end, the mission of this research is to advance the vision of social-aware and trusted D2D connectivity, as well as to facilitate its further adoption. In **Chapter 5** of this thesis, we begin by reviewing the various types of underlying incentives with the emphasis on sociality and trust, discuss these factors specifically for humans and for networked devices (machines), as well as propose a novel framework allowing to construct the much needed incentive-aware D2D applications. The supportive system-level performance evaluations suggest that trusted and social-aware direct connectivity has the potential to decisively augment the network performance. Further, also the future perspectives of its development across research and standardization sectors are provided. The results of this thesis in the field of security and social relationships bridging across D2D communications and cellular systems are published in:

- G. Araniti, A. Orsino, L. Militano, L. Wang, A. Iera, Context-aware Information Diusion for Alerting Messages in 5G Mobile Social Networks, in *IEEE Internet of Things Journal*, vol.PP, no.99, pp.1-1 (reference [25] of this thesis).
- A. Orsino, G. Araniti, L. Wang and A. Iera, Multimedia Content Diusion Approach for Emerging 5G Mobile Social Networks, 2016 *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Nara, 2016, pp. 1-6 (reference [26] of this thesis).
- A. Orsino, L. Militano, G. Araniti and A. Iera, Social-aware Content Delivery with D2D Communications Support for Emergency Scenarios in 5G Systems, *European*

*Wireless 2016; 22th European Wireless Conference*, Oulu, Finland, 2016, pp. 1-6 (reference [27] of this thesis).

- A. Ometov, A. Orsino, L. Militano, G. Araniti, D. Moltchanov, S. Andreev, A Novel Security-centric Framework for D2D Connectivity Based on Spatial and Social Proximity, *Computer Networks*, Volume 107, Part 2, Pages 327-338, October 2016 (reference [28] of this thesis).
- A. Orsino, A. Ometov, Validating Information Security Framework for Ooading from LTE Onto D2D Links," *2016 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT)*, St. Petersburg, 2016, pp. 241-247 (reference [29] of this thesis).
- L. Militano, A. Orsino, G. Araniti, M. Nitti, L. Atzori, A. Iera, Trust-based and Social-aware Coalition Formation Game for Multihop Data Uploading in 5G Systems, *Computer Networks*, Available online 4 August 2016 (reference [30] of this thesis).
- L. Militano, A. Orsino, G. Araniti, M. Nitti, L. Atzori, A. Iera, Trusted D2D-based Data Uploading in In-band Narrowband-IoT with Social Awareness, In *Proceeding of Personal, Indoor, and Mobile Radio Communications (PIMRC), 2016 IEEE 27th Annual International Symposium on*, Valencia, Sept. 2016 (reference [31] of this thesis).
- A. Ometov, A. Orsino, L. Militano, D. Moltchanov, G. Araniti, E. Olshannikova, G. Fodor, S. Andreev, T. Olsson, A. Iera, J. Torsner, Y. Koucheryavy, T. Mikkonen, Toward Trusted, Social-aware D2D Connectivity: Bridging Across the Technology and Sociality Realms, in *IEEE Wireless Communications*, vol. 23, no. 4, pp. 103-111, August 2016 (reference [32] of this thesis).
- A. Ometov, A. Levina, P. Borisenko, R. Mostovoy, A. Orsino, S. Andreev, Mobile Social Networking under Side-Channel Attacks: Practical Security Challenges, Accepted to *IEEE Access*, Jan. 2017 (reference [33] of this thesis).

### 1.2.4 5G Internet of Things

Wireless technology has already become a commodity in our society as a plethora of powerful companion devices facilitate novel user applications and services. This concept is also know in current mobile network system under the name of Internet of Things. In today's 'human-intense' urban locations, people are faced with increasingly more heterogeneous connectivity options, which creates challenges for efficient decision-making to reap the maximum user benefits. On the other hand, service providers are struggling to augment the capacity of their network deployments quickly in response to unpredictable and sporadic traffic loading. Then, in **Chapter**

**6** we envision that *mobile* vehicles and flying robots may be equipped with high-rate radio access capabilities to better accommodate the varying space-time user demand. More specifically, these radio access capabilities are possible due to the use of D2D links and transmissions. Additionally, various user-owned equipment may take a more active part in 5G IoT scenarios service provisioning by sharing wireless connectivity and content with relevant consumers in proximity. However, this emerging vision remains conditional on identifying adequate pragmatic sources of motivation for user involvement, which is aggravated by the unpredictable and heterogeneous mobility. Thus, within this research topic the results are in the form of a novel mobility-centric analytical methodology for 5G multi-connectivity scenarios in the context of truly mobile access (users, car- and drone-mounted small cells, etc.) and couple it with practical user incentivization considerations. In addition, performance evaluation have been performed in order to shown the effective benefits that D2D communications can bring to 5G-grade IoT scenarios. Finally, an investigation on the monetary side either for the network provider and users has been proposed by taking into account novel concepts typical of the forthcoming 5G infrastructure such as aerial access node identified by drones and unmanned vehicles. The results in the field of the integration of D2D into emerging 5G system are listed below:

- A. Samuylov, A. Orsino, D. Moltchanov, S. Andreev, L. Militano, G. Araniti, A. Iera, H. Yanikomeroglu, Y. Koucheryavy, On Tighter User Involvement in Multi-Connectivity 5G Scenarios with Access Infrastructure Mobility, Submitted to *IEEE Journal on Selected Areas in Communications*, Sept. 2016 (under review, reference [34] of this thesis).
- A. Orsino, A. Ometov, G. Fodor, D. Moltchanov, L. Militano, S. Andreev, O.N.C. Yilmaz, T. Tirronen, J. Torsner, G. Araniti, A. Iera, M. Dohler, Y. Koucheryavy, Eects of Heterogeneous Mobility on D2D- and Drone- Assisted Mission-Critical MTC in 5G, in *IEEE Communication Magazine*, (Accepted), Aug. 2016 (in press, reference [35] of this thesis).
- O. Galinina, L. Militano, S. Andreev, A. Pyattaev, K. Johnsson, A. Orsino, G. Araniti, A. Iera, M. Dohler, Y. Koucheryavy, Demystifying Competition and Cooperation Dynamics of the Aerial mmWave Access Market, Submitted to *IEEE/ACM Transaction on Networking*, Aug. 2016 (under review, reference [36] of this thesis).
- A. Orsino, I. Farris, L. Militano, G. Araniti, A. Iera, D2D Communications for Delay-sensitive IoT Mobile Services over Multiple Edge Nodes, Submitted to *IEEE Internet Computing Magazine*, Jan. 2017 (under review, reference [37] of this thesis).

# 2

# Background

## 2.1 Towards 5G systems

The innovative and effective use of information and communication technologies (ICT) is becoming increasingly important to improve the economy of the world [38]. In this scenario, being one of the fastest growing and most dynamic sectors in the world, wireless networks are perhaps the most critical element in the global ICT strategy. This is demonstrated by the quick deployment of the *fourth generation (4G)* wireless communication systems, e.g., Long Term Evolution (LTE) and LTE-Advanced (LTE-A), in many countries. The European Mobile Observatory (EMO) reported that the mobile communication sector had total revenue of €174 billion in 2010. This is due to the fact that the development of wireless technologies and the related introduction of novel services has greatly improved people's ability to interact each other, to communicate and to live in both business operations and social functions [39].

The growth of wireless mobile communications is mirrored by a rapid pace of technology innovation. From the second generation (2G) to the 4G deployments, several enhancements have been introduced to handle an always large and growing set of applications oriented for the users. As it can be noticed in Tab. 2.1, the wireless mobile network has been transformed from a pure telephony system (i.e., circuit switching) to an IP-based network that can transport rich multimedia contents. By focusing on the last deployed 4G systems, such systems were designed to fulfill the requirements of International Mobile Telecommunications-Advanced (IMT-A) using IP for all services to guarantee an easier deployment and to allow a decrease in the cost of deployment [40]. In 4G systems, an advanced radio interface is used with orthogonal frequency-division multiplexing (OFDM), multiple-input multiple-output (MIMO), and link adaptation technologies. 4G wireless networks can support data rates of up to 1 Gbps for low mobility, such as nomadic/local wireless access, and up to 100 Mbps for high mobility, such as mobile access.

**Table 2.1.** Characteristics of wireless cellular systems from 2G to 5G

| | 2G | 3G | 4G | 5G |
|---|---|---|---|---|
| Deployment | 1990-2001 | 2001-2010 | 2011-in progress | 2015-20 onwards |
| Data Rates | 14.4-64 kbps | 2 Mbps | 100 Mbps - 1 Gbps | $\sim$ 10 Gbps |
| Services | Digital voice, SMS, MMS | Enhanced audio/video, web browsing | IP telephony, HD mobile TV | Dynamic Information access, wearable devices with AI capabilities |
| Multiplexing | TDMA, CDMA | CDMA | OFDMA | OFDMA |
| Core network | PSTN | Packet N/W | IP | IP |
| Standards | 2G: GSM, 2.5G: GPRS, 2.75G: EDGE | 3G: IMT-2000, 3.5G: HSDPA, 3.75G: HSUPA | LTE, WiMAX | In progress |
| WEB standard | www | www(IPv4) | www(IPv4) | www(IPv6) |
| Handoff | Horizontal | Horizontal & Vertical | Horizontal & Vertical | Horizontal & Vertical |

However, there is still a dramatic increase in the number of users who subscribe to mobile broadband systems every year. The availability of high data rates and coverage with 4G involved a drastic increase in the request of Internet access on the move. Similarly, more powerful smartphones and laptops are becoming more popular nowadays, demanding advanced multimedia capabilities. The EMO pointed out that there has been a 92% growth in mobile broadband per year since 2006 while Ericsson is predicting 50 billion of connected devices by 2020 [41]. The exponential increase in the number of connected devices involves many research challenges which need to be addressed. 5G networks need to be developed further to support up to 1000 times higher traffic volumes compared to 2010 travel levels over the next 10 years [42].

What will the 5G network, which is expected to be standardized around 2020, look like? It is now too early to define this with any certainty [39], but it starts to become clear the key features and the main expected performance requirements of 5G systems, depicted in Fig. 2.1 and in Fig. 2.2, respectively. It is widely agreed that compared to the 4G network, the 5G network should achieve 1000 times the system capacity, 10 times the spectral efficiency, energy efficiency and data rate (i.e., peak data rate of 10 Gb/s for low mobility and peak data rate of 1 Gb/s for high mobility), and 25 times the

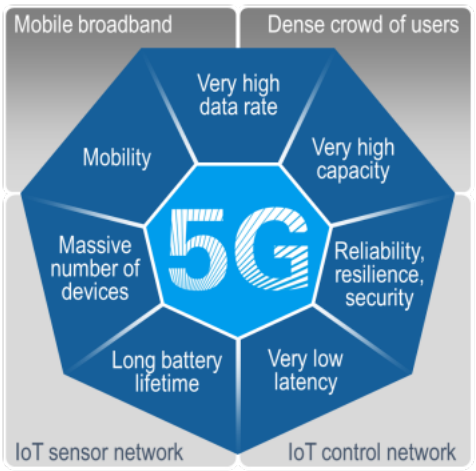



**Fig. 2.1.** Key features of 5G systems.

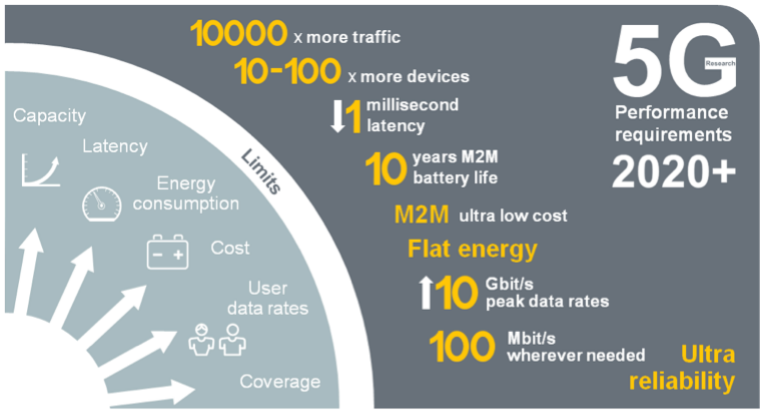



**Fig. 2.2.** Performance requirements of 5G systems.

average cell throughput. The aim is to *connect the entire world*, and achieve seamless and ubiquitous communications between *anybody* (people-to-people), *anything* (people-to-machine), wherever they are (*anywhere*), whenever they need (*anytime*), by whatever electronic devices/services/networks they wish (*anyhow*). This means that 5G networks should be able to support communications for some special scenarios, not supported by 4G networks (e.g., for high-speed train users), which thus pose unprecedented challenges. For the previous considered scenario involving high-speed trains, their mobility speed can easily reach 350-500 km/h, while 4G networks can only support communication scenarios up to 250 km/h. This is only one aspect of the challenges of 5G systems which thus dictate for drastic novelties to be introduced in the design of network architectures and data transmission procedures. An exhaustive overview of the features expected by 5G systems, with a summary of the related benefits and challenges, is given in Sec. 2.1.1.

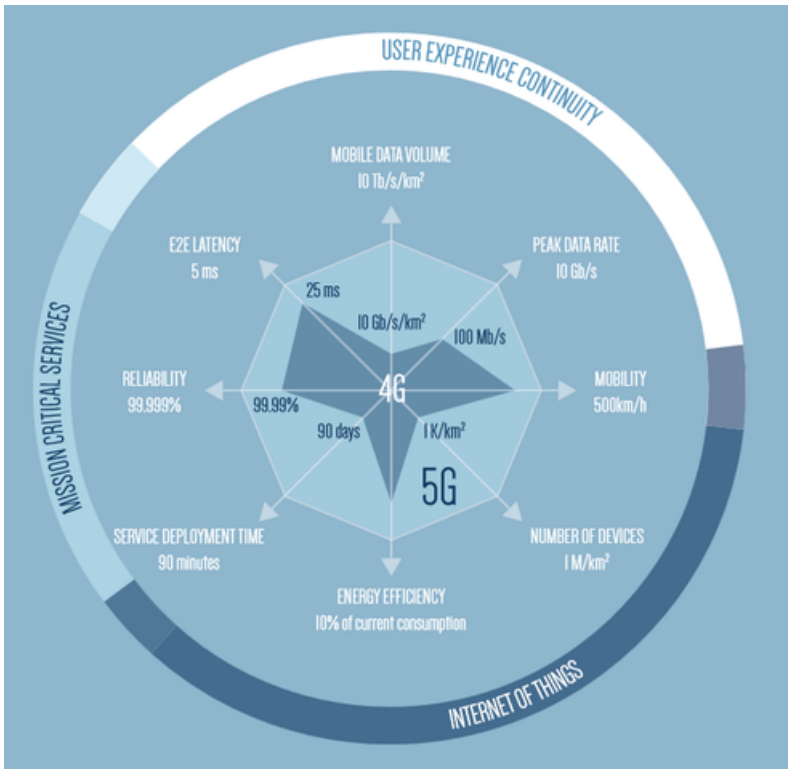



**Fig. 2.3.** Expected benefits for 5G vs. performance of 4G.

### 2.1.1 Features and Expected Benefits

The 5G cellular network is coming, but the definition of the technologies which will define it is still in progress. Will 5G be just an evolution of 4G, or will emerging technologies cause a disruption requiring a wholesale rethinking of entrenched cellular principles? In [43], several potential disruptive technologies and their implications for 5G have been considered. The main 5G technologies can be summarized as follows:

- *Device-centric architectures.* The traditional paradigm for network deployment which is based on the key role of base stations may drastically change in 5G. It may be time to reconsider the concepts of uplink and downlink, as well as control and data channels, to better route information flows with different priorities and purposes toward different sets of nodes within the network. This dictates for novel solutions in terms of network deployment able to bring intelligence close to devices.
- *Millimeter wave (mmWave).* The spectrum has become scarce at microwave frequencies, while it is plentiful in the mmWave realm. Such a spectrum led to a quick increase in the interest of academic and industry research groups in mmWave communications. Although still in a preliminary stage, mmWave technologies have already been standardized for short-range services (IEEE 802.11ad) and deployed for particular applications (e.g., small-cell backhaul).

- *Massive MIMO.* Massive multiple-input multiple-output (MIMO)[1] is based on the concept of utilizing a very high number of antennas to multiplex messages for several devices on each time-frequency resource with the aim of radiating energy toward the intended directions to minimize intra- and inter-cell interference. Massive MIMO may require major architectural changes, particularly in the design of base stations, and it may also lead to new types of deployments.
- *Smarter devices.* The idea behind 2G/3G/4G cellular networks was to have a complete control at the infrastructure side. 5G systems should drop this design assumption and exploit intelligence at the device side within different layers of the protocol stack. For example, by allowing device-to-device (D2D) connectivity or exploiting smart caching at the mobile side. This design philosophy obviously requires a change at the node level (i.e., component change), but it further dictates for drastic changes at the architectural level.
- *Native support of device-to-device communications (D2D) for IoT applications.* A native[2] inclusion of D2D in 5G involves satisfying three fundamentally different requirements associated with different given by the IoT paradigm: *(i)* support of a massive number of low-rate devices, *(ii)* sustaining a minimal data rate in potentially all circumstances, and *(iii)* guaranteeing very-low-latency data transfer. Addressing these requirements in 5G requires new methods and ideas at both the component and architectural levels.

The impact of above considered technologies, leveraging the Henderson-Clark model, can be summarized as follows:

- *Design evolutions.* Changes are needed at both the node and architectural levels. Examples are the introduction of new types of incentives for the end-user and machines to be an active part of the mobile infrastructure.
- *Component changes.* Disruptive changes are needed in the design of a class of network nodes (e.g., the introduction of a new waveform).
- *Infrastructure changes.* As for the "Design evolution", new types of nodes and new features should be considered within the mobile network infrastructure. A practical example is represented by vehicle and drones (i.e., that are gaining momentum in the last year) carrying base stations.

[1] Several works in massive MIMO have assumed a working frequency 5GHz or less. While the same principles may prove useful at millimeter wavelengths, a successful integration of massive MIMO and millimeter waves may take on a considerably different form.

[2] As was learned with MIMO, first introduced in 3G as an add-on and then natively included in 4G, major improvements come from native support (i.e., from a design that is optimized from its inception rather than amended a posteriori).

- *Radical changes.* These disruptive changes may involve and may have a strong impact at both the node and infrastructure levels.

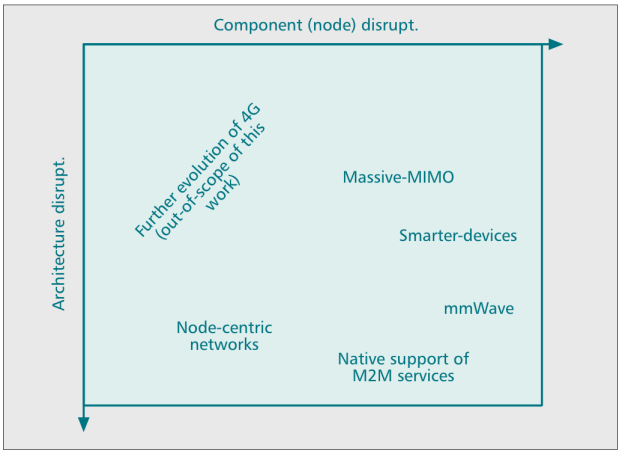



**Fig. 2.4.** The five disruptive directions for 5G, classified according to the Henderson-Clark model.

## 2.2 Capabilities and Technologies of Forthcoming 5G Systems

### 2.2.1 Requirements and capabilities

In order to enable connectivity for a very wide range of applications with new characteristics and requirements, the capabilities of 5G wireless access must extend far beyond those of previous generations of mobile communication. These capabilities will include massive system capacity, very high data rates everywhere, very low latency, ultra-high reliability and availability, very low device cost and energy consumption, and energy-efficient networks.

#### Massive system capacity

To support the dramatically increase of traffic expected in the next years in an affordable way, 5G networks must deliver data with much lower cost per bit compared with the networks of today. Furthermore, the increase in data consumption will result in an increased energy footprint from networks. 5G must therefore consume significantly lower energy per delivered bit than current cellular networks. The exponential increase in connected devices, such as the deployment of billions of wirelessly connected sensors, actuators and similar devices for massive machine connectivity, will place demands on the network to support new paradigms in device and connectivity management that do not compromise security.

#### Very high data rate everywhere

Every generation of mobile communication has been associated with higher data rates compared with the previous generation. However, as already mentioned in the previous section, 5G should support data rates exceeding 10Gbps in specific scenarios such as indoor and dense outdoor environments. Further, data rates of several 100Mbps should generally be achievable in urban and suburban environments, whereas data rates of at least 10Mbps should be accessible almost everywhere, including sparsely-populated rural areas in both developed and developing countries.

#### Very low latency

To support low-latency-critical applications, 5G should allow for an application end-to-end latency of 1ms or less, although application-level framing requirements and codec limitations for media may lead to higher latencies in practice. Many services will distribute computational capacity and storage close to the air interface. This will create new capabilities for real-time communication and will allow ultra-high service reliability in a variety of scenarios, ranging from entertainment to industrial process control.

#### Ultra-high reliability and availability

In addition to very low latency, 5G should also enable connectivity with ultra-high reliability and ultra-high availability. For critical services, such as control of critical infrastructure and traffic safety, connectivity with certain characteristics, such as a specific maximum latency, should not merely be typically available. Rather, loss of connectivity and deviation from quality of service requirements must be extremely rare.

#### Very low device cost and energy efficiency

Low-cost, low-energy mobile devices have been a key market requirement since the early days of mobile communication. However, to enable the vision of billions of wirelessly connected sensors, actuators and similar devices, a further step has to be taken in terms of device cost and energy efficiency. It should be possible for 5G devices to be available at very low cost and with a battery life of several years without recharging.

### 2.2.2 Machine-Type Communications

Fundamentally, applications such as mobile telephony, mobile broadband and media delivery are about information for humans. In contrast, many of the new applications

and use cases that drive the requirements and capabilities of 5G are about end-to-end communication between machines. To distinguish them from the more human-centric wireless-communication use cases, these applications are often termed machine-type communication (MTC). Although spanning a wide range of applications, MTC applications can be divided into two main categories  massive MTC and critical MTC depending on their characteristics and requirements.

Massive MTC refers to services that typically span a very large numbers of devices, usually sensors and actuators whereas Critical MTC refers to applications such as traffic safety/control, control of critical infrastructure and wireless connectivity for industrial processes. Therefore, There is much to gain from a network being able to handle as many different applications as possible, including mobile broadband, media delivery and a wide range of MTC applications by means of the same basic wireless-access technology and within the same spectrum. This avoids spectrum fragmentation and allows operators to offer support for new MTC services for which the business potential is inherently uncertain, without having to deploy a separate network and reassign spectrum specifically for these applications.

### 2.2.3 Spectrum for 5G

In order to support increased traffic capacity and to enable the transmission bandwidths needed to support very high data rates, 5G will extend the range of frequencies used for mobile communication. This includes new spectrum below 6GHz, as well as spectrum in higher frequency bands. Specific candidate spectrum for mobile communication in higher frequency bands is yet to be identified by the ITU-R or by individual regulatory bodies. The mobile industry remains agnostic about particular choices, and the entire frequency range up to approximately 100GHz is under consideration at this stage, although there is significant interest in large contiguous allocations that can provide dedicated and licensed spectrum for use by multiple competing network providers. The lower part of this frequency range, below 30GHz, is preferred from the point of view of propagation properties. At the same time, very large amounts of spectrum and the possibility of wide transmission frequency bands of the order of 1GHz or more are more likely above 30GHz.

Spectrum relevant for 5G wireless access therefore ranges from below 1GHz up to approximately 100GHz. It is important to understand that high frequencies, especially those above 10GHz, can only serve as a complement to lower frequency bands, and will mainly provide additional system capacity and very wide transmission bandwidths for extreme data rates in dense deployments. Spectrum allocations at lower bands will remain the backbone for mobile-communication networks in the 5G era, providing ubiquitous wide-area connectivity.

### 2.2.4 5G Technology Components

#### Access/Backhaul integration

In the future, the access (base-station-to-device) link will also extend to higher frequencies. Furthermore, to support dense low-power deployments, wireless backhaul will have to extend to cover non-line-of-sight conditions, similar to access links. In the 5G era, the wireless-access link and wireless backhaul should not therefore be seen as two separate entities with separate technical solutions. Rather, backhaul and access should be seen as an integrated wireless-access solution able to use the same basic technology and operate using a common spectrum pool.

#### Direct Device-to-Device communications

The possibility of limited direct device-to-device (D2D) communication has recently been introduced as an extension to the LTE specifications. In the 5G era, support for D2D as part of the overall wireless-access solution should be considered from the start. This includes peer-to-peer user-data communication directly between devices, but also, for example, the use of mobile devices as relays to extend network coverage. D2D communication in the context of 5G should be an integral part of the overall wireless-access solution, rather than a stand-alone solution. Direct D2D communication can be used to offload traffic, extend capabilities and enhance the overall efficiency of the wireless-access network. Furthermore, in order to avoid uncontrolled interference to other links, direct D2D communication should be under network control. This is especially important for the case of D2D communication in licensed spectrum.

#### Flexible duplex

In very dense deployments with low-power nodes, the TDD-specific interference scenarios (direct base- station-to-base-station and device-to-device interference) will be similar to the normal base-station-to-device and device-to-base-station interference that also occurs for FDD. To reach its full potential, 5G should therefore allow for very flexible and dynamic assignment of TDD transmission resources. This is in contrast to current TDD-based mobile technologies, including TD-LTE, for which there are restrictions on the downlink/uplink configurations, and for which there typically exist assumptions about the same configuration for neighbor cells and also between neighbor operators.

#### Multi-antenna transmission

Multi-antenna transmission already plays an important role in current generations of mobile communication and will be even more central in the 5G era, due to the

physical limitations of small antennas. The 5G radio will employ hundreds of antenna elements to increase antenna aperture beyond what may be possible with current cellular technology. In addition, the transmitter and receiver will use beamforming to track one another and improve energy transfer over an instantaneously configured link. Beamforming will also improve the radio environment by limiting interference to small fractions of the entire space around a transmitter and likewise limiting the impact of interference on a receiver to infrequent stochastic events. The use of beamforming will also be an important technology for lower frequencies; for example, to extend coverage and to provide higher data rates in sparse deployments.

## 2.3 Device-to-Device Communications

The D2D communications technology has been addressed in 3GPP LTE release 12 system [44]; notwithstanding, it is expected to have a complete standardization of proximity services in next 3GPP releases 13 and 14. As mentioned in the introduction to this chapter, the exploitation of D2D communications between UEs in proximity is expected to achieve improvements in terms of spectrum utilization, overall throughput, energy consumption, and to guarantee better public safety networks management. In what is presented next, a general overview of the current D2D standardization process is provided together with the system architecture proposed to integrate this new technology in the current cellular systems, and a number of possible applications in different scenarios and use cases.

### 2.3.1 Standardization Overview

The standardization process is an aspect of utmost importance to be considered for the commercial feasibility and future deployment of new technologies. In the particular case of D2D communications, although direct communications are already provided by the use of unlicensed Industrial, Scientific and Medical (ISM) bands (e.g., Wi-Fi Direct), its standardization in the context of the cellular system is currently still ongoing. A first example of the introduction of D2D communications into the LTE-Advanced (LTE-A) network is provided by Qualcomm Company, which developed a mobile communication system called FlashLinq [45]. In particular, FlashLinq is a PHY/MAC network architecture, which allows cellular devices automatically and continuously discovering thousands of other FlashLinq enabled devices within 1 kilometer and communicating peer-to-peer, at broadband speeds and without the need of intermediary infrastructures. Moreover, peer-to-peer communications enabled through Qualcomm's FlashLinq can share connectivity with a cellular network technology unlike Wi-Fi Direct's-based peer-to-peer. FlashLinq discovery procedure is carried out

by broadcasting public/private expressions mapped into tiny 128-bit packages of data, which represent basic information of either devices or users.

From a standardization point of view, 3GPP is focusing its efforts on D2D communications (recently begun in release 12 [44]) for public safety Proximity Services (ProSe) [46]. This strategy has been initially targeted to allow LTE becoming a competitive broadband communication technology for public safety networks used by first responders. However, from a technical perspective point of view, the exploitation of the proximity nature of the communicating devices will provide the further performance benefits: *(i)* D2D UEs will be able to exploit high data rate with a low delay due to the short range; *(ii)* compared to traditional downlink/uplink cellular communication, D2D will enable energy savings and improve radio resource utilization; *(iii)* cellular data traffic offloading and, consequently, lower overload in the network. In detail, the 3GPP Radio Access Network (RAN) working group has proposed in TR 36.843 Rel. 12 [46] two basic functions for supporting *ProSe discovery* and *ProSe communications* over the LTE radio interface. ProSe discovery allows an UE using the LTE air interface to identify other UEs in proximity. Two kinds of ProSe discovery exist, namely *restricted* and *open*; the difference consists in whether the permission is necessary or not for the discovery for a UE. ProSe communication instead, is the data communication between two UEs in proximity using the LTE air interface. 3GPP Services working group (SA1) has defined in specification TR 22.803 [6] the use cases and scenarios for ProSe. In the document, conditions for service flows and potential requirements for different use cases are analyzed in order to provide a support for D2D systems design. Some examples of use cases and scenarios identified for general commercial/social use and network offloading are summarized below.

The following terms are defined by 3GPP in the description of D2D use cases:

- *ProSe Discovery*: it is a process that identifies a UE in proximity of another, using EvolvedUMTS Terrestrial Radio Access Network (E-UTRAN).
- *ProSe Communication*: it is a communication between two UEs in proximity through an E-UTRAN communication path established between the UEs. The communication path can for example be established directly between the UEs or routed via local evolved-NodeB (eNB).
- *ProSe-enabled UE*: it is a UE that supports ProSe Discovery and/or ProSe Communication.
- *LTE D2D*: it is a series of technologies characterized by ProSe capability.

**Table 2.2.** Available documents for D2D

| | | |
|---|---|---|
| SA1 | TR 22.803 | Feasibility study for Proximity Services (ProSe) |
| | TS 22.278 | Service requirements for the Evolved Packet System (EPS) |
| | TS 22.115 | Service aspects; changing and billing |
| | TS 21.905 | Vocabulary for 3GPP specifications |
| SA2 | TR 23.703 | Study on architecture enhancements to support Proximity-based Services (ProSe) |
| | TS 23.303 | Proximity-based Services (ProSe); Stage 2 |
| SA3 | TR 33.833 | Study on security issues to support Proximity Services |
| RAN 1 & RAN 2 | TR 36.843 | Study on LTE device to device proximity services - Radio Aspects |
| CT1 | TS 24.333 | Proximity-serices management object (MO) |
| | TS 24.334 | Proximity-services (ProSe) user equipment (UE) to Proximity-services function aspects; Stage 3 |

### 2.3.2 Uses cases and scenarios presented in 3GPP Rel. 12

Some examples of use cases for ProSe Discovery and ProSe Communication scenarios defined by 3GPP SA1 in specification TR 22.803 [6] are given below.

**Restricted/open ProSe Discovery:** these are use cases for a basic ProSe Discovery scenario that can be exploited for any kind of application. In case of restricted ProSe Discovery, a ProSe-enabled UE discovers another UE in proximity only if it has previously achieved the permission; while, in case of open ProSe Discovery, a ProSe-enabled UE is able to discover neighbor devices without the necessity of a permission. An example of restricted use case is the friend discovery in a social network where the discovery is constrained by the UE's privacy settings. While a shop/restaurant advertisement is an example of open use case because shops and restaurants are open to be discovered by all the possible ProSe-enabled UEs in proximity, being free of privacy issues.

**Network ProSe Discovery:** it is a use case for ProSe Discovery scenarios where the Mobile Network Operator (MNO) verifies if a UE has the permission to discover another UE and the proximity. Therefore, in this case the network should be able to determine and provide the ProSe-enabled UEs with their proximity.

**Service continuity between infrastructure and E-UTRA ProSe Communication paths:** this is a use case for a ProSe Communication scenario where

the operator is able to switch user traffic from the initial infrastructure communication path to the ProSe communication one. Then, the traffic can be addressed again towards an infrastructure path, without being perceived by the users. Hence, the operator should be able to dynamically control the proximity criteria (e.g., range, channel conditions, achievable QoS) for switching between the two communication paths.

**ProSe-assisted WLAN Direct Communications:** WLAN direct communication is a use case available between ProSe-enabled UEs with WLAN capability when they are in Wi-Fi Direct communications range. It is based on the ProSe Discovery and the WLAN configuration information from the 3GPP Evolved Packet Core (EPC). In this case the operator is able to switch data session between infrastructure path and WLAN ProSe communication path.

**ProSe Application Provided by the Third-Party Application Developer:** in this case the operator can provide ProSe capability features in a series of APIs to third-party application developers. Through this cooperation between the operator and third-party application developers, the user can download and use a wide variety of new ProSe applications created by third-party developers. In this case the operator's network and the ProSe enabled UE should provide a mechanism that enables to identify, authenticate and authorize the third-party application to use ProSe capability features.

In Table. 2.2 the available specifications together with the corresponding main topics provided by the 3GPP working groups is summarized. It can be noticed the presence of the mentioned SA1 and RAN working group handling, respectively, feasibility study for ProSe and LTE radio interface issues. Other examples of topics under investigation supporting ProSe are the study of the architecture, security issues and Management Objects (MOs) representing parameters that handle the configuration of ProSe-enabled UEs.

### 2.3.3 System Architecture

In order to support the scenarios illustrated earlier in this chapter, the enhancements in the LTE architecture illustrated in Fig. 2.5 have been proposed. In details, this architecture aims at meeting the following requirements introduced by the 3GPP specifications to:

- Allow the operator to control the ProSe discovery feature in its network and authorizing the functionalities required for the ProSe discovery of each UE.
- Allow the ProSe communication or ProSe-assisted WLAN Direct communication and seamless service continuity when switching user traffic between an infrastructure path and a ProSe communication of the ProSe-enabled UEs.

- Allow an authorized third party ProSe application interacting with the 3GPP network in order to use the ProSe services provided by the network.
- Be able to control ProSe communication between ProSe-enabled UEs in case the UEs are served by either the same or different eNBs.
- Handle the ProSe-related security functions that correspond to privacy, support for regulatory functions including Lawful Interception, and authentication upon ProSe discovery and ProSe communication.
- Allow the operator's authorization and authentication of the third-party applications before making use of the ProSe features.

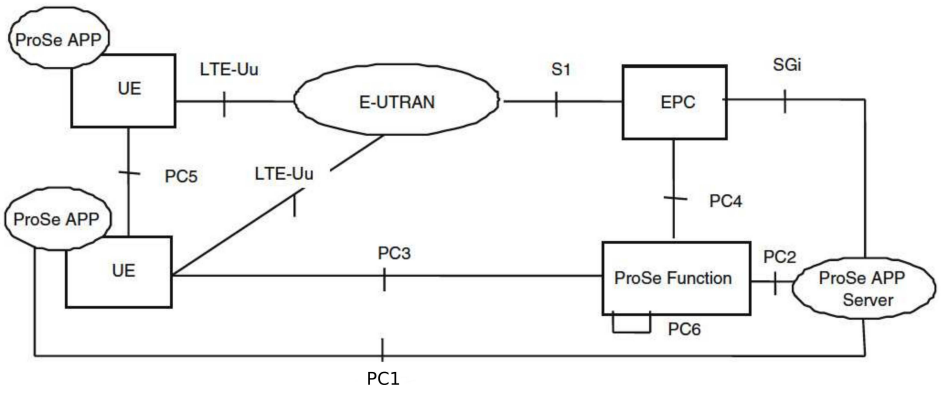



**Fig. 2.5.** 3GPP D2D proposed architecture.

As depicted in Fig. 2.5, in addition to the entities of the conventional LTE architecture (i.e., Evolved UMTS Terrestrial Radio Access Network, E-UTRAN and Evolved Packet Core, EPC), several new entities are required. These new entities are illustrated in the following.

- *Application servers (ProSe App Server)*: include the ProSe capability in order to build the application functionalities, such as specific agencies (PSAP) in case of Public Safety or social media in the commercial cases. These applications are proposed outside the 3GPP architecture; notwithstanding there may be reference points toward 3GPP entities. Moreover, the Application server can communicate with an application in the UE.
- *Applications in the UE (ProSe UEs App)*: exploit the ProSe capability in order to build the application functionality. An example may be the communication between members of Public Safety groups as well as social media application that requires finding other UEs in proximity.
- *ProSe Functions*: are a reference point toward the ProSe App Server, the EPC, and the UE. The functionalities may comprise (but not limited to): *(i)* inter-networking

via a reference point toward the third-party applications; *(ii)* authorization and configuration of UEs for discovery and direct communication; *(iii)* allowing the functionality of the EPC-level ProSe discovery, and charging. Notice that for the interconnection of the new entities and the connection with the conventional LTE ones, seven new interfaces/reference points are illustrated in the figure as PC1, PC2, PC3, PC4, PC5, PC6, and SGi (Fig. 2.5).

### 2.3.4 Application Scenarios

Applications of 5G D2D communications include local service, emergency communication, and the Internet of Things (IoT) enhancement. A brief description of these applications is provided in the following.

#### Local Service

In this scenario, user data is directly transmitted between terminals without being routed through the network side. Local service is usually utilized for social apps that are a basic D2D application based on the proximity feature. Through the D2D discovery and communication functions, a user can find other close users in order to share data or play games with them.

Another basic application of local service is the local data transmission, which exploits the proximity and direct data transmission characteristics of D2D to extend mobile applications while saving spectrum resources and then, making possible a new source of revenue for operators. In fact, local advertising service based on proximity can accurately target people in order to improve its benefits. Some examples of local transmissions conceived to improve commercial benefits are: a shopping mall where discounts and commercial promotions are sent to people walking into or around the mall; a cinema where information about movies and showtimes can be sent to people close by.

A third application of local service is the cellular traffic offloading that can reduce network overloading problems. In fact, consider that nowadays media services are becoming more and more popular; their massive traffic flows cause an extensive pressure on core networks and spectrum resources. In this context, D2D-based local media services allow operators to save spectrum resources in their core networks. In hotspot areas, operators or content providers can exploit media servers storing popular media services and sending them in D2D modality to the users. Alternatively, users can utilize D2D communications to obtain the media content from close terminals which have obtained media services. This enables to optimize the downlink transmission pressure of operator cellular networks. Furthermore, the cellular communication be-

tween short-distance users can be switched to the D2D modality in order to offload cellular traffic.

### Emergency Communications

Natural disasters such as earthquakes can damage traditional communication network infrastructures making networks not available and causing enormous rescue efforts. This problem could be overcome through the introduction of D2D communications. In fact, although the communication network infrastructures may be irremediably affected, a wireless network can still be created between terminals based on the D2D connections. This means that an ad hoc network can be set up based on multi-hop D2D to guarantee smooth wireless communication between users. Moreover, a wireless network affected by terrain or buildings can have blind spots. With single-hop or multi-hop D2D communication, users may be connected in the blind spots to other users, which are in coverage areas and then, be connected to the wireless network.

### IoT Enhancement

One of the main aims of designing new mobile communication technologies is to create an extensive interconnection among different networks involving various types of terminals. This is the motivation, which has pushed forward the development of the Internet of Things (IoT) in the cellular communication framework. The industry forecast says that by 2020 there will be 50 billion cellular access terminals on a global scale and most of them will be devices with the IoT feature. In this context, the connection between D2D with IoT will drive towards a truly interconnected wireless network.

A common application of D2D-based IoT challenge is vehicle-to-vehicle (V2V) communication in the Internet of Vehicles (IoV). For instance, when a vehicle runs at high speeds, it can warn close vehicles in D2D mode before it changes lanes or slows down. According to the received messages, close vehicles warn drivers or even automatically handle the driving in an emergency situation; hence, thanks to this application drivers can react more quickly to diminish the number of traffic accidents. D2D communications provide inherent advantages when they are considered in the context of IoV security issues also thanks to their favorable features in terms of communication delay and neighbor discovery.

As there exist many IoT devices in a 5G network, access load is becoming a serious issue to be taken into account. Nevertheless, D2D-based network access is expected to improve this problem. In a scenario characterized by many terminals, low-cost terminals can access close special terminals in D2D modality instead of direct

connections with BSs. Moreover, if multiple special terminals are isolated, the wireless resources for accessing low-cost terminals may be reutilized by these special terminals. Notice that this not only improves access pressure on BSs, but also optimizes the spectrum efficiency. Furthermore, the D2D-based access modality is more flexible and costs less than the small cell structure of the existing 4G networks.

In a smart home application, a smart terminal may be considered as a special terminal. Wireless appliances in the smart home access the smart terminal in D2D modality; while, the smart terminal may access the BS in a traditional cellular mode. The cellular-based D2D communication can represent a real breakthrough for the development of the smart home industry.

#### Other Applications

D2D communications may also be considered in other potential scenarios, such as multiuser MIMO enhancement, cooperative relaying, and virtual MIMO. In the context of the traditional multiuser MIMO, BSs find precoding weights based on the feedback received by the terminals in the respective channel in order to create nulls and delete interference between users. Through the introduction of D2D communications, paired users may directly exchange information about channel status. Hence, terminals can put together channel status information to be sent to the BSs improving the performance of multi-user MIMO.

D2D communications may also contribute to solve problems in new wireless communication scenarios. For instance, in the indoor positioning terminals may not achieve satellite signals if they are indoors. In this case, the traditional satellite-based positioning cannot work efficiently. In case of D2D-based indoor positioning, either pre-deployed terminals with given location information, or usual outdoor terminals with given position can detect the location of terminals to be localized, and support indoor positioning at a low cost in 5G networks.

## 2.4 State-of-the-art on D2D Communications over Cellular Networks

D2D communications are expected to play a key role in the ecosystem of future 5G cellular networks. This is motivated by two aspects: *(i)* the amount of data traffic exchanged over radio mobile systems is exponentially increasing and this dictates novel communications paradigms for radio mobile networks; *(ii)* use cases for D2D communications presented above represent key 5G services. As a consequence, the natively support of D2D communications becomes crucial in 5G systems.

D2D communication was initially proposed in cellular networks as a new paradigm to enhance network performance. Several studies in the literature have already discussed the improvements in terms of spectral efficiency and reduced communication delay that D2D communication can provide in cellular networks [47, 48, 49, 50, 51, 52]. On the other hand, this new paradigm presents several aspects to be investigated in terms for instance of interference control overhead and network protocols. Therefore, the feasibility of D2D communications in the context of LTE-A is currently a fascinating topic under investigation by academia, industry, and the standardization bodies. A general overview of state-of-the-art applications based on D2D communications for future 5G wireless systems is given next in both, uplink and downlink scenarios. Then, some examples of services where D2D communications have been efficiently exploited in LTE-A networks will be illustrated and assessed through exhaustive performance evaluation.

Several studies addressing D2D communications for downlink services can be found in the literature, covering several aspects and applications as for instance mobile data offloading [53], cell coverage extension [13] or content sharing [54], [55]. Recently, D2D communications have been taken into account also for downloading multicast services with focus on direct device communications over short links of a different technology than the cellular one. To cite some of them, in [56] a subset of mobile devices are considered as anchor points in a cell to forward the multicast data received from the BS to other devices in proximity through multihop ad-hoc Wi-Fi links. In [57] cellular users directly communicate to carry out cooperative retransmissions using generic short-range communication capabilities. However, the use of heterogeneous wireless interfaces introduces several issues in terms of content synchronization that become essential in case of multicast video streaming applications. Moreover, as also stated in [58], the use of cellular D2D links provides several benefits compared to *outband* D2D links, like Wi-Fi, in terms of improved user throughput. Although, the focus of the literature has been mainly on technical issues for downlink services, uplink direction scenarios are of undoubted interest as also witnessed by recent publications, such as [59] where relaying on smartphones is proposed to transmit emergency messages from disconnected areas. Multihop D2D communications have been also investigated in a very few recent works. In [60] and [61] network-assisted D2D communication is addressed with an analysis on power control and mode selection on the direct links. However, the analysis refers to a more traditional two-hop scenario, with a UE or the eNodeB as the last hop node. Similarly, multihop D2D communication is considered in [62] and [15] for end-to-end Machine-to-Machine and human-traffic connectivity.

As an example of D2D communications over cellular LTE-A links a downlink scenario for multicast transmission is considered in order to efficiently overcome the

limitations identified in [56] and [57]. In details, in the scenario proposed in this study a portion of multicast users, which sense poor channel qualities is split into clusters. The members of these clusters are served through cellular D2D transmissions, while the remaining users (i.e., those with better channel quality) are served over cellular transmission from the BS.

All the solutions illustrated in this section exploit D2D communications relying on LTE-A network infrastructure. In LTE-A, Orthogonal Frequency Division Multiple Access (OFDMA) and Single Carrier Frequency Division Multiple Access (SC-FDMA) are considered, respectively, in case of downlink and uplink. The eNodeB handles the spectrum resources by providing the appropriate number of RBs to each scheduled user and by selecting the Modulation and Coding Scheme (MCS) for each RB. Scheduling solutions are based on the Channel Quality Indicator (CQI) feedback, which is sent by a UE to the eNodeB over dedicated control channels. Each CQI value correspond to a given maximum supported MCS as specified in [63]. The MCS parameters can be adapted at every CQI Feedback Cycle (CFC), which can last one or several Transmission Time Intervals (TTIs) where one TTI is 1 ms.

# 3

# Uplink Transmissions

## 3.1 Cellular- vs. D2D-solutions

In the last years, D2D communications is gaining momentum well justified by the promising advantages of this innovative paradigm in terms of improved spectrum utilization, higher data rate, and lower energy consumption. Direct interactions between local devices enable novel applications and services [64] that can have high relevance in critical situations such as public safety and disaster scenarios, where the network resources have to be used efficiently. In fact, the use of D2D links can be substantially more efficient than conventional communications through the eNodeB whenever a communication is inherently local in scope [65], [66]; besides, it can help to either extend the cell coverage, to offload cellular traffic [10], [67], or to support content sharing in a neighbourhood [54], [68].

However, the focus of the D2D related literature at the first stages of this technology has been on downlink communications in most of the cases [69], [70]. Many works have dealt with use cases and expected performance improvements related to D2D, or with specific technical issues such as peer service discovery, D2D link set-up, interference management, and so on. Even if only a few papers have considered D2D communications specifically for the uplink direction, there are several scenarios and services that can benefit from D2D interactions to improve the uplink performance of the Long Term Evolution-Advanced (LTE-A) system. This is the case of disaster scenarios, where updated information from the incident area should be timely and reliably sent to a control center, or also scenarios where several users wish to upload multimedia content to the Cloud. The interest for these scenarios is witnessed by some recent publications, e.g., in [59] D2D relaying by smartphones is used to send out emergency messages from disconnected areas and to support information sharing among people gathered in evacuation centers.

### 3.1.1 The LTE-A Reference System

In order to investigate the goodness and which benefits the D2D paradigm may bring to LTE uplink transmissions, in a first stage a single LTE-A system has been considered. In particular, in Long Term Evolution-Advanced (LTE-A) [71], Orthogonal Frequency Division Multiple Access (OFDMA) and Single Carrier Frequency Division Multiple Access (SC-FDMA) are used to access the downlink and the uplink, respectively. The eNodeB manages the spectrum by assigning the adequate number of RBs[1] to each scheduled user and by selecting the Modulation and Coding Scheme (MCS) for each RB. Scheduling procedures are based on the *Channel Quality Indicator* (CQI) feedback, transmitted by a UE to the eNodeB over dedicated control channels. The CQI is associated to a given maximum supported MCS as specified in [71] (see Table 3.1). The MCS parameters can be adapted at every *CQI Feedback Cycle* (CFC), which can last one or several Transmission Time Intervals (TTIs) (one TTI is 1 ms).

**Table 3.1.** CQI-MCS mapping for D2D and cellular communication links

| CQI index | Modulation Scheme | Efficiency D2D [bit/s/Hz] | Min. Rate D2D [kbps] | Efficiency Cellular [bit/s/Hz] | Min. Rate Cellular [kbps] |
|---|---|---|---|---|---|
| 1 | QPSK | 0.1667 | 28.00 | 0.1523 | 25.59 |
| 2 | QPSK | 0.2222 | 37.33 | 0.2344 | 39.38 |
| 3 | QPSK | 0.3333 | 56.00 | 0.3770 | 63.34 |
| 4 | QPSK | 0.6667 | 112.00 | 0.6016 | 101.07 |
| 5 | QPSK | 1.0000 | 168.00 | 0.8770 | 147.34 |
| 6 | QPSK | 1.2000 | 201.60 | 1.1758 | 197.53 |
| 7 | 16-QAM | 1.3333 | 224.00 | 1.4766 | 248.07 |
| 8 | 16-QAM | 2.0000 | 336.00 | 1.9141 | 321.57 |
| 9 | 16-QAM | 2.4000 | 403.20 | 2.4063 | 404.26 |
| 10 | 64-QAM | 3.0000 | 504.00 | 2.7305 | 458.72 |
| 11 | 64-QAM | 3.0000 | 504.00 | 3.3223 | 558.72 |
| 12 | 64-QAM | 3.6000 | 604.80 | 3.9023 | 655.59 |
| 13 | 64-QAM | 4.5000 | 756.00 | 4.5234 | 759.93 |
| 14 | 64-QAM | 5.0000 | 840.00 | 5.1152 | 859.35 |
| 15 | 64-QAM | 5.5000 | 924.00 | 5.5547 | 933.19 |

[1] The RB corresponds to the smallest time frequency resource that can be allocated to a user (12 sub-carriers) in LTE. For example, a channel bandwidth of 20Mhz corresponds to 100 RB.

Further, it was assumed that in the considered LTE-A scenario a UE can either communicate through the serving eNodeB (*i.e., cellular mode*) or it can bypass the eNodeB and use direct communications over D2D links (*i.e., D2D mode*). The eNodeB is in charge of the D2D session setup (e.g., bearer setup) [49], while power control and resource allocation procedures on the D2D links can be executed either in a distributed or in a centralized (i.e., the approach considered) way [72]. Accordingly, the eNodeB is aware of the current cell load and the user channel conditions and can efficiently allocate dedicated D2D resources in order to improve the session quality and the allocation flexibility. In addition, is assumed that uplink cellular resources are allocated to D2D communications, because *(i)* it guarantees a more efficient reuse of resources compared to downlink allocation, and *(ii)* downlink resources can be made available to other services within the cell.

Going into more details, in the considered system D2D connections can be supported on Frequency Division Duplex (FDD) and Time Division Duplex (TDD) bands. The FDD mode poses additional issues in terms of terminal design, cost and complexity [72]; for this reason, has been considered TDD with a frame structure *type 2 configuration 0* foreseen by 3GPP [71] (see Table 3.2). In Table 3.2, 'D' denotes that the subframe is reserved for downlink transmission, 'U' denotes that the subframe is reserved for uplink transmission, and 'S' denotes a special subframe. The chosen *configuration 0* guarantees the highest number of uplink slots among all the configurations of the type 2 frame. The communication range between nearby devices can reach tens of meters, indeed, the data rate on the D2D link are properly calculated based on the CQI level, the allocated resources and the UE transmitted power.

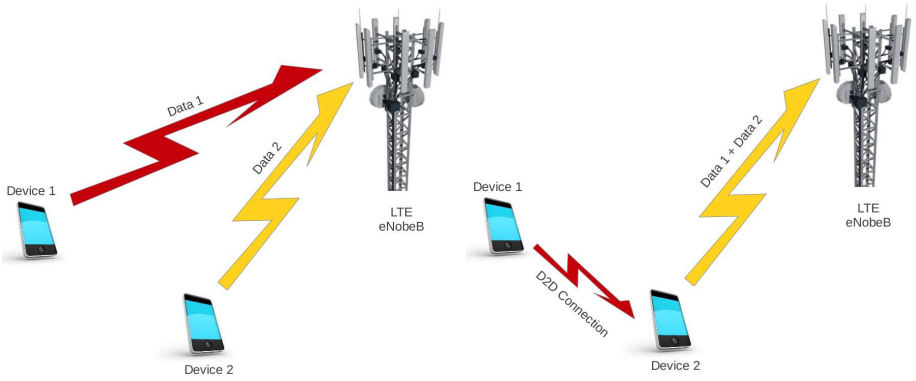
**Table 3.2.** Uplink-downlink configurations for frame structure type 2 (TDD)

| **Uplink-Downlink** | **Downlink-to-Uplink** | **Subframe number** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **configuration** | **Switch-point periodicity** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| 0 | 5 ms | D | S | U | U | U | D | S | U | U | U |
| 1 | 5 ms | D | S | U | U | D | D | S | U | U | D |
| 2 | 5 ms | D | S | U | D | D | D | S | U | D | D |
| 3 | 10 ms | D | S | U | U | U | D | D | D | D | D |
| 4 | 10 ms | D | S | U | U | D | D | D | D | D | D |
| 5 | 10 ms | D | S | U | D | D | D | D | D | D | D |
| 6 | 5 ms | D | S | U | U | U | D | S | U | U | D |

### 3.1.2 Legacy and D2D-based Data Uploading Schemes

In the following subsection is provided a simple comparison among a Legacy LTE solution, where all the users upload their own data directly to the eNodeB, and some basic approach based on D2D, where the users demand the upload of their data to a user they have in proximity. In doing this also a simple analytical evaluation is provided to show the benefits that may be achieved in considering D2D links as a support for the data upload.

Starting with the environment considered, let us refer to the case where multiple users in a single LTE-A cell are interested in uploading some multimedia content to the Cloud or to a central server on the Internet. This may be the case of a disaster scenario, where videos from several devices in the area of interest have to be uploaded timely and reliably. In this case, the *data uploading time* plays a very important role. Similarly, in other scenarios of interest multiple users, for example gathered for a concert or a fair, are willing to upload some multimedia content. More specifically, data uploading in the classic *cellular-mode* occurs through the activation of separate links from each UE to the eNodeB (Fig. 3.1(a)). In this case, the eNodeB measures the uplink channel quality from each UE and decides the MCS and the number of allocated frequency resources (RBs) in the UL slots of the transmission frame. On the other hand, UEs in proximity to each other may establish D2D links, as simplified in Fig. 3.1(b) for the case with two users. The device with a poor uplink to the eNodeB can take advantages from a nearby device with a good channel quality by using it as a relay towards the eNodeB.




(a) Classic Uploading Solution. (b) D2D-based Uploading Solution.

**Fig. 3.1.** Reference D2D-based uploading scenario.

In the following, is assumed that *(i)* the UEs use the decode-and-forward (DF) relaying protocol; i.e., the relaying node decodes the received message before transmitting it to the eNodeB; *(ii)* each UE operates in half-duplex mode, thus, the uplink

slots are in the frame are alternatively used by D2D communication and transmissions toward the eNodeB.

In order to give a clear idea about what may be the benefits and contributions brought by D2D communications, here below we provide some understanding examples of possible cellular and D2D-based uploading strategies that could be used in a LTE-A scenarios.

**Cellular-mode data uploading**

The classic data upload in a *cellular mode* is used as a term of comparison against the designed D2D-based schemes (i.e., described later). Without loss of generality, is assumed that the eNodeB equally divides the available RBs $R$ in the uplink slots of the data frame among all the requesting users. Thus, with two users, each of the two will get $r_1 = r_2 = R/2$ RBs. The data rate over a communication link for a single node is in function of the allocated resources and its CQI level [73]. For the sake of simplicity, in this sample study case is considered $b_c$ (where $c = 1 \dots 15$) the data rate per allocated RB (see minimum rate value in Table 3.1) and compute the obtained data rate on a link linearly with the allocated resources. This simplification will be removed in the performance evaluation section. In such a case, the CQI level of the two users is, e.g., $c_1 = 5$ and $c_2 = 10$, then the data rate per RB will be respectively $b_{c_1} = 147.34kbps$ and $b_{c_2} = 458.72kbps$, and the corresponding uplink data rate $d_1 = b_{c_1} \cdot r_1$ and $d_2 = b_{c_2} \cdot r_2$. In the sample case with $R = 50$, the data rate per user will be $d_1 = 3.68Mbps$ and $d_2 = 11.47Mbps$. Consequently, if the data file of each user has a size of $D = 100MB$, then the uploading time is approximately equal to $t_1 = \frac{D \cdot 8}{d_1} \approx 217s$ and $t_2 = \frac{D \cdot 8}{d_2} \approx 70s$.

**D2D-based uploading (DBU)**

With reference to the sample two-user case discussed in Section 3.1.2, $UE_1$ will transfer its data on a D2D link to $UE_2$, whereas $UE_2$ will transfer both its own data and the data received by $UE_1$ in uplink to the eNodeB. In this case, is assumed that the CQI level and the radio resources available on the D2D link between the two users guarantee a data rate $d_{12}$ from $UE_1$ to $UE_2$ greater than $d_2$ (i.e., representing the data rate in uplink for $UE_2$).In particular, under the assumptions of frequency reuse and all RBs $R$ available for the communication, the data rate on the D2D link is $d_{12} = 46.2Mbps$ (based on the data rate per RB in Table 3.1).

Therefore, with this solution (i.e., hereafter called *D2D-based uploading* – DBU) $UE_2$ first uploads its own data and then takes care of the data received by the peer UE. With the same amount of resources allocated to $UE_2$ as in the cellular-mode case,

i.e., $r_2 = R/2 = 25$, the time $t_2^{'}$ required to transfer its own content remains the same as $t_2 \approx 70s$. Once uploading its own data, it can start receiving the data from $UE_1$. Assuming that all data is first received by the relaying node, a time contribution for data transmission on the D2D link $t_{21} = \frac{D \cdot 8}{d_{12}} \approx 17s$ has to be added. The third time contribution to be considered is equal to $70s$ when $UE_2$ will upload the data for $UE_1$. With this basic configuration, $UE_1$ has a benefit in terms of time for transferring the data, i.e., $t_1^{'} < t_1$ with $t_1^{'} \approx 157s$ and $t_1 \approx 217s$. Moreover, also the network provider has some benefits as it saves resources; indeed, exploiting the DBU approach, 25 RBs are allocated to $UE_2$ for 140 seconds, instead of 50 RBs to $UE_1$ and $UE_2$ for the first 70 seconds and 25RBs to $UE_1$ for the remaining 147 seconds.

#### D2D-based uploading - time minimization (DBU-TM)

The promising results obtained with the DBU scheme can be further enhanced by the *D2D-based uploading - time minimization* (DBU-TM) approach. We consider that the set of resources allocated to the two users separately by the eNodeB in cellular mode, are *pooled together* and allocated to $UE_2$ only by the DBU-TM scheme, i.e., in the reference sample case $r_1^{'} = 0$ and $r_2^{'} = 50$ RBs. As a consequence, $UE_2$ will now experience a data rate in uplink equal to $d_2^{'} = b_{c_2} \cdot r_2^{'} = 22.94Mbps$. Hence, when giving priority to its own traffic, it will be able to upload his data in half of the time $t_2^{'} = \frac{D \cdot 8}{d_2^{'}} \approx 35s$. After this, in the next $17s$ it will receive the data from $UE_1$ and then in $35s$ it will be able to upload the data of $UE_1$. Now $UE_1$ will have a reduced uploading time of $t_1^{'} \approx 87s$, which is way less than the time it would take in the cellular-mode where $t_1 \approx 217s$. Moreover, the network provider has still an advantage in terms of used resources compared to the cellular-mode case; the uplink radio resources used by DBU-TM are equal to 50 RBs for 70s.

## 3.2 Multihop D2D Content Uploading

In a traditional cellular system, end-user devices do not cooperate, so each of them separately uploads its own content to the eNodeB, with the risk of spectrum crunch and poor service quality in crowded places. In this "non-cooperative" case, a UE located far from the eNodeB could suffer from low channel quality and not be able to upload a high-quality video flow within a time frame that is considered as "acceptable". This may be of high concern in an emergency scenario, for example. To cope with this issue, the UE far from the base station may use another UE in the proximity, with a higher-quality uplink, as a *relay*.

Along this line, the basic idea proposed in this Section is that a set of UEs "cooperate" to upload their contents to the eNodeB by forming a "multihop D2D chain",

where the UEs located farther from the base station relay their content to a nearby UE and only the UE at the head of the chain, the so-called *gateway*, is in charge of uploading all the contents received from the other UEs to the eNodeB. The UEs in the chain, then, are all sources of their own content and cooperate to forward the content generated by the preceding nodes in the chain toward the gateway thus benefiting of the higher performance of the D2D links w.r.t. the direct cellular ones. Further, the gateway is the UE with the highest link quality in the chain; it may receive, if needed, all the radio resources that would have been separately allocated by the eNodeB to the UEs in the D2D chain.

### 3.2.1 System Model and Problem Formulation

A single cell in the Long Term Evolution-Advanced (LTE-A) network, with multiple UEs interested in uploading their own content to the Internet is considered. In the traditional *cellular-mode*, separate links are activated from each UE to the eNodeB for content upload over the allocated uplink radio resources. With the proposed *co-operative upload* instead, some UEs in reciprocal proximity may establish D2D links and form what we call a "coalition" so that a UE with a poor uplink channel quality can utilize a nearby UE with a better link conditions as a relay for content upload toward the eNodeB. Under the control of the eNodeB, the UEs in a coalition organize themselves to form a "logical multihop D2D chain" and cooperate in uploading the content generated by *all* of them to the eNodeB. Each UE in the chain, but the last one, behaves as a content *source* and as a *relay*, as illustrated in Fig. 3.2. In particular, the UE at the end of the chain only transmits its own generated content but has no content to forward on behalf of other nodes; all the other nodes in the chain *also* act as relays for the contents received from the upstream UEs. They manage two active D2D links: an *incoming* link to receive data from the previous source in the logical chain and an *outgoing* link to relay data (its own and the received one from the incoming link) to the subsequent UE in the chain. The source at the head of the chain is the UE with the best uplink channel conditions and acts as a *gateway*; it receives all the relayed contents from the chain and is in charge of uploading it to the eNodeB.

Further, a network-assisted D2D chains formation under the control of the eNodeB is assumed. In general, only the UEs in the cell that are in mutual coverage can establish direct links, and this needs to be carefully modelled as a constraint for the multihop D2D chain formation. Uplink resources are allocated to D2D links in Time

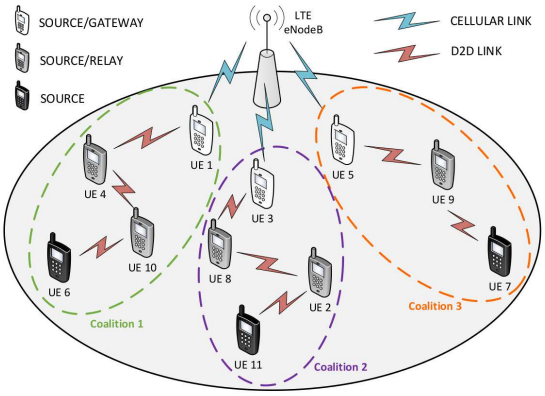



**Fig. 3.2.** Multihop D2D-based content uploading.

Division Duplex (TDD) mode[2]. Devices in the same coalition may share the same resources, whereas devices in different coalitions are always allocated to orthogonal frequency resources by the scheduler at the eNodeB, so that no mutual interference is caused by different coalitions (this is a reasonable assumption used in other works [74]). Cellular links are modeled as a Rayleigh fading channel and D2D links as a Rician fading channel [75]. As for the work illustrate in Section 3.1, is used the *type 2 configuration 0* LTE frame structure [71], composed of six out of ten subframes (or Transmission Time Intervals, TTIs) of 1 ms duration dedicated to uplink (U). At each frame beginning, the eNodeB executes the Radio Resource Management (RRM) policy during the $2ms$ duration of a downlink (D) and a special (S) subframes preceding the first U subframe. Each UE operates in half-duplex mode; thus, it either receives or transmits in a given TTI (U subframe).

A reasonable assumption that has been considered for *rational* self-interested devices, is that each UE uploads its own generated content first, and then the content received by the preceding UEs in the chain, but only after having received the whole content (in other words, UEs use the decode-and-forward relaying protocol). In each U subframe, the half-duplex UEs may either receive from the previous UE in the chain or relay data to the next node in the chain. By numbering the position of the nodes in the chain progressively starting from the gateway, when a generic node $i$ transmits to node $(i-1)$, the nodes $(i-1)$ and $(i+1)$ are in receiving mode. Consequently, in a given subframe, the first UE in a chain can transmit simultaneously with all the UEs in odd positions (i.e., the third, the fifth, the seventh, and so on), while the UEs in an even position (i.e., the second, the fourth, the sixth UE and so on) receive data. Similarly, when the even UEs transmit, the odd UEs receive in a given U subframe.

[2] Assigning uplink resources to D2D links guarantees a more efficient reuse w.r.t. a downlink allocation [50]. This is because the TDD mode poses less issues than the Frequency Division Duplex (FDD) mode in terms of terminal design, cost and complexity [72].

On the RRM point-of-view, simultaneously transmitting UEs within the same coalition can use either the same or different frequencies, based on the decision of the eNodeB according to the interference level experienced on each direct link. In this sense, there are two extreme cases that are of interest in such a case: the *best-case* that corresponds to "no interference", where the same radio resources can be reused on the D2D links, and the *worst-case*, where simultaneous transmissions interfere and so orthogonal resources have to be used. In this latter case, to avoid interference on simultaneous D2D transmissions, the radio resources used on the D2D links are only those ones allocated to the specific D2D pair in the uplink toward the eNodeB by the scheduler.

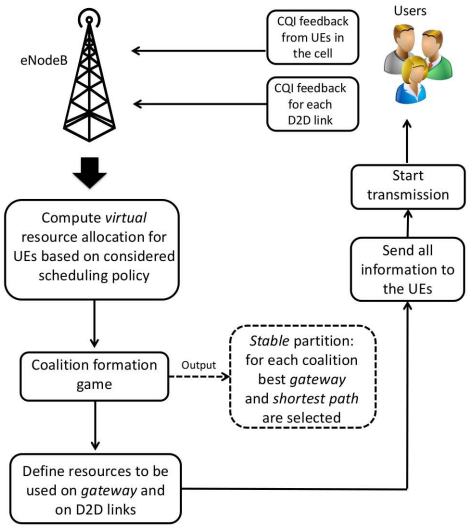



**Fig. 3.3.** Flowchart of the proposed solution.

Given the different schemes that the eNodeB can exploit to share the available radio resources among the users, for the lack of clarify the RRM algorithm implemented by the eNodeB is summarized in the flow chart in Fig. 3.3. Preliminarily, the eNodeB measures the *Channel Quality Indicators* (CQIs) of the uplink from all UEs in the cell and collects from each UE the CQIs relevant to the direct links with all its neighbors[3]. Then, the eNodeB assists the users in the chain formation process by implementing the proposed solution as illustrated in the following. In the first step the eNodeB computes the radio resources allocated to the UEs as if they were transmitting separately on the uplink according to the two scheduling policies detailed above. These resources are "virtual" in the case the UEs will form a coalition and

[3] ideal channel feedback and no errors in the CQI estimation has been assumed.

will be used as the *pool* of resources allocated to the gateway[4]. Based on this initial information, the eNodeB implements the coalition formation algorithm (see Section 3.2.2). As a result, stable coalitions are formed in the cell, the roles of each node in the coalition is identified, and routing path is defined.

Focusing on a feasible coalition, a step-wise decision algorithm determines the *best path* that covers all the nodes in the chain. In particular, the eNodeB first sorts the devices in a decreasing order of uplink CQI (first those with better channel quality) and then selects the first node in the list as the *gateway* for the coalition. This is important, so that the resource pooling will produce the highest throughput toward the eNodeB for the whole multihop chain. Once the gateway is selected, the *best path* over the set of nodes is computed with focus on the D2D link qualities. In this case, a simple *greedy* approach is exploited where the next hop from the gateway is selected as the one in the one-hop vicinity with the best D2D link quality. Similarly, each node in the chain will select its neighbor based on the best CQI of the direct link to the remaining nodes in the coalition. Once the coalitions are formed in the cell, the eNodeB determines the radio resources assigned to the gateway and to each D2D link and transmits all the information to the UEs so that the transmissions can start.

**Virtual resources allocation and data rate computation**

In order to characterize analytically the virtual resource allocation and the data rate computation, the following sets are defined: $\mathcal{N} = \{1, \ldots, n \ldots, N\}$, the set of UEs in the cell; $\mathcal{M} \subseteq \mathcal{N} = \{m_1, \ldots, m_M\}$, the set of UEs operating in cellular mode; $\mathcal{D} \subset \mathcal{N} = \{d_1, \ldots, d_D\}$, the set of UEs operating in D2D mode; $\mathcal{W} = \{1, \ldots, w, \ldots, W\}$, the set of RBs in the system; and $\mathcal{G} = \{1, \ldots, g, \ldots, G\}$ the set of MCSs in the system [73]. To compute the "virtual" radio resource allocation, the eNodeB determines the RB assignment $\rho_{w,n}$ and the power allocation $Pt_{n,w}\, \forall w \in \mathcal{W}$ and $\forall n \in \mathcal{N}$, so that a utility $U$ is maximized:

$$\max_{\rho_{w,n}, Pt_{n,w}} U, \text{ subject to: } \begin{cases} \sum_{w=1}^{W} Pt_{n,w} \leq P_{\max_n},\ Pt_{n,w} \geq 0 & \forall n \in \mathcal{N} \\ \sum_{n=1}^{N} \rho_{w,n} = 1 & \forall w \in \mathcal{W} \\ \rho_{w,n} \in \{0,1\} & \end{cases} \tag{3.1}$$

where the first constraint limits the transmitted power, with $P_{\max_n}$ being the maximum power for UE $n$; the second constraint states that all RBs should be allo-

[4] The eNodeB assigns to the gateway of each coalition a *pool* of uplink resources, which can reach up to the sum of the radio resources separately requested by the UEs in the coalition (if less resources are needed for the coalition, then the eNodeB will not allocate the whole sum of separately allocable resources).

cated, and the last constraint shows that no subcarrier can be allocated for uplink transmission from multiple UEs. Function $U$ can have different definitions in terms of the specific objectives. For instance, for the case of Maximum Throughput (MT) we have: $U = \sum_{n=1}^{N} \sum_{w=1}^{W} \rho_{w,n} log_2(1+\gamma_{n,w})$, for the case of Proportional Fair (PF) we have: $U = \sum_{n=1}^{N} \sum_{r=1}^{W} ln[\rho_{w,n} log_2(1+\gamma_{n,w})]$ where $\gamma_{n,w}$ is the SINR on the cellular link from UE $n$ in RB $w$. In the performance evaluation section, we will compare the results for the MT and the PF schedulers with those achieved by the Round Robin (RR) policy and a MaxMin fair scheduler (MM) proposed in [76].

The computational complexity of the joint subcarrier and power allocation problem for the multi-user Orthogonal Frequency-Division Multiple Access (OFDMA) system as LTE has been demonstrated in [77] to be NP hard, which is computationally prohibitive (the second formulation proposed in the cited paper corresponds to our problem). A possible way to solve it is through a two-step approach with lower complexity as proposed, e.g., in [78], so that an acceptable suboptimal performance can be achieved. In the first step, the RBs are allocated according to the specific scheduler rules (MT, PF, MM, or RR), based on the assumption of *equal power allocation for all UEs, uniformly distributed over the available RBs*. In the second step, the power level is decided on the performed RB assignment [79].

According to the free space propagation loss model, the power received by a generic UE $j$ in RB $w$ on the direct link $i \rightarrow j$ can be written as: $Pr_{j,w} = Pt_{i,w} \cdot |h_{i,j}|^2 = Pt_{i,w} \cdot Pl_{i,j}^{-\alpha} \cdot |h_0|^2$, where $Pt_{i,w}$ is the transmitted power from UE $i$ in RB $w$, $h_{i,j}$ is the length of the link $i \rightarrow j$, $h_0$ is the channel coefficient, $Pl_{i,j}$ is the path loss on the link $i \rightarrow j$, and $\alpha$ is the path loss compensation factor computed by the eNodeB based on the operation environment (in the [0,1] range). Assuming that all subcarriers in an RB experience the same channel conditions, the SINR $\gamma_{j,w}$ in RB $w$ for the generic user $j$ on a link $i \rightarrow j$ is: $\gamma_{j,w} = \frac{Pt_{i,w} \cdot Pl_{i,j}^{-\alpha} \cdot |h_0|^2}{\mathcal{I}_{j,w} + N_0}$, where $N_0$ is the thermal noise density level at the receiver, and $\mathcal{I}_{j,w}$ is the set of interfering signals received by user $j$ on RB $w$, which has different values in case of either a D2D or a cellular transmission. In particular, the interference on a transmission from a cellular user $m$ to the eNodeB denoted by $bs$ is given by the power signals of all the UEs transmitting in D2D mode on the same RB UE $m$ is transmitting in cellular mode [74]. Given the RBs allocated to the single UEs and the transmission power level for each RB, the SINR value determines the MCS $g$ to be used for the uplink and the corresponding spectral efficiency $eff_w$ (*bits/symbol*). Thus, the bit rate $R_{m,g,w}$ and the throughput $TP_{m,g,w}$ of an uplink cellular transmission from UE $m$, when using MCS $g$ in RB $w$, is defined as in [80]:

$$R_{m,g,w} = \Theta \cdot eff_w = \frac{CS_{ofdm} \cdot SY_{ofdm}}{T_{subframe}} \cdot eff_w \tag{3.2}$$

$$TP_{m,g,w} = R_{m,g,w} \cdot (1 - BLER(w, \gamma_{m,w})) \tag{3.3}$$

where $\Theta$ is a fixed parameter depending on the network configuration, $CS_{ofdm}$ and $SY_{ofdm}$ are the number of subcarriers and symbols per RB respectively, $T_{subframe}$ is the time duration of an RB, and $BLER(w, \gamma_{m,w})$ is the BLock Error Rate suffered by RB $w$. When focusing on a D2D link, the interference at the receiver $d'$ is given by the power signals of all the D2D UEs $d^{''} \in \mathcal{D} \setminus \{d, d'\}$ and the cellular user $m$ that are reusing the same RB as the receiver $d'$. Similar to the cellular links, the SINR value determines the MCS level $g$ to be used for the direct link and the corresponding spectral efficiency $eff_w$ (*bits/symbol*) in RB $w$. The bit rate $R_{d,g,w}$ and the throughput $TP_{d,g,w}$ for a D2D link transmission follow the formulations given in (2) and (3.3).

### 3.2.2 Our proposal from the Cooperative D2D Content Uploading

A non transferable utility (NTU) coalitional game in cost form is defined by the pair $(\mathcal{N}, \mathcal{C})$ where $\mathcal{N} = \{p_1, \ldots, p_N\}$ is the set of $N$ players and $\mathcal{C}$ is a set valued function such that for every coalition $\mathcal{S} \subseteq \mathcal{N}$, $\mathcal{C}(\mathcal{S})$ is a closed convex subset of $\mathbb{R}^{|\mathcal{S}|}$ that contains the cost vectors the players in $\mathcal{S}$ can achieve ($|\mathcal{S}|$ is the number of members in coalition $\mathcal{S}$). Concerning the problem that has been investigated, the players are the single UEs forming a cooperative D2D chain. The objective for the players is to minimize their cost which is measured as the time required to upload their content to the eNodeB. Since this cost cannot be arbitrarily apportioned among the players in a coalition, this translates in a NTU game. Moreover, the game is in characteristic form because the cost of each player only depends on the players forming the coalition it is part of and not on the other players in the network. In particular, this is because players in different coalitions are not causing mutual interference as orthogonal RBs are allocated by the scheduler.

A *collection* of coalitions $\mathcal{K}$ is defined as a set $\mathcal{K} = \{\mathcal{S}_1, \ldots, \mathcal{S}_k\}$ of mutually disjoint coalitions $\mathcal{S}_i \subset \mathcal{N}$ such that $\mathcal{S}_i \cap \mathcal{S}_{i'} = \emptyset$ for $i \neq i'$. If the collection contains all players in $\mathcal{N}$, i.e., $\bigcup_{i=1}^{k} \mathcal{S}_i = \mathcal{N}$, then the collection is a *partition* $\Pi$ or *coalition structure (CS)*. The set of all possible *coalition structures* is identified by $\Pi(N)$. A cost game is said *subadditive* when, given any two disjoint coalitions $\mathcal{S}_1$ and $\mathcal{S}_2$, if coalition $\mathcal{S}_1 \cup \mathcal{S}_2$ forms, then it can give its members any allocations they can achieve when acting in $\mathcal{S}_1$ and $\mathcal{S}_2$ separately [81]. Intuitively, if the game is subadditive, then it is always convenient that players cooperate and join larger coalitions. However, in many real problems this is not always true as there may be inherent constraints on *feasible coalitions* which should be taken into consideration. The motivations behind these constraints are linked to the specific problem and can derive from technological, social, historical

or reputation aspects. For instance, from prior experience it may be known that in order to successfully execute a given task certain alliances of players are indispensable, thus the corresponding coalition is useful, or, on the contrary, specific combinations of players are known to under-perform, so they are to be excluded because considered as harmful. Similarly, constraints may exist on the size of the coalitions to be formed.

However, due to the geographically distribution of the users, some pairs of UEs may not be in reciprocal visibility to set up a D2D link. This indeed introduces a constraint on the *feasible* coalitions that can be formed. To characterize the *feasible* coalitions and coalition structures, we formally define the constraints to the problem with a set of *positive constraints* $\mathcal{P} \subseteq 2^N$ such that a coalition $\mathcal{S}$ satisfies a constraint $P \in \mathcal{P}$ if $P \subseteq \mathcal{S}$, a set of *negative constraints* $\mathcal{Q} \subseteq 2^N$ such that a coalition $\mathcal{S}$ satisfies a constraint $Q \in \mathcal{Q}$ if $Q \nsubseteq \mathcal{S}$, and a set of *size constraints* $\mathcal{Z}$ that defines the constraints on the coalitions size [82].

We formally define the *cooperative D2D-uploading game* in cost form as a tuple $G =< \mathcal{N}, \mathcal{P}, \mathcal{Q}, \mathcal{Z}, \mathcal{C} >$ where $\mathcal{N}$ is the set of UEs in the cell and $\mathcal{S} \subseteq \mathcal{N}$ is any multihop D2D chain, $\mathcal{C}$ is the set of cost vectors the players can achieve in all coalitions $\mathcal{S} \subseteq \mathcal{N}$, $\mathcal{P}$ and $\mathcal{Q}$ subsets of $\mathcal{N}$, and $\mathcal{Z} \subseteq \mathbb{N}$. More generally, a coalition $\mathcal{K} \subseteq \mathcal{N}$ is *feasible* for $G =< \mathcal{N}, \mathcal{P}, \mathcal{Q}, \mathcal{Z}, \mathcal{C} >$ if: (i) $P \subseteq \mathcal{K}$ for some $P \in \mathcal{P}$; (ii) $Q \nsubseteq \mathcal{K}$ for all $Q \in \mathcal{Q}$; and (iii) $|\mathcal{K}| \in \mathcal{Z}$. The set all feasible coalitions is denoted by $f(\mathcal{N}, \mathcal{P}, \mathcal{Q}, \mathcal{Z})$. For a locally constrained game, a coalition structure $CS$ is feasible if and only if $CS \subseteq f(\mathcal{N}, \mathcal{P}, \mathcal{Q}, \mathcal{Z})$.

It is worth noticing that for this work have been considered only negative constraints that are a consequence of bad channel conditions on the D2D links and we set $\mathcal{P} = \emptyset$, $\mathcal{Z} = \emptyset$. For the exact definition of $\mathcal{Q}$ the eNodeB considers the D2D CQI feedback from the UEs. In particular, those coalitions for which a path cannot be constructed are considered *not feasible* and thus stored in $\mathcal{Q}$. When $\mathcal{Q} \neq \emptyset$, it is implicitly said that the grand coalition is not formed as it is certainly not a *feasible* coalition. Nevertheless, also when $\mathcal{Q} = \emptyset$, that is when all UEs are in mutual coverage for a D2D link, the game can be demonstrated to be *non-subadditive* in general. To solve the so-defined game has been taken inspiration from traditional coalitional game formation solutions and apply them to the feasible coalitions only.

### Coalition cost for the content uploading game

For the considered game, has been defined $\mathcal{C} : \mathcal{S} \rightarrow \mathbb{R}^{|\mathcal{S}|}$ such that $\mathcal{C}(\emptyset) = \emptyset$, and for any coalition $\mathcal{S} \subseteq \mathcal{N} \neq \emptyset$ it is a singleton set $\mathcal{C}(\mathcal{S}) = \{\mathbf{c}(\mathcal{S}) \in \mathbb{R}^{|\mathcal{S}|}\}$ where each element of the vector $\mathbf{c}(\mathcal{S})$ is the cost $c_i(\mathcal{S})$ associated to each player $i \in \mathcal{S}$. This cost is defined as the uploading time needed for the data generated by node $i$ to reach the eNodeB. Similarly, the cost $c(\mathcal{S})$ of any coalition $\mathcal{S} \subseteq \mathcal{N}$ is computed as the total

uploading time needed for all data generated in the coalition to reach the eNodeB. In particular, the cost for any singleton coalition is equal to the content uploading time in the cellular mode for the single player $i$. This is computed when the UE uploads its content $b_i$ over its cellular link (i.e., directly to the eNodeB) having a data rate $r_i^c$: $c(\{i\}) = \frac{b_i}{r_i^c}$. For any coalition $\mathcal{S} \subseteq \mathcal{N}$ with cardinality $|\mathcal{S}| > 1$ instead, the associated cost is defined as $c(\mathcal{S}) = UT(\mathcal{S})$, where $UT(\mathcal{S})$ is the data uploading time modeled later in the text. If the multihop D2D chain cannot be formed due to coverage constraints, then is defined: $c(\mathcal{S}) = \sum_{i \in \mathcal{S}} c(\{i\})$.

Let each UE $i \in \mathcal{N}$ have a video file of a given time duration and a predetermined quality ready to upload to the eNodeB. These two parameters determine the data size $b_i \neq 0$ of the content to upload. It is known that the "virtual" radio resources for the UEs in a multihop D2D coalition are available to the gateway if needed. As a result of having more resources available, a higher uplink data rate $r_i^{c'}$ is obtained for the gateway-to-eNodeB link. Then, $r_i^d$ represents the data rate for UE $i$ on the D2D outgoing link to the next UE in the multihop chain. To define the $c(\mathcal{S})$ term the number of LTE frames required to transfer all the content from the UEs in the coalition to the eNodeB have to be computed. In doing this, it is necessary to quantify the time intervals and the TTIs needed for data transmission on the D2D links, by following the listed steps:

1. Compute the channel occupation time for a generic UE $i$ in the multihop D2D chain; this is the time spent by the UE to transmit to the next hop its own data and the data received from the previous UE in the chain.
2. Compute the time to upload the contents of the entire chain; to this aim compute the number of U subframes used by the gateway and the second UE in the chain to relay the received data to the eNodeB and to the gateway, respectively.
3. Based on the data frame structure, compute the number of data frames according to the time in terms of TTIs required for uploading all data in the chain.

Regarding the first step, UE $N$ has been considered as the last UE in the chain. For the sake of notation simplicity, once the best path over the UEs in a coalition is computed (i.e., the multihop chain is identified), the N-hop path with $i = 1$ being the gateway and $i = N$ being the last UE in the path is considered. Indeed, UE $N$ will occupy the channel for a time $T_N = b_N / r_N^d$ to forward its data of size $b_N$ to UE $N-1$ over the D2D link having data rate $r_N^d$. Considering UE $N-1$, to send its own data of size $b_{N-1}$ and the data received from the previous UE which is of size $b_N$ the channel will be occupied for a time $T_{N-1} = b_{N-1}/r_{N-1}^d + \left(b_N/r_N^d + b_N/r_{N-1}^d\right)$. By repeating this reasoning for all UEs in the chain, and considering that the gateway, UE 1, transmits to the eNodeB with a data rate $r_1^{c'}$, we compute the channel occupation

time $T_1(N)$ for the gateway to upload all data from the D2D chain to the eNodeB as a function of the number of UEs in the chain:

$$T_1(N) = \frac{b_1}{r_1^{c'}} + \left(\frac{b_2}{r_2^d} + \frac{b_2}{r_1^{c'}}\right) + \left(\frac{b_3}{r_3^d} + \frac{b_3}{r_2^d} + \frac{b_3}{r_1^{c'}}\right) + \ldots + \left(\frac{b_N}{r_N^d} + \frac{b_N}{r_{N-1}^d} + \ldots + \frac{b_N}{r_1^{c'}}\right) = \sum_{i=1}^{N}\left(\frac{b_i}{r_1^{c'}} + \sum_{j=2}^{i}\frac{b_i}{r_j^d}\right). \tag{3.4}$$

The formulation can be generalized to the channel occupation time for any UE $n = \{1, \ldots, N\}$ in the multihop chain. This includes the time to forward to the next hop in the chain all data generated by UE $n$ and the data from its previous UEs in the chain as given below.

$$T_n(N) = \begin{cases} \sum_{i=n}^{N}\left(\frac{b_i}{r_1^{c'}} + \sum_{j=2}^{i}\frac{b_i}{r_j^d}\right) & n = 1 \\ \sum_{i=n}^{N}\sum_{j=n}^{i}\frac{b_i}{r_j^d} & n > 1 \end{cases} \tag{3.5}$$

Considering step 2, since all UEs in a cooperative multihop D2D chain generate data, and given the decode-and-forward relaying assumption, the total uploading time $T(N)$ will be determined by the sum of the occupation time of the first two UEs in the chain: $T_1(N)$ and $T_2(N)$. Hence, the corresponding number of uplink subframes can be computed. Given the *type 2 configuration 0* LTE frame structure, when relaying data from previous UEs toward the eNodeB, the gateway will use all six available U subframes when all the data from the previous UEs have been received. Only at that time, no subframe is used to receive additional data. Before that moment, only three U subframes per LTE data frame can be used by the gateway to transmit to the eNodeB, since the other three are used to receive data from the previous UEs.

On the other hand, on all the D2D links the UEs will use three U subframes per frame to transmit and three subframes per frame to receive data. Based on these considerations, the $T_1(N)$ term is splitted into two contributions (see (3.7)), namely one contribution (i.e., $T_1^{'}(N)$) where three subframes per frame are used by UE 1 to upload data, and a second term (i.e., $T_1^{''}(N)$) where six subframes per frame are used by UE 1 to upload data. In particular, the second term has a different value according to the relation between $\frac{b_{N-1}}{r_1^{c'}}$ and $\frac{b_N}{r_N^d}$. The first case refers to the situation where the data from UE $N$ reaches the gateway only after all data from the other UEs in the chain have already been uploaded to the eNodeB. In this case, six subframes are used to upload the data from UE $N$ only. In the other cases, besides the data from UE $N$, also a portion of data from UE $N-1$ will be uploaded using six subframes.

$$T_1^{'}(N) = T_1(N) - T_1^{''}(N) \tag{3.6}$$

$$T_1^{''}(N) = \begin{cases} \frac{b_N}{r_1^{c'}} & \frac{b_{N-1}}{r_1^{c'}} \leq \frac{b_N}{r_N^d} \\ \frac{b_N}{r_1^{c'}} + \frac{b_{N-1} - (b_N / r_N^d) \cdot r_1^{c'}}{r_1^{c'}} & \frac{b_{N-1}}{r_1^{c'}} > \frac{b_N}{r_N^d} \end{cases} \tag{3.7}$$

The number of U subframes $F(N)$ required to transmit all data is computed by simply dividing the occupation time of the first two UEs in the chain by the U subframe duration (TTI=1ms): $F_1^{'}(N) = T_1^{'}(N)/TTI$, $F_1^{''}(N) = T_1^{''}(N)/TTI$ and $F_2(N) = T_2(N)/TTI$.

In step 3, the total number of data frames required for uploading data from all the UEs in the D2D chain can be determined as: $\left[ \frac{F_1^{'}(N)}{3} + \frac{F_1^{''}(N)}{6} + \frac{F_2(N)}{3} \right]$. Being ten the total number of TTIs in the LTE data frame, the total content uploading time to the eNodeB, i.e., the cost in coalition $\mathcal{N}$, is given by the number of data frames needed:

$$c(\mathcal{N}) = UT(N) = 10 \cdot TTI \cdot \left[ \frac{F_1^{'}(N)}{3} + \frac{F_1^{''}(N)}{6} + \frac{F_2(N)}{3} \right]. \tag{3.8}$$

Finally, the uploading time of a specific UE in the D2D chain can be computed according to the UE position in the chain that determines the priority order for data transmission. The data delivery time from the last UE in the chain, UE $N$, is equal to the total time: $UT^N(N) = UT(N)$. The data delivery time from a generic UE $i$ is computed by repeating the same reasoning on the sub-chain from $n$ to the gateway, so the cost for player $i$ in the coalition is $c_i(\mathcal{N}) = UT^i(N) = UT(i)$.

**Feasible coalition formation algorithm**

The set of all possible partitions of $\mathcal{N}$ has a total number of $B_N$, where $B_N$ is the N-th Bell number [83], and it grows exponentially with the number of UEs $N$. Thus, finding the optimal partition via exhaustive search through all possible partitions is not feasible, as it is an NP-complete problem [84]. To characterize the *feasible coalitional structure* to form for the game, a simple merge-and-split rules [85] has been proposed. The key mechanism is to enable players to join or leave a coalition based on well-defined preferences so that each player is able to compare and order its potential coalitions based on which coalition it prefers to be a member of [86].

**Definition 3.1 (Preference order).** *The preference order $\succ_i$ for any player $p_i \in \mathcal{N}$, is defined as a complete, reflexive, and transitive binary relation over the set of all feasible coalitions that player $p_i$ can possibly form, i.e., the set $\Pi_i$ of coalitions containing $p_i$.*

A UE can decide to join or leave a coalition according to its preference order. In particular, for each player $p_i$, if $C \succ_i C'$, $p_i$ prefers being a member of coalition $C$

more than coalition $C'$. A less restrictive preference order is $C \succeq_i C'$, whereby player $p_i$ prefers coalition $C$ at most as much as coalition $C'$. In the case investigated during this work, the preference order is defined according to its *individual cost*. Thus, for each UE $p_i \in \mathcal{N}$ and for all $C, C' \in \Pi_i$, we say that:

$$C \succ_i C' \Leftrightarrow c_i(C) < c_i(C') \wedge c_j(C') \leq c_j(C' \backslash \{i\}), \forall j \in \{C' \backslash \{i\}\} \wedge \\ c_j(C) \leq c_j(C \backslash \{i\}), \forall j \in \{C \backslash \{i\}\}. \tag{3.9}$$

In words, any UE $i$ prefers being a member of coalition $C$ over $C'$ if it obtains a lower individual cost $c_i(C)$, without causing an increase in the cost for any other player in $C$ and $C'$ (*Pareto order* preference).

The preference order is at the basis of the two rules for the coalition formation game.

**Definition 3.2 (Merge rule).** *Merge any pair of coalitions $C$ and $C'$ into a unique feasible coalition $\{C \bigcup C'\} \Leftrightarrow [(\exists k \in C \text{ s.t. } \{C \bigcup C'\} \succ_k C) \vee (\exists k \in C' \text{ s.t. } \{C \bigcup C'\} \succ_k C')] \wedge \{C \bigcup C'\}$ is feasible.*

**Definition 3.3 (Split Rule).** *Split any coalition $\{C \bigcup C'\}$ in feasible coalitions $\{C, C'\} \Leftrightarrow [(\exists i \in C \text{ s.t. } C \succ_i \{C \bigcup C'\}) \vee (\exists j \in C' \text{ s.t. } C' \succ_j \{C \bigcup C'\})] \wedge \{C, C'\}$ are feasible.*

The merge rule implies that two coalitions join to form a larger *feasible coalition* if operating all together strictly reduces the cost of at least one player, while all the other involved players do not experience a higher cost. The split rule implies that a coalition splits only if there exists at least one player that obtains a lower cost, under the constraint that this has no negative effect on the cost of the other players and the resulting coalitions are both *feasible*.

The game is implemented by the eNodeB, as summarized in Algorithm (1). The objective of a UE is to find a coalition that guarantees the lowest uploading time through an iterative application of the merge and the split rules. By starting from an initial partition $\Pi^{ini}(N) = \mathcal{N} = \{p_1, p_2, \ldots, p_N\}$, the eNodeB iteratively applies the merge and split rules to any pair of coalitions in the partition. In particular, the merging process stops when no couple of coalitions exists in the current partition $\Pi^{cur}(N)$ that can be merged. Thus, the split rule is applied to every coalition in the partition, by updating $\Pi^{cur}(N)$ if a split is applied. When no split occurs, the algorithm considers again the merging function. The algorithm terminates when no merging or splitting occurred in the last iteration. In this case, the final resulting partition $\Pi^{fin}(N)$ will be adopted by the eNodeB. Moreover, the network structure is adapted to environmental changes by periodically repeating the solution computation.

In particular, in a dynamic environment, the period of time for the update should be chosen depending on how rapidly the conditions change.

**Algorithm 1:** Coalition formation for cooperative D2D multihop data uploading

**Data:** Set of UEs $\mathcal{N}$

**Result:** Coalition structure $\Pi^{fin}$

*Phase I - Neighbor Discovery:*

- Each UE discovers neighboring UEs and sends feedback to the eNodeB about the CQI on the corresponding D2D links.
- Partition the network by $\Pi^{ini}(N) = \mathcal{N} = \{p_1, p_2, \ldots, p_N\}$.
- Set the current partition as $\Pi^{cur}(N) = \Pi^{ini}(N)$.

*Phase II - Coalition Formation:*

In this phase the eNodeB performs the coalition formation using merge-and-split.

**repeat**

  **repeat**

    For every UE $i \in \mathcal{N}$ in the current partition $\Pi^{cur}(N)$:

    - UE $i$ investigates possible *merge* operation using the preference order given in (3.9).
    - If a *merge* operation is performed update the current partition $\Pi^{cur}(N)$.

  **until** *no merge occurs*;

  **repeat**

    For every UE $i \in \mathcal{N}$ in the current partition $\Pi^{cur}(N)$:

    - UE $i$ investigates possible *split* operation using the preference order given in (3.9).
    - If a *split* operation is performed update the current partition $\Pi^{cur}(N)$.

  **until** *no split occurs*;

**until** *no merge nor split occur*;

*Phase III - Cooperative content uploading:*

- The network is partitioned using $\Pi^{fin}(N) = \Pi^{cur}(N)$.
- The eNodeB informs the UEs how to operate using the multihop D2D relaying.

*Adaptation to network changes (periodic process):* Periodically the algorithm is repeated to allow the network topology configuration to adapt to environmental changes.

Considering the finite number of partitions, it can be proved by contradiction that the proposed merge and split coalition formation algorithm converges to a *stable* final partition of disjoint coalitions of UEs (for more details see, e.g., [85]).

**Complexity analysis**

The complexity of the proposed algorithm is related to the iterative implementation of merge-and-split operations. Indeed, by considering the worst case for the merge operation, where each coalition needs to make a merge attempt with all the other coalitions in the partition, at the beginning all UEs act non-cooperatively and form $N$ singleton coalitions. In the worst case, the first merge occurs after $\frac{N(N-1)}{2}$ attempts, the second requires $\frac{(N-1)(N-2)}{2}$ attempts and so on [87], [88]. The total worst case number of merge attempts is $O(N^3)$ [89]. However, in practical settings, the merge process requires a significantly lower number of attempts. In fact, after the first run of the algorithm, the initial $N$ singleton coalitions will merge to form larger coalitions.

As regards the split rule, splitting can imply finding all the possible partitions of size two for each coalition in the current network partition. However, the split operation is restricted to the already formed coalitions, which are typically not the grand coalition. Even if this reduces the complexity, this could be further reduced by the fact that, in a practical setting, it is not required to go through all the split forms. As soon as a coalition finds a split form, the UEs in this coalition will split, and the search for further split forms is not required.

In any case, it is important to underline that solving the proposed merge-and-split based algorithm has a complexity far lower than optimally solving the coalition formation problem (which is unfeasible, due to its NP-complete nature [84]). The actual reduction in the merge-and-split complexity w.r.t. to the worst case will also become evident in the performance evaluation in Section 3.2.3. For instance, in case of 26 UEs in the network, the observed average number of coalitions will be in the order of 5, which means a reduction of a factor of 5.2 in the maximum number of the future merge attempts. Similarly, the average number of UEs per coalition is also of a few UEs per coalition, which means that the split attempts are also reduced w.r.t. worst case.

### 3.2.3 Performance Evaluation

A numerical evaluation is conducted by using MATLAB® to assess the performance of the proposed solution. In the considered scenario the end-users are willing to upload a video to Youtube, which is mostly comprised of short video clips and 97.9% video lengths are within 600 seconds [90]. It was assumed that the end-users define the video

quality to upload beforehand according to the MPEG-2 encoding possibilities [91]. The selected video quality implicitly determines also the amount of data to be uploaded. In fact, MPEG-2 supports different video quality levels with a corresponding maximum bitrate and frame size for each video resolution. In such a case, the following bitrate values to characterize the video quality have been considered: $[3, 6, 10, 20]$ $Mbps$. As a side note, differences in the data amount mean also differences in the "acceptable" uploading time for the UEs. This parameter can be tuned according to the constraints set by the specific service scenarios in which the proposed solution is applied.

The assessment campaign has been conducted by following the system model guidelines in [73]. The main simulation parameters are listed in Table 4.3. A single cell with available radio resources $RB = 50$ has been considered, wherein up to 26 UEs are uniformly distributed. Channel conditions for the UEs are measured by the SINR experienced over each sub-carrier [92] when path loss and fading phenomena affect the signal reception. Further, the radio resources that can be used on a single D2D link of the multihop chain depend on the frequency reuse efficiency. In particular, the two extreme cases have been considered: the so-called *best-case*, in which all radio resources can be reused on the D2D links since there is no interference between D2D and uplink transmission, and the *worst-case*, where the transmissions in the multihop D2D chain interfere on all radio resources. In this latter case, the radio resources that can be used on a D2D transmission are limited to the virtual resources allocated by the eNodeB to the involved pairs of UEs. Only results relevant to the two cited cases are reported in the performance analysis, as these represent the lower and upper bounds and all other cases of radio resource re-use on the D2D links fall in-between them. The performance evaluation have focused on: (i) *the UE average data uploading time gain*, (ii) *the multihop D2D chain configuration*, and (iii) *the UE average energy consumption gain*. Here gain is intended as the improvement in the delay and in the energy consumption that is achieved by a cooperative upload w.r.t. a pure cellular upload modality. The analysis also evaluates the effects of the RRM policy implemented by the eNodeB, i.e., *maximum throughput (MT)*, *proportional fair (PF)*, *maxmin fair(MM)*, and *round robin (RR)* schedulers.

Although the main of the system evaluation has been on the data uploading time reduction, also the impact of the proposed scheme on the UEs' energy consumption has been monitored. In the cellular mode, the energy consumption for a generic UE $i$ is a function of the transmitted amount of data $b_i$ and it is equal to the power consumption on the uplink toward the eNodeB multiplied by the time where the UE is active to transmit: $E_i^c(b_i) = (P_{tx}^c + P_0) \cdot \frac{b_i}{r_i^c}$.

In particular, the power consumption of UEs includes two contributions, the transmission power $P_{tx}^c$ and the circuit power $P_0$, being this latter the power consumed by

all the circuit blocks along the signal path that cannot be ignored. When considering the cooperative data uploading, three cases may be present: (1) the UE is the gateway; it consumes energy in receiving data from the second UE and in transmitting data to the eNodeB; (2) the UE is the last UE in the chain; it only consumes energy in transmitting its own data to the next UE in the D2D chain; (3) the UE is an intermediate UE in the chain; it consumes energy to receive data from the previous UE and to transmit data to the next UE in the chain. In all three cases, energy is also spent during the idle times on the channel. However, according to [93] the power consumption in idle times is as low as $-50dbm$; therefore, this contribution can be neglected and only the transmitting and receiving power on the D2D links, $P_{tx}^d$ and $P_{rx}^d$, are considered. The energy consumption for a generic UE $i$ in the D2D chain will be the sum of the energy spent for transmission and for reception: $E_i(N) = Etx_i^d(N) + Erx_i^d(N)$.

$$Etx_i^d(N) = \begin{cases} (P_{tx}^c + P_0) \sum_{j=1}^{N} \frac{b_j}{r_1^{c'}} & i = 1 \\ (P_{tx}^d + P_0) \sum_{j=i}^{N} \frac{b_j}{r_i^d} & 1 < i \leq N \end{cases} \tag{3.10}$$

$$Erx_i^d(N) = \begin{cases} (P_{rx}^d + P_0) \sum_{j=i+1}^{N} \frac{b_j}{r_{i+1}^d} & 1 \leq i < N \\ 0 & i = N \end{cases} \tag{3.11}$$

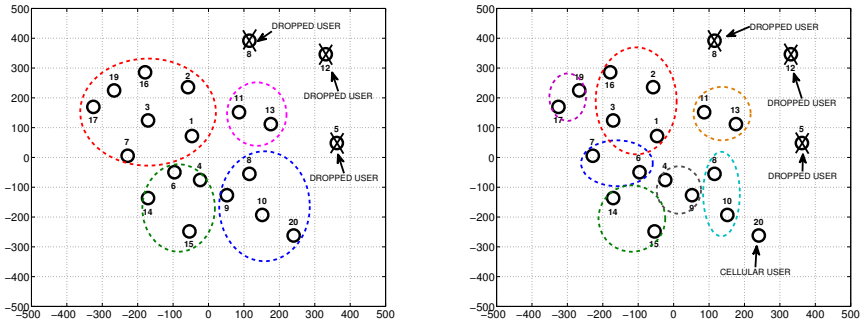
**Analysis of a sample study case**

For increasing the comprehensiveness and the understanding of the proposed game theoretical D2D multihop uploading scheme, a sample study case with the MT resource allocation policy implemented at the eNodeB (similar analysis can be done with the PF, MM and RR schemes) has been taken into account. The objective is to investigate the coalition formation process for the case with $N = 20$ UEs in the cell, and to compute the gains for each UE in the cooperative D2D chain. In Fig. 3.4 the resulting coalitions are shown for the best and worst case analysis. As it can be observed, differences in terms of length and number of coalitions are obtained as a consequence of the resource reuse possibilities on the D2D links. In particular, in the best-case a smaller number of coalitions is formed and longer D2D chains are created (they can reach the length of seven UEs), whereas in the worst-case the longest chain is of four UEs. The motivation for this behavior is related to the lower amount of radio resources available on the D2D links in the worst-case, which reduces the cooperation possibilities and gains. It is worth noticing also that three UEs (i.e., UEs 5, 8 and 12) do not receive radio resources from the MT scheduler as they experience very bad channel conditions. Moreover, in the worst-case analysis, UE 20 is not joining

**Table 3.3.** Main Simulation Parameters

| Parameter | Value |
|---|---|
| Cell radius | 500 m |
| Maximum D2D link coverage | 100 m |
| Frame Structure | Type 2 (TDD) |
| TTI | 1 ms |
| Cyclic prefix/Useful signal frame length | 16.67 $\mu s$ / 66.67 $\mu s$ |
| TDD configuration | 0 |
| Carrier Frequency | 2.5 GHz |
| Cellular transmission power consumption | 23 dBm |
| D2D power consumption | -19 dBm |
| CQI-MCS mapping for D2D links | [94] |
| Noise power | -174 dBm/Hz |
| Path loss (cell link) | 128.1 + 37.6 log(d), d[km] |
| Path loss (D2D link, NLOS) | 40 log(d) + 30 log(f) + 49, d[km], f[Hz] |
| Path loss (D2D link, LOS) | 16.9 log(d) + 20 log (f/5) + 46.8, d[m], f[GHz] |
| Shadowing standard deviation | 10 dB (cell mode); 12 dB (D2D mode) |
| Sub-carrier spacing | 15 kHz |
| BLER target | 1% |
| # of Runs | 500 |

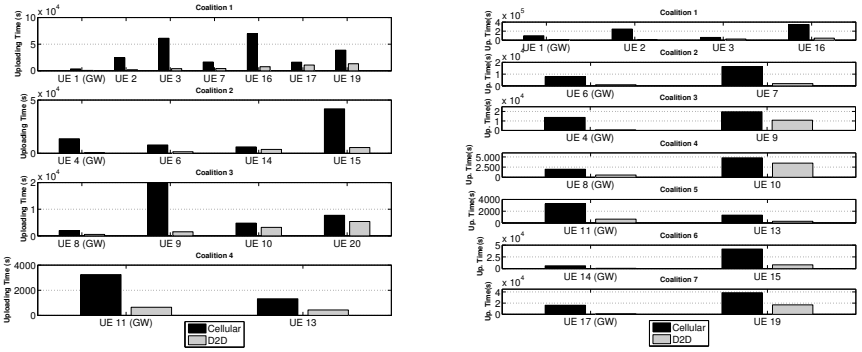
any coalition and will operate in traditional cellular mode, since no other UE finds it advantageous to merge with it in a coalition.




(a) Best-case (b) Worst-case

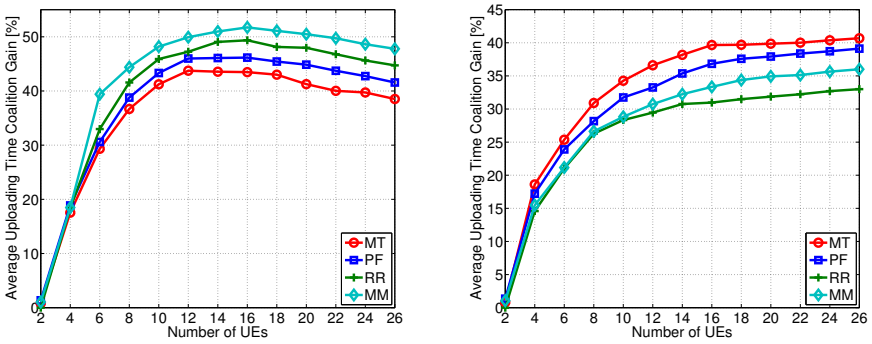
**Fig. 3.4.** Coalitions in a sample study case with $N = 20$, based on the MT radio resource allocation policy.

Further, details of the single coalitions being formed and the uploading time gain for the single UEs have been considered. In the plots in Fig. 3.5, the UE playing the gateway role in each coalition is highlighted with the (GW) notation. The first




(a) Uploading time for best-case (b) Uploading time for worst-case

**Fig. 3.5.** Uploading time for the coalitions in a sample scenario with $N = 20$, based on the MT radio resource allocation policy.




(a) Best-case (b) Worst-case

**Fig. 3.6.** Average data uploading time gain for UEs in the D2D chain.

and most important observation is that all UEs in all the coalitions gain from an uploading time reduction when the cooperative technique is used. This is an expected result according to the individual preference relation set for the single UEs when joining a coalition. In particular, in all coalitions the gateway obtains the highest uploading time gain. This is also an expected result as the radio resources are pooled together and the highest transmission priority is given to the data of the gateway itself.

To complete this first analysis, in Table 3.4 further information has been reported about the energy consumption gain, the allocated RBs, and the content size generated by the UEs. It is particularly interesting to observe that all UEs do not only achieve uploading time gains, but in most of the cases, they also achieve energy consumption gains. Surprisingly, also the gateways (highlighted with bold text in the Table) will save energy in some of the coalitions, e.g., UE 4 in the best-case configuration, and UEs 4, 14 and 17 in the worst-case configuration. Noteworthy, this happens for small coalitions when the total data in the chain is small and the transmission time on the

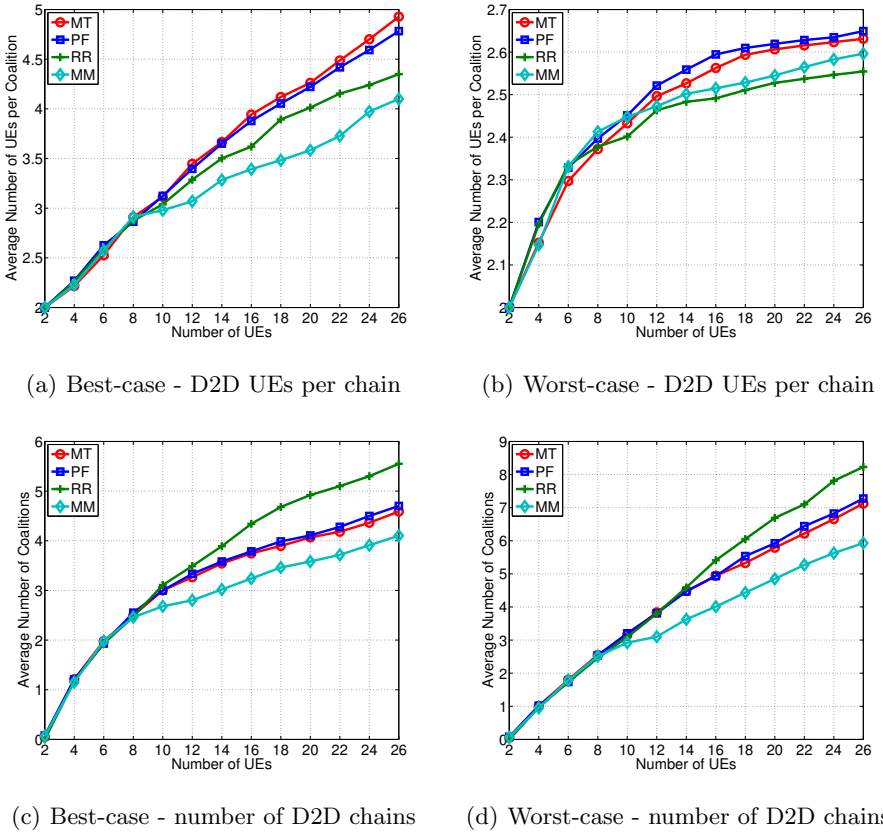
**Table 3.4.** Energy consumption, RBs and data size for a sample case with $N = 20$ and MT radio resource allocation.

| UE ID | Assigned RBs | Data size [MB] | $E^c$ [Joule] | $E^d$ [Joule] (Best-case) | Energy Gain [%] (Best-case) | $E^d$ [Joule] (Worst-case) | Energy Gain [%] (Worst-case) |
|---|---|---|---|---|---|---|---|
| 1 | 5 | 1003 | 423,94 | 663,91 | **-36,14** | 664,40 | **-36,19** |
| 2 | 1 | 1031 | 29414,38 | 0,236 | +99,99 | 1,47 | +99,99 |
| 3 | 6 | 1026 | 7313,37 | 0,0319 | +99,99 | 0,22 | +99,99 |
| 4 | 1 | 777 | 1641,32 | 663,96 | **+59,54** | 663,97 | **+59,54** |
| 5 | X | X | X | X | X | X | X |
| 6 | 3 | 68 | 939,20 | 0,184 | +99,98 | 995,82 | **-5,68** |
| 7 | 2 | 873 | 1964,70 | 0,083 | +99,99 | 0,0018 | +99,99 |
| 8 | 7 | 991 | 237,14 | 597,53 | **-60,31** | 597,53 | **-60,31** |
| 9 | 2 | 868 | 2355,31 | 0,068 | +99,99 | 0,044 | +99,99 |
| 10 | 6 | 613 | 572,77 | 0,071 | +99,98 | 0,020 | +99,99 |
| 11 | 5 | 180 | 388,59 | 995,82 | **-60,97** | 995,85 | **-60,97** |
| 12 | X | X | X | X | X | X | X |
| 13 | 3 | 629 | 158,58 | 41,33 | +99,99 | 0,013 | +99,99 |
| 14 | 3 | 971 | 719,73 | 0,097 | +99,98 | 665,01 | **+7,60** |
| 15 | 1 | 733 | 4995,85 | 0,045 | +99,99 | 0,56 | +99,98 |
| 16 | 1 | 487 | 41538,03 | 0,012 | +99,99 | 0,083 | +99,99 |
| 17 | 1 | 874 | 1936,45 | 0,014 | +99,99 | 829,91 | **+57,14** |
| 18 | X | X | X | X | X | X | X |
| 19 | 1 | 699 | 4630,95 | 0,0012 | +99,99 | 0,032 | +99,99 |
| 20 | 2 | 1304 | 923,46 | 0,031 | +99,99 | X | X |

cellular links is low. This result is interesting, since although the main objective of the proposed solution is to achieve gain in the data uploading time, also energy saving is obtained in small coalitions thanks to the low power consumption on the D2D links.

### Analysis with a variable number of UEs

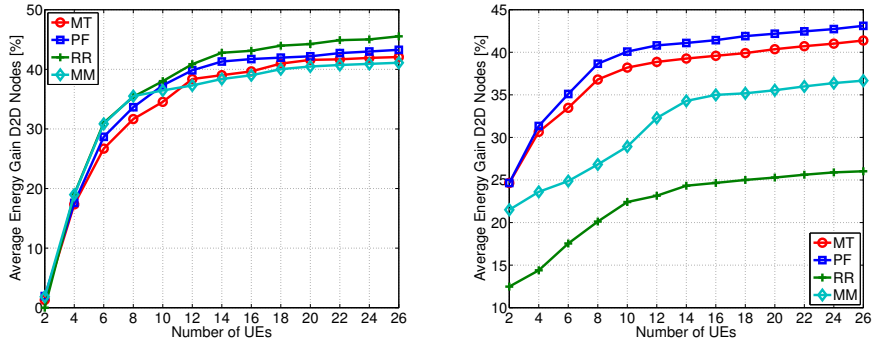
In Fig. 3.6 the average data uploading time gain in the formed coalitions is presented for a variable number of UEs ($N = 2, \ldots, 26$) uniformly deployed in the cell. As expected, lower gain values are obtained in the worst-case analysis, since less resources can be reused on the D2D links. As it can be noticed, the gain increases with the number of UEs reaching a maximum value of $44 - 52\%$ in the best-case (in the MM, RR, PF, MT decreasing order) with 16 UEs, whereas it slightly decreases for larger numbers of UEs. The main motivation for this trend is that with small numbers of UEs a higher number of RBs can be allocated to the single UEs by the scheduler. This increases the potential gains introduced by the resource pooling at the gateway and leads to a higher average uploading time gain for the formed coalitions.




(a) Best-case - D2D UEs per chain (b) Worst-case - D2D UEs per chain

(c) Best-case - number of D2D chains (d) Worst-case - number of D2D chains

**Fig. 3.7.** Configuration of multihop D2D chains as a result of the coalition formation game.




(a) Best-case (b) Worst-case

**Fig. 3.8.** Average energy consumption gain for UEs in the D2D chain.

When observing the worst case the average uploading time gain for the formed coalitions reaches a maximum value of $33-41\%$ with 26 UEs (with the MT, PF, MM, RR decreasing order). Differently from the best case, a peak value with a decreasing trend after that peak is not observed. The reason for this is that since the D2D links are less performing due to the high interference, the average gains over the formed coalitions is lower and smaller coalitions are formed in all cases (this will be shown in

the next Figures). As a consequence the obtained gains are on average lower for any value of UEs in the system.

Further, as observed in the sample study case presented above, when considering the single UEs in the D2D chains, it can be observed that, in almost all cases, the gateway has a much higher gain w.r.t. to the last UE in the chain, whereas the other UEs will have a gain falling in-between the two previous cases (plots are not reported due to length constraints). The reason behind this is again that the first UE will have many more resources to upload its own data first. This is an important observation when also considering that, in general, the first UE will also have higher energy consumption. Noteworthy, also in these cases some of the UEs with bad channel conditions are dropped by the scheduler or are working in cellular mode in singleton coalitions and the number of singleton coalitions is larger in the worst-case analysis.

Finally, the results for the average number of UEs that is joining a multihop D2D chain and length are presented in Fig. 3.7. In particular, in Fig. 3.7(a) and Fig. 3.7(b) the average number of UEs joining a multihop D2D chain is reported for the best-case and the worst-case analysis, respectively. This value increases with the total number of UEs in the cell. In particular, in the best-case the average length ranges between 2 and 5, whereas in the worst-case it ranges between 2 and 2.6. The average number of coalitions being formed, reported in Fig. 3.7(c) and Fig. 3.7(d), increases with the number of UEs with a maximum of 5.6 and of almost 8.2, in the best and worst-case respectively.

To conclude the performance evaluation, the average energy consumption for the UEs in the D2D chain has been considered. Is observed that all the cooperating UEs except the gateway save their energy; see Fig. 3.8 where the average energy consumption gain for the best and worst-case analysis is reported for all the UEs but the gateway. This positive side-effect can reach a 46% gain in the best-case and 43% in the worst-case. Noteworthy, the scheduling policy has a higher impact in the worst-case analysis. Finally, has been observed that in some of the tested cases, in particular with small coalitions, the gateway also achieves energy savings. As an example, the reader can refer to the UE with ID 4 in the sample case reported in Table 3.4. Even if this is not true in general, the strong benefits achieved by the first UE in terms of uploading time (usually much higher than for the other UEs) justify the assumption of its willingness to act as a gateway for the multihop D2D chain.

## 3.3 IoT energy-aware D2D data collection

Fifth Generation (5G) wireless systems are expected to connect an avalanche of "smart" objects disseminated from the largest "Smart City" to the smallest "Smart

Home". In this vision, Long Term Evolution-Advanced (LTE-A) is deemed to play a fundamental role in the Internet of Things (IoT) arena providing a large coherent infrastructure and a wide wireless connectivity to the devices. However, since LTE-A was originally designed to support high data rates and large data size, novel solutions are required to enable an efficient use of radio resources to convey small data packets typically exchanged by IoT applications in "smart" environments. On the other hand, the typically high energy consumption required by cellular communications is a serious obstacle to large scale IoT deployments under cellular connectivity as in the case of Smart City scenarios. Network-assisted Device-to-Device (D2D) communications are considered as a viable solution to reduce the energy consumption for the devices. Taking into account this issues and challenges, in this line of research an *IoT energy-aware D2D data collection* has been proposed that consists in appointing one of the IoT smart devices as a collector of all data from a cluster of objects using D2D links, thus acting as an *aggregator* toward the eNodeB. By smartly adapting the Modulation and Coding Scheme (MCS) on the communication links, it has been shown that is possible to maximize the radio resource utilization as a function of the total amount of data to be sent. A further benefit that has been highlight during this research is the possibility to reduce the transmission power when a more robust MCS is adopted.

In particular, the reference scenario considered within this research topic is a small-scale area belonging to a Smart City environment where a single LTE-A cell with several IoT devices deployed within the cell. A User Equipment (UE) in a LTE-A network can either communicate through the serving eNodeB (*i.e., cellular mode*) or it can bypass the eNodeB and use direct communications over D2D links (*i.e., D2D mode*). In this latter case, the eNodeB is in charge of the D2D session setup (e.g., bearer setup) [49], while power control and resource allocation procedures on the D2D links can be executed either in a distributed or in a centralized way [72]. In this case, it has been assumed a network-assisted D2D communications environment, where the coordination between radio interfaces is controlled by the LTE-A base station (i.e., the eNodeB). In particular, the transmission mode (i.e., either cellular- or D2D-mode), interference management and scheduling tasks are all managed by the eNodeB. Uplink cellular resources are allocated to D2D communications, which is a common choice in the literature [50], because it makes frequency reuse less challenging as the introduced interference is significantly lower w.r.t. the use of downlink resources.

The eNodeB is in charge to manage the spectrum by assigning the adequate number of Resource Block (RB) pairs to each scheduled UE and by selecting the MCS for each RB pair. Scheduling procedures are based on the *Channel Quality Indicator* (CQI) computed by the eNodeB based on the signal-to-interference-plus-noise ratio

(SINR) feedback transmitted by a UE to the eNodeB. The CQI is associated to a maximum supported MCS as specified in [71] (see Table 3.1 within Section 3.1.1). To handle the variations of the radio channel conditions, the Adaptive Modulation and Coding (AMC) mechanism adjusts the transmission rate by selecting the proper MCS. The MCS parameters can be adapted at every *CQI Feedback Cycle* (CFC), which can last one or several Transmission Time Intervals (TTIs) (i.e., 1ms). Each radio resource includes two logical parts: the Transport Block (TB) carrying the Medium Access Control (MAC) header and the Service Data Unit; and the overhead consisting of redundancy bits generated by physical layer processing such as Cyclic Redundancy Check (CRC) insertion and channel coding. The TB Size (TBS) depends on the selected MCS. It is worth noting that when the largest available modulation scheme (64-QAM) is used, i.e., when the channel quality if very good, the largest TBS can convey up to 712 bits of payload [95], which is well beyond the typical data size for most IoT applications, thus leading to low efficiency in the use of radio resources.

### 3.3.1 LTE Standard IoT Data Uploading

For the purpose of the research, an LTE-A eNodeB that receives data from a set of IoT devices within a single TTI has been considered. If the data to be sent to the eNodeB requires multiple TTIs, the same solution is applied in the consecutive TTIs. Data uploading in the standard *cellular mode* occurs through the activation of separate links from each UE to the eNodeB.

Formally, let $\mathcal{K}$ be the set of $K$ LTE-A equipped devices, with each device having some data $d_k$ to upload in a TTI. Let $C$ be the number of available CQI levels and let $c_k \in \{1, 2, \ldots, C\}$ be the CQI reported by device $k \in \mathcal{K}$ in the uplink. Each CQI level is associated to a given supported MCS. For a given MCS value $m$, the bits per RB that can be sent depend on the spectral efficiency for the given MCS, $b_m$ expressed in bit/s/Hz as reported in Table 3.1. Moreover, let $\mathcal{R}$ be the set of $R$ radio resources (the RBs) that can be allocated to the UEs in $\mathcal{K}$.

Further is assumed that the eNodeB implements a simple Round Robin allocation, whereby the whole set of radio resources is equally shared by all *cellular mode* UEs. In addition, the transmission power for a device is equally distributed over the available RBs. Hence, the maximum uplink resources allocated to each UE will be $r_k = \lceil R/K \rceil$, $\forall k \in \mathcal{K}$. The maximum data rate for UE $k$ is proportional to the number of allocated resources $r_k$ and the CQI level $c_k$. However, the IoT data is typically of small size (a few bytes) and typically, one RB per single TTI is enough to upload all the data. On the contrary, as also pointed out in [95], when using the largest available modulation scheme (64-QAM), a payload size of 1 byte leads to a low usage of the RB (39,32% of its capacity).

The low efficient use of the radio resources has, at the same time, an impact on energy efficiency. In particular, with the classic *cellular mode* IoT data uploading, the energy consumption is intrinsically determined by the amount of data to upload and the efficiency $b_m$ for the MCS of each device. In general, the energy efficiency can be defined as the ratio between the amount of data to upload expressed in bits, $D$, and the energy consumption, $E$, $\eta = D/E$. The energy consumption for user $k$ over the LTE-A link can be computed as the product of the transmit power $P_k$ per single RB, the number of allocated RBs $r_k$ and the transmission time, i.e., the TTI: $E_k = P_k \cdot r_k \cdot TTI$. Thus, the overall energy efficiency for the IoT data uploading from all $K$ LTE-A equipped devices in *cellular mode* can be computed as:

$$\eta = \sum_{k \in \mathcal{K}} \frac{d_k}{P_k \cdot r_k \cdot TTI} \tag{3.12}$$

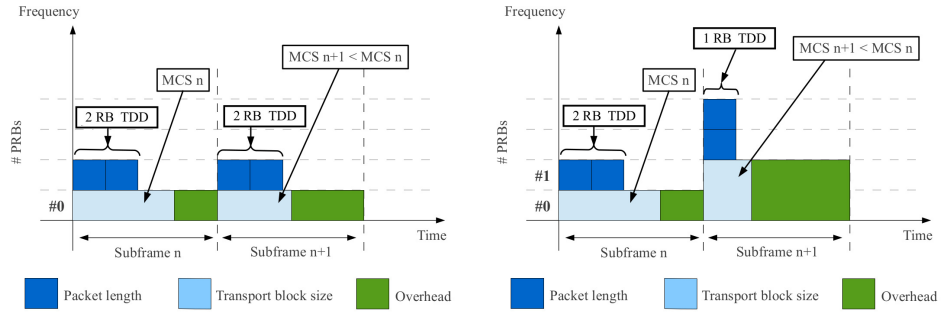
**Energy Efficient IoT Data Uploading**

Since most IoT applications are characterized by transmitting small amounts of data, low transmission efficiency is typically attained. Therefore, the objective of reducing the power consumption in the uplink can be reached by adopting a more robust MCS which requires a lower transmit power for the device. A more robust MCS guarantees a smaller TBS, which is however acceptable as long as it can contain the data to upload. On the other hand, it might also happen that adopting a very robust MCS over multiple RBs is also more energy efficient. For an energy efficient *cellular mode* IoT data uploading, an optimal MCS selection has been proposed in [95] where IoT data are transmitted to the eNodeB with the lowest energy consumption is possible.

To evaluate the power savings with the optimal MCS selection, the standard transmission power $P_{tx}$ formulation [73] for a generic UE in a subframe has been used:

$$P_{tx} = \min(P_{max}, P_0 + 10 \cdot log(r) + \alpha \cdot PL + \delta_{mcs} + f(\Delta_i)) \tag{3.13}$$

where $P_{max}$ is the maximum transmitted power of the UE, $r$ is the number of Physical RBs (PRBs) allocated per user, $P_0$ is the target power in one RB as specified by the eNodeB to reliably demodulate and decode the data, $\alpha$ is the path loss compensation factor specified by the eNodeB in a [0,1] range, $PL$ is the estimated UE Path Loss in uplink, $\delta_{mcs}$ is an MCS dependent offset which can be seen as the ratio between the target MCS and the basic MCS according to the UE feedback, and $f(\Delta_i)$ is the closed loop correction function. In particular, according to the $\delta_{mcs}$ when using a higher/lower MCS level, the corresponding transmit power should be increased/decreased.




(a) Fixed number of RBs. (b) Variable number of RBs.

**Fig. 3.9.** Energy efficient IoT data collection uploading solution.

As the authors in [95] discuss, when the number of RBs is fixed to $r$, the optimal MCS to be adopted for an energy efficient solution, i.e., $MCS_r^*$, is the one maximizing the TBS utilization rate:

$$MCS_r^* = \arg\max_{MCS} \left( \frac{D}{TBS(MCS, r)} \right), \text{ s.t. } D \leq TBS(MCS, r) \tag{3.14}$$

where $TBS(MCS, r)$ is the TBS determined by the MCS and the number of allocated resources $r$. With reference to Fig. 3.9(a)), when the proposed optimization is implemented in subframe $n+1$ a lower MCS is being adopted w.r.t. the MCS adopted in subframe $n$. This maximization guarantees that the data is actually sent with the minimum TBS is needed, and an energy saving can be obtained thanks to a power decrease related to a lower MCS level. However, having a fixed value for $r$ means that a UE needs to necessarily use all the allocated resources $r$. Removing this constraint, so that the number of RBs is not fixed, it is possible to determine the most energy efficient MCS as:

$$MCS^* = \arg\min_{MCS_r^*} (\delta_{mcs}(MCS_r^*) \cdot r) \tag{3.15}$$

where $\delta_{mcs}(MCS_r^*)$ is the power offset of each optimal MCS to the basic MCS. In such a situation, as represented in Fig. 3.9(b) (see final configuration in subframe $n+1$ after the optimization is implemented, compared to subframe $n$), the proposed minimization allows to find the MCS with the smallest transmit power as this is equal to: $P = Pbasic \cdot \delta_{mcs} \cdot r$, where $Pbasic$ is the power per RB for the basic MCS, $\delta_{mcs}$ is the power offset between MCS and the basic MCS.

Thus, a simple though effective solution to find the most energy efficient combination of MCS and number of RBs needed is:

- select for each RB number $n = [1, r]$ the MCS according to equation (3.14);

- from all the resulting MCS values, select the one minimizing the power transmission according to equation (3.15).

In particular, in the problem setting considered in this research, the value for $r$ is the maximum number of allocated resources to the single device according to the radio resource allocation implemented by the eNodeB, i.e. the Round Robin policy in this case. Specifically, when considering the LTE standard IoT data uploading solution illustrated in Section 3.3.1, the value of $r$ is identically equal to $r_k$ (i.e., the amount of RBs available for UE $k$).

### 3.3.2 D2D-Based Energy Efficient IoT Data Collection

The D2D-based solution we propose in this line of research, hereafter also referred to as D2D-EE, is based on the possibility for two or more UEs to cluster together and *cooperatively upload* their data in a unique transmission to the eNodeB. In order to enhance the overall energy efficiency, one of the UEs will act as *aggregator* for the cluster and transmit the whole data bundle.

Since D2D links cover short-distances, the channel quality is typically good even if lower transmission power is used. Consequently, short-rage communications implicitly introduce energy savings. Nonetheless, similarly to a standard uplink transmission to the eNodes, also on a D2D link further energy efficient techniques can be implemented. This means that in our scenario where IoT devices are clustered together, a more robust MCS can be used both on the cellular link from the aggregator to the eNodeB and on the D2D transmissions within the cluster. Based on this observation, the objective of the proposed energy efficient solution for the IoT data uploading is based on the following three aspects: (i) the adoption of low transmission power over short-range D2D links, (ii) an optimal energy efficient MCS selection on every D2D link within a cluster, and (iii) optimal energy efficient MCS selection in the uplink from the aggregator to the eNodeB. Specifically, for the objectives listed in (ii) and (iii), the approach presented in [95] has been extended in order to make it compliant to the specific scenario and data communication adopted in this work.

To this aim, it was assumed that the eNodeB implements the algorithm described in the rest of section. A RRM scheme is implemented to configure *(i)* the set of UEs acting as *aggregator*, *(ii)* the cluster configuration for the D2D data collection, *(iii)* the MCS and the RBs assigned to the aggregators, and *(iv)* the MCS and the RBs for supporting the D2D transmissions in each cluster. In particular, when the data collection in the IoT is triggered in a single TTI, a single execution of the listed steps is executed to collect the data. Whenever significant variations in the channel conditions are observed (e.g., due to UEs' mobility), the algorithm should be repeated to update the service configuration.

**Assumptions for the D2D-EE algorithm implementation**

For the proposed D2D cluster-based IoT data collection, has been assumed a network-assisted D2D communications where the eNodeB knows the current network state and is able to implement the proposed D2D-based IoT data collection. In particular, the eNodeB will be responsible for the allocation of the available radio resources to the cluster head(s) in the network, the so-called *aggregators* that will be in charge of uploading all IoT data from the cluster to the eNodeB. The main advantage of doing this is that the eNodeB has higher computational capabilities w.r.t. to the IoT devices. On the other hand, a completely distributed approach would require high signaling overhead for information sharing among the objects to build a shared knowledge of the network topology and the relevant channel quality information. For the intra-cluster D2D communications, instead, the radio resources are allocated according to a Round Robin policy, where the set of available resources is the set of resources allocated to the respective aggregator. For the D2D communications we foresee a decode-and-forward (DF) relaying configuration operating in half-duplex TDD mode. First, all the data from the cluster is received by the cluster aggregator; then, the aggregator will forward all aggregated data to the eNodeB. Thus, cellular mode and D2D transmissions will never occur in the same TTI (recall that uplink resources are used for the D2D transmissions) and consequently we can assume no interference is to be managed among cellular and D2D links within a cluster. Therefore, the uplink slots are alternatively used by D2D communication and transmissions toward the eNodeB

**Clustering for the D2D-based IoT data collection**

An important step for the implementation of the proposed D2D-based solution has been the clustering of the IoT devices into one or multiple clusters with one aggregator per cluster. Based on the cell-mode CQI values for the devices, the solution proposed for the cluster formation problem is an iterative algorithm based on the following simple steps being implemented by the eNodeB:

1. from the cell-mode CQI list sorted in descending order, select the UEs with highest cell-mode CQI levels as potential aggregators and compute for each of them the number of devices for which a D2D link is feasible;
2. out of the set of potential aggregators, the UE is selected for which the number of devices in coverage for a D2D link is the highest[5];
3. the selected device will act as an aggregator and will form a D2D cluster with the devices in D2D coverage;

[5] Given the small data to be sent, any CQI level greater than zero on the D2D link is assumed to be sufficient to send the data.

4. all devices belonging to the formed cluster are removed from the list; if still devices are present in the ordered list, then repeat the algorithm.

The iterative algorithm is repeated until all devices are part of a cluster (also clusters with only one device can be formed). Noteworthy, these steps make it highly likely that no *mutual inter-cluster interference* will be experienced at the aggregators. In fact, once an aggregator is selected all devices in coverage for a D2D link will be part of the same cluster and all the remaining nodes are excluded.

The output of the clustering algorithm defines the number and the size of the clusters the IoT devices are grouped in and the aggregator for each cluster. Noteworthy, the cluster size could have an influence to the communication efficiency within the cluster itself. Indeed, since a Round Robin scheduling is assumed for the radio resource allocation to the D2D transmitters in each cluster, smaller clusters mean a higher number of RBs available for each D2D transmission. As a consequence, on each D2D link the proposed energy efficient algorithm may introduce higher benefits. The motivation for this is that the algorithm is implemented over a larger number of RBs and has higher margins to optimize the energy and efficiency in the communication. On the contrary, if the cluster size is big, the opposite observations can be made. In particular, the extreme case is represented when only 1 RB per D2D link is available, where the benefits introduced by the algorithm are related exclusively to the use of a more robust MCS (i.e., without decreasing the number of RBs).

### The proposed D2D-EE solution step by step

The proposed D2D-EE solution foresees the implementation of the steps described below and reported in the message diagram in Fig. 3.10.

**Cell-mode CQI collection**: The eNodeB collects the cell-mode CQI feedbacks from all IoT devices willing to upload some data, i.e., $c_k, \forall k \in \mathcal{K}$.

**D2D-mode CQI collection**: The eNodeB collects also the $c_{i,j}$ values from all UEs $i, j \in \mathcal{K},\ i \neq j$; this information is used to discover the UEs in mutual coverage for a D2D link. In particular, the eNodeB computes a D2D CQI matrix (DCM) (an example is reported in Table 3.5) based on the $c_{i,j}$ values for all the links between the devices (we have always $c_{i,i} = 0$). A $c_{i,j} = 0$ value in the DCM indicates that a D2D link cannot be activated between devices $i$ and $j$.

**Aggregator selection and cluster formation**: The information from the DCM, coupled with uplink CQI levels for all devices will be used by the eNodeB to cluster all devices in a set $\mathcal{S} = \{s_1, \ldots, s_S\}$ of mutually disjoint clusters $s_i$ of cardinality $|s_i|$ be equal to $s_i = \left\{s_1^i, \ldots, s_{|s_i|}^i\right\}$, such that $s_i \cap s_{i'} = \emptyset$ for $i \neq i'$ and $\bigcup_{i=1}^{S} s_i = \mathcal{K}$. Let $\mathcal{A} = \{a_1, \ldots, a_S\}$ be the set of *aggregators* in the network. These devices are in charge

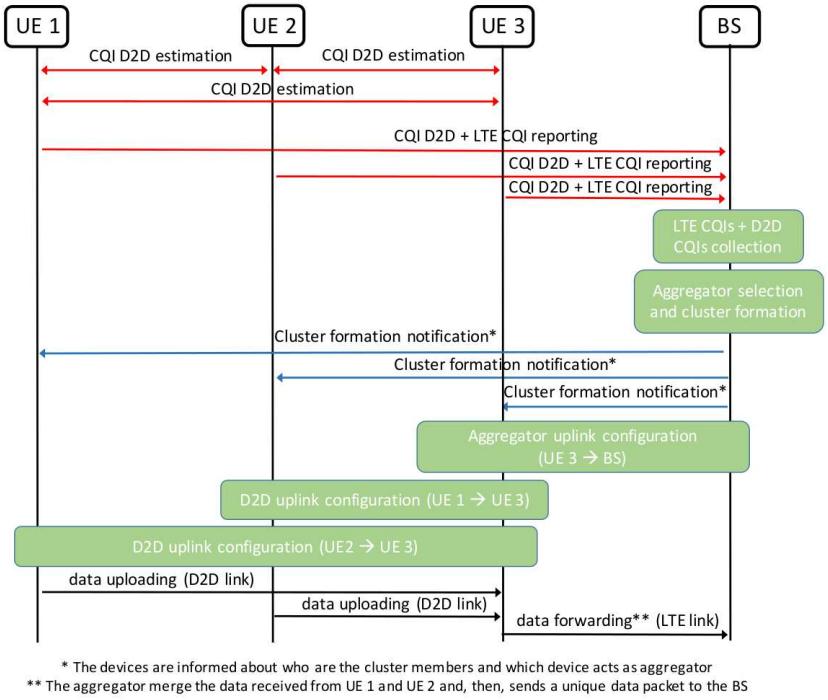



**Fig. 3.10.** Message diagram for the proposed D2D cluster-based IoT data uploading.

**Table 3.5.** LTE-D2D CQI Matrix

| | device 1 | device 2 | device 3 | ... | device $j$ |
|---|---|---|---|---|---|
| device 1 | 0 | $c_{1,2}$ | $c_{1,3}$ | ... | $c_{1,j}$ |
| device 2 | $c_{2,1}$ | 0 | $c_{2,3}$ | ... | $c_{2,j}$ |
| ... | ... | ... | ... | ... | ... |
| device $k$ | $c_{k,1}$ | $c_{k,2}$ | $c_{k,3}$ | ... | $c_{k,j}$ |

of collecting all data within the cluster and upload it to the eNodeB. Obviously, the aggregator will have to be active the time to receive all data in the cluster and upload it to the eNodeB. The association of the devices to a cluster and the aggregator selection follows the solution described in Section 3.3.2. In particular, the proposed solution is defined so that the *mutual inter-cluster interference* is assumed to be negligible.

**D2D link configuration**: For each cluster $s_i = \left\{ s_1^i, \ldots, s_{|s_i|}^i \right\} \in \mathcal{S}$, the eNodeB will define the resources and the MCS level to be used on the D2D links towards the aggregator. The D2D transmitter operates in half-duplex mode in which it cannot transmit and receive at the same time. In addition, all the devices performing a D2D connection have to remain active the time needed to upload their data to the aggregator. When all data from the cluster are received by the aggregator, this will forward the data to the eNodeB. Moreover, the clustering algorithm, discussed in Section 3.3.2, limits the *mutual inter-cluster interference* so that all radio resources can be reused in the clusters within the cell. Under this service configuration, the radio

resources allocated to the D2D links within a cluster can follow a Round Robin policy where the available resources are all the $R$ resources. These resources are equally shared among the devices in a single cluster so that each D2D communication link in the cluster can use no more than $r^d = \lceil R/|s_i| \rceil$ RBs. Based on the $r^d$ RBs available on a single D2D link, the following energy efficient configuration is implemented on each of the D2D links within the cluster:

- select for each RB number $n = [1, r^d]$ the optimal MCS, i.e., $MCS^*_{r^d}$ maximizing the TBS utilization, according to equation (3.16):

$$MCS^*_{r^d} = \arg\max_{MCS} \left( \frac{D}{TBS(MCS, r^d)} \right), \text{ s.t. } D \leq TBS(MCS, r^d) \quad (3.16)$$

  where $TBS(MCS, r^d)$ is the TBS determined by the MCS and the number of allocated resources $r^d$. This maximization guarantees that the data is sent with the minimum TBS is needed, and an energy saving can be obtained thanks to a power decrease related to a lower MCS level;
- from all the resulting MCS values, select the one minimizing the power transmission according to equation (3.17):

$$MCS^* = \arg\min_{MCS^*_{r^d}} (\delta_{mcs}(MCS^*_{r^d}) \cdot r^d) \quad (3.17)$$

  where $\delta_{mcs}(MCS^*_{r^d})$ is the power offset of each optimal MCS to the basic MCS.

This minimization allows to find the MCS with the smallest transmit power as this is equal to: $P^d = Pbasic^d \cdot \delta_{mcs} \cdot r^d$, where $Pbasic^d$ is the power per RB for the basic MCS on the D2D link, $\delta_{mcs}$ is the power offset between MCS and the basic MCS.

**Aggregators uplink configuration**: Once the data within a cluster have reached the aggregator, the uplink radio resources are used in cell-mode transmissions towards the eNodeB. Under the assumption of negligible *mutual inter-cluster interference* according to the clustering algorithm described in Section 3.3.2, each aggregator will be able to use the whole set of radio resource available: $r_a = R$, $\forall a \in \mathcal{A}$. In the uplink transmission, each aggregator will then implement the energy efficient data uploading presented above, to find the optimal MCS and number of RBs to adopt, where the maximum number of RBs the aggregator can use in the algorithm is exactly the $r_a$ value.

### 3.3.3 Performance Evaluation

A simulative analysis has been conducted by using Matlab® to assess the performance of the D2D-based scheme proposed for the IoT data collection in a Smart City

scenario and to show its superior performance compared to the standard operation of LTE-A. In particular, have been compared three alternative solutions: (i) *LTE-A* standard solution, where the devices upload their own data through unicast link toward the LTE eNodeB, (ii) *LTE-EE* solution, where the devices implement the energy efficient solution presented in [95] on standard LTE unicast links toward the eNodeB, and (iii) *D2D-EE* solution, which is the proposed solution on energy efficient D2D communications within clusters of devices and energy efficient unicast cellular transmissions from the cluster aggregator to the eNodeB. The key performance indicators considered in this analysis have been (i) the *Transport Block utilization*, and the (ii) the *energy efficiency*.

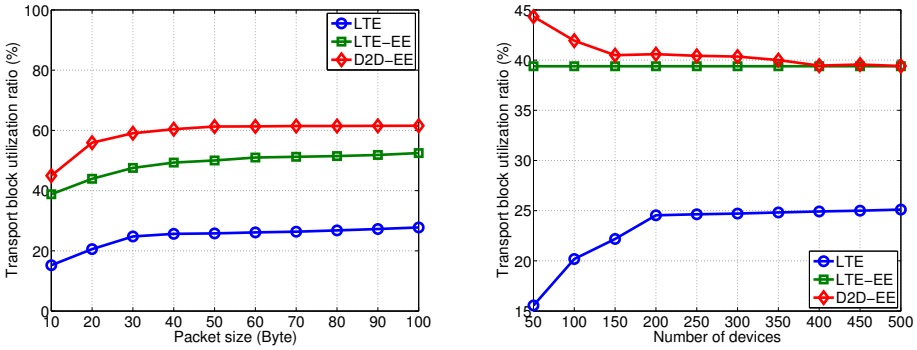
The simulated scenario consists of an LTE eNodeB with a coverage radius equal to 250m. In particular, the IoT devices are uniformly deployed within the LTE coverage and running the same application (i.e., personal e-health, environmental monitoring, intelligent transportation system, assisted leaving and so on). For the sake of simplicity, is assumed that all the devices have to forward the same sensing data to a remote server (i.e, Cloud) through the LTE eNodeB. In addition, different transmission power levels have been considered for the transmission modes used by the devices: (i) a transmitted power of 23 dBm is considered for standard LTE cell-mode uplink transmissions, whereas (ii) a power equal to 10 dBm if the devices use the D2D link. Furthermore, the D2D coverage has been fixed to 50m. The focus is on a single TTI; for data requiring multiple TTIs, the same solution is applied in consecutive TTIs. Channel conditions for the UEs have been evaluated in terms of SINR experienced over each sub-carrier when path loss and fading phenomena affect the signal reception [92]. The performance analysis has been conducted by following the guidelines for the system model defined in [73] and for a number of available RBs $R = 100$ per cell, a varying data size per device to be uploaded in the $[1 - 80]$ byte range (typical values for IoT data) and a varying number of devices in the cell in the range $[50 - 500]$. Finally, also the impact of the devices density in terms of $devices/km^2$ within the cell is evaluated. This last analysis has the objective to show that when higher possibilities for D2D communications between devices exist, then the proposed solution performs better. For an overview of the simulation parameters please refer to Table 3.6.

### Transport Block Utilization Analysis

Firstly, the Transport Block utilization has been investigated. As observed in Fig. 3.11, the proposed D2D-EE always outperforms the other solutions. In particular, as expected the Transport Block utilization increases with the data size for all solutions (the number of devices is set to 50 in this case) until reaching a convergence value (around the 40 bytes value), see 3.11 (a). In particular, with the D2D-EE, a maximum

**Table 3.6.** Main Simulation Parameters

| Parameter | Value |
|---|---|
| Cell radius | 250 m |
| Bandwidth | 20 MHz |
| Frame Structure | Type 2 (TDD) |
| TTI | 1 ms |
| TDD configuration | 0 |
| eNodeB Tx power | 46 dBm |
| $P_{max}$ cell-mode Tx power | 23 dBm |
| $P_{max}$ D2D-mode Tx power | 10 dBm |
| Noise power | -174 dBm/Hz |
| D2D transmission range | 100 m |
| Path loss (cell link) | 128.1 + 37.6 log(d), d[km] |
| Path loss (D2D link, NLOS) | 40 log(d) + 30 log(f) + 49, d[km], f[Hz] |
| Path loss (D2D link, LOS) | 16.9 log(d) + 20 log (f/5) + 46.8, d[m], f[GHz] |
| Shadowing standard deviation | 10 dB (cell mode); 12 dB (D2D mode) |
| Sub-carrier spacing | 15 kHz |
| BLER target | 10% |
| # of Runs | 1500 |




(a) Varying the packet size (number of devices is 50).

(b) Varying the number of devices (data size is 10 byte).

**Fig. 3.11.** Transport Block utilization.

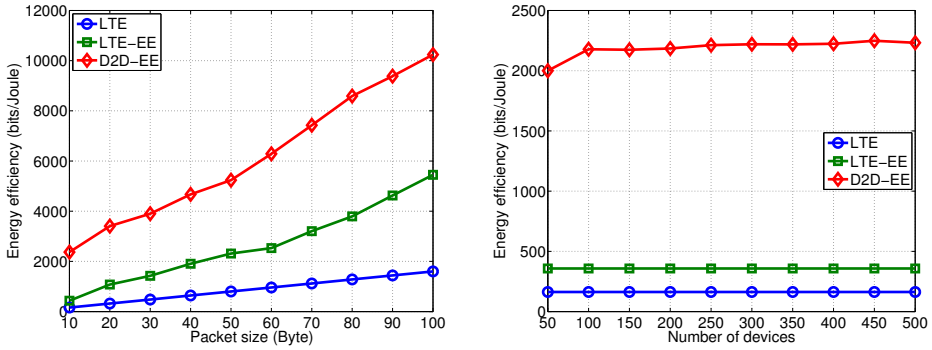
utilization of 62% is reached with 100 bytes to send per device, whereas the LTE-A and LTE-EE reach a maximum of 27% and 52% utilization for the same amount of bytes per device, respectively.

When, instead, the packet size is keeping constant (we set it to 10 byte) and the number of devices in the network vary, the Transport Block utilization for the D2D-EE solution decreases from a maximum of 45% utilization, to a 39% value when considering more than 400 devices. Moreover, the D2D-EE approach converges to

that one of the LTE-EE when the number of devices per TTI is greater than 400. The LTE-EE, instead, shows a constant utilization percentage, 39%, independently of the number of devices, whereas the LTE-A reaches a maximum Transport Block utilization of 25% with 200 devices, starting from about 15% utilization with 50 devices. These very low utilization values are due to the very small data size which causes the cell-mode transmissions to under-utilize the RB used for transmissions. Moreover, it is important to underline that both the LTE and the LTE-EE solutions are actually never serving all the devices in the single TTI. In fact, with 100 RBs per TTI, no more than 100 devices can be served.

### Energy Efficiency Analysis

The second and most interesting result can be found observing the energy efficiency, expressed in bits/Joule, shown in Fig. 3.12(a). The energy efficiency increases with the packet size for all the three solutions (the number of devices is set to 50 in this case) with more emphasis for the D2D-EE and the LTE-EE solutions. In all cases the energy efficiency for the D2D-EE solution is much higher, with the highest data size (i.e., 100 bytes) it is over 5 times more efficient than the LTE-A standard solution and about 2 times more efficient than the LTE-EE solution.




(a) Varying the packet size (number of devices is 50).

(b) Varying the number of devices (data size is 10 bytes).

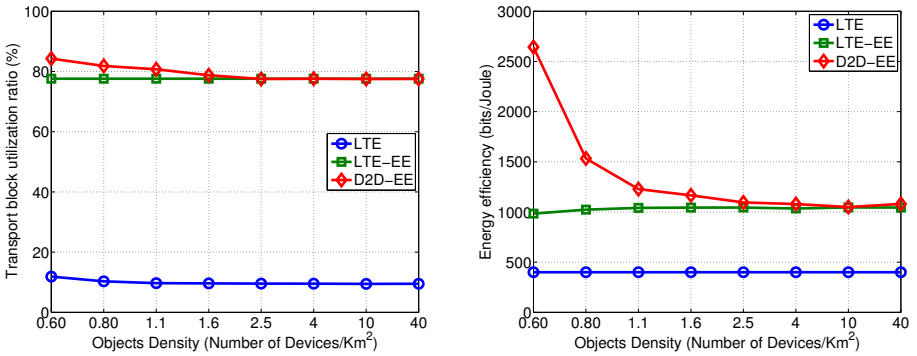
**Fig. 3.12.** Energy efficiency.

When considering, instead, a varying number of devices with a data size set to 10 bytes, is observed that larger number of devices make the energy efficiency increase only for the D2D-EE solution and has no impact on the LTE-A and LTE-EE solutions, see Fig. 3.12(b). At its maximum value, with 500 devices, the D2D-EE is about 6 times more efficient than the LTE-EE and about 11 times more efficient than the LTE-A solution. This very important results derive from the three contributions in the D2D-EE solution: low transmission power on D2D links, optimal energy efficient

MCS selection on every D2D link within a cluster, and optimal energy efficient MCS selection in the uplink from the aggregator to the eNodeB.

### Impact of Devices Density

Interesting is also to understand how the distribution of the devices within the cell has an impact on the performance improvements obtained by the D2D-EE solution. In particular, the density of the devices influences the D2D communication possibilities and has been investigated up to which value of density the D2D-EE solution is still the most convenient solution. In particular, in Fig. 3.13 the Transport Block utilization and the energy efficiency are shown for a varying value of the node density in the cell, the results for a $[0.6 - 40]$ $devices/km^2$ range are reported as this is the range where the convergence of the D2D-EE solution to the LTE-EE is visible. As is observed from the plots, the Transport Block utilization for the D2D-EE and the LTE-EE always outperform the LTE-A solution. Moreover, the D2D-EE solution shows better performances w.r.t. the LTE-EE for values below 2.5 $devices/km^2$. For higher density values, that is when the devices in the cell are very densely distributed, the D2D-EE converges to the LTE-EE, as it can be seen in Fig. 3.13(a).



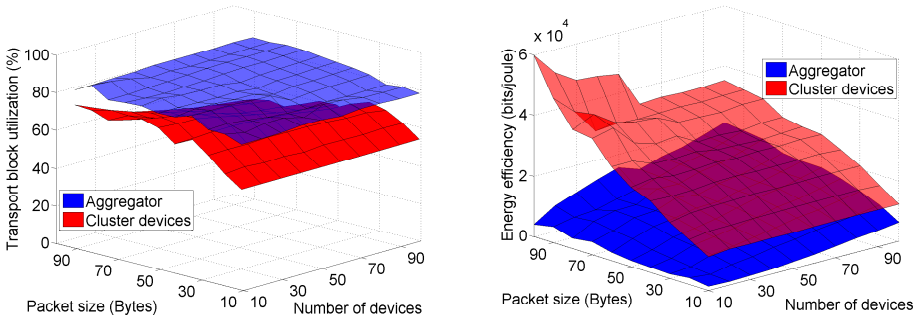

(a) Transport Block utilization. (b) Energy efficiency.

**Fig. 3.13.** Performance for varying device density in the cell.

Also the energy efficiency for the D2D-EE converges to the LTE-EE solution for even more dense distribution of devices in the cell, i.e., 10 $devices/km^2$, see Fig. 3.13(b). In all other cases the D2D-EE solution outperforms the LTE-EE solution from the energy efficient point of view (and the LTE-A solution as well). These results, witness to the fact that the possibility to set up D2D links among the devices is the key feature for the implementation of the D2D-EE. Nevertheless, in modern IoT scenarios the density of devices is typically high which suggests that the proposed D2D-EE solution can successfully be implemented.

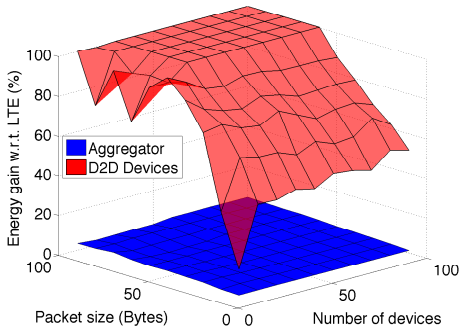
### Analysis on the Clusterization Effects

In the last part of the performance evaluation, it was focused on understanding what the differences are for the performance between the aggregator and the other devices in the cluster. In particular, an aggregator has much more data to send, as it collects all data from the cluster, and at the same time it adopts a higher power in transmission over a greater number of RB pairs. In this case, the attention was focused on a $[10-100]$ range of devices whereas the data size varies in the $[10-100]$ bytes interval.




(a) Transport Block utilization.

(b) Energy efficiency.



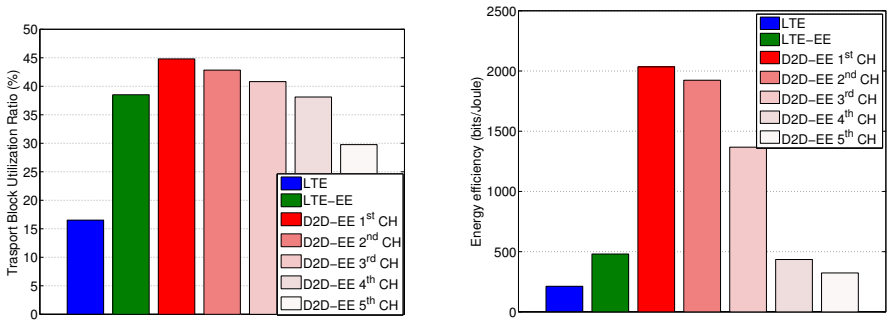

(c) Energy consumption.

**Fig. 3.14.** Performance for varying number of devices and data size: aggregator vs. cluster devices.

In particular, the energy efficiency for the D2D objects decreases with the number of devices and for smaller data sizes. For the aggregator instead, is possible to notice the opposite trend for the energy efficiency. In any case, the energy efficiency for the D2D objects is always higher than the one of the aggregator (see Fig. 3.14(b)). This is mainly due to the lower transmission power adopted on the D2D links. Finally, in Fig. 3.14(c) it was interesting to observe the energy savings for the D2D objects and the aggregator in the cluster w.r.t. the standard LTE-A data uploading. As is observed from the plots, the D2D objects in the cluster obtain always very high savings, reaching up to 99% energy savings, whereas for the aggregator this ranges

between a maximum of 6% with a small number of devices in the network and a small data size, and a minimum close to zero when many devices are involved with large data size to send per device. This shows that *also the aggregator has an energy savings* in some cases, which is mainly due to the implementation of the energy efficient solution in cell-mode uplink transmission, and anyway there is never an energy increase for them. Nevertheless, it has been observed that in the worst case the energy saving for the aggregator compared to the standard LTE solution is close to zero.

To cope with this differentiation in the energy savings among the device, one could design enhanced clustering algorithms that consider an update of the configuration over time in order to share the "burden" of playing the aggregator role. Such a role-shifting approach would also mean different clusters being formed over time which may affect the efficiency of the proposed solution.




(a) Transport Block utilization. (b) Energy efficiency.

**Fig. 3.15.** Performance with aggregator role shifting among the devices (50 devices and 10 Bytes data).

For completeness in the analysis, has been also considered this effect in a specific study case with 50 UEs and 10 bytes of data for each device, as this is representative of the worst case when the lowest gains are obtained (clearly better performances are obtained in the other cases). The analysis is based on a policy whereby the aggregator is either a node with the best CQI level, the second best CQI level, the third best CQI level and so on, and evaluate the resulting energy efficiency and transport block utilization. What has been observed from the resulting plots reported in Fig.3.15, is that indeed a performance reduction is observed, but still better performance figures are obtained when choosing as cluster head the nodes until the third/fourth best CQI level. On the other hand, there are multiple devices with the same CQI level. This means that taking turns in acting as the aggregator over all the devices with the 3-4 highest CQI level towards the eNodeB, may be a good solution to share the burden and thus avoiding the cluster head to run out of its battery.

# 4

# Mobility

Next-generation D2D communications technology is rapidly taking shape today, where a cellular network assists proximal users at all stages of their interaction. To this end, the respective D2D performance aspects have been thoroughly characterized by past research, from discovery to connection establishment, security, and service continuity. However, prospective D2D-enabled applications and services in future 5G systems envision highly opportunistic device contacts as a consequence of unpredictable human user mobility. Therefore, the impact of mobility on D2D communication in typical 5G environments needs a careful investigation to understand the practical operational efficiency of future cellular-assisted D2D systems.

## 4.1 Characterization of User Mobility in D2D Systems

The emerging fifth generation (5G) communication systems are expected to provide 1,000-fold gain in network capacity, massive connectivity for 100 billion devices, and 10 Gbps individual user throughput augmented with extremely low latency and response times. To achieve these ambitious targets, prospective 5G deployments incorporate an increasing number of small cells across different radio access technologies, supplied with diverse power and coverage capabilities, thus enabling high-rate and short-range data transmission.

In this context, 5G Proximity Services (ProSe) (i.e., see 3GPP specification TR 22.803) open rich collaboration opportunities for people and their mobile devices in physical proximity. Importantly, ProSe are built on the underlying D2D communication technology, which allows for direct data transfer between two proximal devices without the need for expensive cellular network resources. Accordingly, D2D technology not only provides high-rate, low-latency traffic offloading with the "personalized" device-centric small cells, but also enables a plethora of proximity-based social networking use cases. However, D2D-based interaction is stochastic by nature and results

in highly opportunistic contacts due to potential mobility of all involved user devices hence it may induce "negative" or "positive" effect on system-level performance.

Therefore, on the way to integrating the native support for D2D communication into the 5G system architecture the effects of user mobility have to be thoroughly characterized as they may have a profound impact on the resulting system performance. In fact, mobility affects the chances that the users meet and establish a D2D connection. Moreover, mobility-related parameters determine: (i) the individual D2D link performance (length, duration, throughput, etc.) and (ii) the overall D2D system performance (offloading gain, device/content availability, etc.). Ultimately, the resulting performance depends on the user movement patterns and other factors, including the type of application running on top of the D2D links.

However, existing literature falls short of quantifying the impact of mobility on proximal communication. In particular, system-wide performance evaluation results have not yet been reported to understand how the D2D operation reacts to frequent and opportunistic contacts due to realistic user mobility. Bridging this glaring gap, the aim of this line of research has been to offer a first-hand tutorial on the effects of mobility on system-level performance an assessment methodology of D2D-enabled cellular networks.

### 4.1.1 Mobility-aware D2D performance assessment

Generally, there are two approaches to comprehensively assess the effects of mobility in D2D-enabled cellular networks. The first one is to conduct implicit mobility modelling, that is, to capture the system-wide metrics by accounting for cluster/link contact times as well as for inter-contact times between consecutive connections. The second approach is based on explicit mobility modelling by incorporating the movement of users directly into the performance evaluation and optimization framework.

#### Metrics of interest

There exist multiple mobility-related parameters characterizing the impact of mobility on the performance of D2D communication. As an example, Fig. 4.1 illustrates the scenario, where a user enters the D2D coverage area of another user at time $t_1$ and leaves it at time $t_2$. The probability that the tagged user is within the D2D range of its neighbour at time $t_1$ is often called the contact probability. Accordingly, the number of contacts per a unit of time determines the availability of D2D connectivity. Further, the time interval $(t_2 - t_1)$ is named the contact time or the D2D link residence time, which is the actual duration of a D2D connection. The above three metrics are labelled as "implicit", given that they are in essence the effects of user mobility.

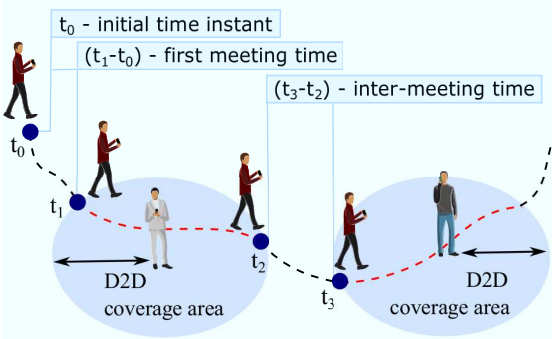



**Fig. 4.1.** A clarification of mobility-related parameters.

The parameters introduced above may still be insufficient to fully characterize the real-time user mobility. For instance, both inter-meeting time $(t_3 - t_2)$, where $t_3$ is the time of a subsequent D2D connection, as well as the first meeting time $(t_1 - t_0)$, where $t_0$ is some initial time instant, could also affect the resulting performance. For the unpredictable user movements, all the discussed metrics are in fact random variables that are completely defined by their distributions.

Finally, it is comprehensive to argue that various mobility models may demonstrate qualitatively different behaviour with respect to the introduced metrics. Moreover, the closed-form expressions are available for the above metrics only in simple cases. Hence, even though mobility effects can be somewhat captured with the discussed implicit parameters, to cope with the limitations listed above it has more value to investigate the impact of mobility patterns on the D2D system performance with "explicit" mobility modelling.

### Suitable mobility models

Five characteristic mobility models have been considered that have the potential to describe the effects of user mobility in crowded pedestrian D2D scenarios. These models differ in their complexity and qualitative response of their related implicit metrics, even when the same average user speed is employed to parameterize them.

•**Random Waypoint (RWP)** is a well-known mobility model, whereby a user moves inside a rectangle of known dimensions. The subsequent user location is chosen uniformly within the same rectangle, whereas its average speed is distributed uniformly between 0 and a given maximum. However, for the model in question it is extremely cumbersome to analyse the implicit mobility metrics of interest. So far, only the steady-state distribution of user location has been obtained.

•**Random Direction Model (RDM)** is an extension of the renowned Pearson-Rayleigh random walk model that may capture the temporal behaviour of users in an area of interest, as well as allow for the closed-form expressions across various

implicit metrics [96]. Although this model is known for more than a century due to the famous correspondence between Karl Pearson and Lord Rayleigh in Nature, it has attracted full attention of the research community only recently. According to RDM, a user begins its movement at some position and selects a direction randomly and uniformly between 0 and $2\phi$. It then moves in the chosen direction for an exponentially-distributed amount of time with the mean $E[\tau]$ at a constant speed of $v$. Upon completion, the user chooses another random direction and continues. Contrarily to the RWP, this model can be defined over the entire two-dimensional plane.

•**Brownian Motion (BM)** is obtained when both the step size and the mobility duration tend to zero, such that their ratio remains constant. More specifically, BM is defined by a stochastic differential equation $dX(t) = \mu X(t)dt + \sigma X(t)dW(t)$, where $W(t)$ is a Wiener process, while $\mu$ and $\sigma$ are the drift and the volatility of the process. BM is the simplest form of random motion over a plane allowing for a number of implicit metrics to be expressed in the closed form. Note that when the number of steps in any unbiased random walk (e.g., RDM) increases, the central limit theorem applies. Then, the spatial location of a user may be closely approximated by the corresponding parameter of the BM process. While the random walk and the Brownian motion are generally not the same, the resulting metrics affecting the performance of a D2D-enabled system can be obtained by using the BM approximation [97], including the contact time, first- and inter-meeting times, etc. Therefore, the role of the BM model in modern network analysis is underestimated and it could be considered as an attractive candidate for characterizing the effects of mobility in future D2D networks.

•**Lèvy Flight (LF)** is different from the aforementioned models, which all focus on capturing the short-term mobility (e.g., up to tens of minutes) of users moving on a plane. The LF process is defined as $dX(t) = \mu X(t)dt + \sigma X(t)dW(t) + kdN(t)$, where $dN(t)$ is a Poisson counter, such that $PrdN(t) = 1 = \lambda dt$ and $k$ follows a stable distribution with the parameter $\alpha$. Recent investigations reveal that movement of people over larger time spans may follow distinct patterns, where multiple short "runs" interchange with occasional long-distance travels [98]. LF is known to be one of the models suitable for representing such patterns. The said process is similar to the RDM except for the distribution of the run length that is heavy-tailed. While the LF model is significantly more complex to analyse as compared to RDM or BM, several important LF characteristics have already been obtained. These include the contact probability, which is shown to be maximal across the entire class of similar random walks.

•**Jump Brownian Motion (JBM)** is a process exhibiting some properties of BM, but with occasional "jumps" of non-infinitesimal size. The key difference of

JBM as compared to LF is in that $k$ follows a distribution with the finite mean and variance, e.g., a Normal distribution. This model inherits several important Markovian properties of the class of unbiased random walks and Brownian motion processes, and has been deeply studied in finance. Conveniently, the JBM patterns may serve as a first-order approximation of the LF process.

Augmenting the considered models, real-world human mobility patterns could be employed to parameterize them and thus predict the resulting system performance in a desired special-case scenario.

### 4.1.2 Mobility Implications on D2D System Design

It is worth noticing that user mobility and D2D system performance have a tight correlation based on the channel quality experienced by the involved devices as well as the movement pattern itself. Hence, each mobility model discussed above may uniquely affect the D2D performance.

#### Implementing mobile D2D environment

Whenever proximal communication takes advantage of cellular network assistance, the neighbouring users can establish a direct connection and exchange information over a D2D link with the help coming from the operators network (e.g., for user and service discovery, secure connection initiation, etc.). Indeed, as we can see in Fig. 4.2, the discovery of users in proximity and the D2D connection establishment functions are completely delegated to the cellular network infrastructure.

As a characteristic practical example, the focus was on a D2D system implementation, where each user is informed by the 3GPP LTE (cellular) base station about other relevant users in proximity. Then, the actual communication process is operated over WiFi-Direct (short-range) links with one or more neighbours that already have the desired content in their memory (cache). Naturally, multiple D2D connections may be established for the same item of content due to the user mobility.

#### Review of potential D2D applications

The ongoing proliferation of mobile devices with advanced computation capabilities and enhanced storage capacities may now support enhanced services for high-quality content sharing and next-generation social networking based on user proximity. Large-scale multimedia exchange, such as video-on-demand, video/image file downloading, and video streaming services are potential applications to be considered. Todays multicast schemes are not able to efficiently manage such asynchronous multimedia transfers, where multiple users may stream the same file at different time instants and

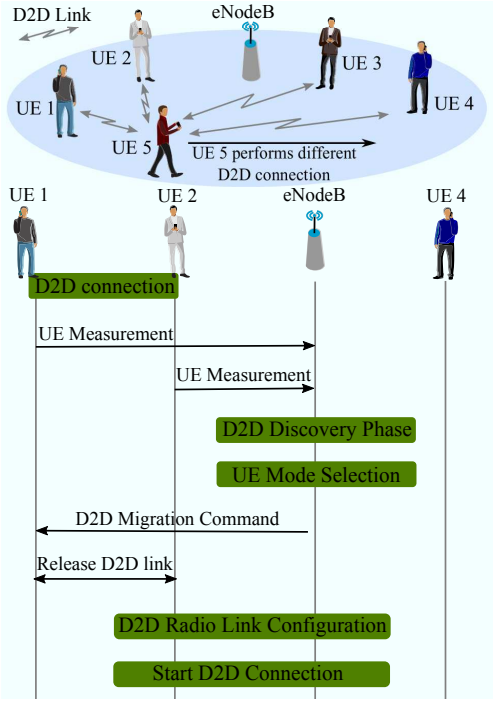



**Fig. 4.2.** Signalling diagram for D2D session continuity.

across a wide geographical area. Therefore, sharing/distributing content over D2D links with mobile peers represents an increasingly attractive solution. In what follows, the prominent D2D applications are divided into two groups, namely, delay-tolerant and delay-sensitive.

**Table 4.1.** Mobility metrics and application-related settings.

| Applications | Metrics of interest | | | Settings |
|---|---|---|---|---|
| | $T_I$ (s) | $T_L$ (s) | $L$ (KB) | ($T_I$, $T_L$, $L$) |
| Warning messaging | - | - | - | (0.1, 0.1, 100) |
| Video streaming | - | + | + | (0.01, 0.01, 10) |
| Video file download | + | + | + | (180, inf, 10e4) |

The motivation behind considering these particular classes stems from their different behaviour with respect to the implicit mobility parameters introduced in Section 4.1.1. Clearly, delay-sensitive applications need to be served immediately, whereas delay-tolerant ones can allow for a reasonable extra latency.

In particular, the following specific D2D applications have been considered: (i) dissemination of warning messages, (ii) video streaming, and (iii) video file downloading. In more detail, the time between two consecutive content arrivals for warning messaging and video streaming is relatively small. For these applications, the contact probability and the D2D link lifetime are thus equally important. Each considered

D2D application is defined by a triplet ($T_I$, $T_L$, $L$), where $T_I$ is the content inter-arrival time (in seconds), $T_L$ is the content lifetime (in seconds), and $L$ is the size of the content arriving at each $T_I$ (in kilobytes, KB). The typical values assumed for the three different considered applications are reported in Table 4.1. As an example, for the dissemination of warning messages the characteristic triplet is (0.1, 0.1, 100), which implies that every 100 ms a single packet of size 100 KB arrives into the D2D system and has to be delivered within 100 ms.

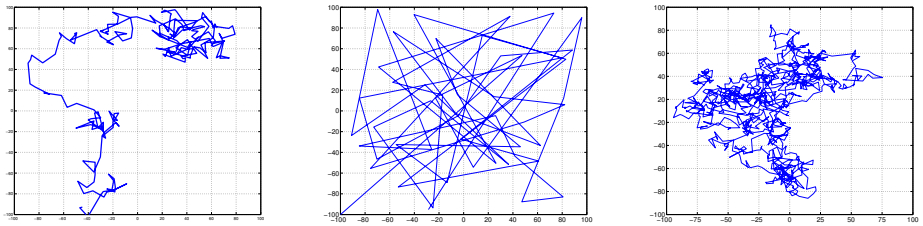
### 4.1.3 Performance Evaluation of Mobile D2D

#### Considered D2D scenario and simulator capabilities

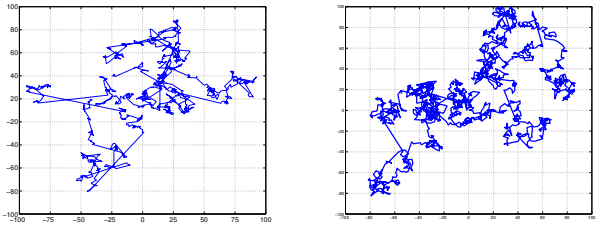
The reported system-level simulation data have been collected by employing our own WINTER-sim evaluation framework. It is a flexible tool designed to support the contemporary D2D deployment strategies, which implements the complete LTE-assisted WiFi-Direct infrastructure together with its most important features, such as cellular and D2D-based "cells", cell boundary effects, uniform and clustered user distributions, etc.

The reference scenario describes a dense environment, such as the one in a shopping mall, stadium, concert hall, or fair. In particular, a total number of 30 cellular/D2D users are uniformly distributed under the coverage of an LTE eNodeB (base station) with a cell radius of 100m. The D2D range is equal to around 30m. The eNodeB manages its spectrum by assigning the necessary amounts of resources to each scheduled user as well as advises on the modulation and coding schemes. Moreover, the eNodeB is also in charge of the D2D session setup (e.g., bearer setup), while power control and resource allocation procedures on the D2D links can be executed in a distributed (i.e., on each user) or a centralized (i.e., on the eNodeB) fashion. Specifically for this analysis, it was assumed a centralized control by the eNodeB (i.e., network-assisted D2D scenario). In addition, user movements are captured by the mobility models discussed above and illustrated in Fig. 4.3.

A new content acquisition session is assigned to each user unless it has obtained its desired content. The interference between the ongoing D2D sessions is modelled after the logic commonly utilized by the WiFi-Direct protocol: built on top of the CSMA/CA function implementing the binary exponential back-off algorithm. As D2D links require some time to setup, the corresponding initial latency is invoked every time a user attempts to request its target content from a D2D neighbour.When acquiring content over the D2D links, the users may alternatively claim the missing fragments from the cellular infrastructure. Further, the content dissemination process is considered asynchronous implying that users do not request the same type of



(a) RWP (b) RDM (c) BM



(d) LF (e) JBM

**Fig. 4.3.** Sample user movement trajectories of the considered mobility models.

information simultaneously. Once some content is requested by a user, it identifies its relevant D2D partners and, if any of those are in proximity, establishes a direct connection over WiFi-Direct. Otherwise, the content is downloaded from the LTE network. The main simulation parameters are summarized by Table 4.2.

**Table 4.2.** Main system-level simulation parameters.

| Parameter | Value |
|---|---|
| Number of content fragments | 1 |
| Number of users | 30 |
| Target data rate on D2D link | 40 Mbps |
| Target data rate on LTE link | 10 Mbps |
| User transmit power | 23 dBm |
| Cellular bandwidth | 5 MHz |
| Cell radius | 100 m |
| Maximum D2D transmission range | 30 m |
| D2D link setup time | 1 s |
| Total simulation time | 15 minutes |

To perform a fair comparison of various mobility models, it has been introduced the notion of mobility intensity by interpreting it as the average speed of a user. For the RWP, this parameter readily follows from the model properties. For the LF process, the step size distribution is drawn from the $\alpha$-stable distribution with $\alpha = 1.5$ ensuring the finite mean of step size distribution. Further, it is known that for the BM

models the average speed is infinite, which may not be realistic for practical mobility patterns. In this case, these processes have been approximated by their discrete-time and discrete-space equivalents on a lattice with minor displacements after small intervals of time. For the JBM process, the jumps occur in arbitrary directions selected uniformly between 0 and $2\phi$, whereas the final destination is rounded up to be the nearest vertex of the lattice.

### Obtained numerical results and discussion

To comprehensively describe how user mobility affects the system-wide performance of typical D2D applications, the focus was on: (i) the average download time and (ii) the total data delivered over the D2D links. In addition, to offer a complete comparison of the effect of different mobility models on the D2D user performance, (iii) the number of contacts and (iv) the contact time have been characterized. The obtained results for the average number of contacts with other D2D users having the requested content as well as the respective average contact time for the considered mobility patterns are reported in Fig. 4.4.

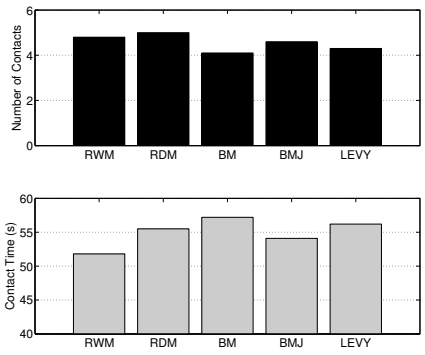
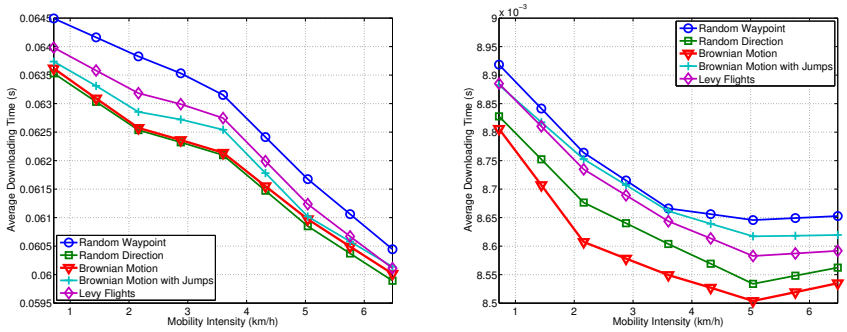



**Fig. 4.4.** Average number of contacts and average contact time.

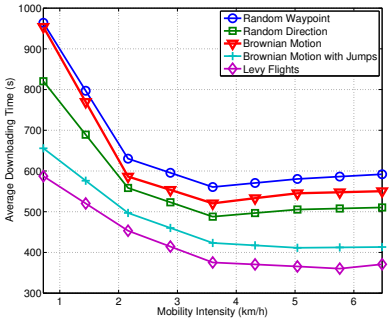
Following the guidelines given in 3GPP specification TS 36.304 (see Section 5.2.4.3 therein), the concentration was on typical pedestrian speeds of 3 km/h. As showed from Fig. 4.1, the RWP model demonstrates the lowest contact time, whereas the corresponding number of contacts remains one of the highest. Interestingly, the BM model is at the other extreme offering the minimal number of contacts while providing with the longest contact time out of all the studied models. Such behaviour is related to smaller covered area for the BM case with the same average user speed. Finally, the LF model strikes a good balance between the number of contacts and the contact time. Indeed, the longer the contacts among the devices last and the higher is the probability to download the content through a D2D link.

Given that the target data rate over the D2D link is 40 Mbps, a user adhering to, e.g., the BM mobility model is able to download about 250 MB of data per a single D2D contact. In addition, having more frequent contacts may be beneficial for delay-sensitive applications, where smaller amounts of data are transferred repeatedly. In contrast, delay-tolerant applications (e.g., delayed file download) are stronger affected by their corresponding D2D contact times. In this case, the parameter that plays an important role is the movement speed of a user.




(a) Warning messages.

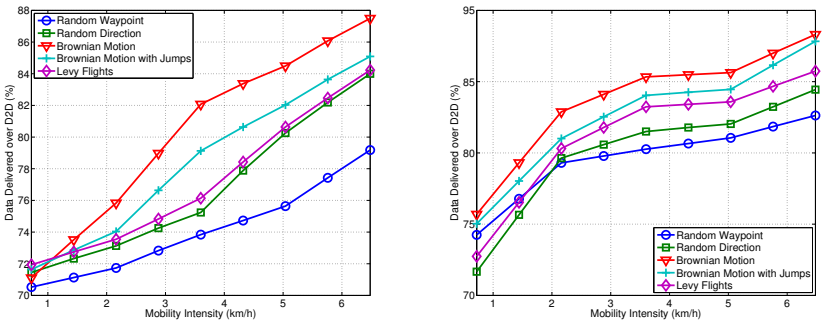
(b) Video streaming.




(c) Video file download.

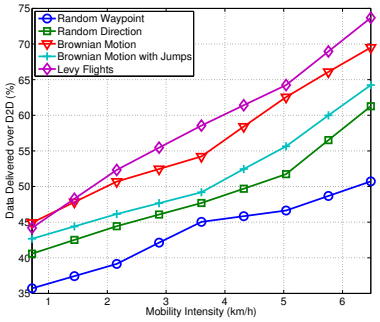
**Fig. 4.5.** Average content download times for different user mobility models and D2D applications.

A further investigation has been how the above mobility-centric metrics translate into the system-level performance indicators, including the average download time and the total data delivered over the D2D links. The former parameter for, e.g., the warning messaging over D2D is demonstrated in Fig. 4.5(a). In this case, the actual type of user mobility does not affect the resulting system operation significantly. Fig. 4.6(a) highlights the volume of D2D data transmitted by the same D2D application. It is possible to witness a linear increase in delivered data with the growing intensity of mobility. Here, the BM model performs better as compared to its counterparts, whereas the RWP model shows the poorest performance. The underlying reason is in that for real-time applications with strict latency requirements the contact time




(a) Warning messages.

(b) Video streaming.




(c) Video file download.

**Fig. 4.6.** Average data delivered over the direct link for different user mobility models and D2D applications.

is the dominating mobility-related factor. As discussed previously on this (see Fig. 4.4), the BM model delivers the longest contact times. Somewhat similar results are observed for video streaming over D2D. In fact, what is possible to learn from Fig. 4.5(b) and Fig. 4.6(b), is that the results follow similar trends as those for the warning messaging.

Further, for streaming services the contact times are of much higher importance, which are in turn affected by the average user speeds. To this end, the relative benefit of the BM model is due to the fact that it guarantees the maximal contact times for the same intensity of mobility. In contrast, since the contact time for the RWP model is minimal out of all the considered patterns, this mobility model results in the worst performance. Drastically different performance is observed in case of the delayed video file downloading. The corresponding trends in the content download times are shown in Fig. 4.5(c), where the mobility intensity of under 3.5 km/h still allows the users to reduce their D2D content transfer delays.

However, a speed beyond 3.5 km/h begins to yield a slight increase in the download time as a consequence of shorter contact time.In Fig. 4.6(c), the respective results in terms of the total data delivered over the D2D links are shown. In particular, focusing on the semi-stationary scenarios (with user speeds of around 0.7 km/h), the users are

not able to complete the full download of their requested content solely on the direct connections (the data delivered is under 100 MB).

As a conclusion, as the intensity of mobility grows, the entire video file may be acquired over a D2D link (with the exception of the RWP model). Recalling the results in Fig. 4.4, it becomes clear that the users moving according to the RWP cannot download their desired video files due to the very short D2D contact times.

**Impact of mobility in D2D small-scale open space scenario: lessons learnt**

Summarizing the main results obtained by the simulation analysis discussed in the previous section, the following findings for the considered small-scale open space environments are listed:

- The performance for various mobility models is tightly coupled with the number of contacts and the contact time for the D2D links. In particular, longer contact times lead to higher robustness of the direct connection as well as better chances to download the entire fragment in a single D2D contact.
- Considering delay-sensitive D2D applications, both warning messaging and video streaming appear to have very limited sensitivity with respect to user mobility models. For such applications, the most influential metric is the contact time.
- For video file downloading applications, the obtained performance are drastically different. In particular, mobility strongly affects the possibility for the users to download the entire content through a D2D link. The greater the mobility intensity and the higher is the possibility to fully download the content solely on the proximity-based communication.
- Finally, as real-time services are characterized by smaller content inter-arrival times and shorter content lifetimes, these services are strongly dependent on the number of contacts among the relevant users. On the contrary, the possibility to use D2D links in delay-tolerant applications is determined by the intricate interplay between the D2D contact time and the number of user contacts.

## 4.2 D2D Handover in 3GPP LTE Systems

Once that the impact of mobility over D2D communications have began more clear, a further analysis on mobility regarded to augment future handover operations by employing proximity-based communications. The underlying rationale behind this research is to equip the mobile users with better-quality direct links and thus improve the resulting service perception under the typical 3GPP LTE handover procedures. Primarily in frequently-visited areas of overlapping cellular coverage, D2D connectivity can readily offer the much needed data relaying capability to proximal devices

performing the cell change. To this end, the proposed *D2D-assisted* handover scheme efficiently delivers the attractive energy efficiency, data rate, and packet delivery ratio benefits. By utilizing the tools from stochastic geometry, the main performance metrics of interest for the purpose have been derived, such as the distribution of signal-to-noise ratio experienced by a user entering the zone of overlapping coverage and the amount of time it remains in contact with a suitable D2D partner.

### 4.2.1 Reference System Model

The reference scenario considered during this research is characteristic of a mobility-centric environment, where multiple UEs participate in downloading of rich multimedia content over the 3GPP LTE cellular network. With the proposed *D2D-assisted* handover, the UEs in need of changing a cell may establish the D2D links with other devices in close proximity (on the order of tens of meters), so that they can take advantage of a better channel quality over D2D to download their desired content from the partner D2D device. For the sake of analytical tractability, it is assumed that the D2D partner in question, acting as a relay, has already downloaded the needed item of content and hence no extra delay is introduced by doing so.

Further, the set of the cellular users is labeled with $C$, the set of the D2D users with $D$, and the set of the available BSs with $B$. The power received by a user $c \in C$ from the BS $b \in B$ can be expressed as $P_{R_c} = P_{T_b} \cdot |h_{b,c}|^2$, where $P_{T_b}$ is the power transmitted by the BS $b$ and $|h_{b,c}|^2$ is the gain on the channel from the BS $b$ to the user $c$. In turn, the channel gain coefficient includes all the corresponding losses due to the path loss (attenuation), shadowing, and other detrimental wireless factors such as fading, multipath, etc. Similarly, for direct connections the received power at a particular D2D user $d \in D$ from the user $i \in D$ can be expressed as $P_{R_d} = P_{T_i} \cdot |h_{i,d}|^2$, where $P_{T_i}$ is the transmit power of the user $i$ and $|h_{i,d}|^2$ is the channel gain on the link from the user $i$ to the user $d$. The SINR expressions for cellular and D2D links are, correspondingly, $\gamma_c = \frac{P_{T_b} \cdot |h_{b,c}|^2}{I_c \cdot \sigma^2}$ and $\gamma_d = \frac{P_{T_d} \cdot |h_{d',d}|^2}{I_d \cdot \sigma^2}$, where $\sigma^2$ is the noise power, whereas $I_c$ and $I_d$ represent the interference on the cellular and the D2D links, respectively.

In the considered system model, the eNodeB manages the radio resources by assigning the needed numbers of RBs to each scheduled user and then selecting the appropriate modulation and coding scheme (MCS) on every such RB. The underlying scheduling procedures are based on the CQI feedback, as transmitted by each UE to the eNodeB over the dedicated control channels. The data rate for the MCS level $q \in Q$ is denoted with $f(q, n_q)$ (where $Q$ is the set of the available MCS schemes), which is a function of $q$ and the actually assigned RBs $n_q \in N$ (the total number of

RBs $N$ strictly depends on the available system bandwidth). Then, $R_{Ch,c}$ and $R_{Ch,d}$ are the effective data rates for cellular and D2D transmissions, respectively.

With regards to the standard 3GPP handover procedures (i.e., see 3GPP specification 3GPP TS 36.331 V8.5.0), their evolution tailored to 4G systems and beyond has been specified by LTE Release 10 and 12. Presently, user handover is triggered by the eNodeB based on periodic channel measurements provided by the UE in question. In particular, the entire operation can be split into three main phases: (i) handover preparation, (ii) handover execution, and (iii) handover completion. During the handover preparation, the channel measurements are communicated by the UE to its serving eNodeB. The latter decides to trigger a handover subject to certain conditions. To this end, the serving and the target BSs exchange a series of control messages via the X2 signaling interface in order to transfer the UE-related parameters and allocate a portion of radio resources for such incoming "new" data session. At the handover completion phase, the target eNodeB informs the LTE Mobility Management Entity (MME) that the user-plane path has been changed successfully, whereas the Serving Gateway (S-GW) is commanded to update the said path.

In more detail, 3GPP specification TS 36.331 also describes the so-called A2, A3, and A4 triggering events. Particularly, A2 is invoked whenever the serving cell performance degrades under a certain threshold, A3 is effective when a neighboring cell becomes an offset better than the serving cell, and A4 is enabled if the neighboring cell becomes better than a given threshold. Correspondingly, for the *"A2-A4 Handover"* procedure, the Reference Signal Received Quality (RSRQ) measurements by the UE from its serving and the neighboring cells are utilized. To this end, a handover is triggered whenever the RSRQ from the serving BS falls below the preset threshold (event "A2") and then executed if the RSRQ for the target cell becomes higher than another threshold (event "A4"). The second standardized LTE handover scheme is named the *"A3 Handover"*. Accordingly, in every Transmission Time Interval (TTI), the eNodeB receives the Reference Signal Received Power (RSRP) measurements by the UE from its serving and the neighboring cells. The respective handover procedure, also known as the *Strongest Cell Handover Algorithm* or the *Traditional Power Budget* algorithm, triggers the change of the cell when the RSRP is an offset better (event "A3") than that in the original serving cell. However, such a handover is only executed if the discussed condition is effective for a certain time, to control the unwanted "ping-pong" effect.

### 4.2.2 Analyzing D2D-Assisted Handover Procedure

#### Proposed Mathematical Model

For the sake of analysis, a simplified scenario is considered as illustrated in Fig. 4.7(a). In particular, there are two BSs whose coverage areas intersect with each other at points $A$ and $B$ having the coordinates $(0, 150)$ and $(0, -150)$, respectively. Accordingly, they form a lens-shaped overlapping zone, which serves as a reasonable first-order abstraction, and has been demonstrated below that it allows for obtaining the basic understanding of performance gains behind the D2D-assisted handover. Further, $UE_1$ is assumed to move at the speed of $v$ m/s from $BS_0$ to $BS_1$ and thus crosses the area of interest at a straight line at points $O$ and $P$ with the coordinates $(-50, 0)$ and $(50, 0)$, respectively. Hence, $UE_1$ crosses the zone of overlap along the line connecting the centers of the BSs.

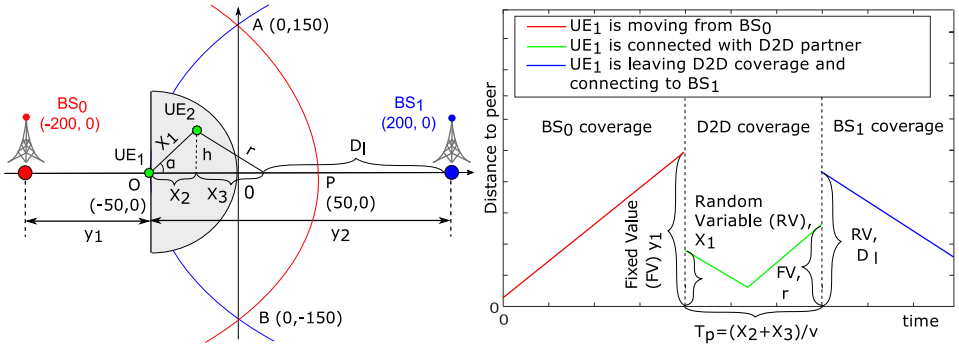



**Fig. 4.7.** A simplified scenario for analytical modeling.

It is additionally assumed that the D2D partners are distributed over the area $\Re^2$ according to a Poisson process with the intensity of $\lambda$ users per a square meter. These devices are considered to have the target content required by $UE_1$ and in practice may actually be a subset of all the UEs in the area of interest. The potential D2D partners are treated as static, while the direct transmission range equals $r = 50$m. Upon entering the zone of overlap, the UE in question selects one of the D2D users at the intersection of its coverage and that of the $BS_1$ to establish a direct link with it. All relevant information on the feasible D2D partners and their current locations is provided by $BS_1$ [10]. When, after some time, $UE_1$ leaves the connection range of a particular D2D partner, the corresponding link is transferred to $BS_1$, that is, the considered handover is completed. Note that for any given intensity $\lambda$, there is always a chance that there are no suitable D2D partners in proximity. The corresponding probability is given by $1 - \sum_{i=1}^{\infty} p_i$, where $p_i$ is the probability to have $i$ D2D users in the area of interest. The latter follows a Poisson distribution with the mean of $\lambda S_1$,

where $S_1$ is the area given by an intersection of the D2D coverage range centered at $O$ and the coverage of $BS_1$. In what follows, performance metrics are considered *conditioned* on the fact that at least one D2D user is available in this area.

The choice of a suitable D2D partner is of particular interest, as it directly affects the performance of $UE_1$. It is assumed here that there is no severe interference, which limits the SNR function to only depend on distance. To this end, Fig. 4.7(b)summarizes the distance-related parameters involved into our model. More specifically, the interest is in the

- SNR when entering the zone of overlap;
- amount of time a link with the D2D partner is active;
- distance to a new BS when leaving the D2D coverage.

Whereas simple, the described scenario allows for highlighting clearly the basic trade-offs behind the aforementioned performance metrics of interest. First, it is easy to see that for $UE_1$ to receive the best possible SNR upon entering the zone of overlap, the corresponding D2D partner has to be as close as possible to the point $O$. However, in order to maximize the D2D connection lifetime, the farthest available D2D user along the path of movement has to be selected. In addition, such a choice should also minimize the distance to the target BS once the D2D connection becomes unavailable. If there is a D2D device exactly at the point $O$, then the connection lifetime equals to $2r/v$, thus delivering the upper bound. However, there might be no users located in the vicinity of $O$. Therefore, during this research, the focus was on the random choice of the D2D partner.

Let $X_1$ be a random variable (RV) characterizing the distance between the point $O$ and a randomly-chosen D2D partner, denoted as $UE_2$. Instead of considering a more complex geometry of the area where this D2D device is selected, a good approximation of it is a semicircle centered at $O$, as shown in Fig. 4.7. This approximation should remain sufficiently accurate given that the BS coverage area is assumed to be significantly larger than that of the D2D users. Since the assumed distribution of suitable D2D UEs is Poisson and there is at least one D2D partner in the considered zone, its position is distributed uniformly inside the semicircle. By defining the radio propagation model as $L(d) = Ad^{-\gamma}$, where $d$ is the distance between the communicating entities and $\gamma$ is a propagation-dependent coefficient, the SNR can be expressed as a function of RV $X_1$ as

$$S_N(x_1) = \frac{P_{Tx}}{B\sigma^2} A x_1^{-\gamma} = K x_1^{-\gamma}, \tag{4.1}$$

where $\sigma^2$ is the power spectral density of noise at the receiver measured in db/Hz, $B$ is the bandwidth in Hz, $P_{Tx}$ is the transmit power, and $K = AP_{Tx}/B\sigma^2$.

It is important to note that the SNR experienced by $UE_1$ upon entering the area of interest does not have the mean value. This is due to the fact that a D2D partner may be located infinitely close to $UE_1$, which can be alleviated by disallowing any D2D user to be positioned in a sufficiently small semicircular restrictive zone of radius $r_\star << r$ around $UE_1$. In this case, the probability density function (pdf) of the distance to this D2D device [99] is produced by

$$f_{X_1}(x_1) = 2x_1/(r^2 - r_\star^2),\ r_\star < x_1 < r, \tag{4.2}$$

and the following integral for the moments of the RV in question converges

$$E[S_N^v] = \int_{r_\star}^{r} \frac{2y}{(r^2 - r_\star^2)} \left(\frac{K}{y^\gamma}\right)^v dy. \tag{4.3}$$

Evaluating (4.3) leads to

$$E[S_N^v] = \frac{2K^v(r^{(2-\gamma v)} - r_\star^{(2-\gamma v)})}{(r^2 - r_\star^2)(2-\gamma v)}, \tag{4.4}$$

and the mean SNR is thus equal to

$$E[S_N] = \frac{2K(r^{(2-\gamma)} - r_\star^{(2-\gamma)})}{(r^2 - r_\star^2)(2-\gamma)}. \tag{4.5}$$

The pdf of SNR upon entering the area of interest can be obtained without the above restrictive circular semicircle by employing the RV transformation technique according to [100]

$$w_{S_N}(y) = \sum_{i=1}^{M} f(\psi(y))|\psi'(y)|, \tag{4.6}$$

where $M$ is the number of branches of the inverse function and $x = \psi_i(y) = \phi^{-1}(x)$ is the $i$th branch.

Omitting the intermediate calculations, we establish the following for the SNR pdf

$$w_{S_N}(y) = \frac{2K^{2/\gamma}}{\gamma r^2} y^{-2/\gamma - 1}. \tag{4.7}$$

Further, the mean time during which a particular D2D connection is active has been also investigated. To this end, it is sufficient to determine the mean of the RV $D_c = X_2 + X_3$. Due to the fact that the chosen D2D partner is located uniformly within the semicircle, the angle $\alpha$ is also uniform in the range $(-\pi/2,\ \pi/2)$. Then, consider $UE_2$ and let $h$ be perpendicular to the trajectory of $UE_1$. Further, let the RV

$X_2$ correspond to the projection of $X_1$ onto the $UE_1$ trajectory. Similarly, the RV $X_3$ denotes the projection of the coverage radius of the D2D user. Hence, the distance up to the point where $UE_1$ leaves the D2D coverage area is $d_c = x_2 + x_3$. Note that the RVs $X_1$ and $X_2$ are mutually dependent as they are both expressed via the same RVs $\alpha$ and $X_1$, for which reason we cannot establish them separately. From the geometric considerations, it has been seen that $x_2 = x_1 \cos\alpha$ and $h = x_1 \sin\alpha$, whereas $x_3$ may be expressed via $r$, $\alpha$, and $x_1$ as $x_3 = \sqrt{r^2 - (x_1 \sin\alpha)^2}$. Therefore, it is possible to derive

$$d_c = x_1 \cos\alpha + \sqrt{r^2 - (x_1 \sin\alpha)^2}. \tag{4.8}$$

By taking the expectation, has been obtained

$$E[D_c] = E[X_1 \cos\alpha] + E[\sqrt{r^2 - (X_1 \sin\alpha)^2}]. \tag{4.9}$$

In more detail, the first component of the right-hand side in (4.9) is

$$\begin{aligned} E[X_1 \cos\alpha] &= E[X_1]E[\cos\alpha] = \\ &= E[X_1] \int_{-\pi/2}^{\pi/2} \frac{1}{\pi} \cos\alpha d\alpha = \\ &= \frac{4(r^3 - r_\star^3)}{3\pi(r^2 - r_\star^2)}, \end{aligned} \tag{4.10}$$

where the first transition holds due to independence of the involved RVs.

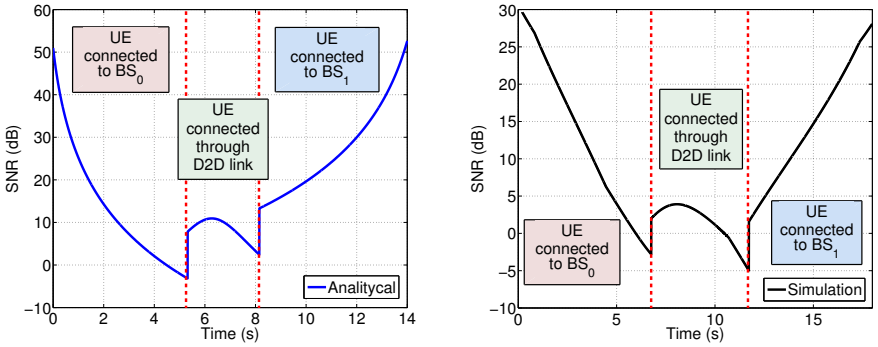
The mean of $X_3$ cannot be obtained directly, as there is no simple expansion of $\sqrt{r^2 - (X_1 \sin\alpha)^2}$. Hence, the fist step was to characterize the pdf of $X_3$, and then proceed with expressing the sought mean as

$$E[X_3] = \int_{r_\star}^{r} x_3 f(x_3) dx_3. \tag{4.11}$$

Accordingly, the pdf of $X_3$ can be derived by utilizing the useful RV transformation technique (see e.g., [100]), similarly to (4.6). Omitting the intermediate calculations, the resulting pdf is given by

$$f(x_3) = \frac{4x_3^2}{\pi(r^2 - r_\star^2)\sqrt{r^2 - x_3^2}}. \tag{4.12}$$

Summarizing, the mean of $X_3$ is delivered by




(a) Analytical results. (b) Simulation results.

**Fig. 4.8.** SNR achieved by $UE_1$, analysis and simulations.

$$E[X_3] = \int_{r_\star}^{r} x_3 \frac{4x_3^2}{\pi(r^2 - r_\star^2)\sqrt{r^2 - x_3^2}} dx_3 = \\ = \frac{4(2r^2 + r_\star^2)}{3\pi\sqrt{r^2 - r_\star^2}}. \tag{4.13}$$

Therefore, the mean distance $E[D_c]$ is expressed as

$$E[D_c] = \frac{4(r^3 - r_\star^3)}{3\pi(r^2 - r_\star^2)} + \frac{4(2r^2 + r_\star^2)}{3\pi\sqrt{r^2 - r_\star^2}}, \tag{4.14}$$

thus leading directly to obtaining the mean D2D contact time.

Finally, consider the distance $D_l$ to the target BS when $UE_1$ leaves the D2D coverage. From the geometrical considerations (see Fig. 4.7) and using (4.14), we establish

$$E[D_l] = y_2 - \frac{4(r^3 - r_\star^3)}{3\pi(r^2 - r_\star^2)} - \frac{4(2r^2 + r_\star^2)}{3\pi\sqrt{r^2 - r_\star^2}}. \tag{4.15}$$

### Validating the Analytical Model

In order to validate the analytical model proposed above, the SNR achieved by the user in the scenario given by Fig. 4.7 has been evaluated. In particular, has been investigated the case when $UE_1$ moves from the point $(-150, 0)$ to the position $(150, 0)$. Here, $UE_2$ is the D2D partner, which location is locked at the coordinates $(-25, 37.5)$. In what follows, the SNR for $UE_1$ is computed both analytically and via simulations – the analytical part has been completed in Matlab, whereas the simulation profile is implemented in the open-source discrete-event NS-3 environment [101]. Correspondingly, the obtained results are illustrated in Fig. 4.8.

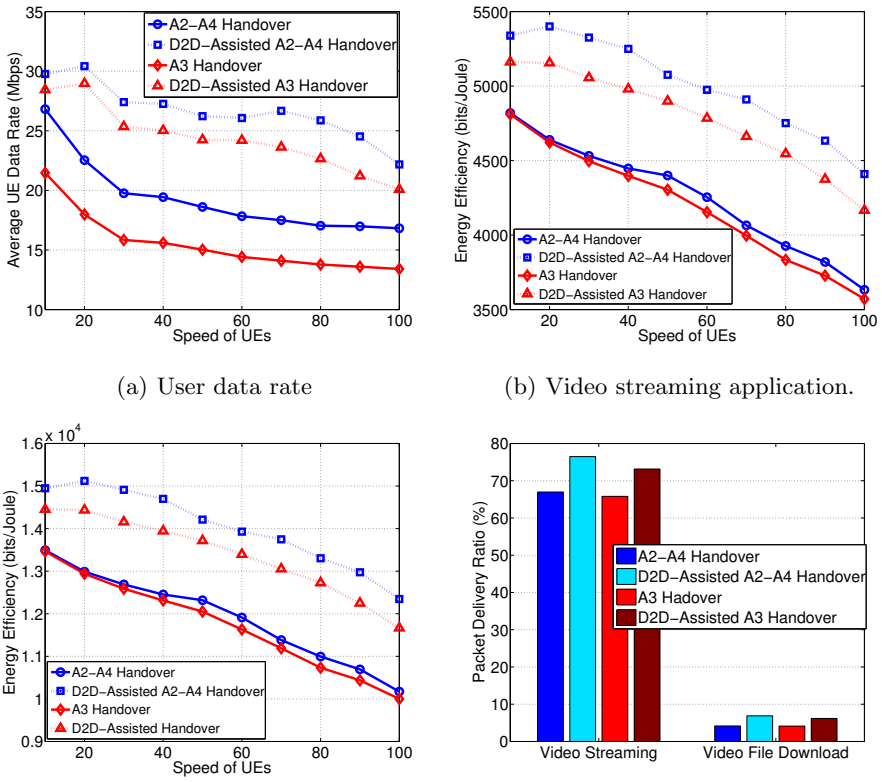
As it is possible to learnt from the plots, both curves indicate a similar trend where the SNR of $UE_1$ decreases with the growing distance to the BS. In particular,

when the user resides in the zone of overlap, the $BS_0$ initiates the handover procedure. Then, a D2D connection is established once the user discovers (as assisted by the cellular network) a suitable D2D partner in proximity (e.g., $UE_2$). Consequently, the SNR achieved by the user increases dramatically in contrast to the conventional cellular case. Naturally, since reliable D2D coverage spans for only some tens of meters, the farther the user moves from the D2D partner the lower the resulting SNR becomes. It is worth noticing that during its D2D connection $UE_1$ executes normally all the required handover procedures. Later, when the D2D connection is no longer reliable due to the lengthening D2D link, $UE_1$ reconnects to $BS_1$, which means that all the necessary handover operations are completed and the user service has been successfully transferred to the target BS, $BS_1$.

### 4.2.3 Numerical Performance Evaluation

To evaluate the effects of the proposed D2D-assisted handover, an extensive simulation campaign has been conducted with the popular NS-3 tool. To this end, the existing NS-3 modules and functionalities, such as the S1 and X2 interfaces, the propagation loss models, and the conventional LTE-based handover algorithms have been updated for the purpose. The simulated scenario consists of two LTE macro BSs deployed within the area of [500m x 500m]: the first BS is centered at $(-200, 0)$, whereas the second BS is positioned at $(200, 0)$. Consequently, the coverage radius of each BS is approximately 250m, which defines the zone of overlap by the points $A = (0, 150)$, $B = (0, -150)$, $O = (-50, 0)$, and $P = (50, 0)$ (see Fig. 4.7).

The performed simulations differentiate between two groups of devices: the *D2D partners* and the *cellular users*. The first group corresponds to the UEs acting as D2D relays and willing to forward the multimedia data to the cellular users during their handover procedures. These are uniformly distributed in the area described by the points $AOBP$ and move with the speed of around 0, thus being relatively static. In contrast, the second group represents the conventional LTE cellular users that are deployed uniformly within their macro cells and have the movement speeds in the range of $[10, 100]$ km/h, with mobility patterns characteristic of the Random Waypoint model. For the sake of an example, the number of cellular users in the system equals to 100 UEs (uniformly shared by the two macro BSs), whereas the number of D2D relays is set to 20 UEs. Further, two different types of over-the-top applications have been considered: (i) a delay-sensitive service, e.g., *video streaming* and (ii) a delay-tolerant service, e.g., *video file downloading*. Each application is defined by its data inter-arrival time and packet length value; and the reader is referred to Table 4.3 for the main employed system parameters.




(a) User data rate

(b) Video streaming application.

(c) Video file downloading application.

(d) Packet delivery ratio

**Fig. 4.9.** Average UE energy efficiency and packet delivery ratio for UE movement speed of 100 km/h.

The performance evaluation concentrates on three main metrics of interest discussed previously: (i) the *UE data rate*, which essentially is the throughput that a user experiences during the download of its desired content; (ii) the *packet delivery ratio (PDR)*, which is the ratio between the packets successfully delivered to the destination over the total number of packets that have been sent; (iii) the *UE energy efficiency*, which is the ratio between the number of bytes received and the total energy consumed to obtain said data. The latter is defined as:

$$EE = \frac{Bits_{received}}{(T_{r_{cell}} \cdot P_{r_{cell}}) + (T_{r_{D2D}} \cdot P_{r_{D2D}})}, \tag{4.16}$$

where $T_{r_{cell}}$ and $T_{r_{D2D}}$ is time required to receive the data over the cellular and the D2D links, respectively, $P_{r_{cell}}$ is the transmit power in the cellular mode, and $P_{r_{cell}}$ is the transmit power in the D2D mode.

First, a preliminary study concentrates on the average UE data rates as reported by Fig. 4.9(a). This performance parameter is relatively independent from the running application, as it simply characterizes the amount of data that a user is able to receive under a certain quality of its current link. In particular, is observed that the data rate in question decreases linearly as the average user speed grows along

**Table 4.3.** Main simulation parameters

| Parameter | Value |
|---|---|
| Cell radius | 500m |
| Frame Structure | Type 2 (TDD) |
| TTI | 1ms |
| TDD configuration | 0 |
| Carrier Frequency | 2GHz |
| eNodeB Tx power | 46dBm |
| D2D user Tx power | 13dBm |
| UE Tx power | 23dBm |
| Noise power | -174dBm/Hz |
| Pathloss (cell link) | 128.1 + 37.6 log(d[km]) |
| D2D pathloss (LOS) | 16.9 log(d[m]) + 20 log(f[GHz]/5) + 46.8 |
| Shadowing deviation | 10dB (cell mode); 12dB (D2D mode) |
| BLER target | 10% |
| UE movement | Random Waypoint model |
| # of UEs | 100 |
| # of D2D relays | 20 |
| Simulation time | 360s |
| Inter-arrival time (video streaming/downloading) | 30ms / 120s |
| Data size (video streaming/downloading) | 100KB / 100MB |

the horizontal axis. This is due to an increase in the channel quality fluctuations at higher speeds. Here, the most important learning is that D2D communications deliver considerable throughput gains as compared to the standard LTE handover procedures. For instance, with the proposed D2D assistance, the A2-A4 handover algorithm enjoys a data rate improvement of over 30% at a speed of 10 km/h. Similarly, for the A3 handover algorithm with D2D assistance, a gain of 37% becomes available. The main reason behind these results is in that a D2D link in the area of overlap allows mobile UEs to maintain higher channel quality levels in the course of their migration between the BSs.

The second useful parameter discussed here is the average energy efficiency of the UEs performing a handover. As we can conclude from Fig. 4.9(b) and Fig. 4.9(b), both considered applications demonstrate a similar decreasing trend with the growth in movement speeds. This is explained by the fact that the only factor varying in (4.16) is the number of bits received by users, whereas the packet size remains higher for the video file downloading application. To this end, an interesting finding is that the use of D2D assistance enables the respective gains of about 15% for both A2-A4 handover and A3 handover schemes. The extent of this improvement is rooted in the

lower power and time needed to acquire the content over a D2D link as compared to the conventional cellular-based multimedia download.

Finally, the PDR values for video streaming and video file downloading are contrasted in Fig. 4.9(d). For clarity, the focus was on the worst-case performance, when users move at the velocities of 100 km/h. What it is possible to learn from the corresponding charts, is that the PDR increases whenever D2D communications are utilized to support the current LTE handover procedures. In case of a video streaming application, around 75% of packets are decoded correctly by their target users as compared to 65% if no D2D assistance has been employed. Similarly, for video file downloading scenario the respective PDR improves by 4 to 8% with D2D-assisted handover. The differences between the two applications are related to the fact that the amount of data transmitted by the LTE eNodeB is significantly higher in case of video file downloading.

## 4.3 Mobility-Aware D2D-Empowered 5G Systems

Within the complex and diversified portfolio of 5G-grade applications and services, multimedia streaming and interactive gaming rapidly gain user popularity, but also require lower battery consumption and higher throughput. This is evidenced by an unprecedented increase in multimedia traffic, which challenges the contemporary content distribution systems. Consider a "fully" mobile scenario where multiple users moving accross a give are of interest request a certain fragment of multimedia content from the network infrastructure asynchronously. Then, concurrent and independent links are typically established, which adds to the network capacity crunch and may produce bottlenecks. Indeed, with the conventional unicast transmissions, the per-user throughput and energy efficiency decrease linearly with the number of users. Moreover, users at the cell edge typically suffer from poor channel conditions due to larger distances to the BS and, potentially, higher interference. To make 5G systems more resource and energy efficient, improved network selection and optimized resource utilization are thus in prompt demand.

### 4.3.1 Resource Allocation and Connectivity Management in 5G

Due to potential availability of multiple radio access technologies (RATs) as well as direct D2D connections, optimizing system-wide performance of a D2D-empowered 5G network with heterogeneous mobility becomes a complex and non-trivial task. Notably, having an opportunity to connect to a multitude of serving cells does not always lead to immediate benefits for the system nor the user [102]. In fact, with the ongoing extreme densification of wireless infrastructure, adequate network selection

and management strategies need to be developed to avoid unnecessary confusion for the users as well as mitigate the extra complexity for the telecom operators. Should users act in an uncontrolled manner when selecting their most suitable connectivity options, such selfish behavior could lead to excess battery consumption. This is due to more frequent handovers between the alternative RATs, which may result in significant quality-of-service (QoS) degradation.

The above accentuates the need for increased network involvement into the resource allocation process, and a differentiation between (i) *network-centric*, (ii) *network-assisted*, and (iii) *user-centric* approaches [102]. The network-centric concepts assume that a dedicated coordinating entity (residing e.g., at the LTE eNodeB) is in charge of collecting the user equipment (UE) state information (channel quality, connectivity options, etc.) as well as responsible for performing network-wide resource optimization. Once computed, resource allocations are enforced by advertising them to the access points and then the users. Such a system assumes full control of all the access points, and thus becomes feasible for the network operators with high-speed backhaul connections (e.g., in the C-RAN settings). In addition, network-centric mechanisms potentially allow for tighter optimization of the radio resource management, also serving as a benchmark for network-assisted and user-centric approaches.

With network-assisted resource management, the central coordinating entity still collects (partial) information about the UE states and communicates it to the users. However, here the UEs make the actual network selection decisions directly, while minding their heterogeneous constraints (throughput, energy, etc.). This method is a distributed control scheme and may not be optimal, but has the potential to approach the network-centric solutions. Finally, user-centric resource allocation is when the UEs make local decisions on their connectivity, since no information is provided by the network. Users measure their own channel state information but are unaware about the states of other users. The resulting decision algorithms are relatively simple: e.g., choose a cell based on certain preset preferences (battery consumption, perceived throughput, loading variations, etc.). While the resultant resource distribution depends heavily on the user decision logic, the resource utilization often remains far from optimal.

### 4.3.2 New Framework to Assess Time-Dependent 5G Behavior

#### Characteristic D2D-Empowered HetNet Deployment

The reference HetNet scenario illustrated in Fig. 4.10 comprises an area of interest served by a macro BS with the coverage radius of $r_M$. There are also $M$ micro BSs deployed in the area, each having the coverage radius of $r_m$ together with a number

of mobile users, $N$, that are distributed uniformly. All of the users are equipped with direct communications capabilities, while some of them act as the D2D *partners*, hence providing with an additional connectivity option. At each instant of time, a user is allowed to only be associated with a single network tier (e.g., macro or micro), or have an active D2D connection to its partner. The users can only establish direct connections if they reside in mutual proximity i.e., within the D2D radius of $r_D$.

Whenever system-wide (global) optimization is conducted, the macro BS follows a *network-centric* approach to perform radio resource allocation aiming to supply every user with the highest possible performance (i.e., using a max throughput or max energy efficiency scheduler). Since the goal of this research has been in the characterization of time-dependent system behavior, in case of any local connectivity changes, the users begin to decide individually as to which tier they are willing to utilize for maximizing their own performance (that is, follow a user-centric approach). Accordingly, it may happen that the selected connection is not feasible due to a lack of radio resources, prohibitive interference, or absence of mutual proximity (in case of a D2D link). When no global optimization is performed, D2D connections are preferred over the small cell links. Here, macro BS tier has the lowest priority. Notably, exploiting a solution that guarantees the highest possible user throughput may translate (under certain conditions) into improved energy efficiency and better network sustainability [103].

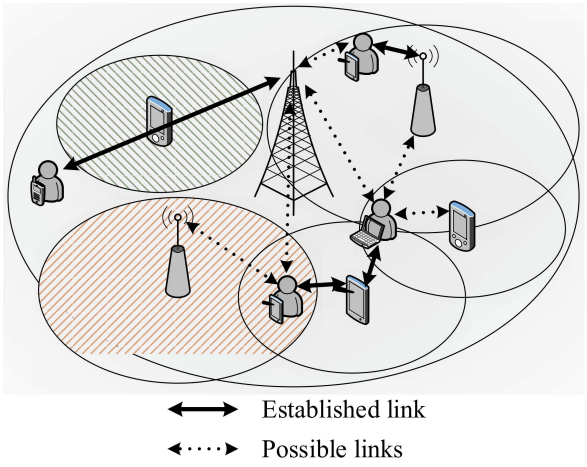



**Fig. 4.10.** Our reference 5G-grade HetNet scenario.

It is assumed that users request certain popular content (e.g., multimedia sharing at a mass event, such as a football match, marathon, or concert). In the general setup, the content is pre-cached by the macro and small cell BSs located within the service area. In addition, is required that such pre-cached content is available at $N - K$ out of $N$ users serving as the D2D partners. Hence, only $K$ users are interested in

acquiring the content. All users move across our area of interest according to e.g., random-direction mobility (RDM). This simple model has been shown in the past to provide qualitatively similar performance as compared to more complex and realistic formulations [22]. Finally, it is assumed that the locations of users, the interference picture, and the connectivity parameters are available to the macro BS, which represents the central *decision-making entity*. Said information is used to calculate the resource allocation shares. These allocations are then advertised to the users enforcing their connectivity options within the area of interest.

**Proposed 5G System Modeling Considerations**

The time-dependent behavior of the above HetNet deployment can be characterized "qualitatively" with our proposed formulations. Consider the time instant $t = 0$, when the network-wide optimization of the 5G system has been performed. The system state is represented as e.g., a vector indicating the numbers of UEs associated with their serving entities, including the micro and macro BSs as well as the D2D partners. For particular deployment parameters (the number of BSs and D2D peers, the area of interest, the user mobility patterns, etc.), temporal system evolution can be characterized by semi-Markov processes, where the time interval between the state changes corresponds to the minimum inter-connectivity change time. Given certain mobility patterns of users and their parameters, and relying on the powerful random walk theory, for sufficiently large numbers of users in the system the latter mainly depends on the average user speed.

Broadly, one could assess the 5G system behavior by representing it as a discrete-time Markov chain with the discretization interval of $\tau$ that corresponds to the minimum inter-connectivity change time, and then solve it for the number of steps required to transition from a globally-optimized solution at $t_0$ to the user-centric result at some $t_1$, $t_1 > t_0$. Denote $\Delta_g$ as the absolute difference between the network-centric and the user-centric performance, whereas $\Delta_t$ is the divergence time from the optimal solution, $\Delta_t = t_1 - t_0$. The parametrization of the model in question is a complex task, since mapping the optimal allocations onto the states of the Markov chain and subsequent derivation of the minimum inter-connectivity change time are both non-trivial. Furthermore, since the UE is allowed to reside within coverage of multiple serving entities at any time instant, the state space of the model explodes. However, the proposed methodology allows to specify the qualitative behavior of the 5G system after the network-wide optimization point.

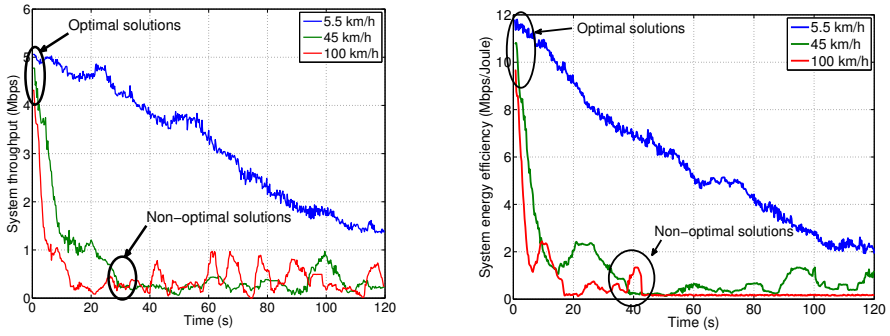
Observe that the overall framework proposed can be regarded as a stochastic system that approaches its steady-state corresponding to the user-centric resource allocation from a certain state corresponding to the network-centric (optimized) allo-

cation. In the theory of Markov chains, this is named the *mixing time* of a chain [104]. Particularly, it is known that the mixing time is proportional to the sum $C\sum_{i=1}^{H-1}\nu_i^n$, where $n$ is the time-step of the chain, $\nu_i < 1$, $i = 1, 2, \ldots, H-1$, are non-unit eigenvalues of the transition probability matrix of a chain, $H$ is the total number of states in the chain, and $C$ is a constant. The transition probability matrix of an irreducible Markov chain always has the dominant eigenvalue, that is, an eigenvalue whose value is close to a unit dominating the sum. By letting $\Delta t \to 0$, the approximating function is $y(t) = Ce^{-\lambda t}$, where $\lambda$ is the divergence rate from the optimal state with respect to a certain metric.

Therefore, is expected the *exponential degradation* of all the performance metrics of interest after the optimization time instant. To confirm this theoretical conclusion as well as provide quantitative evidence on the temporal 5G system behavior, the following section carries out an extensive system-level evaluation.

### 4.3.3 Time-Dependent 5G Performance Evaluation

In this section, is reported an overview on the results of extensive simulation effort conducted to confirm the theoretical conclusions made in the previous section as well as deliver numerical evidence to understand the effects of mobility and the number of users on the system-wide (re-)optimization of both throughput and energy efficiency parameters. Further, by utilizing the state-of-the-art energy-aware BS switch on/off schemes [105], has been demonstrated the applicability of the proposed methodology in practical scenarios, where users exchange rich multimedia content (e.g., pictures or videos) at a mass event, such as the stadium area during a football match.




(a) Aggregate system throughput (b) Aggregate energy efficiency

**Fig. 4.11.** Temporal evolution of system throughput and energy efficiency values with $N = 50$ users.

**5G System-Level Simulation Framework**

The numerical assessment has been conducted within the network simulator 3 (ns-3) environment that is applicable for system performance evaluation across a wide range of use cases. Here, in the considered scenario UEs are uniformly distributed within a multi-RAT 5G deployment and follow the RDM mobility patterns, while acquiring certain heavy content. Three types of HetNet connectivity tiers have been modeled: (i) macro 3GPP LTE eNodeB, (ii) 3GPP LTE small cells (e.g., femto cells), and (iii) WiFi-based D2D "cells". For the latter, is assumed that 20% out of $N$ users act as the D2D cachers for the purposes of proximity-based content distribution. A single macro LTE eNodeB with the coverage radius of 500 m provisions for 100 resource blocks (RBs), out of which 25 RBs are available to 10 deployed femto cells.

Channel conditions of the UEs are evaluated in terms of SINR experienced on each sub-carrier when the path loss and fading effects are accounted for. The D2D connection discovery and establishment are managed directly by the eNodeB (i.e., with 5G-grade ProSe functionality), whereas the D2D links take advantage of WiFi-Direct protocol. The transmitted traffic is modeled after the "Facebook Live" video streaming service with the quality of $720p$ and the maximum bitrate of 4 Mbps. The system-level metrics under consideration are: (i) time-dependent *system throughput* and (ii) time-dependent *system energy efficiency*. More specifically, these parameters are assessed by taking into account (i) their evolution over a particular time window $\Delta_t$ and (ii) the corresponding intensity of user mobility. For further details, refer to Table 4.4 that collects the primary simulation parameters.

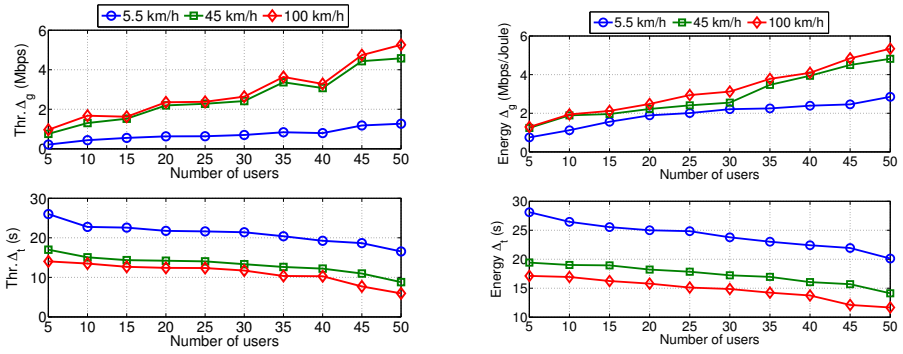
With respect to the employed optimization algorithm, the considered setup manages the network-wide selection of the most appropriate connectivity option via a dedicated utility function that relates the perceived user throughput and the respective transmit power together with the actual traffic loading. Accordingly, the advanced RAT selection method proposed extends simpler past considerations of signal strength or channel quality, as it takes into account the real-time 5G system conditions both at the network- and user-side.

**Representative Performance Results**

In Fig. 4.11, is reported the obtained performance results for three different profiles of user (device) mobility: 5.5 km/h (low), 45 km/h (moderate), and 100 km/h (high). Here, the characteristic density of 10 femto cells and 50 users has been considered, out of which 20% serve as D2D "cells". It is worth observing that the aggregate throughput and energy efficiency values decrease over time after the system-wide optimization instant for all the intensities of mobility. The transition time from the

**Table 4.4.** Key simulation parameters

| **System parameter** | **Value** |
|---|---|
| Macro cell radius | 500 m |
| D2D 'cell' radius | 30 m |
| Femto cell radius | 50 m |
| LTE carrier frequency | 2.6 GHz |
| WiFi-Direct carrier frequency | 2.5 GHz |
| User (UE) transmit power | 23 dBm |
| Macro cell transmit power | 46 dBm |
| Femto cell transmit power | 20 dBm |
| D2D link setup time | 1 s |
| D2D target data rate | 40 Mbps |
| Number of runs | 500 |
| **Application parameter** | **Value** |
| Video resolution | 720p, 30 fps |
| Key-frame interval | 1 every 2 s |
| Max bit rate | 4 Mbps |
| Rate control | CBR |
| Audio sample rate | 44.1 KHz |
| Audio bitrate | 128 Kbps |




(a) Aggregate system throughput (b) Aggregate energy efficiency

**Fig. 4.12.** Temporal evolution of system throughput and energy efficiency values for varying numbers of users.

network- to the user-centric performance is produced by employing the exponentially-weighted moving average (EWMA) scheme with the corresponding weight coefficient for the current observation set to $\gamma = 0.1$.

It is important to note that lower intensity of mobility results in a longer time before transitioning to the user-centric operation, since the environment changes less dynamically. Further, observe that the difference between the network-centric and the user-centric performance levels is dramatic, with the difference on the order of 5 to 11 times. In contrast, as confirmed by Fig. 4.11(a) and Fig. 4.11(b), the experienced throughput and energy efficiency decrease to the sub-optimal levels more rapidly as the user speeds grow higher. This is because faster UEs (e.g., connected cars, drones, etc.) leave their optimal operating conditions earlier as well as have higher chances to encounter alternative (non-optimal) connectivity options on their path, thus in general degrading their aggregate throughput and energy performance.

Further, Fig. 4.12 introduces the corresponding $\Delta_g$ and $\Delta_t$ results (for both throughput and energy efficiency) as the number of users in the system varies from 5 to 55. Consistent with the above observations, for higher UE mobility the system performance degrades faster to the user-centric levels where users have to decide individually on their radio connectivity options, since network-controlled recommendations are lagging behind. In this case, the network-wide re-optimization of the system has to be performed more frequently compared to when the user speeds are lower (i.e., 5.5 km/h).

The discussed behavior is also very visible in Fig. 4.11, where the time interval $t + \Delta t$ representing the deviation from the optimized solution becomes shorter as the intensity of UE mobility grows. A somewhat similar behavior is also observed for the energy efficiency values. it is worth noticing that for a certain user speed, the increase in the number of served devices does not impact the metrics under consideration significantly, while the associated increase in $\Delta_g$ as well as the decrease in $\Delta_t$ remain linear. Here, the curve for $\Delta_g$ grows as the user speeds increase.

Analyzing the obtained results, it is possible to conclude that user mobility has a profound implication on the selection of the appropriate 5G system re-optimization period. While for lower intensity of mobility (i.e., around 5.5 km/h) the performance degradation time is rather long and remains on the order of minutes, in higher mobility scenarios (e.g., vehicular, urban, industrial) said divergence time is much shorter. As a result, the choice of when the 5G system has to be re-optimized becomes a fundamental consideration for the network operators, as well as affects the underlying trade-off between the system performance optimality and the amount of efforts (in terms of computation and signaling overhead) needed to maintain it.

The actual value of the divergence rate, $\lambda$, may be established by solving the equation $y(t) = Ce^{-\lambda t}$ for $C$ and $\lambda$ at the time instants $t = 0$ and $t = \Delta t$. Thus is possible to arrive at $C = G_2$ and $\lambda = (1/\Delta_t)\ln(G_2/G_1)$, where $\Delta_g = G_2 - G_1$. These important findings are reported in Table 4.5. In practice, after estimating the average speed of connected UEs, the gap between the network- and the user-centric operation levels could be controlled, as the proposed approach allows the 5G service providers to cater for their desired trade-off between the signaling/computing load and the resulting network performance by adjusting the re-optimization period.

**Table 4.5.** Divergence exponents $\lambda$ for various user speeds.

| Number of users | **Speed** | | |
|---|---|---|---|
| | 5.5 km/h | 45 km/h | 100 km/h |
| 5 | 0.01022 | 0.09001 | 0.17585 |
| 10 | 0.01233 | 0.08188 | 0.22426 |
| 15 | 0.01625 | 0.12330 | 0.18381 |
| 20 | 0.01235 | 0.13359 | 0.23081 |
| 25 | 0.01207 | 0.13837 | 0.18455 |
| 30 | 0.01279 | 0.13927 | 0.23914 |
| 35 | 0.01172 | 0.15923 | 0.28507 |
| 40 | 0.01255 | 0.16279 | 0.28836 |
| 45 | 0.01400 | 0.21258 | 0.45732 |
| 50 | 0.01600 | 0.24040 | 0.59603 |

### Practical Considerations behind the Divergence Rates

As an important example for the practical applicability of the proposed framework, the effects of the optimized 5G network performance on its energy efficiency operation has been addressed, with the goal to improve the overall system sustainability. Further, a scenario where multiple users assemble together within a confined area of interest has been considered, thus straining the network capacity. This situation is characteristic, for instance, of an important football match when the fluctuations in the instantaneous mobile traffic demand become significant[1]. In this case, the optimality of 5G network performance in terms of efficient resource utilization becomes a critical operating factor that in turn affects the energy efficient operation.

[1] See "Ericsson mobility report"', Fig. 2 – pp. 21. Available at: https://www.ericsson.com/res/docs/2015/ericsson-mobility-report-june-2015.pdf [Accessed 05/2015]

Along these lines, Fig. 4.13 demonstrates the offered traffic loading on the network infrastructure during the entire mass event in question (e.g., a football match). Then, the network-centric optimization is performed attempting to improve system performance at every instant of time chosen by the 5G operator (thus re-optimizing the system every $\Delta_t$), while the BS switching approach from [105] is applied to control the capacity supply by only turning on the small cell BSs that are currently in use. Our implementation of this method employs the divergence time values summarized by Table 4.5.

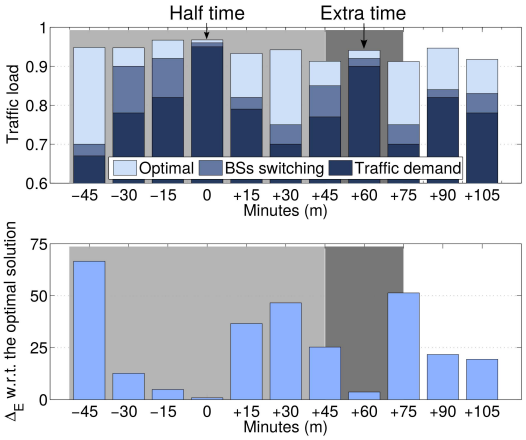



**Fig. 4.13.** Temporal evolution of throughput and fairness for varying user speeds.

Understanding the data collected in Fig. 4.13, one may observe that using the BS switching mechanism brings along significant performance benefits as it is able to meet the actual traffic demand during the entire mass event. On top of this, by switching the small cell BSs on and off across the considered area of interest, the system also achieves considerable energy savings (see $\Delta_E$ in Fig. 4.13) that vary from 68% during intervals, when the spectators are more interested in following the mass event rather than in sharing their multimedia content, and down to 1.8% at times of the peak mobile traffic loading.

# 5

# Safety and Social Relationships

Driven by the unprecedented increase of mobile data traffic, device-to-device (D2D) communications technology is rapidly moving into the mainstream of fifth-generation (5G) networking landscape. While D2D connectivity has originally emerged as a technology enabler for public safety services, it is likely to remain in the heart of the 5G ecosystem by spawning a wide diversity of proximate applications and services. In this Chapter, it is argued that the widespread adoption of the direct communications paradigm is unlikely without embracing the concepts of trust and social-aware cooperation between end users and network operators. However, such adoption remains conditional on identifying adequate incentives that engage humans and their connected devices into a plethora of collective activities. To this end, the mission of this research is to advance the vision of social-aware and trusted D2D connectivity, as well as to facilitate its further adoption. Fist, is provided a review of the various types of underlying incentives with the emphasis on sociality and trust, then, these factors specifically for humans and for networked devices (machines) are discussed as well as novel frameworks allowing to construct the much needed incentive-aware D2D applications. Finally, supportive system-level performance evaluations are performed by suggesting that trusted and social-aware direct connectivity has the potential to decisively augment the network performance. Further, an outline of the future perspectives of development across research and standardization sectors are also illustrated.

## 5.1 Context-Aware Information Diffusion in 5G Mobile Social Networks

### 5.1.1 Reference Scenario and System Model

The reference scenario is an emergency event in which updated alerting messages are sent out to reach the widest set of users within the shortest diffusion time. The alerting message is built based on context-aware information collected from devices and objects scattered in the area of interest. In this research are considered small-scale

areas like a shopping mall, social aggregation places, University campus, where an LTE femtocell is installed to guarantee connectivity. When a sudden emergency occurs, end-users and rescue teams must be promptly provided with up-to-date alerting and context information to best face the situation. The objective is therefore, to define a framework where the devices send updated information to the base station and a composite information is then sent in the downlink direction to reach as many as possible devices in a short time. Within this framework, D2D communications will play a fundamental role not only to enhance the performance in terms of diffusion time (both in the uplink and in the downlink), but also to reach those devices that are not under direct network coverage.

### LTE-A System Background

The main LTE assumptions and D2D connections establishment and management are those already illustrated in Section 3.1.1 and Section 4.2.1.

In reference to this particular research conducted, with $\mathcal{C}$ is indicated the set of $C$ cellular users, with $\mathcal{D}$ the set of $D$ D2D users, and with $\mathcal{L}$ the set of $L$ MCSs in the system [73]. According to the free space propagation loss model, the power received by a generic UE $j$ on the direct link $i \rightarrow j$ can be written as: $Pr_j = Pt_i \cdot |h_{i,j}|^2 = Pt_i \cdot Pl_{i,j}^{-\alpha} \cdot |h_0|^2$, where $Pt_i$ is the transmitted power from UE $i$, $h_{i,j}$ is the channel gain on link $i \rightarrow j$, $h_0$ is the channel coefficient, $Pl_{i,j}$ is the path loss on the link $i \rightarrow j$, and $\alpha$ is the path loss compensation factor computed by the eNodeB based on the operation environment (in the [0,1] range). Assuming that all subcarriers in one RB experience the same channel conditions, the SINR $\gamma_j$ for the generic user $j$ on a link $i \rightarrow j$ is: $\gamma_j = \frac{Pt_i \cdot Pl_{i,j}^{-\alpha} \cdot |h_0|^2}{\mathcal{I}_j + N_0}$, where $N_0$ is the thermal noise density level at the receiver, and $\mathcal{I}_j$ is the set of interfering signals received by user $j$, which takes different values in case of either a D2D or a cellular unicast downlink transmission. Specifically, the power received by user $i$ during a transmission from the BS $b$ is expressed as $Pr_i = Pt_b \cdot |h_{b,i}|^2$, where $Pt_b$ is the power transmitted by the BS.

Similarly, for the D2D communications, the received power for a generic D2D user $d$ during a transmission from a user $i$ can be expressed as $Pr_d = Pt_i \cdot |h_{i,d}|^2$, where $Pt_i$ is the transmission power and $|h_{i,d}|^2$ the channel gain from $i$ to $d$. Cellular and D2D users reuse the same portion of radio spectrum, thus causing mutual interference.

To model the interference on the different links, let $x_{c,d} \in \{0, 1\}$ be a binary variable having value "1" when D2D user $d$ transmits on the same RBs used to transmit from the BS to cellular user $c$ and having value "0" otherwise. Similarly, $y_{d',d} \in \{0, 1\}$ is a binary variable having value "1" when D2D user $d$ transmits on the same RBs used to transmit over a D2D link to $d'$. When considering a downlink transmission from the BS to a cellular user $c$ (similar analysis can be repeated in the

reverse direction), the total interference experienced by cellular user $c$ when receiving data from the BS comes from all D2D links reusing the same RBs that are serving the cellular user $c$:

$$\mathcal{I}_c = \sum_{d \in \mathcal{D}} x_{c,d} \cdot Pt_d \cdot |h_{d,c}|^2 \tag{5.1}$$

Similarly, when focusing on a D2D link $d \rightarrow d'$ , the interference at the receiver $d'$ is given by the power signals of the cellular user $c$ and all the D2D UEs $d^{''} \in \mathcal{D} \backslash \{d, d'\}$ that are transmitting on the same RBs:

$$\mathcal{I}_{d'} = \sum_{c \in \mathcal{C}} x_{c,d} \cdot Pt_b \cdot |h_{b,d'}|^2 + \sum_{d'' \in \mathcal{D} \backslash \{d,d'\}} y_{d,d''} \cdot Pt_{d''} \cdot |h_{d'',d'}|^2 \tag{5.2}$$

Given the noise power $\sigma^2$, the SINR for a cellular transmission from the BS to cellular user $c$ can be written as:

$$\gamma_c = \frac{Pt_b \cdot |h_{b,c}|^2}{\mathcal{I}_c} \tag{5.3}$$

Whereas the SINR on the D2D link from $d$ to the receiver $d'$:

$$\gamma_{d'} = \frac{Pt_d \cdot |h_{d,d'}|^2}{\mathcal{I}_{d'} + \sigma^2} \tag{5.4}$$

The capacity achievable by cellular user $c$ served by BS $b$ with a bandwidth $W$ is given by Shannon's formula:

$$\Phi_c = W \cdot log_2(1 + \gamma_c) \tag{5.5}$$

whereas the capacity for D2D user $d$ is:

$$\Phi_d = W \cdot log_2(1 + \gamma_d). \tag{5.6}$$

As described in the previous Chapters of this thesis, the resulting SINR values defined in equations (5.3) and (5.4) can be mapped to the corresponding CQI values for all users on either cellular and D2D links. In particular, an MCS value corresponds to each CQI value, according to 3GPP 36.213 specifications.

### 5.1.2 The D2D-enhanced Information Diffusion Scheme

#### Downlink Information Diffusion Problem Formulation

The information diffusion process, where the collected context-aware information are transmitted to a set of LTE users, can comply to different technological solutions to minimize the diffusion time. Different from classic multicast solutions, the general idea in this work was that the BS transmits data in unicast modality only to a subset of the users. This set of selected nodes, hereafter called as *primary bridge nodes* (PBNs)

will act as data forwarders in the network to serve all, or part, of the remaining information recipients. The transmission technology adopted by the PBNs to forward data is D2D, which can either exploit unicast or multicast transmissions. In the latter case, a PBN actually forms a cluster of nodes and serves them by acting as the *cluster head*. In those cases where still not all users are served, the possibility for a further transmission hop in D2D modality is considered. In this case, the node acting as forwarder will be referred to as a *secondary bridge node* (SBN). Based on such a hierarchical structure, there is the need to compute the diffusion time defined as the time interval between the start of transmission and the completion of the information diffusion process [106] (i.e., all interested devices have received the data from either the BS, a PBN or a SBN).

To this aim, the D2D-based network is represented by a weighted and undirected graph $G = (V, E)$, where $V$ denotes the set of $n$ nodes, whereas $E$ denotes the set of edges connecting the nodes (D2D links). For a generic link between two nodes $i, j \in V$, $\omega_{i,j}$ is considered as the weight associated to the corresponding edge in the graph. To model this weight networking-related metrics that determine the time needed to transfer some data over the D2D link are also took into consideration.

Further, it has been defined a *social network inter-contact time* $T_{SNIC}$ to correctly determine the time before a user accesses a social media application in its device to receive information. This parameter is aimed at accounting for the realistic human behaviour when using a social media in the information dissemination process. This is particularly important to correctly evaluate the time required to forward the received content/information to a peer whom a user is in direct contact with. Intuitively, the $T_{SNIC}$ term defined above will result in a delay in the information diffusion process. This is why this time term is modelled by an exponential distribution, although other distributions may work as well to the scope.

In particular, in the reference problem the mean value of the exponential distribution characterizes the time (in seconds) that a user waits, on the average, before accessing to the received information over the social media used to disseminate the information. Only after this time the user can use the information and, when required, can forward it over a D2D link to another user it is in contact with.

### The Proposed D2D-enhanced Information Diffusion Scheme

The eNodeB is in charge to: define which unicast transmissions from the eNodeB to enable, identify the nodes acting (if any) as PBN or SBN, and determine the D2D transmission mode (unicast or multicast).

Let $\mathcal{M}$ be the served multicast group formed by $M$ users. Then, let $CQI_m$ $(m = 1, ..., M)$ be the channel quality feedback of the $m_{th}$ user which identifies the

corresponding maximum MCS level supported for UE $m$ [1]. Preliminarily, the eNodeB measures the *Channel Quality Indicator* (CQI) in downlink towards all the interested UEs in the cell. The eNodeB collects from each UE information about the CQI level on the D2D direct links with all its neighbours[2]. The packet scheduler computes (i.e., still on the eNodeB side) the radio resources allocated to the UEs, as if they were receiving the content separately on a unicast link according to the *Round Robin* scheduling policy. These resources are called as "virtual" because they will be actually allocated to any UE only if it will be served in unicast. Otherwise, when one or more nodes are served over a D2D link from a PBN, the "virtual" radio resources can actually be pooled together and allocated to the serving PBN to guarantee a better performing unicast link from the eNodeB to the PBN.

In addition, different terms used in the formulation of the diffusion time have been also introduced. For a generic user $i$ in the system, it is defined the unicast serving time as the combination of two terms: a first networking-related contribution dictated by the link quality that determines the time to receive some data over the LTE downlink direction, and a second social-related contribution related to the frequency a user accesses the social media where the received information is sent to, as defined by the $T_{SNIC_i}$ term. Given the content to be transmitted, the unicast serving time for a generic user $i$ as can be written as follow:

$$T_i^u = T_{b,i} + T_{SNIC_i} \tag{5.7}$$

where $T_{b,i}$ is the BS-to-user $i$ transmission time.

As for the local D2D communications, the weight $\omega_{i,j}$ associated to an edge in the network graph $G = (V, E)$ is introduced. Based on the CQI information for the D2D links in the network, $\omega_{i,j}$ can be computed as the time $T_{i,j}$ to transfer the content from $i$ to $j$ over the connecting D2D link, where the value for $T_{i,j}$ is determined by the channel quality, the transmission mode (i.e., unicast, multicast), and the available radio resources over the D2D link.

Based on the diffusion time information for the unicast mode and the weights associated to the D2D graph, the eNodeB evaluates the best content diffusion configuration to serve all the nodes; the main steps of the proposed algorithm are reported in Fig. 5.1 and detailed next:

• *Unicast and D2D transmission time estimation*: The eNodeB evaluates for each node $i$ the expected unicast serving time $T_i^u$ as in equation (5.7) and the diffusion time

[1] This is defined to successfully decode the received signal with a Bit Error Rate (BER) smaller than a predefined target value (in our case is 10%).

[2] It has been assume ideal channel feedback and do not study the impact of errors on the CQI estimation.

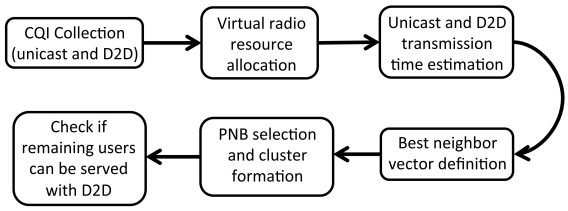



**Fig. 5.1.** Flow diagram for the proposed scheme.

over D2D links to all the remaining nodes $j$, $T_{i,j}$. At this stage, the radio resources considered over the corresponding D2D link are only those allocated to user $i$ on the unicast link from the eNodeB;

• *Best neighbor vector definition*: The eNodeB creates a vector, named *best neighbor vector*, where for each user $i$ the best neighbor $j$ is stored, that is the neighbor node offering the lowest value for $T_{j,i}$. Also eNodeB is considered as a potential "neighbor". Thus, each node will choose as its best serving node either the eNodeB (served over unicast transmissions) or any of its D2D neighbors (served over unicast D2D links).

• *PBN selection and cluster formation*: Based on the *best neighbor vector*, if a UE is the best neighbor of at least one UE in the graph, then it is selected as a PBN. If more than one UE is identified as potential PBN of the same set of nodes, then the one with the highest CQI value (unicast transmission from the BS) is selected. If a single PBN is selected by multiple UEs, a *multicast social cluster* (MSC) is formed, composed by the PBN and all the nodes for which the PBN is the best neighbor. The information the PBN receives from the BS is disseminated to the other nodes in the cluster through a multicast D2D transmission (in this case, the PBN uses the MCS corresponding to the worst CQI value). In this case a larger amount of radio resources can be used since the PBN exploits the pool of radio resources reserved to all the devices in the cluster both to receive data over a unicast link and for the multicast D2D transmission.

• *Update the best neighbor vector*: The UEs that have been clustered are deleted from the *best neighbor vector* and the algorithm is repeated until no more clusters can be formed. The remaining users can be connected through a D2D link with one of the members belonging to one of the formed MSCs. In this case, the cluster member that disseminates the information is identified as a *secondary bridge node* (SBN). However, this choice is performed only if the corresponding diffusion time for the interested UE is lower than the time required by a direct LTE unicast transmission.

• *Information dissemination*: If still some nodes are not being served, then the BS transmits the information to these users directly in unicast. Then the content dissemination can be started by the BS according to the selected configuration and transmission mode.

### 5.1.3 Performance Evaluation

A simulation campaign is conducted by using MATLAB® to evaluate the performance of the proposed algorithm in terms of (i) total information diffusion time, (ii) information diffusion time per UE, (iii) data-rate per UE, (iv) energy efficiency, and (v) Jain's fairness index [107] computed over the information diffusion time for the users. It has been considered a scenario with an LTE femtocell where the network wants to disseminate a context-aware alerting message to all mobile social users (radius of 500m). The amount of bandwidth resources is set to 50 RBs (10 MHz) and the considered range of the D2D communication is of about 100m [46]. The packet scheduler used to manage the radio spectrum is Round Robin and the number of users is fixed to 100 whereas the packet size 30MB. The main simulation system parameters can be found in Table 5.1.

In particular, the proposed solution has been compared with two alternative approaches available in literature that exploit either multicast and D2D transmissions. The two approaches considered are named: (i) Conventional Multicast Scheme (CMS) [108], and (ii) D2D-enhanced CMS ($D^2CMS$) [68].

**Table 5.1.** Main Simulation Parameters

| Parameter | Value |
|---|---|
| Cell radius | 500 m |
| D2D radius | 100 m |
| eNodeB Tx power | 46 dBm |
| Cellular UE Tx power | 23 dBm |
| D2D UE Tx power | 23 dBm |
| Noise power | -174 dBm/Hz |
| Path loss (cell link) | 128.1 + 37.6 log(d), d[km] |
| Path loss (D2D link) | 16.9 log(d) + 46.8, d[m] |
| Shadowing standard deviation | 10 dB (cell mode); 12 dB (D2D mode) |
| BLER target | 10% |
| # of runs | 1000 |

The first shown results illustrate the total diffusion time achieved to disseminate the content to all the considered users and the average diffusion time per single UE (Fig. 5.2 and Fig. 5.3). In both plots, the proposed solution is outperforming all alternative solutions. As expected, the worst performing solution is CMS, due to the conservative approach to accommodate the transmission to the worst case, whereas the $D^2CMS$ approach slightly improves the performance thanks to the use of D2D links to serve the users with the worst channel conditions. The gain achieved by the

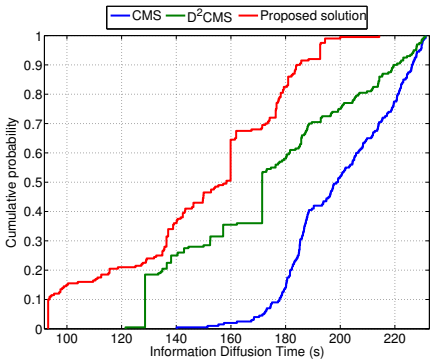



**Fig. 5.2.** Total information diffusion time.

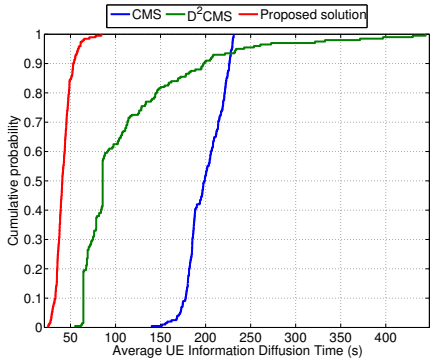



**Fig. 5.3.** Average UEs information diffusion time.

proposed algorithm w.r.t. the alternative solutions is of up to 50% with respect to the CMS approach and a gain up to 40% with respect to the $D^2CMS$ approach.

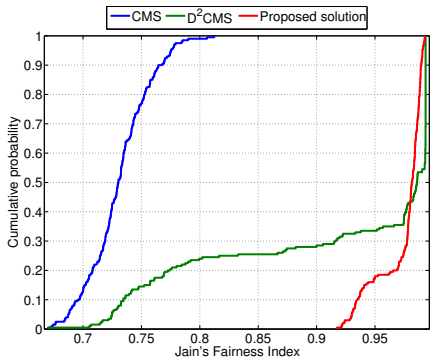



**Fig. 5.4.** Jain's fairness index.

The subsequent analysis addresses the Jain's fairness index for the UEs in the multicast group as shown in Fig. 5.4. This metric is generally in the range of $[0-1]$, where the value of 1 corresponds to all users being served with the same information

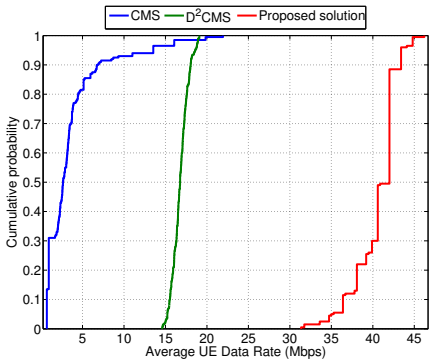



**Fig. 5.5.** Average UE data rate.

diffusion time (maximum fairness). The resulting fairness distribution for the proposed approach almost always outperforms the alternative solutions. The low value of Jain's index for the CMS solution is related to the social network inter-contact time which may be very variable over the set of all multicast users. In the other cases, this term has a lower impact as not all the users are served over the social media they are connected to (directly from the eNodeB), but a subset of users is served over a direct D2D link. When considering the $D^2CMS$ scheme, it is possible to observe that in about 45% of the cases the proposed approach performs way better (i.e., 27% gain on average), whereas for the remaining cases the behaviour is very similar, with the $D^2CMS$ solution offering a slightly better Jain's index (about 1.5% better).

Further, the focus was on the average data-rate per UE. As plotted in Fig. 5.5, the data-rate achieved by a single mobile social user for the proposed algorithm is higher w.r.t. the other solutions. This result is the direct consequence of our algorithm objective of selecting the most suitable transmission mode (i.e., unicast, multicast, D2D) for each user. In fact, on the network side, the data-content is disseminated at high transmission rates even if social metrics influence the entire process.

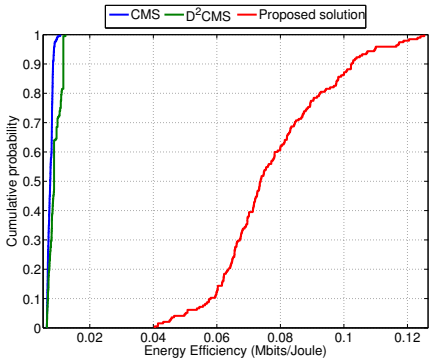



**Fig. 5.6.** Energy efficiency.

Finally, an energy efficiency analysis for the problem investigated is shown in Fig. 5.6. Also in this case the proposed solution overcomes the CMS and the $D^2CMS$ solutions. The reason for this is that our solution efficiently exploits the available transmission modes (i.e., multicast, unicast, and D2D) and, at the same time, drastically decreases the overall information diffusion time and the average diffusion time per UE. This translates in higher data rates and, as a consequence, in an energy efficient content dissemination.

## 5.2 Security-Centric Framework for D2D Connectivity Based on Social Proximity

Device-to-device (D2D) communication is one of the most promising innovations in the next-generation wireless ecosystem, which improves the degrees of spatial reuse and creates novel social opportunities for users in proximity. As standardization behind network-assisted D2D technology takes shape, it becomes clear that security of direct connectivity is one of the key concerns on the way to its ultimate user adoption. This is especially true when a personal user cluster (that is, a smartphone and associated wearable devices) does not have a reliable connection to the cellular infrastructure.

### 5.2.1 Considered system model

The target scenario is a set of wearable devices and each of these has a wireless connection via a certain radio technology to a more powerful aggregating device. Further, the user smartphone is assumed as the said aggregator that transmits data from wearable devices to the application server in the operator's network [109]. Practically, the mobile smartphone in question may have a number of radio interfaces, including short-range (e.g., BLE, WiFi) and cellular (LTE). In addition, this device is assumed to have a possibility to connect directly to another smartphone over a D2D link. In other words, the second level of abstraction is considered – a type of an *ad hoc* network topology between user mobile phones. Finally, at the highest level of abstraction, there is an infrastructure-based cellular network with all the smartphones connected to it. Detailed overview of the considered architecture may be found in [14].

A mobile smartphone with its associated wearable devices is named as a body area network or a user personal cloud. To this end, user devices belonging to an individual person are assumed to all be trusted nodes. The data circulating between wearables may then be forwarded over the mobile phone's cellular link to the operator's network and further on to the corresponding application cloud. However, no restrictions is yielded on the specific locations of users and some of them might end up being

out of cellular coverage. In case of unreliable cellular connection, the needed data can be relayed by other proximate users, whereas the users themselves may move around according to a certain mobility model. It is important to note that in the envisioned scenario the smartphone represents the bottleneck in providing connectivity to the body area network (or user personal cloud). The devices forming the body area network typically have very short-range connectivity (e.g., Bluetooth low-energy) and connect to the Internet through a gateway node, such as the user smartphone in our case.

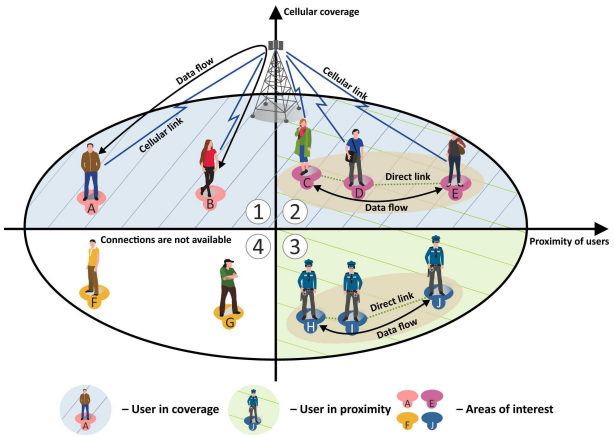



**Fig. 5.7.** Available D2D system operation modes.

Let us then concentrate on an arbitrary collection of proximate users in our network (i.e., a cluster). Depending on its location, there could be a number of special cases of interest, see Fig. 5.7. First, the cluster could be fully under the coverage of a cellular BS and conventional information security procedures may be employed to protect data transmitted over the cellular connection to the infrastructure network. In more detail, the first case in Fig. 5.7 suggests that both security procedures and data flows travel through the base station (BS), while for the second case only security procedures are enabled by the BS (data is exchanged directly between smartphones). In the third case, both security procedures and data flows utilize a direct link among users. Although the proposed framework is designed to embrace all the discussed use cases, the last of the three is of particular interest as it has not been addressed comprehensively in past literature. Enabling proximate users to not only communicate directly in a secure fashion, but also validate their data exchange as they leave and return under the cellular coverage, has been one of the main targets of this present research. As a last possible case, the cluster could be fully out of the cellular network's coverage. In this case, existing ad hoc specific solutions may be utilized to provide continuous secure connectivity for users over their direct links. However, according to the network-assisted D2D concept in beyond 4G systems, the management of the

direct link initialization, operation, and destruction is orchestrated by the cellular infrastructure.

Within our proposed framework, depending on the specific application running on top of user personal networks, the resulting clusters are based on two types of proximity-related parameters. First, there is *spatial proximity* of mobile users, which affects the optimal configuration of clusters with respect to wireless channel quality criteria. Optimizing this metric across all the mobile devices it may be possible to improve the data rate performance of the system. The other type of proximity is so-called *social proximity* of users. A mobile device can be aware of its previous contacts with other mobile users, or alternatively this information can be obtained from the contacts already stored on the smartphone. In what follows, it is shown how this information can be efficiently exploited to improve the performance of the security algorithm introduced later. To this end, the initial clustering of nodes is conducted by utilizing game-theoretic approaches – a subset of classical optimization theory – by efficiently exploiting both spatial and social notions of proximity.

Importantly, the proposed framework takes into account the effects of user mobility. The classical methods of optimization theory consider a *snapshot* of a network at a certain instant of time $t$ and then aim at developing practical algorithms for the optimized system operation with respect to a certain metric of interest. Clearly, such an approach cannot directly incorporate the mobility of users as it may cause significant deviations from the optimal solution at some other time $t + \Delta t$. However, enabling a particular mobility model and performing respective optimization at discrete instants of time, translates in implicitly capture the effects of mobility. Finally, the reason behind the use of game theoretic approaches in our mobile user environment is due to the complexity of keeping track of the past device behaviour resulting from the high dynamics in these networks [110]. In particular, coalitional game theory is applied to model the cooperative behaviour among network devices focusing on the payoff groups of devices, rather than individual devices.

**Game-theoretic clustering procedure**

The selection of a preferred cluster configuration is modelled by following the assumption and procedures described in details in Section 3.2.2 of Chapter 3. Specifically for this research, the players are user smartphones forming a cluster. The game is given in its characteristic form, as the achievable utility within a coalition only depends on the players forming the coalition and not on other players in the network. The objective for the players is to maximize the value of the coalition that is defined as the degree of geographical proximity and social relationship for the formed cluster. Hence, the coalitional game is an NTU game, since this value cannot be arbitrarily

apportioned among players. We define $\mathcal{V} : \mathcal{S} \rightarrow \mathbb{R}^{|\mathcal{S}|}$, such that $\mathcal{V}(\emptyset) = \emptyset$, and for any coalition $\mathcal{S} \subseteq \mathcal{N} \neq \emptyset$ it is a singleton set $\mathcal{V}(\mathcal{S}) = \{\mathbf{v}(\mathcal{S}) \in \mathbb{R}^{|\mathcal{S}|}\}$, where each element of the vector $\mathbf{v}(\mathcal{S})$ is the value $v_i(\mathcal{S})$ associated with each player $i \in \mathcal{S}$. The latter is defined as:

$$v_i(\mathcal{S}) = \frac{\sum_{j=1}^{|\mathcal{S}|} s_{i,j} \cdot d_{i,j}}{|\mathcal{S}|}, \tag{5.8}$$

where $s_{i,j} \rightarrow [0, 1]$ is an asymmetric function (i.e., $s_{i,j} \neq s_{j,i}$) measuring the social relationship or the degree of *friendship* between two devices. In particular, $s_{i,i}$ is a measure of the willingness of a device to acquire the content over a D2D link from a "friend" instead of directly downloading it from the cellular BS. The second term $d_{i,j}$ is a binary function taking the value of 0 whenever the devices $i$ and $j$ are not in proximity, and the value of 1 otherwise (we set $d_{i,i} = 1$ by construction). The result of the product of these two functions is averaged across the number of players in a given coalition $\mathcal{S}$, which always results in a value within the range $[0, 1]$.

It is possible also to define the value $v(\mathcal{S})$ associated to a coalition $\mathcal{S}$ as the average spatial and social proximity strength of the devices in a cluster:

$$v(\mathcal{S}) = \frac{\sum_{i=1}^{|\mathcal{S}|} v_i(\mathcal{S})}{|\mathcal{S}|}. \tag{5.9}$$

In particular, a value $v(\mathcal{S}) = 1$ is obtained when all the devices are within mutual D2D coverage and have the maximum degree of "friendship", so that they are all willing to acquire their desired content from a D2D partner. This seldom happens in larger coalitions, hence smaller independent coalitions are typically formed. Consequently, the proposed approach is modelled after a coalition formation game, with the aim of revealing the network's coalitional structure.

Further, it is assumed that all considered devices are rational and autonomous, which substantiates the design of an iterative algorithm to form the network coalition structure that improves both spatial and social proximity of the formed clusters. With respect to alternative scenarios illustrated in Fig. 5.7, the coalition formation algorithm may be implemented either in a *centralized* or a *distributed* manner. In particular, for *study case 2* represented in Fig. 5.7, the algorithm will be implemented by the BS (i.e., centralized approach), whereas in *study case 3* the involved devices implement the proposed algorithm autonomously and then synchronize over time by using the beaconing messages to obtain the up-to-date information (i.e., distributed approach). Another alternative for this latter case may become available when at least one of the involved devices is under the network coverage. In such a case, the BS may still be in charge of the solution implementation, whereas the node under coverage

acts as a signalling gateway to the other nodes. However, this latter option may cause some additional signalling overhead.

### 5.2.2 Information security considerations

#### Securing D2D communication

When designing the security proposed solution, it was assumed that the cellular network coverage is imperfect and sometimes users can face situations of unreliable cellular connectivity due to natural obstacles, tunnels, planes or other issues. However, while using proximity-based services, such as games, file sharing, and data exchange, the users are assumed to have continuous support for those applications over a secure channel. In order to understand what kind of new functionality is needed for the discussed security procedures, consider the connectivity cases demonstrated in Fig. 5.7 in more detail. All of the possible scenarios that may appear in a network-assisted D2D system can in principle be reduced to the four cases discussed below.

- *Case 1.* Here, users $A$ and $B$ grouped together have already established their own secure group (i.e., *coalition*) based on their area of interest and are using the cellular connection to the operator's network, the application server, and the PKI. The coalition secret has already been generated at the server side, and the users have all received the corresponding credentials and certificates of each other – they remain connected to the cellular network that orchestrates their data exchange. As a result, the data flows are running over cellular links due to the absence of proximity between the devices.
- *Case 2.* Here, the focus is on another set of devices consisting of $C$ and $D$, as well as $E$ that all have already established a coalition. Then, a *heavy* data flow may be running on the direct link between the devices that does not affect the cellular network capacity. All the needed information security procedures for the coalition establishment and key exchange are performed similarly to *Case 1.*
- *Case 3.* In this case, the coalition does not have an active connection to the cellular network. Hence, all the required key generation and distribution procedures are conducted over the direct D2D connections, by contrast to the previous cases. These procedures require higher involvement of the participating devices. The coalition secret is kept unchanged until the tagged group of the devices regains cellular network coverage.
- *Case 4.* In this case, the users are neither in the cellular coverage nor have a possibility to communicate directly. As a result, no security algorithm needs to be executed and users are waiting for the cellular coverage or direct connection to (re)appear.

**Proposed information security procedures**

For the purposes of the security protocol, is assumed that the cellular network is a trusted authority (TA) that is responsible for the root certificate generation and validation. Moreover, cellular operators are assumed to be responsible for security, anonymity, and privacy aspects of their users. Each user device thus obtains its own certificate signed by TA as soon as it connects to the cellular network for the first time. This step is required to ensure the validity of other users and prevent from the subsequent person-in-the-middle types of attacks on the direct link. The classification of the users is based on their cellular connection availability as well as the fact of their association to a certain secure group: a *light* device has an active, reliable cellular connection; a *dark* device does not have a reliable cellular connection, but used to have it in the past; a *blank* device is that wishing to join the coalition for the first time. In what follows, we address the crucial procedures of coalition initialization and formation.

The procedure of *coalition initialization* may only be executed when connected to the TA, i.e., having a reliable cellular connection. Accordingly, when the $i^{th}$ user receives its initial certificate ($PK_i$) signed by the root certificate ($PK_{TA}, N_{TA}$) and is supplied with a unique device identifier, the corresponding secret ($SK_i$) is generated on the user side. If a group of *light* users is willing to create/initialize a coalition, one of the devices is sending a request to the TA over its cellular link. The request contains the set of device identifiers to be grouped. When the request is processed, a unicast polling procedure is initialized, that is, all of the devices are contacted as to whether they would like to join the coalition. Then, cellular network proceeds with the initial setup of the coalition based on the received responses and according to classical PKI mechanisms. For each initialized secure group, its own coalition certificate ($PK_c, PK_{TA}$) is generated with the corresponding signature by each device's certificate in the group ($PK_i, PK_c$). After these initial steps, secure direct communication becomes possible over any IP-ready network. However, the above coalition establishment procedure may only be executed when all of the devices have reliable cellular connectivity due to the protocol constraints.

After the secure coalition has been established, users need not rely on continuous cellular connectivity and may communicate directly over a secure channel even if the cellular link becomes unavailable. However, this type of connectivity can be significantly augmented by offering a possibility to include new users and exclude existing ones from the tagged coalition. Such scenarios may appear in both considered cases: (i) when all the users are *light* – they have cellular connectivity and (ii) when at least one user is *dark* – does not have a reliable cellular connection. These cases correspond to two distinct network operation modes (namely, infrastructure and ad hoc), and the

respective security enablers for both of them need to be different. The information security procedures for these two scenarios are described as follows.

•*Reliable cellular connectivity.* First,is described how the initialization of the coalition is performed. All of the devices have a pre-generated set of parameters after their initial network entry: (i) own secret $SK_i$, (ii) own certificate signed by the TA certificate $PK_i, PK_{TA}$, and (iii) own unique identifier $ID_i$. Further, after the TA polls the involved devices and receives a list of users to be grouped, it generates a polynomial $f(x) = a_{k-1}x^{k-1} + a_{k-2}x^{k-2} + ... + a_1x + SK_c, f(0) = SK_c$, where $k$ is a threshold value calculated based on the number of devices in the planned coalition, $x_i$ is the device identifier, and $a_i$ is the corresponding device coefficient. Therefore, the RSA-like certificate component for the $j^{th}$ device is calculated as $cert_j = \overline{PK_i}^{f(0)} mod\, N_c$, where $\overline{PK_i}$ is generated by the device, $f(0)$ is the coalition secret, and $N_c$ is generated at the coalition initialization stage as well. Finally, all the certificates are distributed to the devices, and the algorithm proceeds to the phase of direct communication.

•*Unreliable cellular connectivity.* Focusing on the worst-case scenario, when none of the devices have an active cellular connection, the users should rely only on the coalition itself, when admitting an additional user. To solve this issue, a dedicated parameter included into the coalition certificate $PK_c$ has been employed, which is a threshold value of $k$ that characterizes the number of devices in coalition needed to collectively allow for a new device to join in. The value of $k$ is first set at the coalition initialization stage and may then be altered based on the number of involved devices $n$. Originally, for each coalition, the TA generates a Lagrange polynomial sequence with $k$ coefficients and a coalition secret share $SK_c$ stored at the cellular network side. Note that for the considered ad hoc scenario, a modification of the polynomial and its associated secret is not possible. Therefore, a group of devices forming the existing coalition should convene together and reconstruct $SK_c$ (without disclosing it) in order to admit the new device. Clearly, the same procedure executed without cellular network assistance would cause users to exchange excessive amounts of signalling messages in addition to running computationally intensive information security primitives. On the other hand, with the proposed procedure, secure direct connectivity enjoys higher flexibility and has lower overhead.

### 5.2.3 System-level performance evaluation

In this section, are evaluated the performance of the proposed framework. In particular, a large-scale system-level simulator built during the Ph.D period has been employed in order to yield numerical conclusions on the operation of the complete system.

First, a simulation-based campaign has been conducted using the WINTERsim tool available in [111]. The reference scenario consists of a 3GPP LTE BS (termed eNodeB) with the radius of 100m, where users are uniformly distributed within its coverage in the range $[10, 100]$. The movements of the users are characterized by a *Levy Flight* mobility model with an $\alpha$-value equal to 1.5 and the user speed varying in the range $[0.2, 2.0]$m/s. The reason for choosing the Levy Flight mobility model is that recent investigations reveal that movement of people may follow characteristic patterns, where numerous short runs are interchanged with occasional long-distance travels [112, 98, 113]. The parameter $\alpha$ allows adjusting the form of the step-size distributions.

Importantly, in the reference scenario the connection between the smartphone and the devices within the user personal cloud is assumed to be trusted and stable. In particular, with the simulation-based evaluation the focus is on the smartphone which represents a bottleneck for providing stable and secure communication to the entire personal cloud (wearables). Indeed, whenever the cellular connection becomes unavailable (unreliable), the proposed solution is able to offer a connection also to the device that is not in network coverage when in proximity to another device.

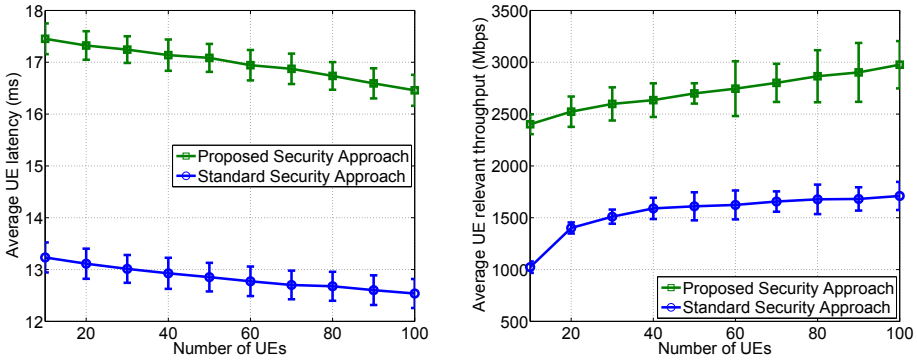
The simulation environment thus translates into a typical pedestrian scenario, as standardized in the 3GPP specification TS 36.304 (see Section 5.2.4.3 therein). In addition, the multimedia traffic within the considered scenario is modelled after a video download application with relatively long inter-arrival time and the packet size of 100MB. The main system parameters are summarized in Table 5.2. The two performance metrics that investigated are: *user latency*, that is, the end-to-end delay to download the multimedia content, *average user relevant throughput*, that is, the throughput achieved by the UE when it downloads the desired content either over the LTE or the WiFi-Direct link, and *blocking probability*, that is, the number of interruptions experienced by the user during a download session. The conventional network operation is compared against the security-centric approach outlined in Section 5.2.2.

First, consider the effects of user mobility on the average latency in the proposed framework (see Fig. 5.8(a) and Fig. 5.9(a)). As is observed, the latency decreases linearly with the growing intensity of mobility either by varying the number of users or the mobility intensity. The reason is that the increase in the user speed translates into higher number of contacts among them. This way, users can download the content over the WiFi-Direct link with higher data rates. However, the conventional security approach performs better compared to the proposed solution. This is due to the fact that the security scheme introduces an additional delay when users are in proximity (can establish a direct D2D connection), but not under the network coverage, i.e., *Case 3* in Fig. 5.7. This effect is particularly visible when the number of users is

**Table 5.2.** The main simulation parameters.

| Parameter | Value |
|---|---|
| Cell radius | 100 m |
| Maximum D2D range | 30 m |
| # of users | 20 |
| Target data rate on LTE link | 10 Mbps |
| Target data rate on D2D link | 40 Mbps |
| eNodeB Tx power | 46 dBm |
| UE Tx power | 23 dBm |
| D2D link setup | 1 s |
| Cellular bandwidth | 5 MHz |
| Mobility model | Levy Flight |
| Simulation time | 15 min |
| Number of simulation runs | 300 |

high (i.e., 100), because the opportunities to establish direct connections become more abundant. However, the advantage of using the proposed approach is in that, generally, conventional systems are unable to provide any type of secure connectivity when there is a lack of cellular coverage.

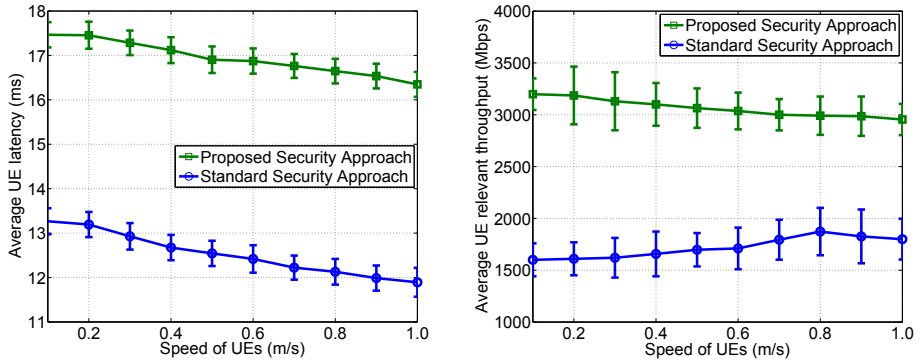



(a) Average user latency (b) Average user relevant throughput

**Fig. 5.8.** Latency and throughput for varying number of UEs (speed is 1 m/s).

The average throughput experienced by the users as a function of the number of UEs and their mobility intensity is shown in Fig. 5.8(b) and Fig. 5.9(b). It is important to note that the proposed security algorithm demonstrates better performance compared to the conventional solution. The reason is that the proposed approach delivers connectivity to users that are in a D2D transmission range, but not under cellular coverage, *Case 3* in Fig. 5.7. In this case, the *extra* throughput is obtained at a cost of additional delay to establish a direct D2D connection and execute all the

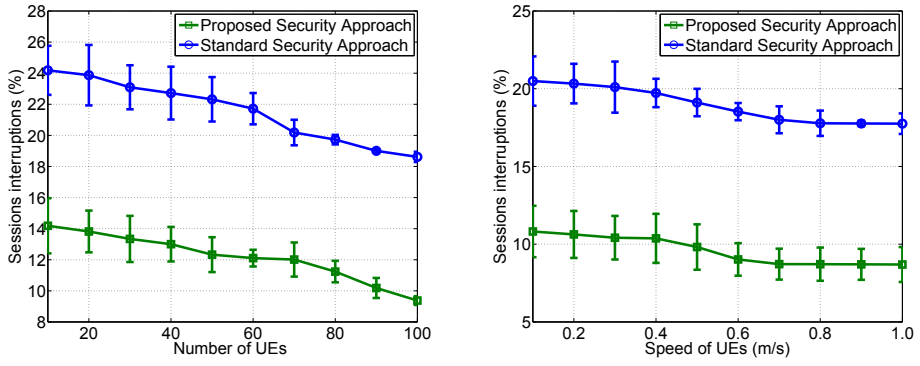
needed security procedures. The amount of the additional delay is due to execution of the security primitives that have to be run among the D2D users.




(a) Average user latency (b) Average user relevant throughput

**Fig. 5.9.** Latency and throughput for varying UE speeds (number of UEs is 100).

Finally, the blocking probability as a function of the number of interrupted download sessions experienced by the users is summarized in Fig. 5.10(a) and Fig. 5.10(b). As is possible to learn from the plots, the proposed security approach performs better compared to the conventional security solution. The explanation is again that the proposed framework is able to guarantee connectivity even if the users are not under network coverage (i.e., *Case 3* in Fig. 5.7). As a consequence, at the cost of extra delay, the users enjoy longer download sessions and increase their chances to obtain the desired multimedia content.




(a) Blocking probability for varying the number of UEs (b) Blocking probability for varying the speed of UEs

**Fig. 5.10.** Blocking probability.

## 5.3 Towards Trusted, Social-Aware D2D Connectivity

### 5.3.1 Bridging Across Technology and Sociality

It is evident that sociality has the potential to become a core incentive across a wide range of applications and services wherein D2D communications may demonstrate non-incremental benefits. However, the social domain should not be considered as a standalone enabling factor for proximate connectivity (see Fig. 5.11). By contrast, it needs to carefully match the respective technology constraints and features of the physical communications domain (such as the utilized spectrum, radio technology, battery/power resources, etc.).

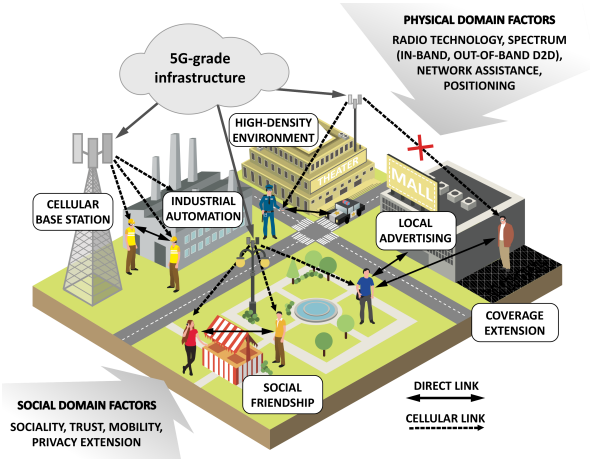



**Fig. 5.11.** Urban network-assisted D2D applications.

In this regard, the overall vision is in that not only human users and their social interactions are to be accounted for, but also the associated interactions between the user devices with their specific notion of sociality. This expectation is well supported by the recent research developments within the IoT community, which target to embrace the social networking concepts [114] to build trustworthy relationships among the devices [115]. In the present research line (see Table 5.3), we thus consider the two distinct types of sociality as described below.

- *User-driven sociality*: in this case, humans are willing to interact and are directly controlling their social activities. The degree of how much two users are interested in exchanging data is characterized by a so-called *Human Social Relationship* (HSR) factor, which may be linked to a social media tie, a family tie, etc. This measure is directly related to the level of familiarity and trust, according to which friends, relatives, or colleagues are likely to connect and share their content more frequently than the unfamiliar users. Within the same class of sociality, it is also possible to consider the relationships based on the *market pricing relational*

**Table 5.3.** Social relationship factors between devices, possible applications, and the associated trust value.

| Relationship | Typology | Description | Applications | Trust value |
|---|---|---|---|---|
| Human social relationship (HSR) | User-driven | Familiarity degree with friends/relatives/colleagues | Leisure applications, confidential data, eHealth, mission-critical communication | [0-1] |
| Market pricing relationship (MPR) | User-driven | Cooperative interactions with services triggered by the environment | Proximate marketing, proximity gaming, advertising | 0.2 |
| Ownership object relationship (OOR) | Device-driven | Relationship between objects owned by the same person | Personal cloud, smart home | 1 |
| Co-location object relationship (C-LOR) | Device-driven | Objects sharing personal experiences (e.g., cohabitation) | Information/data exchange at social aggregation points (concerts, sport events) | 0.8 |
| Co-work object relationship (C-WOR) | Device-driven | Objects sharing public experiences (e.g., work) | Information/data exchange at work aggregation points (e.g., fairs, workshops) | 0.6 |

(MPR) model. The founding principle behind the MPR model is proportionality, as well as knowledge of how the relevant interactions are organized with respect to a common scale of values. In other words, the relationships established among people are driven by their willingness to interact or cooperate only in the light of achieving mutual benefits. In the literature, there are several examples that focus on smart surrounding scenarios for context-aware applications. For instance, triggers from the environment may invite and motivate people to socialize and/or cooperate, and thus take advantage of services within coverage (proximity market, gaming, advertising, etc.).

- *Device-driven sociality*: in this case, devices may autonomously interact according to the specific rules present by the device owners or manufacturers – without an explicit user intervention during such interaction. Social relationships among the device owners are not necessarily required to foster this type of cooperation. To construct this sociality level, mobility patterns and relevant context can be considered to configure the appropriate forms of socialization [114]. Among these, the so-called *co-location object relationships* (C-LOR) and *co-work object relationships* (C-WOR) are established between devices in a similar manner as among humans, when they share personal (e.g., cohabitation) or public (e.g., work) experiences. Another type of relationships may be defined for the objects owned by a single user, which is named *ownership object relationship* (OOR) and may be of interest, for instance, when a number of devices belong to the same personal cloud.

Bridging across the realm of social-awareness and real world D2D-based implementations, a factor of particular importance is dual mobility of the communicating entities. D2D application developers need to extend support for trust and confidence management to ultimately enable secure proximate communications that are aware of unrestricted human/device mobility. In this regard, the most challenging use cases are those, in which the out-of-coverage cellular devices are also becoming involved

into the network-assisted D2D data exchange in the absence of a reliable link to the central trusted authority (residing e.g., in the operator cloud). In order to effectively address this and other aforementioned scenarios, the present research investigates how human- and device-centric social relationships can achieve trusted connectivity in relevant D2D groups under realistic mobility as well as, possibly, intermittent cellular network coverage. In particular, the focus was on three insightful study cases:

- *Trust-based human applications (Case A).* Interactions among humans with tight trust requirements are included here. In these cases, the end-user is willing to reliably know which person the data are exchanged with. To this end, user-driven sociality is of paramount importance and sometimes even becomes the only acceptable enabler. Examples of such applications are found in work-related environments, such as construction sites as well as transport and cargo handling facilities in harbours or airports, where stringent safety regulations dictate increased levels of trust. Other applications may include confidential and mission-critical data collection, such as that for eHealth and safety applications.
- *Leisure and entertainment applications (Case B).* Connectivity between proximate devices supports applications for users at leisure, such as entertainment and gaming, non-confidential information sharing, and similar non-critical services (e.g., map sharing for intelligent transportation systems). These applications do not necessarily need an explicit social relationship between the device owners, and trusted communications may rather be driven by the sociality of devices. Typical scenarios of interest in this category may consider users distributed in a certain area and sharing similar interests, such as content dissemination in a stadium, a university campus, or a pub, where matching people (in terms of interests, age, familiarity, etc.) interact by employing their devices.
- *Critical machine-to-machine (M2M) applications (Case C).* In the situations where, by definition, there is no (or, very limited) human intervention, automated device connectivity may still benefit from some form of social awareness. One may consider hazardous working environments, such as those often met in industrial automation scenarios, where large numbers of machines, sensors, actuators, or robots communicate mission-critical data. To facilitate such information exchange, trust can be delivered by operator-enforced incentives and policies, leading to optimized communications performance with higher degrees of security.

### 5.3.2 Social-Aware Framework for Trusted D2D

The proposed social-aware framework aims at enabling trusted D2D-centric data delivery for proximate users in mobile environments. In these situations, direct links

may (temporarily) extend or substitute cellular network connections, when the operator services become unavailable to (some of) the customers. Relevant clustering of the D2D devices can be conveniently modelled as a non-transferable utility (NTU) coalitional game $(\mathcal{N}, \mathcal{V})$, where $\mathcal{N}$ is a set of $N$ players and $\mathcal{V}$ is a function, such that for every coalition $\mathcal{S} \subseteq \mathcal{N}$, $\mathcal{V}(\mathcal{S})$ is a closed convex subset of $\mathbb{R}^{|\mathcal{S}|}$. The latter contains the payoff vectors that the players in $\mathcal{S}$ can achieve, and $|\mathcal{S}|$ is the number of members in the coalition $\mathcal{S}$. The objective for the players in this NTU game is to maximize the value of the coalition they belong to. In the proposed framework, the utility for a coalition is defined as the degree of proximity and the strength of social relationships for the corresponding D2D-based cluster. To this aim, we define an NTU game, where for any coalition $\mathcal{S} \subseteq \mathcal{N}$ the value $v_i(\mathcal{S})$ associated with each player $i \in \mathcal{S}$ is determined as:

$$v_i(\mathcal{S}) = \sum_{j=1}^{|\mathcal{S}|} s_{i,j} \cdot p_{i,j} / |\mathcal{S}|, \tag{5.10}$$

where $s_{i,j} \rightarrow [0,1]$ is a function measuring the level of social relationships (or *friendship*) between a pair of communicating entities, whereas the second term $p_{i,j}$ is a binary function taking the value of 0 if the users $i$ and $j$ are not in proximity, and taking the value of 1 otherwise (by construction, we set $p_{i,i} = 1$). The resulting product of these two functions is then averaged across the players in a given coalition $\mathcal{S}$, thus always yielding a value within the range of $[0,1]$.

The actual definition of the social relationship level between the devices $s_{i,j}$ needs to allow for appropriate weighting of the contributions coming from human relationships and device sociality. Therefore, it may be defined as a weighted function $s_{i,j} = \alpha \cdot H_{i,j} + (1-\alpha) \cdot D_{i,j}$, where $H_{i,j} \in [0,1]$ is the degree of human-to-human sociality and $D_{i,j} \in [0,1]$ is the degree of device-to-device sociality. The social relationships between humans and devices are modelled based on the values shown in Table 5.3. In such a Table, the typology field identifies to which class the social relationship refers to. A "User driven" typology stands for a relationship that is being used to determine the value of $H_{i,j}$; the *HSR* and *MPR* relationships belong to this class. A "Device driven" typology instead, stands for a relationship that is used to determine the value of $D_{i,j}$ and are identified with the *OOR*, *C-LOR*, and *C-WOR* relationships. Whenever two entities can be associated to more types of relationships of the same class, the strongest tie having the highest value is selected. The motivation for this is that a stronger social relationship leads to higher probability of "trusted" connection thus providing improved performance.

Further, in the model the weighting term $\alpha \in [0,1]$ has been introduced to adjust the influence of the two contributions described above according to a specific appli-

cation and scenario. Thus, the role played by alpha is to enrich the model with a weighting factor that may be tuned according to the scenario. Specifically, a value equal to "1" to $\alpha$ is assigned when considering the CASE A (i.e., trust-based human scenario) and a value equal to "0" for CASE C (i.e., critical machine-to-machine scenario). In particular, the CASE A and CASE B scenarios discussed in this research represent two illustrative examples of the extreme cases with only human-driven sociality and only device-driven sociality.

The third scenario investigated in the paper refers to applications for users at leisure where both human- and device-driven types of sociality are considered (i.e., study case B). In such a case, the importance of the human-driven and device-driven sociality is the same and for this reason it has been chosen to assign the same influence in the evaluation of the $s_{i,j}$ term. Nonetheless, other values of alpha may be more appropriate based on the scenario taken into consideration and the application considered. However, a thorough analysis of all possible scenarios is out of the scope of this research, where the aim is to propose a model that allows to explore how the human social awareness and the D2D-enabled proximate connectivity may interact to improve the resulting communications performance and service quality.

Now, it is possible to define the value of $v(\mathcal{S})$ for a coalition $\mathcal{S}$ as the average degree of proximity and strength of social relationships for the users in the cluster: $v(\mathcal{S}) = \sum_{i=1}^{|\mathcal{S}|} v_i(\mathcal{S})/|\mathcal{S}|$. Importantly, the highest possible value associated with a certain coalition $v(\mathcal{S}) = 1$ is achieved if all of the devices are located in their mutual D2D coverage, as well as all of them enjoy the maximum level of friendship. In practice, the latter seldom happens in the *grand coalition* incorporating all the networked devices, and thus independent and disjoint coalitions are typically formed. To control the resulting stability problems, existing solutions proposed in recent literature can be adopted [15]. For instance, an iterative application of the merge and split rules enables the much needed convergence to a stable coalitional structure of the network.

Once stable D2D-clusters are formed, the D2D connectivity within them should be secured both in the cases of full and partial cellular coverage. Whenever connected reliably to the centralized network infrastructure, the D2D clusters can establish their information security rules by employing the conventional methods, hence relying on the operator infrastructure acting as a trusted authority. However, when cellular connection becomes unavailable, secure associations between D2D partners may benefit from solutions in [116] and [115], which enforce trustworthiness of human- and device-driven interactions, respectively.

### 5.3.3 Performance Evaluation Campaign

To validate the envisioned D2D framework and quantify the benefits of the proposed social-aware, secure clustering solution, a supportive system-level performance assessment has been conducted by utilizing our custom-made simulation environment, named WINTERsim [111]. Due to the need to model full-scale user mobility and application-level traffic, the underlying system-level evaluation methodology had to be streamlined, by simplifying the propagation and interference conditions, and thus employing the parameters summarized in Table 5.4. The output metrics of interest are aggregate effective throughput and corresponding device energy efficiency, as well as degree of connectivity, which indicates the proportion of users covered by cellular and/or direct links.

The reference scenario features a tagged cellular BS (running the contemporary 3GPP LTE technology) deployed within a [150m×150m] area of interest, and having the coverage range of 100m, resulting in around 70% of reliable cellular coverage available to the users. For the sake of completeness, also several alternative values for the LTE coverage range are considered – in order to understand the effects that it has on the degree of connectivity. Further, the communicating entities (humans and their connected devices) are allowed to freely move across the considered area of interest according to the characteristic "Levy flight" mobility pattern [112]. More specifically, the performance of a multimedia application with the packet size of 100 KB and the packet inter-arrival time of 10 s (e.g., video dissemination, e-health, etc.) are investigated. As for the D2D communications technology, discovery and connection setup functions are managed directly by the LTE BS with the appropriate network assistance protocols, whereas the actual direct data transmission is performed out-of-band (e.g., over WiFi-Direct links that can operate in parallel with LTE assistance, as they utilize the unlicensed spectrum).

The following alternative communications options are compared in the system-level study:

- *Cellular (LTE) solution.* A benchmark setup, where the connectivity is available only over the conventional cellular links, without any D2D-based transmission or coverage extension possibilities;
- *Simple D2D solution.* Only mobile devices under the reliable cellular network coverage may connect directly to form the D2D pairs according to the shortest distance between them. The BS is acting as the conventional trusted authority by guaranteeing trustworthy connectivity for all in-coverage D2D partners;
- *Advanced (social-aware) D2D solution.* Users may cluster together according to the proposed social-aware D2D framework. This may also happen under partial

cellular network coverage, thus leading to D2D-based coverage extension. All connectivity (including the out-of-coverage links) is made trusted by taking advantage of the distributed information-security solution without a central trusted authority [116]. To further visualize the effects of both human- and device-driven sociality, the three reference study cases and the associated $\alpha$ values as defined in Section 5.3.2 are considered: 1, 0.5, and 0 for study cases A, B, and C, respectively.

To ease further exposition, for the baseline LTE solution and the Simple D2D schemes it has been taken into account only the portion of data transmitted by the users within the reliable cellular network coverage (by aggregating these effective values across individual users). In case of the Advanced D2D solution, is also considered the traffic of the out-of-coverage users enabled by the proposed trusted, social-aware framework.

**Table 5.4.** Core simulation parameters.

| **Application parameter** | **Value** |
|---|---|
| Packet size | 100 KB |
| Inter-arrival time | 10 s |
| **System parameter** | **Value** |
| Cell radius | 100 m |
| Maximum D2D range | 30 m |
| WiFi-Direct target data rate | 40 Mbps |
| LTE target data rate | 10 Mbps |
| LTE BS Tx power | 46 dBm |
| UE Tx power | 23 dBm |
| Machine Tx power | 0 dBm |
| D2D link setup time | 1 s |
| Mobility model | Levy flight (with parameter 1.5 [112]) |
| Number of UEs | [10-100] |
| $H_{i,j}$ | [0-1] |
| $D_{i,j}$ | [0.6,0.8,1] |

First, Fig. 5.12 indicates the achievable aggregate effective throughput as a function of the number of networked devices. Hence, is noticeable that at all times the proposed *social-aware* D2D solution outperforms the LTE-only alternative considered in this study, as well as the Simple D2D solution. In particular, the case of $\alpha = 0$ (study case C, when only device-driven sociality is considered) achieves the best performance, followed by the cases when $\alpha = 0.5$ and $\alpha = 1$ (study case A, when only human-driven sociality is considered). This result suggests that the interactions based

on the second level of sociality – those accounting for the relationships between the devices – may introduce significant benefits to the system operation, whenever the trust requirements of a running application allow for this.

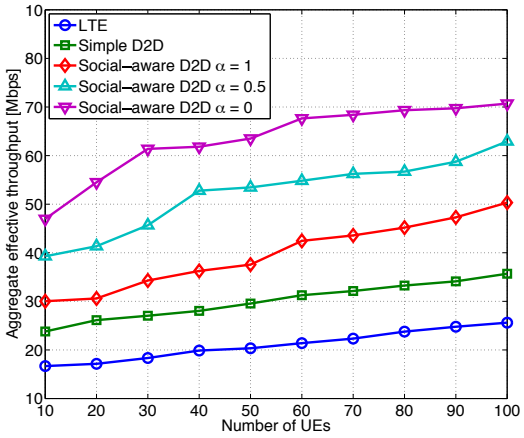



**Fig. 5.12.** Impact of social relationships on the system throughput.

Further, Fig. 5.13 illustrates the degrees of connectivity offered within the area of interest, when a varying percentage of such area is covered by the LTE cellular BS. In particular, the left subplot of Fig. 5.13 illustrates the overall number of users that are served in a given area of interest by considering both the *simple D2D* and the *social-aware D2D* solutions in study case B (where $\alpha = 0.5$). What we can observe from this subplot is that a higher percentage of users are served when we consider a social layer of awareness among the devices and humans. Moreover, is observed that this positive effect is higher for lower values of LTE coverage. Further, in the right subplot of the figure is reported on the amount of users that are served through a D2D connection. Clearly, this is a subset of the whole set of served users as it represents the portion of users that either prefer to establish a proximity-based link instead to download the content directly from the LTE eNodeB, or can be served only over D2D when there is no LTE coverage.

As noticed, when the available cellular coverage area is particularly small, in case of *simple D2D* solution the number of users that establish a D2D connection decreases drastically. This is given by the fact that users are in proximity of the BS and thus achieve a higher channel quality with respect to the one on the D2D link. On the contrary, the percentage of users served via D2D connections is three times higher for the proposed *social-aware D2D* solution. The explanation of this result is in that the proposed solution is able to provide connectivity also to those users that are outside of the cellular coverage (i.e., with D2D clusters). Note that this important outcome is achieved owing to the operation of our social-based, secure cluster formation scheme.

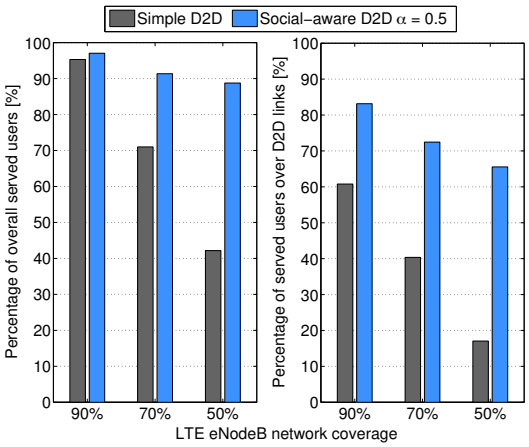



**Fig. 5.13.** Impact of LTE coverage on the degree of connectivity in the system.

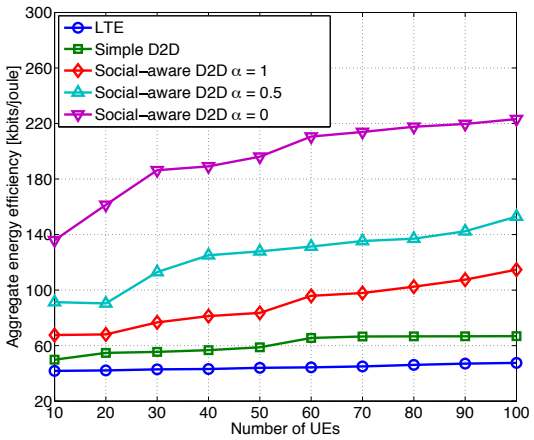



**Fig. 5.14.** Impact of social relationships on the user energy efficiency.

Finally, performance results for the aggregate energy efficiency of user data transmissions are reported in Fig. 5.14. This metric has been evaluated by taking into account the relevant transmission power for each network node (refer to the values reported in Table 5.4). Again, the *social-aware* D2D approach outperforms both the considered LTE and the Simple D2D alternatives. In particular, for the case of $\alpha = 0$ the proposed solution reaches its highest gain with respect to the benchmark LTE operation. This is due to lower transmit power of small-scale devices (i.e., connected machines) as compared to more power-hungry handheld UEs.

To conclude, the proposed analysis indicates that social ties among both humans and their devices impact the ultimate performance of the proposed social-aware scheme that forms the trusted D2D clusters. In particular, with higher levels of social relationships, the resulting effective throughput grows, also yielding positive effects on the energy consumption of the devices and their degrees of connectivity. The key reason is that having better social relationships plays in favour of having larger coalitions

between proximal humans/devices, even in the cases of intermittent cellular network coverage. Clearly, the improved throughput performance of the social-aware solution is achieved at the cost of somewhat increased latency, as compared to the Simple D2D solution. Indeed, to deliver reliable connectivity to proximate humans/devices, especially outside of LTE coverage, more time-consuming security procedures are required to be executed in the UE. For instance, handheld devices can execute the security methods from [116] within about 0.6 s, which leads to slightly higher latencies with the growing number of communicating entities. However, the implementation efficiency of said security mechanisms can be optimized further to reduce the computation time, which is left for subsequent researches.

# 6

# D2D in 5G Internet of Things

The proliferation of heterogeneous devices connected through large-scale networks is a clear sign that the vision of the Internet of Things (IoT) is getting closer to becoming a reality. Many researchers and experts in the field share the opinion that the next-to-come fifth generation (5G) cellular systems will be a strong boost for the IoT deployment. Device-to-Device (D2D) appears as a key communication paradigm to support heterogeneous objects interconnection and to guarantee important benefits. In this line of research, is thoroughly discussed the added-value features introduced by cellular/non-cellular D2D communications and its potential in efficiently fulfilling IoT requirements in 5G networks.

## 6.1 User-in-the-Loop: User Involvement in Multi-Connectivity 5G Scenarios

Wireless technology has already become a commodity in our society as a plethora of powerful companion devices facilitate novel user applications and services. In today's "human-intense" urban locations, people are faced with increasingly more heterogeneous connectivity options, which creates challenges for efficient decision-making to reap the maximum user benefits. On the other hand, service providers are struggling to augment the capacity of their network deployments quickly in response to unpredictable and sporadic traffic loading. In this work, is envisioned that *mobile* vehicles and flying robots may be equipped with high-rate radio access capabilities to better accommodate the varying space-time user demand. Additionally, various user-owned equipment may take a more active part in fifth-generation (5G) service provisioning by sharing wireless connectivity and content with relevant consumers in proximity. However, this emerging vision remains conditional on identifying adequate pragmatic sources of motivation for user involvement, which is aggravated by the unpredictable and heterogeneous mobility. Therefore the contribution here is with a novel mobility-centric analytical methodology for 5G multi-connectivity scenarios in

the context of truly mobile access (users, car- and drone-mounted small cells, etc.) and couple it with practical user incentivization considerations. The achieved findings show improvements in the degrees of availability and reliability of system-level wireless connectivity as well as help employ user-owned devices as an important asset in future networks to opportunistically shape the dynamic traffic demand.

### 6.1.1 Proposed multi-connectivity system model

As shown in Fig. 6.1, a multi-tier 5G-grade heterogeneous network is considered with $K+1, K \geq 1$, diverse connectivity *classes*, represented by a variety of operator-owned and user-owned entities, e.g., the baseline macro LTE cell, small cells, WiFi and mmWave access points, D2D relays, and highly mobile LTE relay cars and drones, operating in both unlicensed and licensed bands. The users willing to acquire some content may connect to any available association point (or a point designated by the network), given that a small cell or a user-owned device is incentivized appropriately to provide connectivity (or content) according to the needs of the proximate users. The notations used in the remainder of this Section are summarized in Table 6.1.

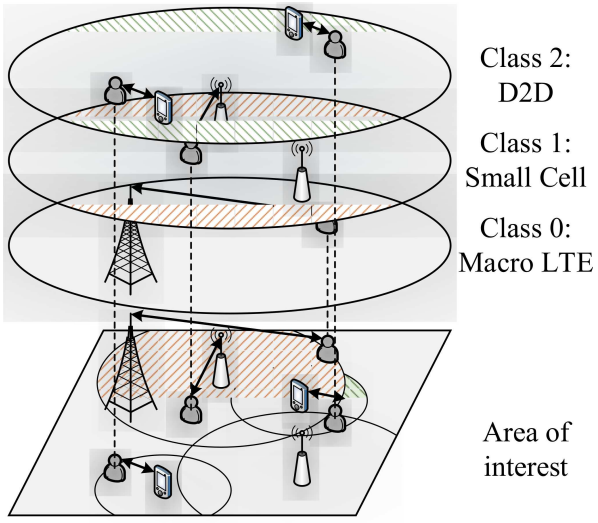



**Fig. 6.1.** Considered system model with $K+1$ classes of small cells.

#### Content supply

Different classes of devices are considered all of which are the content suppliers. Each class is defined as a collection of $N_i$, $i = 1, 2, \ldots, K$, entities logically grouped by a set of key parameters $< P_i, W_i, h_i, v_i, \tau_i >$, where $P_i$ and $W_i$ denotes the transmit power and bandwidth, $h_i$ is the height, $v_i$ is the speed, and $\tau_i$ describes the mobility of a class entity, i.e., $E[\tau_i]$ is the mean "run" time for the random direction model

(RDM). The $P_i$, $W_i$, and $h_i$ parameters are used to specify the content acquisition rate $B_i$, while $v_i$ and $\tau_i$ parametrize the mobility model.

The area of interest is denoted as an *arbitrary* convex shape $A$. It is assumed that the conventional macro LTE service is available everywhere across the area at any time, providing an aggregate content acquisition rate of $B_0$. The content acquisition rate is assumed to be equally divided between the users associated with the macro LTE service. Hence, the baseline LTE is considered as the default connectivity class 0, with a single association point defined as $< P_0, B_0, 0, 0, 0 >$.

The remaining $K$ connectivity classes (hereafter referred to as small-cells[1]) contain entities with similar characteristics: e.g., a class of LTE relay drones, a class of connected vehicles, or a class of D2D-capable devices, etc. The coverage radius of class-$i$ is calculated based on the height $h_i$, the propagation model, and a set of modulation and coding schemes (MCS) for the wireless technology used by class $i$, whereas the content acquisition rate of $B_i$ is equally divided among the associated users. The number of entities in class $i$ is fixed to $N_i$, while the total number of association points is $N = \sum_{i=1}^{K} N_i$. If $v_i = 0$.

**Content demand**

A space-time model is developed to capture the user dynamics. The users attempting to acquire some content arrive into the system according to a Poisson process with intensity $\lambda_U$. This assumption holds for dense urban scenarios as a consequence of superposition of point processes. The arrival positions of the users are assumed to be uniformly distributed in the area of interest $A$. Upon arrival, a user may appear in the coverage area of $j$, $j < K$, classes of small cells. It is associated with a small cell of class $i$ with probability $\alpha_i$, $i = 1, 2, \ldots, K$, based on an appropriate incentivization model and such that $\sum_{i=1}^{j} \alpha_i = 1$.

Since small cells are preferable, only when no small cell coverage is available at its arrival, the user connects to the macro LTE service. The session duration is defined by an exponentially distributed time interval with parameter $\mu_U$, during which a user may (re-)associate with different small cells of various classes as a consequence of mobility of users and small cells. In particular, while connected to a class-$k$ small cell, it connects to a new encountered small cell of class-$i$ with probability $\alpha_{ki}$.

The velocity of users is assumed to be $v_U$ and their height is $h_U$. The movement of users and small cells is described by the RDM according to which a device chooses its movement direction uniformly between 0 and $2\pi$, then moving at a fixed speed along the straight line for an exponentially distributed time with mean $E[\tau]$ before

[1] We collectively refer to LTE small cells, WiFi and mmWave access points, D2D relays, and highly mobile LTE relay cars and drones as *small cells* for the ease of exposition.

choosing a new direction. The motivation for choosing this model is that it is analytically tractable and permits to obtain a number of important metrics in the closed form, while still capturing the key properties of a random movement over the landscape [117]. The parameters of the RDM model are allowed to be different for various classes of small cells and below we parametrize them based on the realistic movement of users and typical classes of small cells.

It has been consider a specific *tagged* user that arrives into the system at time $t_0$ and is associated to a class-$i$ small cell. The time it spends in the coverage area of this small cell depends on the contact time (CT). During this time, the tagged user shares the resources of its connectivity point with other associated users and leaves the small cell coverage by having accumulated a certain amount of data. When leaving the coverage area of a small cell, the tagged user may enter the coverage of another small cell of class $j$, $j = 1, 2, \ldots, K$ or as a fall-back option connect to the macro LTE service, i.e. $j = 0$.

The time that the tagged user spends being associated with the baseline LTE service is named the inter-contact time (ICT). Due to the mobility of the small cells and the tagged user, the ICT and CT are random variables (RVs). Similarly, the amounts of data accumulated during the CTs and ICTs are RVs as well. The connectivity cases that we consider are:

*Partially-dynamic connectivity case.* This case models a user-centric network selection policy [118], where the user decides locally on which association point to choose and thus has to disconnect from its current content supplier to perform measurements and select another. The users are not allowed to change their association point while being connected to a small cell: only when leaving the coverage area, then they are allowed to connect to another class of available small cells or to the macro LTE service.

*Fully-dynamic connectivity case.* This more advanced connectivity option corresponds to a network-centric (assisted) network selection policy [118], in which the intelligent 5G system advises the user timely on its best connectivity choices preventing the user from disconnecting from its current content supplier to select another alternative option. Accordingly, the users may change their association point at any time whenever they encounter a "better" small cell of any class.

**Model for connectivity selection**

When considering a dense urban heterogeneous scenario, where multiple connectivity classes are available, the user attempting to acquire some content needs to select its preferred connectivity class to associate with. This choice has to be related to the *utility* of the decision maker, be it the user itself or the network (that has the

system-wide knowledge), to estimate the benefits of connecting to any of the content suppliers.

The *user-involvement framework* (UIF) detailed in Section 6.1.3 will model this aspect characterizing the following decision-related factors: (i) the intrinsic relationships between the connectivity and the services provided by the content suppliers together with the monetary price that the users are willing to pay in order to achieve their target content acquisition session; (ii) the differentiation of users based on the profiles given by the network marketing theory and their social relationships to engage them with the HITL mechanisms; (iii) the management of user decision strategies to select the most suitable class based on the connectivity features and service costs.

Importantly, the characterization of monetary price and energy efficiency has been attempted in the past with linear expressions, but estimating the connection reliability parameters requires a more in-depth analysis. To this aim, we will model the *expected reliability* with a weighted function of *a-priori* advertised connectivity and *a-posteriori* perceived performance.

**Connecting mobility-centric and user-involvement models**

To fully leverage the predictable and repeating human behavior, the proposed HITL architecture involves the users by actively learning, predicting, adapting to, and steering their behavior – to decisively improve system efficiency and provide superior levels of QoE. In particular, the *mobility-centric framework* (MCF) that is detailed in Section 6.1.2 is capable of taking into account the multi-connectivity 5G environment together with the potential access infrastructure mobility and by doing so provide the users with the relevant information they need to decide on the "best" (most reliable, affordable, etc.) connectivity option for acquiring their desired data.

The developed analytical framework acts as a "predictor" for the users that initiate new communication sessions, but do not have complete knowledge on the multi-connectivity system. In the face of uncertainty, the MCF aids users in making "smarter" decisions regarding the available connectivity options. Accordingly, it first collects information on the classes of the deployed small cells, the number of users and their content sessions, as well as the mobility patterns. Then, it supplies the users with the a-priori connectivity reliability information. Utilizing this knowledge, the users make their initial connectivity choices and update their preferences in the course of time. In doing so, users follow their personal profile with certain priorities (e.g., target price, desired data rate, battery consumption, etc.).

The way users build their personal perspective is captured with the UIF detailed in Section 6.1.3. Correspondingly, users employ the advertised (a-priori) connectivity reliability at the output of the MCF and gradually build their preferences regard-

ing the connectivity options based on individual past experiences in terms of energy efficiency, connection availability, and monetary price. In particular, to capture the dynamic performance fluctuations experienced by the users moving within the area of interest, an exponentially weighted moving average (EWMA) linear predictor is adopted, which takes at its input both the analytical expected connectivity as produced by the MCF, and the real network-wide performance modeled in a system-level simulation (SLS) tool[111].

While the MCF is crucial to assist the users in selecting the appropriate proximate small cell when they begin their content acquisition, with time the UIF gains power to understand the user-driven behavior due to mobility and personal human preferences. In addition, the UIF helps incentivize user-owned equipment across the system to share its connectivity/content for improved coverage and higher connection reliability.

### 6.1.2 Mobility-centric analytical methodology

#### Queuing network model

The content acquisition process in modeled in both partially-dynamic and fully-dynamic connectivity cases by constructing a framework of queuing networks with $2K + 1$ layers, where each layer consists of $N_i$, $i = 1, 2, \ldots, K$ queuing systems. A particular queuing system represents a certain small cell. Each queue in said network is of M/M/$\infty$ type capturing the time when a user is associated with a given small cell. The choice of the queues with an exponential inter-arrival time is facilitated by the Poisson structure of the arrival process and the independence of movement in the area of interest. The exponential service times are dictated by the properties of the encounter process of the tagged user with the content suppliers, which is discussed next.

As follows from the system model, the content consumers arrive into the network with intensity $\lambda_U$ and leave it with intensity $\mu_U$. Once arrived, the users may (re-)associate with the small cells of different classes and such dynamics is captured by the proposed model. The steady-state distribution in the network in question determines the number of users simultaneously sharing resources of the small cells and the macro LTE baseline. It is important to note that the tagged user is not included into this model. The system state observed by the tagged user specifies its service parameters including the mean data rate provided by the system in presence of other users sharing the resources.

The overall number of M/M/$\infty$ queuing systems is $2\sum_{i=1}^{K} N_i + 1$ and they are logically separated into two layers, named the *main* and the *complementary*. A single queue of the main layer completely characterizes the time that a user is associated

**Table 6.1.** Parameters employed by this work

| Parameter | Definition |
|---|---|
| **General parameters** | |
| $A, S_A$ | Convex region of interest and its area |
| $N$ | Total number of small cells |
| $K$ | Number of small cell classes |
| $N_i$ | Number of class-$i$ small cells |
| $P_i, W_i$ | Transmit power and bandwidth of class-$i$ small cells |
| $h_i, v_i$ | Height and speed of class-$i$ small cells |
| $E[\tau_i]$ | Mean 'run' time of class-$i$ small cell in RDM model |
| $A_i, S_{A_i}$ | Coverage of a class-$i$ small cell and its area |
| $D$ | User demand volume |
| $\lambda_U, \mu_U$ | Arrival and departure intensity of users |
| $\rho_U$ | Offered load of users |
| $h_U, v_U$ | Height and speed of the tagged user |
| $B_L$ | Macro LTE data rate |
| $B_i$ | Rate of a class-$i$ small cell |
| **Performance evaluation and optimization model parameters** | |
| $a_{ij}$ | Re-association probability to a class-$i$ small cell |
| $b_i$ | Probability of choosing a class-$i$ small cell |
| $p_i$ | Probability of 'seeing' $i$ classes of small cells upon arrival |
| $\gamma_i$ | Intensity of encounters with class-$i$ small cells |
| $\Gamma$ | Intensity of encounters with any class of small cells |
| $v^\star$ | Relative speed of a user and a class-$i$ small cell |
| $f_{v^\star}(x)$ | Probability density function (pdf) of relative speed |
| $f(x, y)$ | Two-dimensional density of user position |
| $T_I$ | Inter-contact time |
| $T_i, f_{T_i}(x)$ | Class-$i$ small cell contact time and its pdf |
| $\tilde{T}_i, f_{\tilde{T}_i}(x)$ | Recurrence time of a class-$i$ small cell and its pdf |
| $T_{F_i}, f_{T_{F_i}}(x)$ | First class-$i$ small cell sojourn time and its pdf |
| $T_{I_i}, f_{T_{I_i}}(x)$ | Class-$i$ small cell sojourn time and its pdf |
| $\theta_F$ | First re-association probability |
| $\phi_i(T_i)$ | Branches of CT inverse function |
| $\theta_I$ | Re-association probability |
| $E[R]$ | Data rate (throughput) available to the tagged user |
| $m(\cdot)$ | Kinematic measure |
| $\alpha, \beta$ | Relative arriving angles |
| $\zeta_i$ | Probability of being within $i$ small cell classes upon arrival |
| $Q = (q_{i,j})$ | Routing matrix of the queuing network |
| $\lambda_i, \mu_i$ | Arrival and service rates at queue $i$ |
| $\lambda_{ij}$ | Internal rates from queue $i$ to queue $j$ |
| $\pi(\mathbf{x})$ | Steady-state distribution of the queuing network |
| **Incentivization model parameters** | |
| $\upsilon$ | Utility function for incentivization model |
| $\Phi_i$ | Evaluation attribute for a class-$i$ small cell |
| $\varphi$ | Risk tolerance |
| $\eta$ | Weighting factor of incentivization model |
| $\omega$ | Weighting factor for a-priori estimated reliability |
| $C_i, C_{\max}$ | Normalized cost of a class-$i$ small cell and its maximum |
| $G_i, G_{\max}$ | Normalized energy of a class-$i$ small cell and its maximum |
| $\hat{R}_i$ | Estimated connectivity reliability for a class-$i$ small cell |
| $D_i$ | Amount of acquired data over a class-$i$ small cell |
| $\delta_k$ | Proportion of acquired data over $t$-th instant of time |

with the macro LTE class. Each small cell is modeled by two queuing systems: (i) at the main layer for the users entering the small cell coverage after (re-)association from another small cell, and (ii) at the complementary layer for the users encountering a small cell from within the macro LTE coverage. The need for the complementary layer is dictated by the fact that a user – while being connected to a certain small cell – may or may not meet another small cell during its CT, which results in different association times.

All of $2N + 1$ queues can be arranged into $2K + 1$ layers in accordance with the introduced small cell classes, which implies that utilization of queues of the same type will display similar properties. We enumerate the queues as follows: (i) $1, ..., N$ represent the main layer class-$i$ small cells; (ii) $N + 1, ..., 2N$ are the complementary layer class-$i$ small cells; and (iii) $2N + 1$ is the 'macro LTE' queue.

Below, first is considered the fully-dynamic connectivity case, where the users are allowed to change their connectivity options "on-the-fly" whenever a new small cell is encountered. Then, is discussed the modifications required for this more general model to capture the inferior partially-dynamic connectivity case. It is also briefly addressed the case of less dense (sparse) systems, i.e. where either the area of interest is large or the overall number of small cells is small (or both of these considerations). To parametrize the proposed model, it is needed to derive (i) the service times of users at the main layers; (ii) the service time of users at the complementary layers; and (iii) the routing matrix.

**Macro LTE queue service time**

The macro LTE queue service time is the duration that a user spends connected to the baseline cellular service. Define the ICT (inter-contact time) as the interval between the instant of time when a user connects to the macro LTE upon leaving the coverage area of a small cell and until it connects to another small cell of any class. The first contact time (FCT) is defined as the interval between when a user enters the system and finds itself outside the coverage areas of all the small cells and until it connects to a small cell of any class for the first time. In general, the FCT and the ICT may not coincide. The macro LTE service time coincides with the FCT duration when a user enters the system and connects to the baseline class and with the ICT in the rest of the cases.

Let $T_I$ be the RV corresponding to the FCT. Assuming the independence of the user location at the time moment $t$ from its location at the time moment $t + \Delta t$, it has been shown in [119] that the FCT between two mobile nodes moving at speeds $v_1$ and $v_2$ according to the RDM (with the same coverage range $r$) is exponential with the rate of

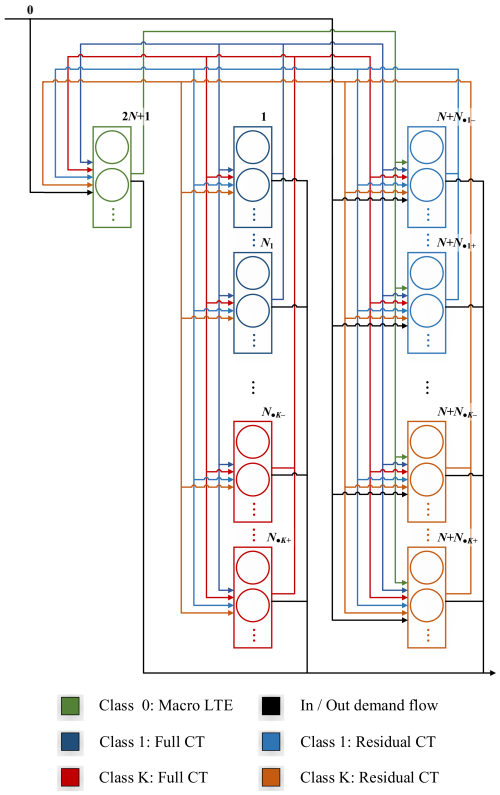



**Fig. 6.2.** An illustration of the queuing network under consideration.

$$\gamma_i^\star = 2rE[v^\star] \iint_{S_{A_i}} f^2(x,y)dxdy,\ \gamma_i^\star > 0, \tag{6.1}$$

where $E[v^\star]$ is the average relative speed between the users, $f(x,y)$ is the stationary distribution of moving users in the area of interest, $S_A$ is the area of interest. The pdf (probability density function) of the relative speed has been obtained in [119] and is given by

$$f_{v^\star}(x) = \frac{x}{v_1 v_2 \sqrt{1 - \left(\dfrac{v_1^2 + v_2^2 + x}{2v_1 v_2}\right)^2}},\ x > 0, \tag{6.2}$$

that is, non-zero for $|v_1 - v_2| < x < v_1 + v_2$.

In the considered case, the coverage radii of class-$i$ small cells and the user are $r_i$ and $r_U$, respectively. Also, recall that the density of the RDM model over any convex area $A$ is known to be uniform, i.e. $f(x,y) = 1/S_A$, where $S_A$ is the area of $A$. Therefore, is possible to obtain

$$\gamma_i^\star = \frac{(r_i + r_U)E[v^\star]}{S_A}. \tag{6.3}$$

Observe that when $v_i = 0$, i.e. the small cells of class-$i$ are stationary, while the tagged user moves at the speed of $v_U$, is obtained $E[v^\star] = v_U$. In this case, (6.3) reduces to

$$\gamma_i^\star = \frac{(r_i + r_U)v_U}{S_A}, \tag{6.4}$$

which implies that the mean ICT is $E[T_I] = S_A/2r_U v_U$.

When $N_i$ small cells of class $i$ are uniformly distributed over the area, the FCT distribution is the minimum of $N_i$ exponential distributions with the same parameter, which is again exponential with the parameter $\gamma_i = N_i\gamma_i^\star$. It has also been shown in [119] that under the aforementioned conditions the ICT also approximately follows the exponential distribution with the same parameter. Hence, it is sufficient to use a single queuing system of M/M/$\infty$ type to represent the macro LTE sojourn time.

An important assumption for (6.1) to hold is that the user moves fast enough. It has been experimentally clarified in [120] that the exponential distribution provides a sensible approximation when the dimension of the area of interest is at most four times greater than the average "run" length in the RDM. If this condition is not met, the distribution in question is a mixture of exponential and power-law components in the form

$$Ct^{-\alpha}e^{-\beta t},\ t \geq 0, \alpha > 0, \beta > 0, \tag{6.5}$$

which is inherent for random walks in any dimension. Recalling that the typical pedestrian speed specified by 3GPP in TS 36.304 document is 3.6km/h and assuming the mean straight run of 10s, the discussed approximation is applicable for $X < 100$m, which is sufficient for the scenario of interest. For larger areas, one could parametrize (6.5) experimentally by utilizing system-wide simulations. Observing the general form of the FCT, one could e.g., use Gamma distribution.

**Small cell service times**

Recall that according to the fully-dynamic connectivity case, a user may change its association even when it is currently connected to a small cell. A user remains associated with a given small cell until it either encounters another small cell and decides to re-connect to it or until it leaves the coverage of its current small cell. Additionally, to determine the small cell sojourn times, is needed to differentiate between the case when (i) a user arrives into the system, finds itself in the coverage area of a certain small cell, and decides to connect to it and the case when (ii) a user connects to the small cell upon meeting it. All these quantities can be expressed by using the CT, which is defined as the amount of time that a user spends in the coverage area of a small cell upon meeting it.

Let $T_i$ be the RV denoting the CT of class-$i$ small cells. Observe that the CT is the first passage time in a circle starting at a random point on the circumference that does not necessarily coincide with the stopping epoch of the RDM. The general

form of this metric is available for simpler walks only, especially for the dimensions of higher than one [97]. However, taking into account rather small coverage areas of the small cells, it is possible to obtain the CT explicitly by approximating the movement of the user and the class-$i$ small cell when they are in contact with each other by straight trajectories.

Observe that a user and a small cell come in contact with each other only when the distance between them is less than the minimum coverage area of both. One possible case is illustrated in Fig. 6.3. Even though both the user and the small cell are randomly directed on a plane, this without loss of generality can be transformed into a relative direction of the user towards the small cell. Accordingly, given that at the initial moment of time the user is located on the circumference of the small cell and is directed towards it, the establish of the CT distribution is needed, i.e. the time that the user spends before coming in contact with the circumference again.

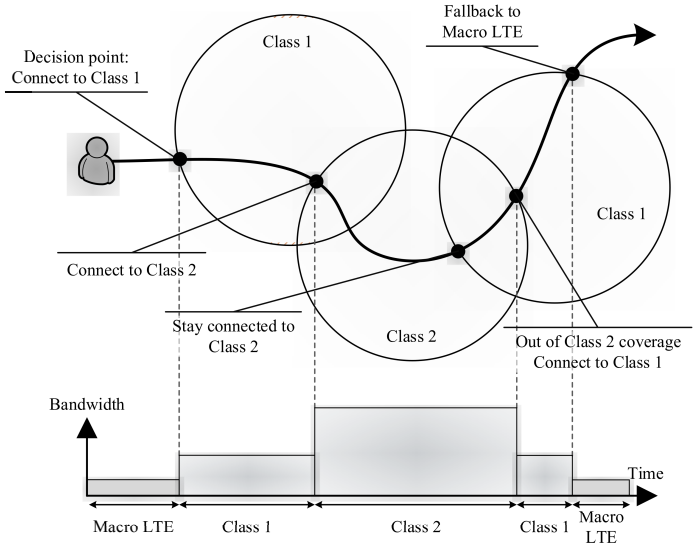



**Fig. 6.3.** An example fragment of a user content acquisition session.

Let the small cell at the beginning of the CT, i.e. at $t_0$ be located at the origin, $x_0 = 0$ and $y_0 = 0$, and assume that the user is at the distance $r$ from it, at the coordinates $x_U = 0$ and $y_U = r$. The small cell is moving with speed $v_i$ along the horizontal axis and the user is moving with the speed $v_U$ making an angle $\alpha$ with the $x$-axis, as shown in Fig. 6.3. In such a case, the interested is in the CT $T_i$, which satisfies the circle equation

$$(x - x_0)^2 + (y - y_0)^2 = r^2, \tag{6.6}$$

where $x$ and $y$ are the coordinates of the user location.

Further, the coordinates of the user dynamics over time are

$$x = x_U + T_i v_U \cos(\alpha), y = y_U + T_i v_U \sin(\alpha). \tag{6.7}$$

Substituting (6.7) into (6.6) is possible to arrive at

$$(T_i v_U \cos(\alpha) - T_i v_i)^2 + (r + T_i v_U \sin(\alpha))^2 = r^2. \tag{6.8}$$

Solving (6.8) with respect to the CT gives

$$T_i = \frac{2 r v_U \sin(\alpha)}{2 v_i v_U \cos(\alpha) - v_i^2 - v_U^2}. \tag{6.9}$$

As the user comes in contact with the small cell at an arbitrary random angle $\alpha$, the latter follows the uniform distribution in $(0, \pi)$ with the associated density $f_\alpha(x) = 1/\pi$. Employing the RV transformation technique in [100], the density of the CT can be provided in the explicit form. Denote the direct transformation $T_i = f(\alpha)$, where $f(\alpha)$ is the right-hand side of (6.9). Now, the inverse transform $\alpha = \phi(T_i)$ is needed, which is the solution of the following quadratic equation with respect to $\cos(\alpha)$

$$\begin{aligned} &4 v_U^2 \left(T_i^2 v_i^2 + r^2\right) \cos^2(\alpha) - 4 r v_U T_i \left(v_U^2 + v_i^2\right) \cos(\alpha) - \\ &- 4 v_U^2 v_i^2 T_i^2 - T_i \left(v_U^2 + v_i^2\right) = 0. \end{aligned} \tag{6.10}$$

$$\phi_i(T_i) = \begin{cases} \frac{1}{\pi} \arccos\left( \frac{-2 r v_U \left(v_i^2 + v_U^2\right) T_i + 2 v_i v_U T_i \sqrt{4 r^2 v_U^2 - \left(v_i^2 - v_U^2\right)^2 T_i^2}}{4 v_U^2 \left(r^2 + v_i^2 T_i^2\right)} \right), 0 \le T_i \le \left| \frac{2 r v_U}{v_U^2 - v_i^2} \right| \\ \frac{1}{\pi} \arccos\left( \frac{-2 r v_U \left(v_i^2 + v_U^2\right) T_i - 2 v_i v_U T_i \sqrt{4 r^2 v_U^2 - \left(v_i^2 - v_U^2\right)^2 T_i^2}}{4 v_U^2 \left(r^2 + v_i^2 T_i^2\right)} \right), 0 \le T_i \le \frac{2 r v_U}{v_U^2 + v_i^2} \\ -\frac{1}{\pi} \arccos\left( \frac{-2 r v_U \left(v_i^2 + v_U^2\right) T_i - 2 v_i v_U T_i \sqrt{4 r^2 v_U^2 - \left(v_i^2 - v_U^2\right)^2 T_i^2}}{4 v_U^2 \left(r^2 + v_i^2 T_i^2\right)} \right), \frac{2 r v_U}{v_U^2 + v_i^2} \le T_i \le \left| \frac{2 r v_U}{v_U^2 - v_i^2} \right| \end{cases} . \tag{6.11}$$

$$\begin{aligned} f_{T_i}(t) = &\frac{1}{\pi} \frac{d \cos^{-1}\left[ \frac{-2 r v_U \left(v_i^2 + v_U^2\right) t + 2 v_i v_U t \sqrt{4 r^2 v_U^2 - \left(v_i^2 - v_U^2\right)^2 t^2}}{4 v_U^2 \left(r^2 + v_i^2 t^2\right)} \right]}{dt} + \\ &+ \frac{\mathbf{1}(t)}{\pi} \frac{d \cos^{-1}\left[ \frac{-2 r v_U \left(v_i^2 + v_U^2\right) t - 2 v_i v_U t \sqrt{4 r^2 v_U^2 - \left(v_i^2 - v_U^2\right)^2 t^2}}{4 v_U^2 \left(r^2 + v_i^2 t^2\right)} \right]}{dt}. \end{aligned} \tag{6.12}$$

Observing that $0 \le \alpha \le \pi$, the inverse transform in question will have three branches as in (6.11), which leads to the following pdf of the transformation $T_i = f(\alpha)$

$$f_{T_i}(y) = \sum_{1 \le i \le 3} f_\alpha\left(\phi_i(y)\right) \left| \frac{d \phi_i(y)}{dy} \right|. \tag{6.13}$$

Due to the huge amount of calculations, it is intentionally offered the resulting pdf of $T_i$ (6.12) in the differential form, where $0 \leq t \leq \left|2rv_U/v_U^2 - v_i^2\right|$. In the special case of $v_i = v_U = v$, the CT reduces to

$$f_{T_i}(x) = \frac{2rv}{r^2 + v^2t^2},\ 0 \leq t \leq \infty. \tag{6.14}$$

Another special case of interest is when $v_i = 0$, that is, the class-$i$ small cells are all stationary, which corresponds to the case of, e.g., conventional micro/femto deployment. In this situation, the CT reduces to

$$f_{T_i}(x) = \frac{2v}{\pi\sqrt{4r^2 - (xv)^2}},\ 0 \leq x \leq 2rv, \tag{6.15}$$

which coincides with the result provided in [121].

The backward recurrence CT is then obtained as

$$F_{\tilde{T}_i}(t) = \frac{1}{E\left[T_i\right]} \int_0^t \left(1 - \int_0^x f_{T_i}(y)dy\right) dx. \tag{6.16}$$

Consider now the amount of time $T_{F_i}$ that a user (upon its arrival into the system) spends connected to a small cell, given that it re-associates to another small cell. Observe that this may happen while still being in coverage of the initial small cell if another small cell is encountered during the backward recurrence time. The probability of the considered event is

$$\theta_F = \sum_{k=1}^{\infty} \int_0^{\infty} \frac{(\Gamma x)^k}{k!} e^{-\Gamma x} f_{\tilde{T}_i}(x)dx = 1 - \int_0^{\infty} e^{-\Gamma x} f_{\tilde{T}_i}(x)dx, \tag{6.17}$$

where $\Gamma$ is the aggregate rate of encountering the small cells, i.e.,

$$\Gamma = (N_i - 1)\gamma_i^{\star} + \sum_{\forall j, j \neq i}^{K} N_j \gamma_j^{\star} \approx \sum_{i=1}^{K} \gamma_i, \tag{6.18}$$

where $\gamma_i^{\star}$ is the rate of encountering a single class-$i$ small cell and $f_{\tilde{T}_i}(t) = F'_{\tilde{T}_i}(t)$ is the pdf of the backward recurrence time. With the complementary probability $1 - \theta_F$, the user leaves the small cell and connects to the macro LTE class.

Since the process of encountering small cells is Poisson in nature, with the exponentially-distributed ICT, $T_{F_i}$ follows an exponential distribution with the parameter $\Gamma$. Similarly, by defining $T_{I_i}$ to be the time that a user spends connected to a small cell given that it re-associates to another small cell, we establish that $T_{I_i}$ also follows an exponential distribution with the same parameter $\Gamma$.

The corresponding probability that a user re-associates while still being in the coverage of a small cell, $\theta_I$, is given by (6.17), thus replacing the backward recurrence time with the CT. Therefore, the service time in M/M/$\infty$ queue of the main layer that models the small cell of class $i$ is the minimum of two exponentially-distributed RVs, each with the parameter $\Gamma$, and hence also follows an exponential distribution with the parameter $2\Gamma$.

When either the total number of the small cells is higher or the area of interest is larger, the probabilities $\theta_F$ and $\theta_I$ may not be negligible. This important situation is then modeled by adding the complementary layers to the queuing network. Accordingly, we approximate the service time in each queue of the complementary layers by using exponential distributions with the mean produced by (6.15).

**Routing matrix**

Recall that new users arrive into the system with intensity $\lambda_U$. Depending on the geometrical position of the user upon entering the system, it may connect (i) to the macro LTE service if it finds itself outside the coverage areas of all the small cells or (ii) to a certain class-$i$ small cell. Consider first the former event and denote its probability by $\zeta_0$. Observe that when the density of the small cells of all classes is relatively high, the probability that their coverage areas overlap is non-negligible. Hence, the approximation of $\zeta$ is given in the form

$$\zeta_0 = 1 - \frac{\sum_{i=1}^{K} N_i S_{A_i}}{S_A}, \tag{6.19}$$

where $S_A$ is the area of interest and $S_{A_i}$ is the coverage area of a class-$i$ small cell, which may actually result in significant errors. Therefore, $\zeta$ is determined by following the lines of *integral geometry* (see, e.g., [122] for further reading).

Consider first a single small cell of an arbitrary class with the coverage $A_1$ in the area of interest. It is possible to write

$$\zeta_0 = 1 - \frac{Pr\{P \in A \cap A_1\}}{Pr\{A_1 \cup A \neq 0\}}, \tag{6.20}$$

where $P$ is the point of interest. Recall that according to [117], $P$ is uniformly distributed in $A$.

Using the notion of *kinematic measure* [122]

$$\begin{aligned} &Pr\{P \in A \cap A_1\} = m(A : P \in A \cap A_1), \\ &Pr\{A_1 \cap A \neq 0\} = m(A_1 : A_1 \cap A \neq 0), \end{aligned} \tag{6.21}$$

where the first expression above is the kinematic measure for the set of motions of $A$, such that $P \in A$, while the second one provides the measure for all the motions of $A$, such that $A_1$ intersects $A$ (for additional details, refer to [122]).

The first measure is immediately computed to be

$$\begin{aligned} m(P \in A \cap A_1) &= \int_{P \in A_1} dx \wedge dy \wedge d\phi = \\ &= \int_{P \in A_1} dx \wedge dy \int_0^{2\pi} d\phi = 2\pi S_{A_1}, \end{aligned} \tag{6.22}$$

where $S_{A_1}$ is the area of $A_1$.

The measure of all the motions of $A$, such that $A_1 \subset A$ [122], is in the form of

$$m(A_1 \cap A \neq 0) = \int_{A_1 \subset A} dx \wedge dy \wedge d\phi = \\ = 2\pi[S_A + S_{A_1}] + L_A L_{A_1}, \tag{6.23}$$

where $L_A$ is the perimeter of $A$. The latter holds for any convex sets $A_1$ and $A$, as long as the curvature of $A$ is not greater than that of $A_1$, while (6.22) is true for general convex $A$ and $A_1$.

Since the coverage areas of the small cells are assumed to all be of circular shape, by substituting $S_{A_1} = \pi r_i^2$ and $L_{A_1} = 2\pi r_i$ into (6.22) and (6.23), and then to (6.20), is obtained

$$\zeta_0 = 1 - \frac{2\pi^2 r_i^2}{2\pi[S_A + \pi r_i^2] + 2\pi r_i L_A}. \tag{6.24}$$

Consider now the case of $N_i$ small cells each having the coverage $S_{A_i}$ and let $\zeta_i$, $i = 1, 2, \ldots, N_i$ be the probability that the tagged user arrives within the coverage of exactly $i$ small cells. The kinematic measure for the set of motions of $A$, such that $P$ is covered by exactly $m$ our of $N_i$ small cells, is given by

$$m(P \text{ in } m \text{ small cells}) = \int dP \wedge dA_1 \cdots \wedge dA_{N_i} = \\ = \binom{m}{N_i} \int_{P \in A} (2\pi S_{A_i})^m (2\pi S_A + L_A + L_{A_i})^{N_i - m} = \\ = \binom{m}{N_i} (2\pi S_{A_i})^m (2\pi S_A + L_A + L_{A_i})^{N_i - m} S_A. \tag{6.25}$$

The measure for all the motions of $A$, so that $A_i \cup A \neq 0$, is

$$m(\forall A_i \cap A \neq 0) = \int_{\forall A_i \cap A \neq 0} dA_1 \cdots \wedge dA_{N_i} = \\ = [2\pi[S_A + S_{A_i}] + L_{A_i} L_{A_1}]^{N_i}. \tag{6.26}$$

After combing the latter two results, is possible to arrive at

$$z_i = \frac{m(P \text{ in } m \text{ small cells})}{m(\forall A_i \cap A \neq 0)} = \\ = \frac{\binom{m}{N_i} (2\pi S_{A_i})^m (2\pi S_A + L_A + L_{A_i})^{N_i - m} S_A}{[2\pi[S_A + S_{A_i}] + L_{A_i} L_{A_1}]^{N_i}}, \tag{6.27}$$

which leads to the following expression for $z_0$

$$z_0 = \frac{(2\pi S_A + L_A + L_{A_i})^{N_i}}{[2\pi[S_A + S_{A_i}] + L_{A_i} L_{A_1}]^{N_i}}. \tag{6.28}$$

Extending the above for $K$ classes of small cells,

$$z_0 = \prod_{i=1}^{K} \left(1 - \frac{(2\pi S_A + L_A + L_{A_i})^{N_i}}{[2\pi[S_A + S_{A_i}] + L_{A_i} L_{A_1}]^{N_i}}\right). \tag{6.29}$$

Further, let $Q = (q_{ij})$, $0 \leq i \leq 2N+1, 0 \leq j \leq 2N+1$, be the routing matrix associated with the network in question. With the probability $1-\zeta_0$, an arriving user observes at least one small cell to connect to. Let $p_i$, $i = 1, 2 \ldots, K$, be the probability that at least one small cell of class $i$ is available at the time of user arrival. These probabilities are given by

$$p_i = 1 - \frac{(2\pi S_A + L_A + L_{A_i})^{N_i}}{[2\pi[S_A + S_{A_i}] + L_{A_i} L_{A_1}]^{N_i}}. \tag{6.30}$$

Therefore, the overall arrival flow is divided between the layers as

$$\begin{aligned}
&q_{00} = 0,\\
&q_{0j} = \frac{p_k b_k}{N_k} \sum_{\kappa_1 \cup \kappa_2 = \mathcal{K}_k} \frac{\prod\limits_{n\in\kappa_1} p_n \prod\limits_{m\in\kappa_2} (1-p_m)}{1 - \sum\limits_{m\neq k} b_m}, j = N_{\bullet k-}, .., N_{\bullet k+}, k = 1, .., K,\\
&q_{0j} = 0,\ j = N+1, N+2, \ldots, 2N,\\
&q_{0(2N+1)} = \prod_{n=1}^{K} (1-p_n)\,,
\end{aligned} \tag{6.31}$$

where $N_{\bullet k-} = N_1 + \cdots + N_{k-1} + 1$, $N_{\bullet k+} = N_1 + \cdots + N_k$, $\mathcal{K}_k = \{1, \ldots, k-1, k+1, \ldots, K\}$, and $b_k$, $k = 1, 2, \ldots, K$, such that $\sum_{k=1}^{K} b_k = 1$, are the priority coefficients that denote the probabilities of connecting to a class-$k$ small cell, if the small cells of multiple classes are available. By setting one of these to 1 and imposing the rest to 0, the absolute deterministic priority between the layers can be captured.

Note that a user remains in the queue until when it leaves the coverage of a corresponding small cell or encounters another small cell and re-associates with it. In the proposed framework, the re-association is dynamic and is controlled by the coefficients $a_{ij}$, $i, j = 1, 2, \ldots, K$, that is, upon encountering a small cell of class $j$ while being connected to a small cell of class-$j$, the user connects to the new cell with the probability $a_{ij}$. In other words, $a_{ij}$ are the system-wide load balancing coefficients that enable the network to coordinate the load across multiple classes of small cells. Said coefficients affect the service rates of the queues, which are now characterized by

$$\begin{aligned}
&\mu_i = \tilde{\Gamma}_i + \frac{1}{E\,[T_i]} + \mu_U,\ i = 1, 2, \ldots, N,\\
&\mu_i = \tilde{\Gamma}_{i-N} + \frac{1}{E\left[\tilde{T}_{i-N}\right]} + \mu_U,\ i = N+1, N+2, \ldots 2N,\\
&\mu_{2N+1} = \Gamma + \mu_U,
\end{aligned} \tag{6.32}$$

where $\tilde{\Gamma}_i = \sum_{j=1}^{K} a_{ij}\gamma_j$, $i = 1, 2, \ldots, K$.

The routing probabilities between the main and the complementary layers are provided in (6.33). Finally, the routing probabilities from the macro LTE queue are

$$
\begin{aligned}
&q_{i0} = \frac{\mu_U}{\mu_k}, i = N_{\bullet k-}, .., N_{\bullet k+}, k = 1, 2, \ldots, K, \\
&q_{ij} = \frac{a_{k_1 k_2} \gamma_{k_2}}{N_{k_2} \mu_{k_1}}, \; N_{\bullet k_1-} \leq i \leq N_{\bullet k_1+}, k_1 = 1, 2, \ldots, K, j = N_{\bullet k_2-}, .., N_{\bullet k_2+}, k_2 = 1, 2, \ldots, K \\
&q_{ij} = \frac{p_{k_2} b_{k_2} \sum\limits_{\kappa_1 \cup \kappa_2 = K_{k_2}} \prod\limits_{n \in \kappa_1} p_n \prod\limits_{m \in \kappa_2} (1 - p_m)}{N_{k_2} E\left[T_{k_1}\right] \mu_{k_1} \left(1 - \sum\limits_{n \in \kappa_2} b_n\right)}, i = N_{\bullet k_1-}, .., N_{\bullet k_1+}, k_1 = 1, 2, \ldots, K, j = N + N_{\bullet k_2-}, .., N + N_{\bullet k_2+}, k_2 = 1, 2, \ldots, K, \\
&q_{i(2N+1)} = \frac{1}{E\left[T_{k_1}\right] \mu_k} \prod_{n=1}^{K} (1 - p_n), i = N_{\bullet k-}, .., N_{\bullet k+}, k = 1, 2, \ldots, K, \\
&q_{i0} = \frac{\mu_U}{\mu_k}, i = N + N_{\bullet k-}, .., N + N_{\bullet k+}, k = 1, 2, \ldots, K, \\
&q_{ij} = \frac{a_{k_1 k_2} \gamma_{k_2}}{N_{k_2} \mu_{k_1}}, i = N + N_{\bullet k_1-}, .., N + N_{\bullet k_1-}, k_1 = 1, 2, \ldots, K, \; j = N_{\bullet k_2-}, .., N_{\bullet k_2+}, k_2 = 1, 2, \ldots, K, \\
&q_{ij} = \frac{p_{k_2} b_{k_2} \sum\limits_{\kappa_1 \cup \kappa_2 = K_{k_2}} \prod\limits_{n \in \kappa_1} p_n \prod\limits_{m \in \kappa_2} (1 - p_m)}{N_{k_2} E\left[\tilde{T}_{k_1}\right] \mu_{k_1} \left(1 - \sum\limits_{n \in \kappa_2} b_n\right)}, i = N + N_{\bullet k_1-}, .., N + N_{\bullet k_1+}, k_1 = 1, 2, \ldots, K, \; j = N + N_{\bullet k_2-}, .., N + N_{\bullet k_2+}, k_2 = 1, 2, \ldots, K, \\
&q_{i(2N+1)} = \frac{1}{E\left[\tilde{T}_{k_1}\right] \mu_k} \prod_{n=1}^{K} (1 - p_n), i = N + N_{\bullet k-}, .., N + N_{\bullet k+}, k = 1, 2, \ldots, K.
\end{aligned}
\tag{6.33}
$$

$$
\begin{aligned}
q_{(2N+1)0} &= \frac{\mu_U}{\mu_{2N+1}}, \\
q_{(2N+1)j} &= \frac{\gamma_k}{N_k \mu_{2N+1}}, j = N_{\bullet k-}, .., N_{\bullet k+}, k = 1, 2, \ldots, K, \\
q_{(2N+1)j} &= 0, j = N + N_{\bullet k-}, .., N + N_{\bullet k+}, k = 1, 2, \ldots, K, \\
q_{(2N+1)(2N+1)} &= 0,
\end{aligned}
\tag{6.34}
$$

which concludes the parametrization of the routing matrix $Q$.

### Analysis and optimization

In the proposed queuing network framework, the rate of external user arrivals into the queue $i$ is

$$
\lambda_{0i} = \lambda_U Q_{0i}, \; i = 1, 2, \ldots, 2N + 1. \tag{6.35}
$$

The intensities of the flows from the queue $i$ to the queue $j$ are then

$$
\lambda_{00} = 0, \; \lambda_i = \sum_{j=0}^{2N+1} \lambda_{ji}, i = 1, 2, \ldots, 2N + 1. \tag{6.36}
$$

Hence, the arrival intensities are

$$
\begin{aligned}
\lambda_{ij} &= \lambda_i Q_{ij}, i, j = 1, 2, \ldots, 2N + 1, \\
\lambda_j &= \sum_{i=1}^{2N+1} \lambda_i Q_{ij}, j = 1, 2 \ldots, 2N + 1.
\end{aligned}
\tag{6.37}
$$

Now, let $\pi(\mathbf{x}), \mathbf{x} = (x_1, ..., x_{2N+1})$ be the stationary distribution of the numbers of users in the individual queues. The corresponding solution in the product form is adapted from [123] as follows

$$\pi(\mathbf{x}) = \prod_{i=1}^{2N+1} \pi_i(x_i),$$
$$\pi_i(x_i) = \frac{\rho_i^{x_i}}{x_i!} p_i(\mathbf{0}), \tag{6.38}$$

where $\rho_i = \lambda_i/\mu_i$, $i = 1, 2, \ldots, 2N+1$.

Even though the proposed solution is available in the product form thus allowing for efficient implementations, for larger areas of interest and denser cell deployments of various classes the computational complexity of the straightforward procedure may be rather high. However, this complexity can be reduced by benefiting from the special structure of the considered network model.

Indeed, the queues associated with each class of small cells behave similarly as the fraction of load is balanced between them by only using the coefficients $a_{ij}$. Employing this property and also accounting for the open nature of the network, the system can be efficiently decomposed into separate queues and further analyze them in isolation. Recall that the steady-state distribution of the number of users in M/M/$\infty$ queue follows Poisson distribution with the parameter $\rho_i = \lambda_i/\mu_i$.

Consequently, the mean rate available to the tagged user is determined by

$$E[R] = \sum_{\forall \mathbf{x}} \frac{\pi(\mathbf{x}) \sum_{i:x_i+x_{i+N}\neq 0} B_i}{x_\bullet}, \; x_\bullet = \sum_{i=1}^{2N+1} x_i, \tag{6.39}$$

which can be used to estimate the time to acquire the demand of size $D$.

Finally, observe that the coefficients $a_{ij}$, $i, j = 1, 2, \ldots, K$, responsible for the (re-)association with a class-$j$ small cell while being connected to a class-$i$ small cell are essentially the parameters of the system that need to be adjusted appropriately to establish the a-priori probabilities of re-association, such that a certain metric of interest is optimized. Therefore, for the proposed framework the following unconstrained optimization problem is formulated

$$\begin{aligned}
&\textbf{Maximize:} \sum_{\forall \mathbf{x}} \frac{\pi(\mathbf{x}) \sum_{i:x_i+x_{i+N}\neq 0} B_i}{x_\bullet},\\
&\textbf{Over:} \quad a_{ij} \text{ and } b_i,\ i, j = 1, 2, \ldots, K,\\
&\textbf{Subject to:}\ a_{ij} \in [0,1],\ i, j = 1, 2, \ldots, K,\\
&\qquad\qquad b_i \in [0,1], i = 1, 2, \ldots, K,
\end{aligned} \tag{6.40}$$

which can be solved e.g., numerically. Note that the priority coefficients $b_i$, $i = 1, 2, \ldots, K$, responsible for selecting a class-$i$ small cell when multiple choices are available, can be excluded from the optimization problem formulation in case when additional, operator-driven factors also affect this selection.

### 6.1.3 Practical user involvement considerations

In this section, the system-wide analytical methodology is complemented with a discussion on user-specific involvement. This will include (i) the sources of motivation to incentivize the user to lend its personal networked devices for collective tasks (sharing computational resources and connectivity, providing content, etc.) as well as (ii) the strategies to manage the choice of the connectivity class efficiently. Regarding the former, example forms of incentives are trust relationships between communicating humans and their owned devices or certain mechanisms of compensation to encourage cooperative interaction. Analyzing past user-centric networking research [124], the following incentivization factors for the HITL systems are identified: (i) *willingness to share*, (ii) *trust*, (iii) *intrinsic/extrinsic motivations for cooperation*, and (iv) *reciprocity*.

The willingness to share can apply actively or passively, i.e., the device/content owner may or may not be aware that the connectivity is being shared. Further, trust models for supporting cooperation typically consider social interaction and human interests as the basis for building trust, and integrate reputation mechanisms so that misbehavior in the network is reduced. The intrinsic motivations for cooperation are: fairness, sense of community, synergy, and personal interests matching the public interests. Whenever intrinsic motivations are not sufficient, further incentivization is required to provide additional extrinsic motivation.

The latter can include rewards for realigning the utility of an individual towards the public utility and is possible in three forms: reputation, reciprocity, and monetization. Even though reputation and reciprocity are feasible factors, selfish users may still refuse to be cooperative, since they are constrained by limited computation, energy, and/or communication resources. Hence, a combination of reputation and reward-based considerations may be more appropriate. Indeed, as it is shown in the following section, the small cell equipment is incentivized to offer connectivity to other users, since monetary cost savings and energy efficiency gains are obtained.

In the considered reference multi-connectivity 5G scenario, are modelled the meaningful incentives that explicitly capture the user preferences across the available connectivity options, that is, class-$i$ small cells present across the area of interest (e.g., 3GPP LTE, D2D-, car-, or drone-cell). Specifically, whenever the tagged user makes a choice on which class of small cells to connect to, it faces a problem of decision-making *under uncertainty*. In particular, said uncertainty is related to the question whether the selected small cell can actually provide the desired/advertised connectivity performance. Accordingly, the willingness to connect to a class-$i$ small cell and thus pay a specified amount of money for the received service has to account for the *utility* perception of the decision maker.

The proposed user involvement model includes three key parameters that impact the utility function and hence the user decisions. These allow to characterize the advertised/expected QoE for each class-$i$ small cell, while adjusting the service price accordingly. To this aim, for each class-$i$ small cell in the system is considered: (i) the *perceived connectivity reliability* level associated with each class-$i$ small cell (i.e., the probability measure of acquiring data from the content supplier within a certain delay deadline); (ii) the *nominal monetary cost* related to the exploitation of service provided by a given class-$i$ small cell, and (iii) the *energy efficiency* of the user when connected to a particular small cell class.

On the other hand, when modeling the user profiles in the HITL context, it is crucial to account for the decision maker's attitude towards risk taking. The latter determines the shape of the utility function [125] when connecting to a class-$i$ small cell. The risk taking attitude can be regarded as a parameter within the user profile, where new technologies and uncertain performance levels are offered. The user willing to receive certain expected values of QoE rather than any uncertain alternative is named as *risk averse*. For instance, such user may cover the segment of the market that includes elderly people who prefer the 'good old known' service to the new 'unknown' opportunities.

Conversely, the user willing to receive similar expected values for certain and uncertain alternatives is called *risk neutral*, whereas the user preferring the uncertain alternative is named *risk seeking*. The latter category corresponds to the "early adopters" in the market, since e.g., younger generation is more interested in exploring new technologies and services. To explicitly capture the user preferences, is adopted a monotonically increasing exponential function over a single evaluation measure (or, evaluation attribute) $\Phi$ (i.e., larger values of $\Phi$ are preferred to its smaller values) in the form

$$\upsilon(\Phi) = \begin{cases} \dfrac{\exp[-(\Phi - I_L)/\varphi] - 1}{\exp[-(I_H - I_L)/\varphi] - 1}, & \varphi \neq \infty \\ \dfrac{\Phi - I_L}{I_H - I_L}, & \text{otherwise,} \end{cases} \tag{6.41}$$

where $I_H$ and $I_L$ are the highest and the lowest levels of interest for $\Phi$, respectively, and $\varphi$ is the risk tolerance.

The considered utility function is scaled so that $\upsilon(I_L) = 0$ and $\upsilon(I_H) = 1$. The $\varphi$ parameter choices may be modeled after the customer profiles and network marketing categories as e.g., detailed in [126]. For the value of $\Phi_i$ associated with each class-$i$ small cell,a weighted function is considered over (i) the nominal cost parameter $C_i$ (normalized over the maximum cost $C_{\max}$), (ii) the estimated connectivity reliability $\hat{R}_i$, and (iii) the energy efficiency $G_i$ (normalized over the maximum cost $G_{\max}$) as

$$\Phi_i = \eta_c \frac{C_i}{C_{\max}} + \eta_g \frac{G_i}{G_{\max}} + \eta_r \hat{R}_i, \tag{6.42}$$

where $\eta_c + \eta_e + \eta_r = 1$ and each term is a weighting factor to offer higher or lower importance to the individual parameters (is assumed for simplicity $\eta_c = \eta_e = \eta_r$)[2].

All the considered parameters (including $\Phi_i$) have their values within $[0, 1]$, as well as $I_H$ and $I_L$ in (6.41) assume the values "1" and "0", respectively. Further, $C_i$ is a constant nominal cost associated with the class-$i$ small cell, while $C_{\max}$ is the maximum cost across all the possible small cells. Then, $G_i$ is the energy efficiency figure for a class-$i$ small cell, whereas $G_{\max}$ is the maximum value of the energy efficiency among the available small cells of the same class.

Further, the energy efficiency of the class-$i$ small cells is defined as the ratio between the amount of acquired data $D_i$ (expressed in bits) and the corresponding energy consumption (i.e., the product of the transmit power $P_i$ and the active time $T_i$). Hence, the overall energy efficiency of connecting to class-$i$ small cells can be computed as

$$G_i = \frac{D_i}{P_i T_i}. \tag{6.43}$$

Note that the last term in (6.42) is the estimated connectivity reliability $\hat{R}_i$ for class-$i$ small cell, which is computed according to a weighted function of the *a-priori* (advertised) connectivity reliability and the *a-posteriori* (perceived) performance. Any control scheme simply averaging the instantaneous samples over a certain time interval may not be suitable here due to high radio channel fluctuations. This is why an EWMA (exponentially weighted moving average) linear predictor is employed, which takes at its input both the advertised probabilistic reliability and the real perceived performance provided as the output of our designed system-level evaluation tool.

To better follow the performance variations, the EWMA is enhanced with a simple heuristic functionality to detect both persistent and spurious performance variations in the time series. In particular, a *level-shift* is a type of non-stationarity causing a significant and sudden change in the mean value of the observed time series. An *outlier* is a measurement that is significantly different from others, beyond the typical level of statistical variations (e.g., due to channel fading). The way to control the level-shifts – after they are detected – is to restart the predictor by ignoring all of the previous history, whereas the outliers may as well be ignored.

At the initial state of the time series in question (when no history exists, that is, at $t_0 = 0$), the only information for deciding on the connection reliability is the a-priori connectivity estimates $b_i$ produced by our system-wide modeling framework in Section 6.1.2 and advertised to the new users via the cellular assistance mechanisms. Then, with every $t$-th new content acquisition connection, a new a-posteriori sample

[2] Note that the value of $\eta$ similarly to the value of $\varphi$ in (6.41) may be derived according to the customer profiles and network marketing categories.

is registered so that the estimated reliability $\hat{R}_i$ could be updated. To this aim, is defined a $\delta_k$ to be the outcome of the $t$-th session in terms of the proportion of the successfully acquired data.

The sample $\delta_k$ is considered to be an outlier, if it differs from the median of $j$ past samples by more than the threshold value $FT$ (feasibility threshold). If the above condition is met several times consecutively, then the outlier is considered to be a level-shift. Accordingly, the estimated connectivity reliability $\hat{R}_i$ for the class-$i$ small cell is calculated as

$$\hat{R}_i(t) = \begin{cases} b_i & t = t_0 \\ \omega\delta_k + (1-\omega)\hat{R}_i(t-1) & t > 1, \end{cases} \tag{6.44}$$

where $\omega \in [0, 1]$ is a weighting factor to offer higher or lower importance to the a-priori estimated reliability.

### 6.1.4 System-wide numerical evaluation

In this section, is reported a system-wide performance characterization that brings together (i) the analytical mobility-centric framework contributed in Section 6.1.2 and (ii) the user involvement aspects discussed in Section 6.1.3. To this end, the evaluation principles as well as detail the relevant technology solutions are outlined and offer performance insights that the proposed framework brings into the HITL context.

#### Multi-connectivity system setup

**Available connectivity options**: To provide an unbiased perspective on the benefits of the proposed methodology in different contexts, the focus is on the system-level investigation on two dissimilar network deployments. The first one corresponds to a contemporary *4G+ setup*, where the macro cellular connectivity is coupled with a static layout of small cells (e.g., femto cells, pico cells, and micro cells) and mobile D2D-ready users. The second deployment option is representative of the future (beyond-)*5G setup*, where small cells may be mobile (e.g., mounted on vehicles), including aerial access points (drone cells), which provide dynamic and readily available multi-connectivity to the customers.

Specifically, is assumed that communication between the users and the mobile small cells (D2D peers, cars, and drones) is facilitated by the proximity services (ProSe) 3GPP function. With this technology, a D2D link (e.g., WiFi-Direct based) may be established between a user and a relevant peer in proximity. Is also assumed that the multi-radio connection setup is managed by the 3GPP LTE infrastructure

that offers assistance for the purposes of device discovery and session management.

**Characteristic mobility models**: To comprehensively assess the effects of realistic user and small cell mobility, representative mobility models and their various combinations based on the class of the small cells are considered. In particular, each model is different in terms of its complexity, internal structure, trajectories, and thus also the number of contacts and contact times even when the same speed is considered. While some of the models initially come from the realm of human mobility, some of them have been adapted here to become representative of mobile network infrastructure.

In more detail, the mobility behavior of human users and D2D relays is characterized with the Levy flight (LF) model. Differently from many conventional mobility models, the LF is able to characterize user movements over larger time spans, where different effects may be experienced. To this aim, multiple short 'runs' within a restricted area interchange with long-distance travels in a random direction, which is characteristic of human mobility in the daily-life situations, such as walking in a shopping mall, moving along urban streets, or attending a sports event in a stadium.

Mobile small cells (mounted on vehicles) move around according to the Manhattan mobility model, which is widely adopted for representing vehicles that drive in urban settings. At each decision point, a vehicle uses a probabilistic selection of its further movement by choosing to continue in the same direction with the probability of 0.5, while turning left or right with equal probabilities of 0.25.

Finally, 3D movements of drone cells follow a modified version of the reference point group (RPG) mobility model. More specifically, in the classical RPG all users are grouped together and each of them moves after the logical center. To make this formulation more compliant to the drone-based movements, has been assumed that they follow a reference point identified by the zone within the area of interest, where the density of the users is the highest (providing additional capacity and connectivity in the locations where the crowd may cause network overloads).

### Parameters and metrics of interest

In all study cases,the proposed HITL solution that integrates the above MCF and UIF frameworks is compared against the baseline decision strategy, where the users always acquire their desired content from the small cell that offers the highest data rate. The radio access technologies under consideration are the macro LTE cellular service with static small cells (e.g., femto cells), and WiFi-Direct based D2D relays or mobile small cells (e.g., car and drone cells).

**Table 6.2.** Simulation setup and parameters

| Parameter | Value | |
|---|---|---|
| | 4G+ deployment | 5G deployment |
| Area of interest | [300,300]m | |
| # of static small cells | 15 | X |
| # of mobile small cells | X | 50 |
| # of D2D 'cells' | 45 | X |
| # of drone cells | X | 25 |
| # of users | 100 | |
| Height, static small cells | 5 | X |
| Height, mobile small cells | X | 1.5 |
| Height, D2D 'cells' | 1.5 | X |
| Height, drone cells | X | min = 10m<br>max = 15m |
| TH, macro cell | 10 Mbps | |
| TH, static small cells | 20 Mbps | X |
| TH, mobile small cells | X | 30 Mbps |
| TH, D2D 'cells' | 40 Mbps | X |
| TH, drone cells | X | 54 Mbps |
| Price, macro cell[1] | 10$ per connection<br>0.05$ per Kbyte | |
| Price, static small cells[1] | 5$ per connection<br>0.025$ per Kbyte | X |
| Price, mobile small cells | X | 15$ per connection<br>0.075$ per Kbyte |
| Price, D2D 'cells' | 20$ per connection<br>0.1$ per Kbyte | X |
| Price, drone cells | X | 25$ per connection<br>0.125$ per Kbyte |
| $P_{tx}$, macro cell | 43 dbm | |
| $P_{tx}$, static small cells | 30 dBm | X |
| $P_{tx}$, mobile small cells | X | 23 dBm |
| $P_{tx}$, D2D 'cells' | 23 dBm | X |
| $P_{tx}$, drone cells | X | 23 dBm |
| Coverage, static small cells | 100m | X |
| Coverage, mobile small cells | X | 50m |
| Coverage, D2D 'cells' | 40m | X |
| Coverage, drone cells | X | 75m |
| Simulation time | 1 hour (3600 s) | |
| # Runs | 300 | |

[1] AT&T: https://www.att.com/shop/wireless/plans/planconfigurator.html

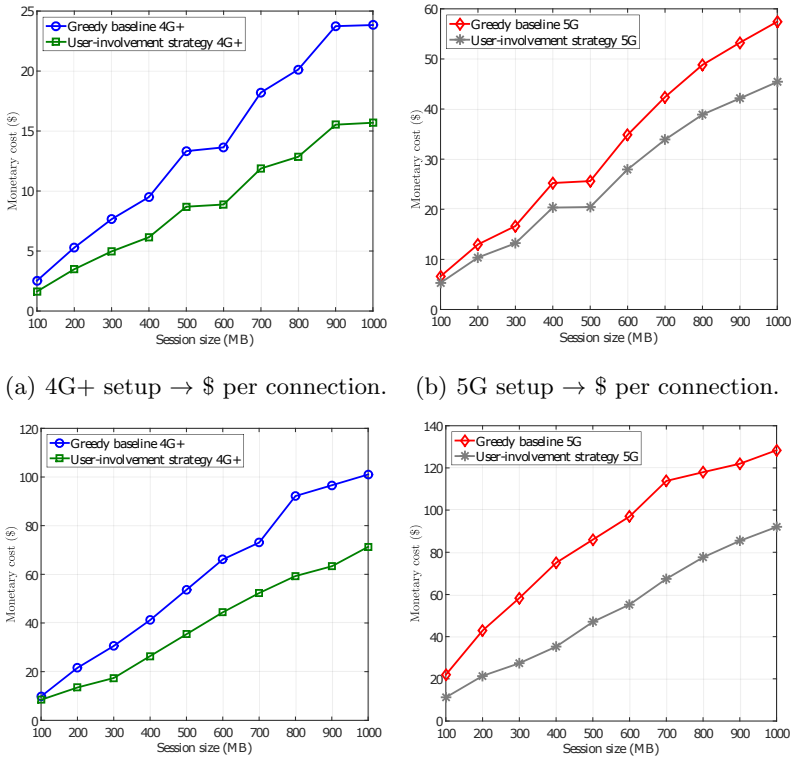
[*] TH = target throughput, $P_{tx}$ = transmit power.

The performance indicators considered are: (i) *energy efficiency*, (ii) *monetary cost* (either per connection or per Kbyte), and (iii) *impact of connectivity options*. In particular, the presented results have been obtained by varying the size of the content acquisition session in the range [100-1000] Mbytes to cover a broad scope of prospective applications.

The users are allowed to move freely within a given area of [300,300]m according to the LF mobility model with the $\alpha$-factor of 1.5. For the "4G+" deployment, the

dense setup is represented by one macro LTE cell for basic coverage, 15 static femto cells, and 75 D2D relays. In the ultra-dense "5G" deployment, in addition to basic macro LTE coverage, other mobile connectivity options are 50 car cells and 25 drone cells.

The car cells move according to the Manhattan mobility model with the average speed of 45 km/h, whereas the drone cells move in 3D space by following our modified RPG mobility model with the average speed of 10 km/h. Each connectivity class has its own settings for available bandwidth, transmit power, height, provided coverage, and monetary cost for offering connectivity. All of these many parameters are conveniently collected in Table 6.2.




(a) 4G+ setup → $ per connection. (b) 5G setup → $ per connection.

(c) 4G+ setup → $ per Kbyte. (d) 5G setup → $ per Kbyte.

**Fig. 6.4.** Average cost to be paid by the users.

### Some representative results

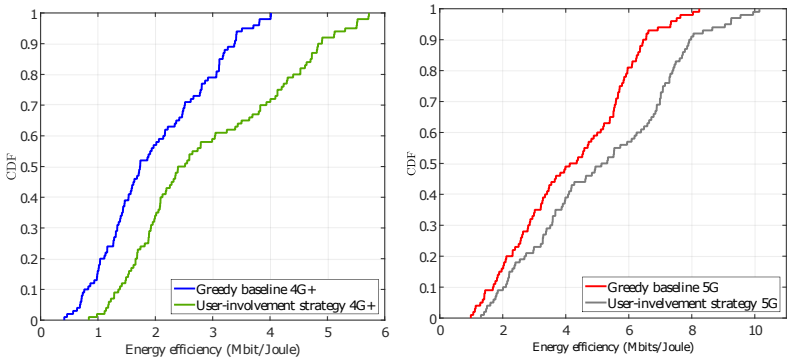
First, the focus was on the average monetary cost for the users to acquire the content (see Fig. 6.4). The actual costs comprise the basic price that is charged for using the conventional macro LTE service plus additional costs arising from the usage of different classes of small cells across the deployment (both operator- and user-owned). Overall, the conclusion is that the average price paid by the users grows linearly with

**Table 6.3.** Maximum savings for the customers

| | Price per connection | Price per Kbyte |
|---|---|---|
| "4G+" deployment | 30$ | 8$ |
| "5G" deployment | 37$ | 12$ |

the amount of data that they acquire. Whether considering the price per Kbyte or the price per connection, the considered HITL approach exceeds the greedy baseline strategy significantly.

This highlights how the developed UIF (as captured by (6.42) of Section 6.1.3) is able to help users in performing "smarter" decisions regarding the best connectivity options. In doing this, an important component is the nominal price offered by each class of the small cells for utilizing their connectivity. Arguing that the price per Kbyte is a more viable market strategy for the service providers, while the price per connection is more favorable for the end-users, Table 6.3 summarizes the maximum amounts of money that the users can save by following our proposed HITL framework.

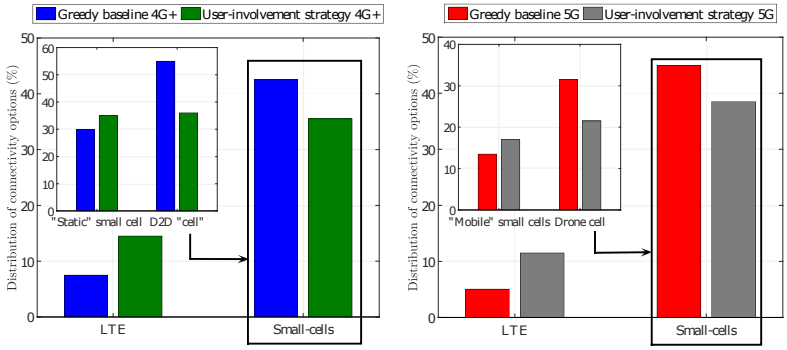



(a) 4G+ deployment. (b) 5G deployment.

**Fig. 6.5.** Energy efficiency.

Next, the focus was on the average energy efficiency achieved by the users when acquiring their desired data versus the variable content size. As observed in Fig. 6.5, the proposed HITL solution outperforms the baseline greedy approach where the users attempt to maximize their perceived throughput. For the "4G+" scenario the average gains are about 56%, whereas in case of the "5G" scenario the benefits increase up to 85%. The explanation behind these results is that with the HITL solution the users obtain comprehensive knowledge about the system and thus are able to adjust their connectivity options to the networking and monetary preferences.

Finally, Fig. 6.6 elaborates on the distribution of connectivity options for a heavy 1-Gbyte content acquisition session. It is remarkable that the users prefer to receive data via one of the classes of small cells rather than directly from the macro LTE




(a) 4G+ deployment. (b) 5G deployment.

**Fig. 6.6.** Distribution of connectivity options for 1 Gbyte session size.

class. This is because both mobile access infrastructure nodes and user devices are well-incentivized to share their connectivity, thus improving the data rates and energy efficiency at affordable costs. Also, Fig. 6.6 reveals that simply selecting the best performing connection in terms of throughput is not always desirable. For instance, in both considered scenarios the greedy user behavior translates into higher prices that need to be paid as well as into lower energy efficiency experienced (hence, shorter battery life).

## 6.2 D2D- and Drone-Assisted Mission-Critical MTC in Multi-Connectivity 5G Scenarios

Mission-critical machine-type communications (mcMTC) are starting to play a central role in the industrial Internet of Things ecosystem and have the potential to create high-revenue businesses, including intelligent transportation systems, energy/smart grid control, public safety services, and high-end wearable applications. Consequently, in the fifth generation (5G) of wireless networks, mcMTC have imposed a wide range of requirements on the enabling technology, such as low power, high reliability, and low latency connectivity. Recognizing these challenges, the recent and ongoing releases of the Long-Term Evolution systems incorporate support for low-cost and enhanced coverage, reduced latency, and high reliability for devices at varying levels of mobility. In this line of research, have been examined the effects of *heterogeneous* user and device mobility – produced by a mixture of various mobility patterns – on the performance of mcMTC across three representative scenarios within a multi-connectivity 5G network. Further, it was established that the availability of alternative connectivity options, such as device-to-device (D2D) links and drone-assisted access, helps meeting the requirements of mcMTC applications in a wide range of scenarios, including industrial automation, vehicular connectivity, and urban communications.

### 6.2.1 Towards A Converged 5G-IoT Ecosystem

#### MTC Requirements and Challenges in 5G Standardization

The rapid proliferation in numbers and functionalities of IoT devices has meant that the standards community is decisively advancing to outline the novel 5G mobile technology. To this end, the vision for the future development of International Mobile Telecommunications (IMT) and beyond was published [127], which presents the overall objectives and requirements for such next-generation systems. That document introduces three broad classes of usage scenarios with very different performance requirements: (i) enhanced mobile broadband, (ii) massive machine-type communications, and (iii) ultra-reliable and low-latency communications.

The 3GPP is eagerly responding to this initiative by starting to ratify a new, non-backward-compatible radio technology in centimeter- and millimeter-wave spectra, and the early commercial deployments of this *new radio* technology are planned for 2018-2020. It is expected that 3GPP's *new radio* will be accompanied by further LTE evolution in parallel. Recognizing the benefits of cellular networks built around a global standards suite, the work in 3GPP includes technology components such as LTE Wi-Fi Link Aggregation (LWA), Licenses Assisted Access (LAA), D2D communications to support smart phone relaying for wearables, power saving for MTC devices, MTC service enabling layers (oneM2M), as well as support for low-throughput and low-complexity MTC devices, realized both as a new LTE UE category (Cat-M1) and a new Narrowband IoT (NB-IoT) radio interface in LTE Release 13 [128]. For MTC, these developments primarily mean a clear distinction between "massive" and "critical" usage scenarios, even though certain IoT applications may simultaneously belong to both categories (for example, critical industrial alarms).

Given the decisive past progress in 3GPP to support MTC requirements (which started as early as in 2005), and the ongoing efforts to define the NB-IoT radio interface, many massive MTC usage scenarios can already be accommodated by existing LTE releases. However, the enhancements promised by the *new radio* are needed in order to enable large-scale deployments of mission-critical MTC (mcMTC) applications, as their main demands are aligned along the lines of mobility (with speeds of up to 500 km/h) and latency (with end-to-end delays of under 1 ms) [127]. This support is crucial in order to promptly leverage the rich business opportunities around mcMTC as part of 5G landscape, but it also poses many important system design questions.

In order to support more reliable consumer and industrial IoT applications [129], we envision that leveraging and integrating across the available heterogeneous access options, such as multi-radio uplink, downlink, and direct D2D link, will be crucial.

The latter is particularly attractive, as the distinction between the network and the user equipment is becoming blurred, which offers excellent opportunities to utilize specialized UE as part of increasingly complex network tasks. This trend goes hand in hand with improved degrees of intelligence in the networked devices, from sensors, wearables, and UE to connected cars and mobile robots, which require very different levels of support for mobility, reliability, and spectrum management. As cooperation between network and user equipment is becoming essential to improve performance of future IoT applications, the impact of more frequent handovers becomes a growing concern.

The promising market of connected – and, soon, self-driven – cars imposes unprecedented requirements in the form of extreme latency and reliability of data delivery at very high-speed mobility. This is particularly challenging given the fundamental fact that higher mobility naturally contradicts better reliability. As the respective operational models in the emerging vehicle-to-everything (V2X) business are taking shape, there is the need of a comprehensive set of tools to handle the unconstrained mobility. This demand is particularly pressing since the impact of mobility has not been revisited in network architectures for a decade or so and now – as we are entering a new era of converged 5G-IoT – is an appropriate time to understand and analyze the various implications of mobility on system performance, as well as to possibly rethink the ways of managing it in 5G networks.

**Representative 5G-grade mcMTC scenarios**

Today, the landscape of the global consumer and industrial IoT business is already extremely broad, stretching from wearable fitness trackers and health care devices to consumer electronics and connected cars. The most challenging study cases emerge in the form of crowded urban scenarios with very high connectivity demands, possibly under unreliable network coverage [130]. In addition to this, in environments with high-speed unrestricted mobility, the availability and reliability of wireless link are of primary importance to ensure strict service-level agreements in new markets around 5G-grade applications and services.

To address the performance of these connected machine-centric networks, the relevant mixtures of realistic mobility models and study their effects on the availability and reliability metrics under partial cellular network coverage have been evaluated. In particular, the focus was on three reference study cases that constitute representative 5G-grade mcMTC scenarios with very diverse application requirements, see Fig. 6.7.

*Industrial automation (CASE A).* Factories of the future will be something more than the standalone "connectivity islands". There is, in fact, an ongoing trend to connect them as part of a broader industrial ecosystem. Accordingly, the typical mobility

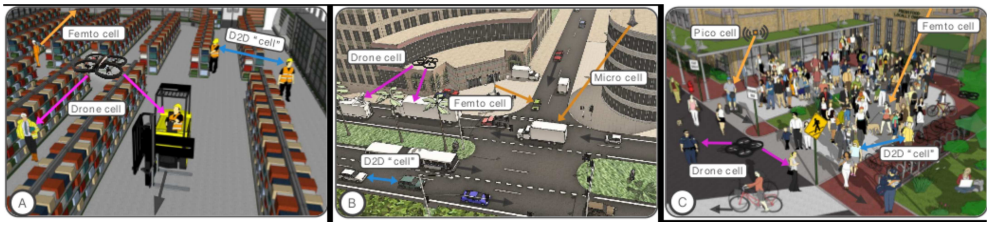



**Fig. 6.7.** Characteristic 5G-grade IoT study cases

aspects related to the supply chain processes within the factory itself or in proximity to its buildings have been considered. The focus was on time constrained communications for the management of assets and goods as part of the on-site production and logistics sectors. This scenario becomes of interest since reliable management of the entire supply chain is crucial to avoid faults and improve the overall factory automation efficiency[3].

*Vehicular connectivity (CASE B).* Communications in the V2X study case comprise data exchange between a connected vehicle and: (i) another vehicle (i.e., vehicle-to-vehicle – V2V); (ii) a road infrastructure (i.e., vehicle-to-infrastructure – V2I); (iii) a personal device moving with pedestrian speeds (i.e., vehicle-to-pedestrian – V2P). Here, the transmitted information can be periodic messages such as speed, positioning, and time related data needed to support critical safety and best effort entertainment applications, as well as offer efficient and comfortable driving experience. In this context, D2D interaction and "mobile" access points, including drones, may be of particular interest to achieve higher communications reliability and improved connection availability.

*Urban communications (CASE C).* This study case covers a set of practical situations where a very large number of mobile end users, potentially carrying several wearable devices, are crowded together in locations such as stadiums, shopping malls, open air festivals, and other public events. The network infrastructure should in these cases be ready to accommodate high densities of active users and their connected devices with large amounts of aggregated traffic. Consequently, the key challenge in these environments is the provision of relatively reliable and available wireless connections to people moving according to certain pedestrian patterns and, likely, crossing areas with partial connectivity from the pre-deployed infrastructure network.

[3] Note that connectivity between the factory entities (e.g., robots, sensors, vehicles, workers) does not necessarily require ultra-low latencies, as response times are typically less constrained for humans than for machines.

### 6.2.2 Mobility-Centric Perspective on mcMTC

#### Multi-Connectivity System Setup

**Available Connectivity Options**: In the subsequent evaluation, is adopted the *legacy LTE* solution as a benchmark where cellular infrastructure serves the mcMTC devices of interest. In addition, proximity-based D2D communications between the involved devices and drone-assisted mobile access are considered to augment the connectivity experience. This set of technologies, referred to as ProSe-based LTE solution, leverages on D2D links whenever mcMTC devices have an opportunity to establish them in proximity (assuming a partner with the desired content), to improve the chances of reliably acquiring the relevant data.

In particular, drones that carry radio transceivers (i.e., drone small cells) are essentially *mobile* access points that provide better network coverage and bring higher data rates to the challenging locations where LTE layout may be under-provisioned. Further, is assumed that the D2D connection setup is managed by the LTE infrastructure for device discovery, session continuity, and security arbitration, whereas Wi-Fi Direct links in unlicensed spectrum are selected as the actual D2D technology. Ultimately, mcMTC connectivity is considered to be *unreliable* in the situations when: 1) the device is outside of cellular LTE coverage; 2) it has no opportunity to establish a D2D link with a relevant partner; and 3) the device cannot be served by a neighboring drone small cell.

**Characteristic Mobility Models**: To comprehensively assess the effects of realistic mobility in the three mcMTC study cases, four mobility models and their heterogeneous combinations have been considered. These models have been carefully selected to capture both short- and long-term time scales of mobility as appropriate for the chosen study cases. While some of the models originally come from the realm of human mobility, some of them have been adapted to become representative of mcMTC moving patterns.The selected models are summarized in in Table 6.4 and highlight their main features.

First, with the Random waypoint (RW) model, the devices move from their current to the new target position randomly by appropriately choosing their speed and travel direction; this behavior aims to capture the short-term mobility on the scale of tens of minutes. Further, the Levy Flight (LF) model is able to mimic movement patterns over a larger time span where mixed effects may be experienced. Another consideration is the Manhattan model that is widely used to follow the mobility of vehicles in urban settings. Finally, the Reference Point Group (RPG) model is particularly suitable to track the mobility of drones. To make it compliant with our scenarios, is assumed that the drones follow a reference point, which is identified by the zone within the

**Table 6.4.** Utilized mobility models

| Mobility model | Corresponding application | Brief description |
|---|---|---|
| Random Walk | Short time scale movement of humans and vehicles | The random travel direction is uniformly distributed in $[0, 2\pi]$. The speed follows the distribution between the boundary values. After a constant time interval, a node computes new direction and speed for future movement. |
| Levy Flight | Long time scale movement of humans and vehicles | Multiple short "runs" within a restricted area are interchanged with long-distance travels in a random direction. |
| Manhattan | Movement of vehicles and pedestrians in urban environment | At each cross-road intersection, a node chooses to continue in the same direction with probability of 0.5, while turns left or right with equal probabilities of 0.25. |
| Reference | Mobility of drones | Models group behavior of a set of nodes where each one Point Group follows the logical center (identified by the *group leader*). The nodes additionally have their own short time-scale Random Walk mobility within the group. |

area of interest where the density of users is the highest. This setup allows the drone small cell to provide additional capacity and coverage in locations where large user densities may cause congestion and network overloads.

**Deployment Parameters**: Three mcMTC scenarios are considered that reflect the industrial automation ($A$), vehicular connectivity ($B$), and urban communications ($C$) applications. In all study cases, the concerned devices acquire information over the link that offers them the highest data rate. The following access technologies are compared: the legacy LTE cellular, Wi-Fi Direct for the D2D links, and millimeter-wave (mmWave) over licensed operator bands for drone small cells. For the mmWave technology, are selected the 28 GHz frequencies as a viable candidate for the 5G *new radio* where the channel propagation, building penetration, and reflection parameters are adopted from [131].

In the three scenarios, is assumed that the LTE coverage is *partial* within the modeled area, which corresponds to when the network is either under-provisioned (e.g., in rural regions) or serves challenging environments (e.g., with obstacles for signal propagation, such as walls, in the basements, etc.). Therefore, is considered that reliable cellular connectivity for the mcMTC devices in all study cases is only available over about 70% of the total area of interest based on deterministic modeling, since network coverage may be intermittent at the cell edges and beyond.

Further, the human users and networked mcMTC devices are allowed to move freely within the considered location according to their specific mobility patterns. For CASE $A$, the setup is represented as an indoor/outdoor area of [200,200]m where industrial robots and machines (i.e., in the indoor part) are first deployed uniformly and then move around within a range of two meters at low speeds (i.e., around 1 km/h). The logistics related procedures are carried out by humans and vehicles where the corresponding mobility is modeled according to the RW model.

For CASE $B$, the setting is the area of [500,500]m where connected vehicles drive according to the Manhattan model. For more realistic simulations, it has been also added some background data traffic from pedestrian users. The latter are characterized by the LF mobility with $\alpha$ factor of 1.5. Finally, CASE $C$ represents a crowded urban scenario where vehicles and users that carry a number of wearable devices are initially deployed within the area of [1000,1000]m and then move around. The respective mobility models are the Manhattan model for vehicles and the LF or the RW models for humans (i.e., people prefer LF or RW pattern probabilistically) where the maximum speed of the nodes is limited by 20 km/h. For further information and details on the simulation settings refer to Table 6.5.

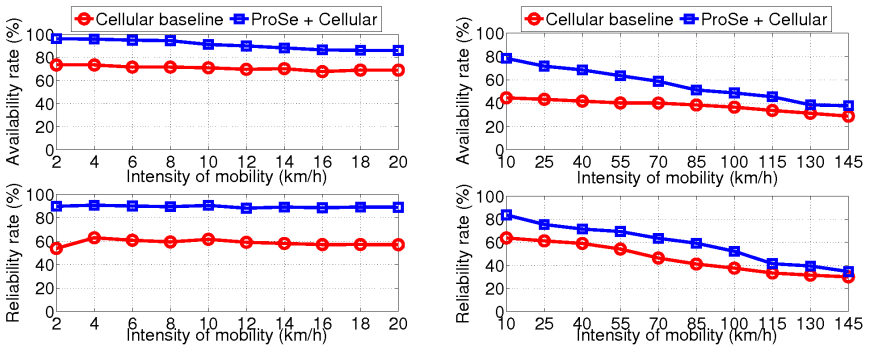
**Table 6.5.** Simulation setup and parameters

| | | CASE $A$ | CASE $B$ | CASE $C$ |
|---|---|---|---|---|
| Application parameter | Amount of data | 300KB | 1500B | 20MB |
| | Inter-arrival time | 10s | 100ms | 1s |
| System parameter | Cell radius | 100m | 250m | 500m |
| | Number of nodes* | 100M/30H/20V | 450V/50H | 300H/650M/50V |
| | Density of nodes | 0.75 node/$m^2$ | 1.0 node/$m^2$ | 1.0 node/$m^2$ |
| | Mobility model | RW/Manhattan | Manhattan/LF | Manhattan/LF/RW |
| | Number of drones | 5 | 10 | 10 |
| | D2D range | 50m | | |
| | D2D link setup | 1s | | |
| | D2D target data rate | 40Mbps | | |
| | Drone altitude | [10-20]m | | |
| | Drone speed | 10km/h | | |
| | Drone mobility | RPG | | |
| | Simulation time | 30 minutes | | |
| | Number of simulation runs | 1000 | | |

* M = Machines, H = Humans, V = Vehicles.

### Selected Numerical Results

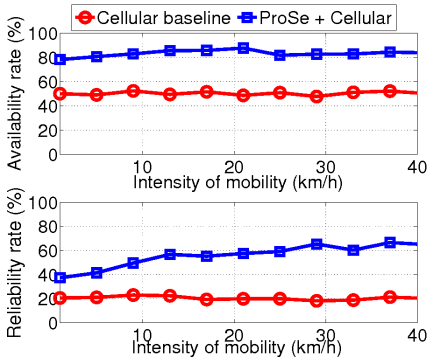
The reported performance assessment has been conducted with a custom-made simulator, named WINTERsim[4]. The main objective of this system-level analysis is to reveal the effects of heterogeneous device mobility in the 5G-grade mcMTC scenarios as well as to quantify the contributions of various multi-connectivity options to the overall communications reliability.




(a) Industrial automation.




(b) Vehicular connectivity.



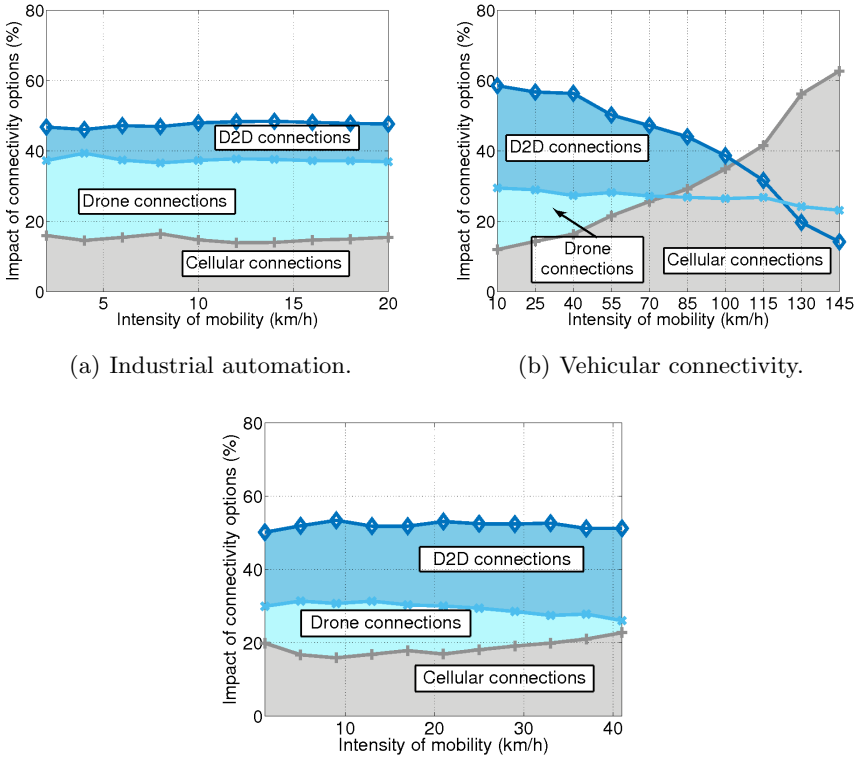

(c) Urban communications.

**Fig. 6.8.** Analysis of system performance in terms of availability and reliability rate as a function of the average device speed in the considered study cases.

Hence, the output metrics of this evaluation are: (i) the *availability rate*, that is, the proportion of users that experience certain connectivity, even though successful acquisition of all the desired data may not be guaranteed; (ii) the *reliability rate* defined as the actual data acquisition probability with which the mcMTC devices are able to successfully receive their data of interest; and (iii) the *impact of connectivity options* characterizing the relative shares of different multi-connectivity links, including cellular-, D2D-, and drone-based alternatives. With this system-level analysis, it is also possible to evaluate other metrics of interest, such as the number of handovers

[4] WINTERsim system-level simulator: http://winter-group.net/downloads/ [Accessed on 08/2016]

between the available connectivity options, the handover delay, and the signaling load caused by unnecessary handovers.




(a) Industrial automation.

(b) Vehicular connectivity.

(c) Urban communications.

**Fig. 6.9.** Impact of available radio access technologies on overall connectivity. The vertical axes display the contribution of each connectivity option.

The availability and reliability rates in the three scenarios under investigation have been simulated over a period of 30 minutes, and are summarized in Fig. 6.8. As is possible to learn from these curves, for CASE $B$ higher mobility speeds affect the system-wide performance considerably. By contrast, in case of low mobility, the results do not vary dramatically, which holds for CASES $A$ and $C$. For the vehicular scenario, D2D communications and drone small cells demonstrate diminishing benefits with the growing intensity of mobility (i.e., 100 km/h and beyond). However, the ProSe-based solution still offers consistent improvements on top of the legacy LTE baseline in all of the study cases. In particular, the gains in terms of the data acquisition rate vary from 25% to 35% for CASES $A$ and $C$, as well as from 5% to 40% for CASE $B$. With respect to the reliability rate, is evident an increase of 25% and 20% on average when considering CASES $A$ and $B$, whereas the improvements for the CASE $C$ reach 40%.

The impact of alternative connectivity options in the studied scenarios is reported in Fig. 6.9. Interestingly, is possible to conclude that D2D connections are utilized the most for mcMTC data acquisition. The explanation behind this fact is in the

large number of potential contact opportunities for proximate users. However, with the growing intensity of mobility, the number of feasible contacts drops and hence contact duration becomes the dominant factor that determines the chances of receiving the relevant content successfully. A similar trend is observed in Fig. 6.9(b) where at the speeds of above 85 km/h the devices prefer – by attempting to maximize their throughput – the more stable cellular LTE connections to any proximate links. In contrast, the impact of mobility is not as severe in Fig. 6.9(a) and Fig. 6.9(c) where D2D- and drone-based links are used more often than the infrastructure-based connections.

In summary, for the scenarios with low (CASE *A*) and limited (CASE *C*) mobility, link availability and reliability may not be affected dramatically by the device mobility. However, this situation could change for other types of similar mcMTC applications having different packet sizes [22]. In these study cases, exploiting D2D- and drone-assisted communications leads to a significant improvement in the data acquisition rates as well as brings along higher reliability. On the contrary, in the very different vehicular scenario where the intensity of mobility is typically higher (CASE B), is observed a considerable system-level performance degradation at the speeds of above 85 km/h. This is because proximity-based communications and drone small cells gradually lose their efficiency to provide additional capacity and coverage, while the only viable alternative remains to acquire data through the cellular infrastructure.

# 7

# Conclusions

The aim of this thesis has been to address the ambitious challenge of delivering uniform connectivity and enhanced quality of experience to both human- and machine-type proximate users by taking into account the rapid growth of bandwidth-hungry application, service, and device volumes. To cope with these ambitious goals, since existing wireless deployments are unable to deliver the desired ubiquitous connectivity experience due to the shortage of available capacity and the lack of service uniformity, the integration of device-to-device (D2D) communications into the emerging cellular systems is envisioned to enable non-incremental performance improvements and increased levels of cooperation. The proposed solutions in the form of communication algorithms, system architectures, and performance evaluation frameworks may therefore become of significant value toward native integration of direct short-range communications into the fifth-generation (5G) system and architecture. The overall research performed on D2D, indeed, promises both theoretical and practical innovations in order to enhance future wireless network connectivity and, as a consequence, augment end-user services experience. In addition, the objectives and the breakthrough targets of my work are aligned with the key strengths and demands of the research environment in Europe as well as internationally.

Going into details, the goals of this thesis have been a rich set of innovations in order to (i) improve the wireless network connectivity, (ii) increase the perceived satisfaction and Quality of Experience (QoE) of the users, and (iii) deliver new 5G-grade broadcast and multimedia services with high-data rate and low delays. In order to cope with these goals, i believe that emerging concept of D2D communications has to be further comprehensively explored. Such a concept is driven by the understanding of the mobility effects in D2D-based scenarios, combining caching of files in the user devices with D2D communications, and accounting for the interplay between the social sphere and the communication properties in the D2D communications systems. During my research activities, I have found that these directions will significantly increase the degrees of connectivity and service experience in the face of growing

application, multimedia service, and device volume. The complex research aimed at by this study converged in both theoretical innovations and practical applications, as the topic itself leads to a rethinking of the architecture of contemporary wireless networks. In particular, the overall research within the context of D2D communications targeted at the following four major objectives:

**Mobility in D2D-based Scenarios**: In summary, I investigated the effects of mobility given by the user movement patterns and other factors, such as the type of application running on top of the D2D links, that are envisioned to have a dramatic impact on the resulting performance. The motivation behind this activity, it was that this important research direction was insufficiently addressed and existing literature falls short of quantifying the impact of mobility on proximal communication. In particular, my current focus was on system-wide performance evaluation results that, nowadays, have not yet been well investigated to understand how the D2D operation reacts to frequent and opportunistic contacts due to realistic user mobility. As the result, I found that the output of conventional models and architecture provides inadequate insight into the effects of mobility in characteristic D2D-based scenarios.

**D2D-aware Content Dissemination**: Contemporary consumer electronics is spawning an explosion of advanced multi-radio devices with the consequent rapid growth of bandwidth-hungry multimedia services. Considering such proliferating user equipment, the past concepts of cellular traffic offloading and content dissemination are not able to fulfill the stringent requirements of "anywhere" and "anytime" connectivity. To overcome this issues, the aim of my research in this field was to exploit innovative solution where the usage of the D2D connections and "D2D Cachers" play a pivotal role. What pushed to investigate these aspects was the relatively limited research attention dedicated to the device caching and the usage of devices as possible "D2D Cachers" in an asynchronous content dissemination scenario. In fact, in most of the published works, much effort has been invested into optimizing the performance of individual radio technologies. The results on this topic, showed that both short- and long-range technologies need to work cooperatively to realize the desired uniform user experience in 5G networks.

**Social and "Secure" D2D Communications**: Regarding the scope of the D2D technology with social awareness and security procedures, it might be useful to recall that D2D communications have been recently considered as enabler for a kind of "wireless sense" or 6th sense" that would facilitate social proximity services. Indeed, it was extended the notion of "wireless sense" in order to increase also the "awareness" in order to facilitates the development of cognitive and self organizing networks (SON) employing artificial intelligence and learning schemes. However, new algorithm and procedures for increasing the safety not only of the network infrastructure, but

also of the devices are crucial for the native support of D2D into the forthcoming 5G systems. For this reason, once that the integration of social aspects in wireless network has been consolidated, a join use of proximity services and social relationships for enhancing security and privacy mechanisms has been also addressed.

**D2D Integration Into 5G-Grade IoT Scenarios**: During this line of research, the potentialities of D2D communications for the Internet of Things have been investigated. A broad overview of ongoing research and standardization activities for D2D communications technology in future generation systems have been given. Further, particular attention has been devoted to possible use cases (e.g., factory automation, vehicle-to-everything, and urban dense scenarios) and benefits this technology may introduce to meet the manifold key requirements and open issues in the future 5G IoT ecosystem. In addition, a look into the novel and futuristic visions (i.e., the usage of drone/vehicles as a part of the network infrastructure and the incentivation of the user through the human-in-the-loop paradigm) of the IoT have been proposed and tested with system-level simulation. This highlighted the manifold challenges ahead of us and research directions that need further investigation to realize the full convergence of D2D and IoT in next-to-come 5G systems, where a device-oriented Anything-as-a-Service ecosystem is expected to be the reality.

A fundamental aspect that was the main boost for looking at the native integration into the forthcoming 5G systems, was that network base stations are becoming comparable to the modern devices with increasing computational power (i.e., smartphone, tablet, etc.). In fact, there is a convergence in performance of all the network "actors" (i.e., base stations and devices). Given this unprecedented device and network evolution, the differences between devices and network nodes will become marginal soon. As a natural consequence, is expected that it will become unclear what to define as "base station" and "device". For these reason, the research conducted during this thesis considered devices acting as a D2D relay for an out of coverage device to forward data and service. In addition, it was assumed that future devices effectively become part of the service deployment and the infrastructure rather than being a classical "user" that receives cellular service. The results obtained by combining the above two aspects led to a socially aware, self organizing network-assisted D2D infrastructure, in which there are interesting business object relationships between the nodes (i.e., base stations and devices).

In conclusion, this thesis aimed to integrate both fundamental and applied research, as well as relied on extensive simulation studies to quantify the ultimate D2D performance gains. Firstly, analytical techniques have been exploited in order to provide a complete set of models, frameworks, schemes, and algorithms useful for the purpose of integration D2D into 5G systems. In particular, well-known and

novel methods from Theory of Probability, Statistics, Matrix Algebra, Optimization Theory, Queuing Theory, Game Theory, and Theory of Stochastic Processes, have been used to provide a complete characterization of the D2D-enhanced system behavior. Hence, new advanced models to evaluate emerging solutions for improving next-generation D2D connectivity, technology integration, and user experience have been proposed and developed. Secondly, the results obtained with analytical techniques have been validated by performing an exhaustive set of simulations by taking into account numerous factors expected to influence the behavior of the users as well as that of network operators. In particular, these aspects are represented by traffic arrival patterns, user mobility behavior, tight coupling between communicating devices and collocated radio technologies, application service requirements, wireless channel degradation factors, social relationships and behavior, etc. In selected special cases, the simulation results converged with those obtained analytically, which ensures the adequateness of the constructed models.

In addition to this, an advanced and in-depth characterization of D2D mobility patterns within the context of future 5G cellular networks has been also investigated. The starting point was with the investigation of suitable mobility models (i.e., already available in literature) for short-range transmission and then with the addressing of the fundamental issues that mobility could introduce to future D2D communications. After that, the different types of behavior of various mobility models by arriving to an understandable picture with a more applied research have been also took into consideration. Further, the potential of increasing device computing power and storage in order to augment the D2D system capacity and improve user connectivity experience have been also explored. For instance, the analysis of the fundamental limitations of content caching and the corresponding data dissemination protocols in multi-tier networks was evaluated. This includes exploring the limits of stability with practical network scaling, as well as the potential benefits of cooperation between devices and network infrastructure. Importantly, centralized vs. fully distributed control options were investigated to understand the impact of network assistance by contrast to device-centric schemes. Finally, it was conducted also the investigation of the influence of social relationships and behavior by focusing on the combined effect of direct communications, social sphere, advanced wireless services, and incentives for the users. This research started by including the investigation of trustworthy schemes involving friends in the vicinity, social network location-based applications, as well as intrinsic relationships between proximate humans using ”online” and ”offline” social networks. Then, the aim was to address novel metrics, frameworks, and protocols for next generation, social-aware D2D communications.

In summary, the work illustrated in this thesis focused on deep and systematic investigations targeting improved network connectivity, user quality of experience/service, device mobility characterization, and outlining novel uses cases enabled by advanced network-assisted D2D connectivity. The proposed solutions have been disseminated through high technical papers in famous international conferences and journals in the form of communication algorithms, system architectures, and performance evaluation frameworks. In fact, the aim was to provide a significant added value toward the integration of proximity services and short-range transmissions into the future 5G cellular systems. These solutions are primarily intended for, but not limited to, cellular operators, telecommunication research companies, equipment vendors, and mobile software companies thus resulting in the considerable benefits for the entire international community.

# References


1. A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G Mobile and Wireless Communications: the Vision of the METIS Project," *Communications Magazine, IEEE*, vol. 52, pp. 26–35, May 2014.
2. J. Monserrat, H. Droste, O. Bulakci, J. Eichinger, O. Queseth, M. Stamatelatos, H. Tullberg, V. Venkatkumar, G. Zimmermann, U. Dotsch, and A. Osseiran, "Rethinking the Mobile and Wireless Network Architecture: The METIS Research into 5G," in *Networks and Communications (EuCNC), 2014 European Conference on*, pp. 1–5, June 2014.
3. S. Chen and J. Zhao, "The Requirements, Challenges, and Technologies for 5G of Terrestrial Mobile Telecommunication," *Communications Magazine, IEEE*, vol. 52, pp. 36–43, May 2014.
4. J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What Will 5G Be?," *Selected Areas in Communications, IEEE Journal on*, vol. 32, pp. 1065–1082, June 2014.
5. N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network Densification: the Dominant Theme for Wireless Evolution into 5G," *Communications Magazine, IEEE*, vol. 52, pp. 82–89, February 2014.
6. 3GPP, "TS 22.803, Feasibility study for Proximity Services (ProSe), Rel. 12," tech. rep., Jun 2013.
7. E. Hossain, M. Rasti, H. Tabassum, and A. Abdelnasser, "Evolution Toward 5G Multi-Tier Cellular Wireless Networks: An Interference Management Perspective," *Wireless Communications, IEEE*, vol. 21, pp. 118–127, June 2014.
8. S. Mumtaz and J. Rodriguez, "Introduction to D2D Communication," in *Smart Device to Smart Device Communication* (S. Mumtaz and J. Rodriguez, eds.), pp. 1–22, Springer International Publishing, 2014.
9. R. Q. Hu, S. Talwar, and P. Zong, "Cooperative, Green and Mobile Heterogeneous Wireless Networks," in *23rd International Teletraffic Congress (ITC 23)*, 2011.
10. S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular traffic offloading onto network-assisted device-to-device connections," *Communications*

*Magazine, IEEE*, vol. 52, pp. 20–31, April 2014.

11. D. Raychaudhuri and N. B. Mandayam, "Frontiers of Wireless and Mobile Communications," *Proceedings of the IEEE*, vol. 100, pp. 824–840, April 2012.
12. X. Tao, X. Xu, and Q. Cui, "An Overview of Cooperative Communications," *Communications Magazine, IEEE*, vol. 50, pp. 65–71, June 2012.
13. G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, and Z. Turanyi, "Design Aspects of Network Assisted Device-to-Device Communications," *Communications Magazine, IEEE*, vol. 50, pp. 170–177, March 2012.
14. S. Andreev, D. Moltchanov, O. Galinina, A. Pyattaev, A. Ometov, and Y. Koucheryavy, "Network-Assisted Device-to-Device Connectivity: Contemporary Vision and Open Challenges," in *European Wireless 2015; 21th European Wireless Conference; Proceedings of*, pp. 1–8, May 2015.
15. L. Militano, A. Orsino, G. Araniti, A. Molinaro, and A. Iera, "A Constrained Coalition Formation Game for Multihop D2D Content Uploading," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 2012–2024, March 2016.
16. A. Orsino, G. Araniti, L. Militano, J. Alonso-Zarate, A. Molinaro, and A. Iera, "Energy Efficient IoT Data Collection in Smart Cities Exploiting D2D Communications," *Sensors*, vol. 16, no. 6, p. 836, 2016.
17. L. Militano, A. Orsino, G. Araniti, A. Molinaro, and A. Iera, "Overlapping coalitions for D2D-supported data uploading in LTE-A systems," in *Personal, Indoor, and Mobile Radio Communications (PIMRC), 2015 IEEE 26th Annual International Symposium on*, pp. 1526–1530, Aug 2015.
18. M. Condoluci, L. Militano, A. Orsino, J. Alonso-Zarate, and G. Araniti, "LTE-direct vs. WiFi-direct for machine-type communications over LTE-A systems," in *Personal, Indoor, and Mobile Radio Communications (PIMRC), 2015 IEEE 26th Annual International Symposium on*, pp. 2298–2302, Aug 2015.
19. A. Orsino, L. Militano, G. Araniti, A. Molinaro, and A. Iera, "Efficient Data Uploading Supported by D2D Communications in LTE-A Systems," in *European Wireless 2015; 21th European Wireless Conference; Proceedings of*, pp. 1–6, May 2015.
20. L. Militano, A. Orsino, G. Araniti, A. Molinaro, A. Iera, and L. Wang, "Efficient spectrum management exploiting D2D communication in 5G systems," in *2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1–5, June 2015.
21. A. Orsino, G. Weisi, and G. Araniti, "Multi-Scale Mobility Models in the Forthcoming 5G Era: A General Overview," *Submitted to IEEE Vehicular Technology Magazine*, July 2016.
22. A. Orsino, D. Moltchanov, M. Gapeyenko, A. Samuylov, S. Andreev, L. Militano, G. Araniti, and Y. Koucheryavy, "Direct Connection on the Move: Characterization of User Mobility in Cellular-Assisted D2D Systems," *IEEE Vehicular Technology Magazine*, vol. 11, pp. 38–48, Sept 2016.
23. A. Orsino, M. Gapeyenko, L. Militano, D. Moltchanov, S. Andreev, Y. Koucheryavy, and G. Araniti, "Assisted Handover Based on Device-to-Device Communications in

3GPP LTE Systems," in *2015 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, Dec 2015.

24. A. Orsino, A. Samuylov, D. Moltchanov, S. Andreev, L. Militano, G. Araniti, and Y. Koucheryavy, "Time-Dependent Energy and Resource Management in Mobility-Aware D2D-Empowered 5G Systems," Oct. 2016.
25. G. Araniti, A. Orsino, L. Militano, L. Wang, and A. Iera, "Context-aware Information Diffusion for Alerting Messages in 5G Mobile Social Networks," *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–1, 2016.
26. A. Orsino, G. Araniti, L. Wang, and A. Iera, "Multimedia content diffusion approach for emerging 5G mobile social networks," in *2016 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, June 2016.
27. A. Orsino, L. Militano, G. Araniti, and A. Iera, "Social-aware Content Delivery with D2D Communications Support for Emergency Scenarios in 5G Systems," in *European Wireless 2016; 22th European Wireless Conference*, pp. 1–6, May 2016.
28. A. Ometov, A. Orsino, L. Militano, G. Araniti, D. Moltchanov, and S. Andreev, "A novel security-centric framework for D2D connectivity based on spatial and social proximity," *Computer Networks*, vol. 107, Part 2, pp. 327 – 338, 2016. Mobile Wireless Networks.
29. A. Orsino and A. Ometov, "Validating information security framework for offloading from LTE onto D2D links," in *2016 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT)*, pp. 241–247, April 2016.
30. L. Militano, A. Orsino, G. Araniti, M. Nitti, L. Atzori, and A. Iera, "Trust-based and social-aware coalition formation game for multihop data uploading in 5G systems," *Computer Networks*, pp. –, 2016.
31. L. Militano, A. Orsino, G. Araniti, M. Nitti, L. Atzori, and A. Iera, "Trusted D2D-based Data Uploading in In-band Narrowband-IoT with Social Awareness," in *Submitted to Personal, Indoor, and Mobile Radio Communications (PIMRC), 2016 IEEE 27th Annual International Symposium on*, Sep. 2016.
32. A. Ometov, A. Orsino, L. Militano, D. Moltchanov, G. Araniti, E. Olshannikova, G. Fodor, S. Andreev, T. Olsson, A. Iera, J. Torsner, Y. Koucheryavy, and T. Mikkonen, "Toward trusted, social-aware D2D connectivity: bridging across the technology and sociality realms," *IEEE Wireless Communications*, vol. 23, pp. 103–111, August 2016.
33. A. Ometov, A. Levina, P. Borisenko, R. Mostovoy, A. Orsino, and S. Andreev, "Mobile Social Networking under Side-Channel Attacks: Practical Security Challenges," *Accepted to IEEE Access*, Jan. 2017.
34. A. Samuylov, A. Orsino, D. Moltchanov, S. Andreev, L. Militano, G. Araniti, A. Iera, H. Yanikomeroglu, and Y. Koucheryavy, "On Tighter User Involvement in Multi-Connectivity 5G Scenarios with Access Infrastructure Mobility," *Submitted to IEEE Journal on Selected Areas in Communications*, Sep. 2016.

35. A. Orsino, A. Ometov, G. Fodor, D. Moltchanov, L. Militano, S. Andreev, O. Yilmaz, T. Tirronen, J. Torsner, G. Araniti, A. Iera, M. Dohler, and Y. Koucheryavy, "Effects of Heterogeneous Mobility on D2D- and Drone- Assisted Mission-Critical MTC in 5G," *Accepted to IEEE Communications Magazine*, Oct. 2016.
36. O. Galinina, L. Militano, S. Andreev, A. Pyattaev, K. Johnsson, A. Orsino, G. Araniti, A. Iera, M. Dohler, and Y. Koucheryavy, "Demystifying Competition and Cooperation Dynamics of the Aerial mmWave Access Market," *Submitted to IEEE/ACM Transaction on Networking*, Aug. 2016.
37. A. Orsino, I. Farris, L. Militano, G. Araniti, and A. Iera, "D2D Communications for Delay-sensitive IoT Mobile Services over Multiple Edge Nodes," *Submitted to IEEE Internet Computing Magazine*, Jan. 2017.
38. C. of the European Communities, "Exploiting the Employment Potential of ICTs." Staff Working Document, 2012.
39. C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *Communications Magazine, IEEE*, vol. 52, pp. 122–130, February 2014.
40. A. Hashimoto, H. Yoshino, and H. Atarashi, "Roadmap of IMT-advanced development," *Microwave Magazine, IEEE*, vol. 9, pp. 80–88, Aug 2008.
41. Ericsson, "More than 50 Billion Connected Devices." White Paper, 2011.
42. N. S. Networks, "2020: Beyond 4G: Radio Evolution for the Gigabit Experience." White Paper, 2011.
43. F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *Communications Magazine, IEEE*, vol. 52, pp. 74–80, February 2014.
44. D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, "LTE release 12 and beyond [Accepted From Open Call]," *IEEE Communications Magazine*, vol. 51, pp. 154–160, July 2013.
45. X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "FlashLinQ: A Synchronous Distributed Scheduler for Peer-to-peer Ad Hoc Networks," *IEEE/ACM Trans. Netw.*, vol. 21, pp. 1215–1228, Aug. 2013.
46. X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Communications Magazine*, vol. 52, pp. 40–48, April 2014.
47. "IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1," *IEEE Std 802.16e-2005 and IEEE Std 802.16-2004/Cor 1-2005 (Amendment and Corrigendum to IEEE Std 802.16-2004)*, pp. 1–822, 2006.
48. B. Kaufman and B. Aazhang, "Cellular networks with an overlaid device to device network," in *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pp. 1537–1541, Oct 2008.

49. K. Doppler, M. Rinne, C. Wijting, C. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to lte-advanced networks," *Communications Magazine, IEEE*, vol. 47, pp. 42–49, Dec 2009.

50. K. Doppler, M. Rinne, P. Janis, C. Ribeiro, and K. Hugl, "Device-to-device communications; functional prospects for lte-advanced networks," in *Communications Workshops, 2009. ICC Workshops 2009. IEEE International Conference on*, pp. 1–6, June 2009.

51. A. Osseiran, K. Doppler, C. Ribeiro, M. Xiao, M. Skoglund, and J. Manssour, "Advances in device-to-device communications and network coding for IMT-advanced," *ICT Mobile Summit*, 2009.

52. T. Peng, Q. Lu, H. Wang, S. Xu, and W. Wang, "Interference avoidance mechanisms in the hybrid cellular and device-to-device systems," *IEEE 20th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pp. 617–621, Sept. 2009.

53. S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy, "Analyzing assisted offloading of cellular user sessions onto d2d links in unlicensed bands," *Selected Areas in Communications, IEEE Journal on*, vol. 33, pp. 67–80, Jan 2015.

54. J.-B. Seo, T. Kwon, and V. Leung, "Social groupcasting algorithm for wireless cellular multicast services," *Communications Letters, IEEE*, vol. 17, pp. 47–50, January 2013.

55. L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G. M. Muntean, "Single Frequency-Based Device-to-Device-Enhanced Video Delivery for Evolved Multimedia Broadcast and Multicast Services," *IEEE Transactions on Broadcasting*, vol. 61, pp. 263–278, June 2015.

56. S. C. Spinella, G. Araniti, A. Iera, and A. Molinaro, "Integration of ad-hoc networks with infrastructured systems for multicast services provisioning," *Int. Conf. Ultra Modern Telecommunications and Workshops, St.Petersburg, Russia*, pp. 1–6, Oct. 2009.

57. Q. Zhang, F. H. P. Fitzek, and V. B. Iversen, "Design and Performance Evaluation of Cooperative Retransmission Scheme for Reliable Multicast Services in Cellular Controlled P2P Networks," in *2007 IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–5, Sept 2007.

58. L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and F. Fitzek, "Wi-Fi cooperation or D2D-based multicast content distribution in LTE-A: A comparative analysis," in *Communications Workshops (ICC), 2014 IEEE International Conference on*, pp. 296–301, June 2014.

59. H. Nishiyama, M. Ito, and N. Kato, "Relay-by-smartphone: realizing multihop device-to-device communications," *Communications Magazine, IEEE*, vol. 52, pp. 56–65, April 2014.

60. J. da Silva, G. Fodor, and T. Maciel, "Performance analysis of network-assisted two-hop d2d communications," in *Globecom Workshops (GC Wkshps), 2014*, pp. 1050–1056, Dec 2014.

61. L. Lei, X. . Shen, M. Dohler, C. Lin, and Z. Zhong, "Queuing Models With Applications to Mode Selection in Device-to-Device Communications Underlaying Cellular

Networks," *IEEE Transactions on Wireless Communications*, vol. 13, pp. 6697–6715, Dec 2014.

62. G. Rigazzi, F. Chiti, R. Fantacci, and C. Carlini, "Multi-hop D2D networking and resource management scheme for M2M communications over LTE-A systems," in *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 973–978, Aug 2014.
63. X. Lu, P. Wang, and D. Niyato, "A layered coalitional game framework of wireless relay network," *Vehicular Technology, IEEE Transactions on*, vol. 63, pp. 472–478, Jan 2014.
64. P. Pahlevani, M. Hundebll, M. Pedersen, D. Lucani, H. Charaf, F. Fitzek, H. Bagheri, and M. Katz, "Novel concepts for device-to-device communication using network coding," *Communications Magazine, IEEE*, vol. 52, pp. 32–39, April 2014.
65. C.-H. Yu, K. Doppler, C. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for Device-to-Device communication underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, pp. 2752–2763, Aug. 2011.
66. M. Condoluci, L. Militano, G. Araniti, A. Molinaro, and A. Iera, "Multicasting in LTE-A networks enhanced by device-to-device communications," in *Globecom Workshops (GC Wkshps), 2013 IEEE*, pp. 567–572, Dec 2013.
67. A. Pyattaev, O. Galinina, S. Andreev, M. Katz, and Y. Koucheryavy, "Understanding practical limitations of network coding for assisted proximate communication," *Selected Areas in Communications, IEEE Journal on*, vol. PP, no. 99, pp. 1–1, 2014.
68. L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G.-M. Muntean, "Single frequency-based device-to-device-enhanced video delivery for evolved multimedia broadcast and multicast services," *Broadcasting, IEEE Transactions on*, vol. 61, pp. 263–278, June 2015.
69. S. Mumtaz, L.-L. Yang, C. Wang, F. Adachi, and N. Ali, "Smart-device-to-smart-device communications," *Communications Magazine, IEEE*, vol. 52, no. 4, pp. 18–19, 2014.
70. L. Militano, M. Condoluci, G. Araniti, A. Molinaro, and A. Iera, "When D2D communication improves group oriented services in beyond 4G networks," *Wireless Networks*, pp. 1–15, 2014.
71. 3GPP, "TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Rel. 11," tech. rep., Sept. 2012.
72. L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in lte-advanced networks," *Wireless Communications, IEEE*, vol. 19, pp. 96–104, June 2012.
73. 3GPP, "TS 36.213 Evolved Universal Terrestrial Radio Access (E-UTRA): Physical layer procedures, Rel. 11," tech. rep., Dec. 2012.
74. Y. Li, D. Jin, J. Yuan, and Z. Han, "Coalitional games for resource allocation in the device-to-device uplink underlaying cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 13, pp. 3965–3977, July 2014.

75. M. Lin, J. Ouyang, and W. Zhu, "Joint beamforming and power control for device-to-device communications underlaying cellular networks," *Selected Areas in Communications, IEEE Journal on*, 2015.

76. L. Tassiulas and S. Sarkar, "Maxmin fair scheduling in wireless networks," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, pp. 763–772 vol.2, 2002.

77. Y.-F. Liu and Y.-H. Dai, "On the complexity of joint subcarrier and power allocation for multi-user ofdma systems," *Signal Processing, IEEE Transactions on*, vol. 62, pp. 583–596, Feb 2014.

78. C. Tan, T. C. Chuah, and S. Tan, "Adaptive multicast scheme for OFDMA-based multicast wireless systems," *Electronics Letters*, vol. 47, no. 9, pp. 570–572, 2011.

79. J. Liu, W. Chen, Z. Cao, and K. Letaief, "Dynamic power and sub-carrier allocation for ofdma-based wireless multicast systems," in *Communications, 2008. ICC '08. IEEE International Conference on*, pp. 2607–2611, May 2008.

80. D. Lopez-Perez, A. Ladanyi, A. Juttner, H. Rivano, and J. Zhang, "Optimization method for the joint allocation of modulation schemes, coding rates, resource blocks and power in self-organizing lte networks," in *INFOCOM, 2011 Proceedings IEEE*, pp. 111–115, April 2011.

81. W. Saad, Z. Han, M. Debbah, A. Hjorungnes, and T. Basar, "Coalitional game theory for communication networks," *Signal Processing Magazine, IEEE*, vol. 26, pp. 77–97, September 2009.

82. T. Rahwan, T. P. Michalak, E. Elkind, P. Faliszewski, J. Sroka, M. Wooldridge, and N. R. Jennings, "Constrained coalition formation," *25th Conference on Artificial Intelligence (AAAI)*, pp. 719–725, 2011.

83. D. E. Knuth, *The Art of Computer Programming, Volume 4, Combinatorial Algorithms, Part 1.* Addison-Wesley Professional, 2011.

84. T. Sandholm, K. Larson, M. Andersson, O. Shehory, and F. Tohmé, "Coalition structure generation with worst case guarantees," *Artificial Intelligence*, vol. 111, no. 1, pp. 209–238, 1999.

85. K. R. Apt and A. Witzel, "A generic approach to coalition formation," *International Game Theory Review*, vol. 11, no. 03, pp. 347–367, 2009.

86. A. Bogomolnaia and M. O. Jackson, "The stability of hedonic coalition structures," *Games and Economic Behavior*, vol. 38, pp. 201–230, 2002.

87. L. Mashayekhy and D. Grosu, "A merge-and-split mechanism for dynamic virtual organization formation in grids," in *Performance Computing and Communications Conference (IPCCC), 2011 IEEE 30th International*, pp. 1–8, Nov 2011.

88. W. Saad, Z. Han, M. Debbah, and A. Hjorungnes, "A distributed coalition formation framework for fair user cooperation in wireless networks," *Wireless Communications, IEEE Transactions on*, vol. 8, pp. 4580–4593, September 2009.

89. L. Mashayekhy, M. Nejad, and D. Grosu, "Cloud federations in the sky: Formation game and mechanism," *Cloud Computing, IEEE Transactions on*, vol. 3, pp. 14–27, Jan 2015.

90. X. Cheng, C. Dale, and J. Liu, "Statistics and Social Network of YouTube Videos," in *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pp. 229–238, June 2008.
91. ISO, "ISO/IEC 13818-2:2013, Information technology - Generic coding of moving pictures and associated audio information: Video," tech. rep., July 2014.
92. C. Mehlfuhrer, M. Wrulich, J. Ikuno, D. Bosanska, and M. Rupp, "Simulating the long term evolution physical layer," in *Signal Processing Conference, 2009 17th European*, pp. 1471–1478, Aug 2009.
93. 3GPP, "TS 36.101, LTE Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception, Rel. 10," tech. rep., June 2011.
94. M. Iturralde, T. Yahiya, A. Wei, and A. Beylot, "Interference mitigation by dynamic self-power control in femtocell scenarios in lte networks," in *Global Communications Conference (GLOBECOM), 2012 IEEE*, pp. 4810–4815, Dec 2012.
95. K. Wang, J. Alonso-Zarate, and M. Dohler, "Energy-efficiency of lte for small data machine-to-machine communications," in *Communications (ICC), 2013 IEEE International Conference on*, pp. 4120–4124, June 2013.
96. M. Franceschetti, "When a Random Walk of Fixed Length can Lead Uniformly Anywhere Inside a Hypersphere," *Journal of Statistical Physics*, vol. 127, no. 4, pp. 813–823, 2007.
97. S. Redner, *A Guide to First-Passage Processes:*. Cambridge University Press, Aug 2001.
98. M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, pp. 779–782, 06 2008.
99. D. Moltchanov, "Survey Paper: Distance Distributions in Random Networks," *Ad Hoc Netw.*, vol. 10, pp. 1146–1166, Aug. 2012.
100. S. M. Ross, *Introduction to probability models*. Academic press, 2014.
101. N. Baldo, "The ns-3 LTE module by the LENA project," 2011.
102. J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in HetNets: old myths and open problems," *IEEE Wireless Communications*, vol. 21, pp. 18–25, April 2014.
103. M. Gerasimenko, D. Moltchanov, R. Florea, S. Andreev, Y. Koucheryavy, N. Himayat, S. P. Yeh, and S. Talwar, "Cooperative Radio Resource Management in Heterogeneous Cloud Radio Access Networks," *IEEE Access*, vol. 3, pp. 397–406, 2015.
104. D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. American Mathematical Soc., 2009.
105. A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis, "Multiobjective Auction-based Switching Off Scheme in Heterogeneous Networks To Bid or Not To Bid?," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2016.
106. I. Taxidou and P. Fischer, "Realtime Analysis of Information Diffusion in Social Media," *Proc. VLDB Endow.*, vol. 6, pp. 1416–1421, Aug. 2013.

107. R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared systems," *tech. rep., Digital Equipment Corporation, DEC-TR-301*, 1984.

108. R. O. Afolabi, A. Dadlani, and K. Kim, "Multicast Scheduling and Resource Allocation Algorithms for OFDMA-Based Systems: A Survey," *IEEE Communications Surveys Tutorials*, vol. 15, pp. 240–254, First 2013.

109. A. Ometov, P. Masek, L. Malina, R. Florea, J. Hosek, S. Andreev, J. Hajny, J. Niutanen, and Y. Koucheryavy, "Feasibility Characterization of Cryptographic Primitives for Constrained (Wearable) IoT Devices," in *Proc. of International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–23, IEEE, 2016.

110. D. Moltchanov, Y. Koucheryavy, and J. Harju, "Simple, accurate and computationally efficient wireless channel modeling algorithm," in *Wired/Wireless Internet Communications*, pp. 234–245, Springer, 2005.

111. "WINTERsim system-level simulator description." `http://winter-group.net/downloads/`, January 2016.

112. D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, pp. 462—465, 2006.

113. I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong, "On the Levy walk nature of human mobility," in *Proc. of INCOCOM 2008, Phoenix, AZ, April 2008*, pp. 276–279, 2008.

114. L. Atzori, A. Iera, G. Morabito, and M. Nitti, "The Social Internet of Things (SIoT) – When social networks meet the Internet of Things: Concept, architecture and network characterization," *Computer Networks*, vol. 56, no. 16, pp. 3594–3608, 2012.

115. M. Nitti, R. Girau, and L. Atzori, "Trustworthiness Management in the Social Internet of Things," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1253–1266, 2014.

116. A. Ometov, K. Zhidanov, S. Bezzateev, R. Florea, S. Andreev, and Y. Koucheryavy, "Securing Network-Assisted Direct Communication: The Case of Unreliable Cellular Connectivity," in *Proc. of IEEE Trustcom/BigDataSE/ISPA*, pp. 826–833, 2015.

117. P. Nain, D. Towsley, B. Liu, and Z. Liu, "Properties of random direction models," in *Proc. IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 3, pp. 1897–1907, 2005.

118. S. Andreev, M. Gerasimenko, O. Galinina, Y. Koucheryavy, N. Himayat, S. P. Yeh, and S. Talwar, "Intelligent access network selection in converged multi-radio heterogeneous networks," *IEEE Wireless Communications*, vol. 21, pp. 86–96, December 2014.

119. R. Groenevelt, "Stochastic Models for Mobile Ad Hoc Networks," PdD thesis, INRIA Sophia-Antipolis, 2005.

120. H. Cai and D.-Y. Eun, "Crossing Over the Bounded Domain: From Exponential to Power-Law Intermeeting Time in Mobile Ad Hoc Networks," *IEEE/ACM Trans. Netw.*, vol. 17, pp. 1578–1591, Oct. 2009.

121. A. M. Mathai, *An introduction to geometrical probability: distributional aspects with applications*, vol. 1. CRC Press, 1999.

122. L. Santalo, *Integral Geometry and Geometric Probability*. Addison-Wesley, 1976.

123. E. Gelenbe, G. Pujolle, and J. Nelson, *Introduction to queueing networks*. Wiley Chichester, 1998.

124. P. Mendes, W. Moreira, T. Jamal, H. Haci, and H. Zhu, "Cooperative networking in user-centric wireless networks," in *User-Centric Networking*, pp. 135–157, Springer, 2014.

125. C. W. Kirkwood, "Notes on attitude toward risk taking and the exponential utility function," *Department of Management, Arizona State University, January*, 1997.

126. C. V. Shawndra Hill, Foster Provost, "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks," *Statistical Science*, vol. 21, no. 2, pp. 256–276, 2006.

127. ITU-R M.2083-0, "IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond," tech. rep., September 2015.

128. H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-Type Communications: Current Status and Future Perspectives Toward 5G Systems," *IEEE Communications Magazine*, pp. 10–17, September 2015.

129. G. Fodor, S. Parkvall, S. Sorrentino, P. Wallentin, Q. Lu, and N. Brahmi, "Device-to-Device Communications for National Security and Public Safety," *IEEE Access*, pp. 1510–1520, 2014.

130. METIS II Deliverable D1.1, "Refined scenarios and requirements, consolidated use cases, and qualitative techno-economic feasibility assessment," tech. rep., [Accessed on January 2016].

131. T. S. Rappaport and S. Sun and R. Mayzus and H. Zhao and Y. Azar and K. Wang and G. N. Wong and J. K. Schulz and M. Samimi and F. Gutierrez, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," *IEEE Access*, vol. 1, pp. 335–349, 2013.