Device-to-device data Offloading : from model to implementation

Filippo Rebecchi

UPMC Sorbonne Universités
Paris, France
18 September 2015

Délestage de Données en D2D : De la Modélisation à la Mise en Œuvre

présentée
par
Filippo Rebecchi

pour obtenir
le grade de docteur
specialité
Informatique, Télécommunications et Électronique

Jury:

Aline Carneiro Viana	Rapporteur	Chargée de Recherche, Inria Saclay – Ile-de-France
André-Luc Beylot	Rapporteur	Professeur, ENSEEIHT
Nathalie Mitton	Examinateur	Chargée de Recherche, Inria Lille – Nord Europe
Chris Blondia	Examinateur	Professeur, University of Antwerp
Serge Fdida	Examinateur	Professeur, UPMC Sorbonne Universités
Vania Conan	Directeur	Directeur de Recherche, Thales Communications
Marcelo Dias de Amorim	Directeur	Directeur de Recherche, CNRS and UPMC Sorbonne Universités

UPMC Sorbonne Universités
Paris, France
18 September 2015

Device-to-Device Data Offloading: From Model to Implementation

A DISSERTATION PRESENTED
BY
FILIPPO REBECCHI

IN FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
COMPUTER SCIENCE, TELECOMMUNICATIONS, AND ELECTRONICS

COMMITTEE:

ALINE CARNEIRO VIANA	REVIEWER	RESEARCH SCIENTIST, INRIA SACLAY – ILE-DE-FRANCE
ANDRÉ-LUC BEYLOT	REVIEWER	PROFESSOR, ENSEEIHT
NATHALIE MITTON	EXAMINER	RESEARCH SCIENTIST, INRIA LILLE – NORD EUROPE
CHRIS BLONDIA	EXAMINER	PROFESSOR, UNIVERSITY OF ANTWERP
SERGE FDIDA	EXAMINER	PROFESSOR, UPMC SORBONNE UNIVERSITÉS
VANIA CONAN	ADVISOR	RESEARCH DIRECTOR, THALES COMMUNICATIONS
MARCELO DIAS DE AMORIM	ADVISOR	RESEARCH DIRECTOR, CNRS AND UPMC SORBONNE UNIVERSITÉS

Acknowledgments

The end of a journey often coincides with a novel start. However, this was a long journey, and it is difficult to realize that a page lasting three years of your life is turning. The duration of a PhD thesis may seem like a long time from the outside. It turns into a such short time when you look at the amount of information you receive and the work you have to do. Thankfully, I was well accompanied. In the end, I will never forget how enriching and challenging were those three last years, and I am sure that the lessons I learned during this experience will accompany me for life.

In the first place, I want to express my sincere gratitude to Marcelo and Vania – my two advisors – without their help and perseverance these three years would have not been the same. They acted at the same time as psychologists, coaches, and technical gurus, always making me look at the big picture rather than drowning in the details. I view them as the real drivers for this thesis defense. I have also spent with them some memorable times, during several trips for conferences and project meetings. Their example will last long with me.

I want also to thank Aline Carneiro Viana and André-Luc Beylot, who accepted to review my thesis. They spent a considerable amount of their precious time to read this manuscript, and their pertinent remarks allowed me to improve the overall quality of the manuscript and the presentation. I want also to thank Nathalie Mitton, Chris Blondia, and Serge Fdida that accepted to be in the jury. I know that their expertise on the domain will improve for sure my reflexion.

Thanks to all the smart people I collaborated with during these three years. All the members of the MOTO project taught me something, and I think in particular to Andrea Passarella and Raffaele Bruno with whom it has been a real privilege to collaborate.

My gratitude goes also to all my kind colleagues and collaborators from Thales and LIP6. Actual and former members of the TAI laboratory at Thales have been like a second family to me in these three years and contributed significantly to the success of this thesis. In particular, I wish to thank Corinne, Jeremie, and Yoann who welcomed me in the first place at TAI. Paolo, Davide and all the Italian people in Thales who allowed me to exchange some words in my native language from time to time. Farid, who drove me to work endless times and with whom I shared most of the time when traveling. Hicham, Mathieu, Mario, Nico, Lionel, Julien, David, Marie-Noelle and Catherine who are the references for the lab, but also people

with whom is extremely pleasant to discuss about everything. FX, Remy, Mathis, K, Jawad, Romain, Bruno, Flo, Mathias, and all the present and past members of TAI.

During the weekly day that I spent at LIP6 and LINCS, I had the opportunity to meet and exchange with Promethé, Sebastian, Giovanni, Marguerite that have all contributed to my experience. In particular, it has been a pleasure to exchange ideas with old and new PhD students who I had the opportunity to meet: Tiph, Matteo, Alex, John, Nadjet, Ahlem, Mehdi, Adisorn, Giulio, Davide, Fadwa, Ben, Salah, Quentin, Alexandre. I am forgetting someone for sure, so I apologize in advance.

I think also to all my friends who have always followed me from far away. Thank you to my whole family, who always backed me. Especially, my parents, Sandro and Sandra that have always encouraged, allowing me to live this experience. I hope they are proud of me, because I am really proud of them. *Grazie*.

In the end, a huge thanks to my dear Anne, which has supported me for the three past years. I know it was not simple. You have been able to understand me, accept me, and I have often been given the chance to get back into question. Thank you for this, and for much more.

Délestage de Données en D2D : De la Modélisation à la Mise en Œuvre

Resumé

La disponibilitè de connectivitè omniprèsente et l'explosion du marchè des appareils mobiles ont entraînè une croissance fulgurante de l'utilisation de donnèes en mobilitè en 2014. Le trafic mobile global atteindra 24,3 exa-octets en 2019. Aujourd'hui, les rèseaux cellulaires sont sous pression, essayant de faire face à cette surcharge de donnèes sans prècèdent. Accueillir cette croissance de façon traditionnelle exigerait des investissements importants dans le rèseau d'accès radio. Des approches alternatives, et plus èconomiques, ont èmergè pour faire face à ce problème. Dans cette thèse, nous tournons notre attention vers l'une de ces solutions, pour laquelle la communautè de recherche porte un intèrêt croissant : le dèlestage (couramment appelèe offloading) grâce à des communications de dispositif à dispositif (D2D). Cette approche non conventionnelle exploite la bande passante inutilisèe dans des technologies sans fil diffèrentes. Il diffère sensiblement du dèlestage à travers des points d'accès Wi-Fi et smallcells, car les utilisateurs font partie du rèseau et sont capables de transmettre des donnèes pour le compte de l'infrastructure. En effet, en dèplaçant la partie du trafic tolèrant aux dèlais sur des connexions directes entre terminaux, on peut apporter un grand bènèfice aux opèrateurs.

Dans cette thèse, nous nous sommes intèressès à l'offloading D2D sous des angles diffèrents. Initialement, nous avons ètudiè des propositions existantes dans la littèrature afin d'identifier les fonctionnalitès communes nècessaires pour le dèlestage dans une architecture rèseau mobile. De cette analyse, nous avons obtenu un signe tangible de la nècessitè pour les opèrateurs de mettre en œuvre des stratègies d'offloading efficaces. Cependant, de nombreuses ètudes proposent de prè-calculer un sous-ensemble optimal de nœuds initiaux à qui transmettre les donnèes au dèpart. Ce choix est fait en utilisant la structure du rèseau opportuniste sous-jacent, ainsi que les motifs de contact entre les utilisateurs. Bien que cette approche soit sans doute èlègante, gènèralement elle ne rèagit pas bien à la variabilitè des conditions, souffrant ègalement d'une application limitèe en raison de la spècificitè et des hypothèses sous-jacentes parfois irrèalistes. Notre première contribution est donc DROiD, une stratègie de dèlestage D2D qui exploite la disponibilitè de l'infrastructure cellulaire comme un canal de retour afin de suivre l'èvolution de la diffusion. DROiD adapte la stratègie d'injection au rythme de la diffusion, à la fois de manière rèactive et simple, et permettant d'èconomiser une quantitè èlevèe de donnèes cellulaires, même dans le cas de contraintes de rèception très serrèes.

Ensuite, nous mettons l'accent sur les gains que les communications D2D pourraient apporter si elles ètaient couplèes avec les transmissions multicast. Le multicast reprèsente, en principe, un moyen très efficace de distribution de contenu à une multitude d'utilisateurs. Nous èvaluons d'abord les lacunes de la mise en œuvre du multicast dans la dernière technologie de rèseau cellulaire (par exemple, dans l'LTE), où l'efficacitè globale est dictèe par le

nœud avec la pire qualitè de canal parmi tous les rècepteurs de la transmission. Nous dèmontrons que par l'utilisation èquilibrèe d'un mix de multicast et communications D2D nous pouvons amèliorer à la fois l'efficacitè spectrale mais aussi la charge des rèseaux cellulaires. Afin de permettre l'adaptation aux conditions actuelles (ex., le nombre de requêtes, la densitè et la structure de la mobilitè des utilisateurs), nous èlaborons une stratègie d'apprentissage basèe sur l'algorithme de "bandit manchot" pour identifier la meilleure combinaison de communications multicast et D2D.

Enfin, nous ètudions des modèles de coûts pour les opèrateurs dèsireux de rècompenser et de stimuler les utilisateurs qui coopèrent dans la diffusion D2D, les reconnaissant pour la transmission de contenu au nom de l'infrastructure. Ces hypothèses dèterminent donc que, la diffusion opportuniste non contr□lèe est coûteuse pour les opèrateurs, car elle gènère des coûts supplèmentaires en raison de chaque transmission D2D. Dans ce cas, nous proposons de sèparer la notion de seeders (utilisateurs qui transportent le contenu, mais ne le distribuent pas) et les forwarders (utilisateurs qui sont chargès de distribuer le contenu). En faisant cela, les opèrateurs bènèficient d'une meilleure souplesse dans la gestion des opèrations de dèlestage. Avec l'aide d'un outil analytique basè sur le principe maximal de Pontryagin, nous dèveloppons une stratègie optimale de dèlestage. L'analyse des rèsultats nous fournit un aperçu sur les interactions entre les seeders, les forwarders, et l'èvolution de la diffusion des donnèes, rèvèlant que la dècision de renvoi pourrait être tout aussi critique que le choix des nœuds.

Device-to-Device Data Offloading: From Model to Implementation

Abstract

The combined availability of pervasive connectivity and the explosion in the smart mobile devices market resulted is a stunning 69% growth in global mobile data usage in 2014. Overall mobile traffic is expected to reach 24.3 exabytes by 2019. As today's most common data access method for users on the move, cellular networks are under pressure trying to cope with this unprecedented data overload. Accommodating this growth in a traditional way would require major investments in the radio access network.

Alternative approaches emerged to deal with this problem. In this thesis, we turn our attention to one of these solutions, recently attracting increasing interest by the research community: *mobile data offloading through device-to-device (D2D) communications*. This unconventional approach leverages the unused bandwidth across different wireless technologies. It differs substantially from regular offloading over Wi-Fi and small-cell networks. Users become part of the network and are capable of transmitting data on behalf of the cellular infrastructure. Indeed, shifting away the delay-tolerant part of the traffic can bring great benefit to operators, provided that the choice of seeder nodes is correct.

In this thesis, we tackle data offloading under different angles. Initially, we study existing propositions in the literature in order to identify the common functionalities needed in an offloading architecture. From this analysis, we get a tangible sign of the need for operators to implement efficient offloading strategies. However, many studies pre-compute a set of optimal *seeders* using the structure of the underlying opportunistic network, as well as the contact patterns among users. Although this approach is undoubtedly elegant, it typically does not react well to changing conditions, suffering also from a limited applicability due to the specific and sometimes unrealistic underlying assumptions. Our first contribution is DROiD, an offloading strategy that exploits the availability of the cellular infrastructure as a feedback channel in order to track the dissemination evolution. DROiD adapts the injection strategy to the pace of the dissemination, resulting at the same time reactive and relatively simple, allowing to save a relevant amount of data traffic even in the case of tight delivery delay constraints.

Then, we shift the focus to the gains that D2D communications could bring if coupled with multicast wireless networks. Multicast represents, in line of principle, a very efficient way of distributing content to a multitude of users. We first assess the shortcomings of the implementation of multicast in the latest cellular network technology (e.g., eMBMS in LTE), where the global efficiency is dictated by the node with the worst channel quality among all the multicast receivers. We demonstrate that by employing a wise balance of multicast and D2D communications we can improve both the spectral efficiency and the network load in cellular networks. In order to let the network adapt to current conditions (number of re-

quests, density and mobility pattern of users), we devise a learning strategy based on the multi-armed bandit algorithm to identify the best mix of multicast and D2D communications.

Finally, we investigate the cost models for operators wanting to reward and stimulate users who cooperate in D2D diffusion, acknowledging them for transmitting content on behalf of the infrastructure. Under these assumptions, uncontrolled opportunistic diffusion is expensive for operators because it generates additional costs owing to D2D transmissions. In this case, we propose separating the notion of *seeders* (users that carry content but do not distribute it) and *forwarders* (users that are tasked to distribute content). By doing so, operators benefit from a better flexibility in the management of offloading operations. With the aid of the analytic framework based on Pontryagin's Maximum Principle, we develop an optimal offloading strategy. Results provide us with an insight on the interactions between seeders, forwarders, and the evolution of data dissemination, revealing that the *forwarding* decision could be just as critical as the choice of *seeders*.

Contents

Listing of figures

1 Introduction

1.1 Context and Motivation

Mobile networks are an integral part of our everyday life. Their capabilities allow experiencing pervasive high data rates everywhere, enabling applications that would have been inconceivable only 10 years ago. Driven by the increasing popularity of smart mobile devices and the introduction of affordable data plans by cellular operators, global mobile traffic is booming. Data-hungry mobile applications, such as audio and video streaming, social sharing, or cloud-based services, are more and more popular among users. In the time lapse of this thesis (2012 – 2015), data traffic grew nearly three times, being expected to grow additionally 10-fold between 2014 and 2019, three times faster than the overall fixed traffic in the same period [1]. It is also anticipated that two-thirds of this traffic will be video related (with or without real-time requirements) by 2017. As today's most common data access method for nodes on the move, cellular networks are under heavy pressure trying to cope with this unprecedented data overload. Accommodating this growth requires major investments both in the radio access network (RAN) and the core infrastructures. Upgrading the RAN is very expensive, since it requires more infrastructure equipment and thus more investment.

Scarce licensed spectrum hinders the RAN enhancements. Regulations allow mobile operators to use only a small portion of the overall radio spectrum, which is also extremely expensive. Users must share the same limited wireless resources, which poses a capacity problem. Considerable progress is constantly made at the physical layer to increase raw bit rates, but this is neither sufficient nor cost-efficient to accommodate all the increase in data service demand [1]. Adding traffic beyond a certain limit mines the performance and the quality of

service (QoS) perceived by the users. During peak times in crowded metropolitan environments, users experience long latencies, low throughput, and outages due to congestion and overload at RAN level [2]. Unfortunately, this trend can only exacerbate in the future due to the predicted mobile data explosion. The problem concerns primarily network operators because they have *to trade-off customer satisfaction with business profitability*, given the trend toward nearly flat rate business models. In other words, the exponential increase in traffic flowing in their RAN does not generate enough additional revenues to be allocated into further RAN upgrades. This creates what Mölleryd et al. call the *revenue gap* [3].

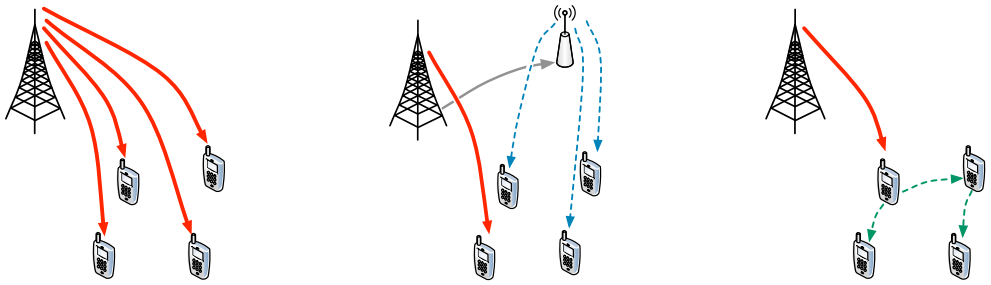
1.2 Data Offloading: a Solution to the Bandwidth Crunch

The aforementioned circumstances fostered the interest in alternative methods to mitigate the pressure on the cellular network. As a first option, mobile operators solved this contingency by throttling connection speed and capping data usage [4]. However, these practices negatively affect the customer satisfaction. Alternative and more disruptive innovations in the architecture of cellular networks have to be explored to help carriers meet the exponential growth in mobile data traffic demand. In this dissertation, we turn our attention to one of these solutions, recently attracting increasing interest by the research community: *mobile data offloading*. An intuitive approach is to leverage the unused bandwidth across different wireless technologies. We consider mobile data offloading as the use of *a complementary wireless technology to transfer data originally targeted to flow through the cellular network*, in order to improve some key performance indicators.

Although the concept of offloading may apply to any network, current academic and industrial research mostly concerns with offloading data from cellular networks. Those are the type of networks that would benefit most from this technique. Besides the obvious benefit of relieving the infrastructure network load, shifting data to a complementary wireless technology leads to a number of other improvements, including: the increase of the overall throughput, the reduction of content delivery time, the extension of network coverage, the increase of network availability, and better energy efficiency.[*] These improvements hit both cellular operators and users; therefore, offloading is often described in the literature as a *win-win* strategy [5]. Unfortunately, this does not come for free, and a number of challenges need to be addressed, mainly related to infrastructure coordination, mobility of users, service continuity, pricing, business models, and lack of standards.

For the reader's convenience, we depict in Fig. 1.1 the two main approaches to offloading in cellular networks when compared with the traditional infrastructure-only mode (Fig. 1.1a). Diverting traffic through fixed WiFi Access Points (AP), as in Fig. 1.1b, represents a conventional solution to reduce traffic on cellular networks. End-users located inside a hot-spot coverage area (typically much smaller than the one of a cellular macrocell) might use it as a

[*] Please note that despite being highly beneficial, implementing offloading capabilities is not a requirement for operators. Offloading rather represents extra available capacity, which can be used whenever appropriate.



(a) Infrastructure only.

(b) AP-based offloading.

(c) Terminal-to-terminal offloading.

Figure 1.1: The two major approaches to cellular data offloading compared to a) the baseline traditional infrastructure-only system. b) Offloading through a wireless Access Point. c) Offloading through terminal-to-terminal transmissions

.

worthwhile alternative to the cellular network when they need to exchange data. Hot-spots generally provide better connection speed and throughput than cellular networks [6]. However, coverage is limited and mobility is in general constrained within the cell. Since the monetary cost of deploying an array of fixed APs is far lower than deploying a single cellular base station, the major worldwide cellular providers have started integrating an increasing number of wireless APs in their cellular networks to encourage data offloading [7]. Meanwhile, a growing number of applications that automatize the offloading process are proposed for popular mobile devices (mainly iPhone and Android based) [8, 9].[†]

The increasing popularity of smart mobile devices proposing several alternative communication options makes it possible to deploy a device-to-device (D2D) network that relies on direct short-range communication between mobile users, without any need for an infrastructure backbone (Fig. 1.1c). This innovative approach has intrinsic properties that can be employed to offload traffic. D2D offloading represents a vibrant research topic at the core of this dissertation. Benefiting from shared interests among co-located users, a cellular provider may decide to send popular content only to a small subset of users via the cellular network, and let them spread the information through D2D communications and opportunistic contacts. This new paradigm is applicable in dense areas (urban and large event scenarios) where moving people always carrying connected devices. When two or more of these users are close enough, they have the potential to exchange data. Moving information hop-by-hop between users can provide an alternative means of distributing the information, which could relieve

[†] The ability to switch seamlessly between heterogeneous networks is referred to as vertical handover [10].

the load on the infrastructure. Note also that these two forms of offloading (AP and D2D based) may be employed concurrently, enabling users to retrieve data in a hybrid mode.

1.3 Novelty and Challenges

The dissertation explores the combination between opportunistic and cellular (infrastructure) networks. We propose a synergic use of a diverse set of complementary offloading techniques, by adding direct hop-by-hop communications between terminals. Such an approach is enabled by the fact that current mobile terminals are equipped with a range of complementary wireless communication technologies, allowing them not only to easily and dynamically connect to different wireless infrastructures, but also to establish direct connections with each other.

We adopt a novel approach, whereby the cellular infrastructures implements a control loop on the status of the dissemination process. Our approach differs with respect to conventional opportunistic data dissemination schemes. In particular, opportunistic networking, while highly regarded in the last few decades as a promising alternative to standard infrastructure-based networks, betrayed expectations to emerge as an autonomous networking paradigm. This is primarily due to its truly distributed nature that makes it extremely hard to offer any performance guarantees on the available bandwidth and data dissemination delay. However, its synergy with an infrastructure network could prove extremely advantageous to lower the load and increase performance of this latter. In this case, the cellular network can be employed to monitor the status of the dissemination process (e.g., in terms of the fraction of users that have received contents by a certain time). The cellular network can intervene when necessary, assuring data reception within guaranteed delays. This centralized offloading strategy has two beneficial effects: 1) mobile operators are actively involved in the offloading process and everything is under their control, and 2) it is possible to offer a QoS guarantee to users at any time.

We devote a large part of this thesis in designing and realizing the interplay between D2D and cellular network, evaluating also the case of cellular multicast. The resulting architecture requires reconsidering existing several wireless network paradigms. Future cellular architectures should intelligently support the distribution of heterogeneous classes of services, including the requirement of the application in terms of delivery guarantees, to face an overall traffic increase of several orders of magnitude. In this context, user mobility results particularly challenging as it can prove at the same time favorable and catastrophic for data forwarding. Finally, without users' cooperation opportunistic offloading strategies are doomed to failure. It is essential to consider an incentive scheme to promote content sharing among neighbor peers in opportunistic fashion. Additional fundamental questions in mobile data offloading concern the role of the network operator and the degree of freedom to let to users.

1.4 Contributions and Thesis Outline

This focus of this thesis is on evaluating opportunistic offloading strategies under multiple angles. The large part of the results of this work has been developed in the framework of the European Project FP7-MOTO [11]. The architecture and protocols developed along this dissertation have been evaluated analytically, or by simulations using different tools. Moreover, a small-scale prototype was built to validate experimentally the findings of simulation. We provide below an outline of the dissertation, summarizing also the contributions for each chapter.

Contribution 1 – A Survey on Data Offloading Techniques in Cellular Networks (Chapter 2)

We begin the dissertation by offering in Chapter 2 an exhaustive survey of the literature in the broad area of data offloading. We categorize existing techniques based on their requirements in terms of content delivery guarantee. Despite the positioning with respect to the state of the art is necessary, this is not the sole purpose of the chapter. By analyzing the literature we come out with a functional architecture to enable mobile data offloading with tight or loose delay guarantees. This architecture will be the base of the following chapters. Finally, we discuss open research and implementation issues. The work related to this chapter is:

- *F. Rebecchi, M Dias de Amorim, V. Conan, A. Passarella, R. Bruno, M. Conti, "Data Offloading Techniques in Cellular Networks: A Survey", in IEEE Communications Surveys & Tutorials, 2015.*

Contribution 2 – DROiD: Adapting to Individual Mobility Pays Off in Mobile Data Offloading (Chapter 3)

Based on the previously defined architecture, we propose exploiting the availability of the cellular infrastructure as a feedback channel to track the dissemination evolution. In Chapter 3, we develop DROiD, a low-complexity offloading strategy that relies both on D2D and AP-based communications. The observation that content dissemination in opportunistic networks follows a stepwise pattern is at the base of our strategy. Our proposal better adapts to the contact patterns between nodes to offer enhanced offloading efficiency. We also built a demonstrator that showcases DROiD, integrating D2D data offloading capabilities into a cellular infrastructure. Published and submitted works related to this chapter are:

- *F. Rebecchi, M Dias de Amorim, V. Conan, "Circumventing Plateaux in Cellular Data Offloading using Adaptive Content Reinjection", under major revision, submitted to IEEE Transactions on Network and Service Management, 2015.*

- *F. Rebecchi, M Dias de Amorim, V. Conan, "DROiD: Adapting to Individual Mobility Pays Off in Mobile Data Offloading", in IFIP Networking, Trondheim, Norway, 2014.*
- *F. Rebecchi, M Dias de Amorim, V. Conan, "Adaptive Mobile Data Offloading", in Algotel, Le-Bois-Plage-en-Ré, France, 2014.*
- *F. Benbadis, F. Rebecchi, F. Cosnier, M. Sammarco M Dias de Amorim, V. Conan, "Demo: opportunistic communications to alleviate cellular infrastructures: the FP7-moto approach", in ACM CHANTS, Maui, HI, 2014.*
- *F. Benbadis, F. Rebecchi, F. Cosnier, M. Sammarco M Dias de Amorim, V. Conan, "Demo: D2D Rescue of Overloaded Cellular Channels", in ACM MobiSys, Florence, Italy, 2015.*

Contribution 3 – Offloading LTE Multicast Data Dissemination through D2D communications (Chapter 4)

Multicast represents, in line of principle, a very efficient way of distributing content to a multitude of users. Chapter 4 shifts the focus to the gains that D2D communications could bring if coupled with multicast wireless networks. We first assess the shortcomings of the implementation of multicast in the latest cellular network technology (e.g., eMBMS in LTE). A wise balance of multicast and D2D communications can improve both the spectral efficiency and the network load in cellular networks. In order to let the network adapt to current conditions (number of requests, density and mobility pattern of users), we devise a learning algorithm to identify the proper mix of multicast and D2D communications. Published and submitted works related to this chapter are:

- *F. Rebecchi, M Dias de Amorim, V. Conan, "Flooding Data in a Cell: Is Cellular Multicast Better than Device-to-Device Communications?", in ACM CHANTS, Maui, HI, 2014.*
- *F. Rebecchi, L. Valerio, R. Bruno, V. Conan, M Dias de Amorim, V. Conan, A. Passarella "A Multi-Armed Bandit Resource Allocation Scheme for D2D-aided Cellular Multicast", submitted to Elsevier Computer Communications, 2015.*

Contribution 4 – Incentives for D2D Offloading: Discussion and Models (Chapter 5)

In D2D offloading it is impossible to consider data dissemination and collaboration without addressing incentives. Chapter 5 investigates a cost model for operators wanting to reward users who cooperate in D2D diffusion. Under these assumptions, uncontrolled opportunistic diffusion is expensive for operators because it generates additional costs owing to D2D

transmissions. We propose separating the notion of seeders (users that carry content but do not distribute it) and forwarders (users that are tasked to distribute content). By doing so, operators benefit from a better flexibility in the management of offloading operations. Published and submitted works related to this chapter are:

- *F. Rebecchi, M Dias de Amorim, V. Conan, "Seeders vs. Forwarders: Optimal Control of D2D Offloading under Rewarding Conditions", submitted to IEEE MASS, Dallas, TX, 2015.*

- *F. Rebecchi, M Dias de Amorim, V. Conan, "The Cost of Being Altruistic: Optimal D2D Offloading under Rewarding Conditions", in Algotel, Beaune, France, 2015.*

1.5 Industrial Exploitation

The contributions in this thesis perfectly fit in the scope of a branch of applied research carried out at Thales. While not directly targeting the market of commercial cellular networks, Thales is particularly active in the field of public safety communications (PSC). Although public safety users are an important community both economically and socially, the PSC market is much smaller than the commercial cellular one. Therefore, specialized public safety technologies cannot attract the level of investment and global R&D that goes in to commercial cellular networks. A strong standards-based approach will ensure interoperability between different vendors leading to a competitive equipment market. Thales vision is that common technical standards for commercial cellular (4G) and public safety networks offers advantages to both communities:

- The public safety community gets access to the economic and technical advantages generated by the scale of commercial cellular networks.

- The commercial cellular community gets the opportunity to address parts of the public safety market as well as gaining enhancements to their systems that have interesting applications to consumers and businesses.

Thales is making a significant effort to develop solutions for mobile broadband in PSC sector, currently building new communication infrastructures and dedicated mobile terminals towards the procurement of national security police, firefighters, and emergency organizations. This new infrastructure will rely on both Professional Mobile Radio (PMR) and emerging 4G technologies: Long Term Evolution (LTE) and WiMax, both of them supporting TETRA (Terrestrial Trunked Radio), the European Standard for public safety. This future solution intends to improve existing public communication services, enriching them with additional capabilities (e.g., video, multimedia, and data services), while maintaining compatibility with existing PMR networks.

Interestingly, several Public Safety agencies are already experimenting with mobile broadband (LTE) technologies for certain use-cases and scenarios. For example, ad-hoc mesh networks, direct communications (D2D), and LTE are being considered and evaluated worldwide. The high flexibility and performance of these technologies enable the deployment of new services and applications. Namely, LTE enables public safety agencies to be more responsive, to increase situational awareness and to coordinate with other agencies by providing faster data sharing, through new applications, real-time video, as well as D2D and group-based communications. In short, LTE is the enabler of next generation communications for public safety. Services to be offered for Public Safety that can be impacted by the outcomes of this thesis, and by data offloading techniques in general, include:

- Support of *Push-to-X* (PTX) services for fleets of machines and PMR (Private Mobile Radio) users. While LTE systems offer high capacity links, they should also support efficiently the typical group communications required in PMR deployments. PTX services call for voice and data-centric group communications. Currently, push-to-talk allows a PMR user to reach an active talk group with a single button press. This voice service can benefit from Push-over-Cellular services and should be extended with publish-subscribe data-centric communications over a multitude of transmission techniques to support additional usages and leverage network capacity.

- D2D services, a.k.a. DMO (Direct Mode Operations) of legacy TETRA-like systems. In DMO mode the Land Mobile Radio (LMR) terminals communicate directly with each other without using the existing infrastructure. For instance, DMO is often used in situations where the devices lie outside the coverage area of TETRA networks. While DMO currently supports only voice communications, it is widely agreed that data services must be also included in the next LTE-based version. D2D services are one of the area where 3GPP has agreed to enhance LTE, permitting to identify mobiles in physical proximity and optimizing direct communication links.

In effect, a considerable part of this thesis targets one of the typical scenario of utilization of public safety networks, where massive data distribution is required in order to support *Push-to-X* application offering optimal situation awareness for first responders in a field of operation. This massive data distribution application scenario has its origin from new and, somehow, visionary data collection services that could be enabled by exploiting the increasing ability of current personal mobile devices to monitor the physical world. Indeed, the more sophisticated the smartphone, the longer its list of embedded information sinks is. For example, a standard rugged secure first responder smart-handled device now ships with a compass, a gyroscope, an accelerometer, a GPS, two microphones, various cameras, and others users generated data applications. First responders will have access to high-speed connectivity and the ability to receive large amounts of data and video to and from the command center, from patrol car to patrol car, and from smart handheld device to smart handheld device in the field.

Those who don't know history are destined to repeat it.
Edmund Burke

2 Data Offloading Techniques in Cellular Networks

In this chapter, we review the main offloading strategies providing a comprehensive categorization of existing solutions. The purpose of this chapter goes beyond the simple positioning of our work with respect to the state of the art. The ultimate goal is to detail a set of features necessary for the functioning of an offloading architecture. The above network architecture will be reused many times over the course of the dissertation and will serve as the basis for our subsequent contributions.

2.1 Classification

Mobile data offloading can be – at a very high level – categorized according to the presence or not of the infrastructure (as we showed in the introduction). However, a more refined classification is required to provide a comprehensive picture. It is important to pinpoint that mobile data offloading techniques can be classed depending on the assumptions one can make on the level of synergy between cellular and unlicensed wireless networks, as well as the involvement of user terminals in the offloading process. Beyond the obvious distinction between AP-based and D2D approaches already mentioned in the previous chapter, another aspect plays a major role in the categorization. In particular, we take into consideration the requirements of the applications generating the traffic in terms of delivery guarantees. For this reason, we also consider a temporal dimension in our classification, depending on the

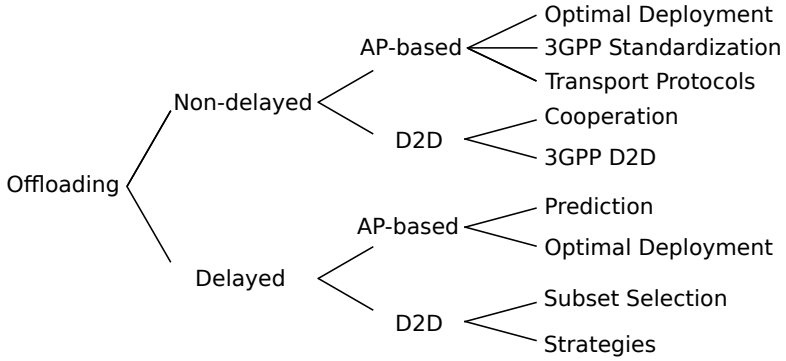


Figure 2.1: Offloading directions in the literature.

delay that the data we want to offload may tolerate upon delivery. This translates into two additional categories: (i) non-delayed offloading and (ii) delayed offloading.

We consider these two orthogonal dimensions (delivery delay guarantees and offloading approach), which correspond to four possible combinations as shown in Fig. 2.1. The biggest difference between non-delayed and delayed offloading mechanisms lies in the way the timeliness of content reception is handled. In fact, in non-delayed offloading we do not have any *extra* delay on the "secondary" interface (considering cellular the "primary"), while in delayed offloading the network adds some delay (either associated to the fact that the user has to wait until it gets close enough to a WiFi AP, or to get messages through opportunistic contacts).

2.2 Taxonomy: Non-delayed offloading

Non-delayed offloading is the most straightforward and experimented class of offloading. Data may be real-time and interactive, thereby enabling the fruition of services such as video streaming and VoIP. So far, WiFi hot-spots have represented the most logical solution due to their widespread diffusion, acceptable performance, and low cost. Nevertheless, we can find in the literature many approaches that exploit D2D content sharing between neighboring nodes. In non-delayed offloading, each packet presents a hard delivery delay constraint defined by the application, which in general is independent of the network. No extra delay is added to data reception in order to preserve QoS requirements (other than the delay due to packet processing, physical transmission, and radio access).

This requirement puts a strain on the network that should meet this deadline to ensure the proper functioning of the application. It turns out that non-delayed offloading is essentially unfeasible in opportunistic networks, since the accumulated end-to-end delay over the transmission path may be too high with respect to the strict delivery requirements. However, if we restrict the analysis to low mobility scenarios, it is still possible to deliver data with strict delay guarantees using D2D transmissions or with the aid of a fixed infrastructure. Non-delayed offloading in most cases may be difficult to implement if one considers that users are mobile

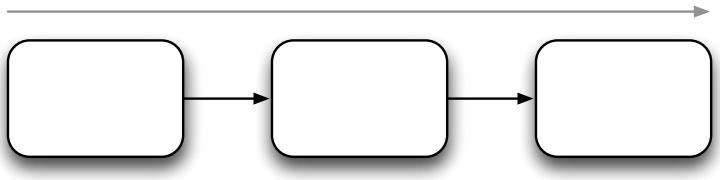


Figure 2.2: Steps toward the integration of alternative access networks and the cellular infrastructure.

and able to switch between various access technologies. If operators want to allow users to be truly mobile and not only nomadic inside the coverage area, they should focus on issues such as transparent handover and interoperability between the alternative access technologies and the existing cellular infrastructure. For instance, this aspect is not granted when one considers a basic offloading implementation through IEEE 802.11 APs. On the other hand, this commitment allows offloading data such as voice over IP (VoIP) or interactive applications, obtaining a nearly transparent offloading process.

2.2.1 AP-based

The prevailing AP-based offloading model today is *user-driven*, meaning that users must explicitly enable the alternative access network in order to benefit from an enhanced experience.[*] This approach is appealing at first, as it requires no modifications in the network infrastructure; however, common limitations such as constrained mobility and lack of session continuity hinder its mass adoption. To pave the way for better cross-resource utilization and improved customer experience, the current trend is to let operators have a deeper control of the offloading process. This eventually raises the question of how a cellular operator can run a profitable business by shifting off-network large parts of its traffic.

Providers are more and more looking toward a tighter integration of alternative access networks and their cellular infrastructure, as depicted in Fig. 2.2. The integration process concerns partnerships between cellular and wireless providers, common billing and accounting policies, shared subscriber databases for authentication, authorization, accounting (AAA), and security provisioning. Two possible network architectures to date are envisioned to integrate cellular and WiFi access: *loose coupling* and *tight coupling*. In loose coupling, the two networks are independent and are interconnected indirectly through an external IP network. Service continuity is provided by roaming between the two networks. In tight coupling instead, the two networks share a common core and many functions, such as vertical and horizontal handover, integrated management of resources, and common AAA.

[*] Smart mobile devices already give priority by default to WiFi when a wireless network results available and WiFi interface is enabled.

Optimal Deployment

Several trace-based analyses demonstrate that the deployment of fixed APs is a viable method to reduce congestion in cellular networks [12, 13, 14, 15]. These studies motivate the increasing interest in data offloading, providing an experimental upper bound on how much data it is possible to offload given an existing AP deployment. However, measurements based only on signal strength do not consider the issues related to higher-layer protocols, which may influence the theoretical possibility of offloading as perceived from a pure signal strength analysis.

An interesting strategy to boost offloading performance is to place optimally the APs in order to maximize the traffic that flows through the alternative channel [15, 16, 17, 18]. Since the optimal positioning problem becomes quickly intractable (NP-hard) as the amount of APs increase, a number of sub-optimal algorithms have been developed to this extent. The basic idea is to place the APs close to the locations with the highest density of mobile data requests (or the number of users). Simulation results show that it is possible to shrink cellular traffic by $20 - 70\%$, depending on the AP density. Besides simulation results, analytical models help to derive theoretical bounds on performance, based on queuing theory [19, 20].

Optimal deployment might be a short-term solution for improving performance of real-time data offloading. Ideally, up to 70% of traffic could be offloaded through a carefully planned deployment. Indeed, if the pattern of requests changes, the selected deployment might not be optimal anymore. Furthermore, all reviewed works assume perfect vertical handover mechanisms, which is an over simplification. Counter-intuitively, adding too many APs could worsen the situation due to mutual interference. An interesting future research area concerns the selection of the optimal AP when multiple APs are simultaneously available. This shares some similarities with the problem of deciding which terminal and which traffic flow to move to a different communication channel [21]. Furthermore, it is related to the Access Network Discovery and Selection Function (ANDSF) mechanism introduced later. On the other hand, analytical models help understand the optimal fraction of data to shift on the alternate channel to maximize the overall data rate and the amount of cellular savings.

3GPP Standardization: ANDSF, IFOM, LIPA, SIPTO

The LTE network proposes an Evolved Packet Core (EPC) flat architecture, fulfilling the requirements for an integrated hybrid network. The EPC is an access-independent all-IP based architecture, capable of providing the handover between IP-based services across a broad range of access technologies (e.g., cellular, WiFi, and WiMAX). Both 3rd-Generation Partnership Project (3GPP) radio access networks and non-3GPP technologies are supported. 3GPP considers data offloading as a key option to tackle the cellular overload problem, proposing the ANDSF mechanism to trigger the handoff between different access technologies [22]. It also proposes three alternative offloading mechanisms that take advantage of the hybrid ar-

chitecture of the EPC: Local IP access (LIPA), selected IP traffic offload (SIPTO) [23], and IP Flow Mobility (IFOM) [24].

ANDSF is a framework for communicating to the mobile devices the policies for network selection and traffic routing, assisting them in the discovery and handover process [25]. Three different access selection strategies are evaluated, based on coverage, SNR, and system load. A congestion control mechanism to assist ANDSF is also proposed. [26]. LIPA is part of the femtocell architecture and allows a mobile terminal to transfer data directly to a local device connected to the same cell without passing through the cellular access network. SIPTO, instead, attempts to offload the core of the network, balancing data flows to selected IP gateways at core level. Note that these solutions (e.g., LIPA/SIPTO) offload the core cellular network and do not relieve bandwidth crunch in the access network. Therefore, they are not the focus of this dissertation. We suggest interested readers to refer to Samdanis et al. [27] and Sankaran [28].

IP Flow Mobility (IFOM) implements offloading at RAN level, allowing providers to move selected IP data-flows between different access technologies without disrupting ongoing communications [29]. Conversely to ANDSF, which is utilized to discover, connect, and manage handover between neighboring APs, IFOM provides offloading capabilities in terms of moving data-flows between access networks. IFOM allows terminals to bind multiple local addresses (CoAs) to a single permanent home IP address (HoA), and to bind distinct IP flows (e.g., HTTP, Video, VoIP) to different CoA. This feature allows different flows related to the same connection to be routed over different radio access technologies based on operator-defined policy. Sometimes IFOM involves a total switchover of all traffic from one access technology to another. In other cases, the network allocates only "best effort" data to the complementary access, while keeping delay-sensitive flows on the cellular network. IFOM allows users benefiting from high bandwidth connections when at least one complementary network is available. At the same time, operators are able to manage the radio access resources optimally, reducing the network overload and providing different QoS levels to distinct data-flows. Drawbacks of IFOM reside in the additional modifications needed both at terminal and network levels to manage the heterogeneity of access technologies. In addition, in very dense wireless environments the management of user mobility should adapt to very challenging conditions, such as interference and dynamic terminal reconfiguration.

3GPP standardized the ability to perform offloading through a variety of access methods in the LTE network architecture. New protocols, such ANDSF and IFOM, transform offloading into a nearly transparent mechanism for end-users. Operators are able to shift selected data-flows between different access technologies without any disruption. This concurs in lowering the network congestion. As of today, no commercial deployments of ANDSF and IFOM exist, though trials are undergoing to understand the feasibility of these solutions. The widespread adoption of these techniques is one of the keys to enable effective operator-driven offloading strategies. The mechanisms presented in this section are, as today, the standard frameworks in which forthcoming AP-based offloading strategies need to be integrated.

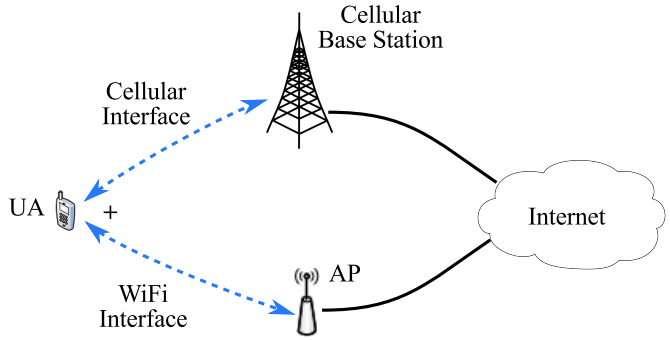


Figure 2.3: Multiple interfaces can be exploited simultaneously by users to increase the throughput, and improve data coverage. Transmission protocols should be capable of handling efficiently such situations.

On the other hand, these solutions could be significantly improved by considering delayed reception and opportunistic transmissions.

Transport Protocols

The development of a novel IP-based transport protocol is an essential prerequisite to enable future offloading capabilities to mobile smart devices. This new transport protocol should be able to cope with seamless switch overs, different simultaneous connections and aggregation between multiple access technologies, as explained in Fig. 2.3. These functionalities cannot be implemented on top of current standard Internet protocols, so we must consider extensions to existing ones.

Possible transport protocols are developed on top of SCTP (Stream Control Transmission Protocol) [30] by striping and transmitting data across multiple network interfaces at the same time [31], or MPTCP (MultiPath Transmission Control Protocol) to use simultaneously several networks to transmit [32]. An advantage of MPTCP is that it does not need any additional requirements on the network side, being entirely implemented at end-hosts. MPTCP has also several working implementations, notably on Android smartphones [33], and a large scale commercial deployment inside Apple iOS 7 operating system [34].

The transition toward the simultaneous use of multiple access technologies brings a number of issues. In order to benefit the most from non-delayed offloading, it becomes mandatory to develop innovative communication stacks beyond classic IP-protocol, capable of supporting advanced features (e.g., multiple instantaneous connections, data aggregation, and inter-technology switchovers). Extensions to standard protocols started to appear to cope with these issues (e.g., SCTP, MPTCP), enabling to aggregate together the bandwidth offered by different technologies, and allowing seamless handover between distinct access technologies. Nevertheless, a widely accepted transport protocol to handle transparently several flows in parallel on separate interfaces has not yet been standardized.

2.2.2 D2D

Real-time D2D offloading is often associated with cooperative strategies to exploit concurrently the availability of multiple interfaces. Thus, the network should be capable of coordinating data retrieval in a distributed fashion. Initially, only out-of-band transmissions were considered. However, the latest developments in the 3GPP LTE Standard (Rel-12) propose integrating direct in-band communication capabilities into the future cellular architecture [35]. This provides additional flexibility to the network but raises issues such as mutual interference and resource allocation, since D2D transmissions take place in the same band as the cellular transmissions. Cooperative data retrieval is shown to improve the spectral efficiency of the network [36]. While classical studies show that the theoretical transport capacity of multi-hop ad hoc networks scales sub-linearly as $\Theta(\sqrt{n})$ [37, 38] with the number n of users, cooperation among nodes brings linear scaling law of $\Theta(n)$ [39].

If peers are not stationary, link quality may suddenly change, making it difficult to guarantee QoS. If data delivery can be deferred, a better candidate for data distribution is delayed offloading (see Section 2.3.2). To guarantee real-time requirements, most architectures assume low mobility and co-located peers interested in receiving a common content [40].

Cooperation

The basic idea of cooperative downloading, is to retrieve each data chunk only once through the cellular channel, and to share it through short-range links [41]. Cellular accesses have to be coordinated among willing nodes in order to relieve the cellular infrastructure. One node can act as the central controller, estimating the available throughput on the cellular link for each other node, so to coordinate content retrieval among peers. Data can be sent using the cellular infrastructure only to those users with the best channel quality, and subsequently relayed to all other nodes by means of D2D transmissions [42]

Cooperative downloading of real-time data can also be achieved through turn-based strategies and successive broadcasts on the WiFi interface to other interested nodes [43, 44, 45]. Like a peer-to-peer network, these strategies are fair and resilient to node failures. The successive broadcasters follow the density of nodes involved in data distribution, to trade off collisions on the wireless medium and data redundancy.

Additional works propose a scalable video multicast solution that jointly exploits cellular broadcast, network coding and D2D transmissions [46]. The base video layer is broadcasted to all the users within the cell. The enhancement layers, instead, are transmitted only to a subset of users. Modulation and coding schemes employed for transmission are the outcome of a joint optimization problem involving D2D transmissions and cellular coverage. The enhancement layers are then forwarded to remaining users through D2D transmissions. Testbeds for cooperative real-time video streaming among mobile nodes are proposed in [47, 48, 49].

Finally, the classical problems of cooperative offloading (i.e., neighbor discovery, connection establishment, and service continuity) can be resolved by adopting a network driven approach with a cellular architecture intelligent enough to assist connected users in the content discovery and connection establishment phases [50].

The complexity of cooperative content distribution is high, involving the joint optimization of different access technologies, interference, transmission rates, scheduling and energy efficiency. Centralized or distributed solutions have been developed and tested through simulation, theoretical analysis and real test beds. Optimal solutions are NP-hard, so heuristics need to be adopted. Most of the papers focus on how to achieve enhanced data rates, saving at the same time battery. In this context, an energy consumption model is provided in [40]. Security and trust considerations concur in making the problem even more complex. A novel approach, involving a continuous control wielded by the network, could possibly simplify the problem.

3GPP D2D

Recent developments in the 3GPP LTE Standard (Rel-12) propose integrating direct in-band communication capabilities into future cellular architectures [35], often also referred to as cellular network underlay [51] or device-to-device (D2D), rather than using traditional technologies working on unlicensed bands (mainly IEEE 802.11 and Bluetooth). This paves the way for a combined use of cellular and short-range transmissions, offering users various degrees of freedom for transmission and a network-assisted environment. End-users discover each other in proximity through explicit probing [51] or via the access network guidance [52]. Upon discovery, nodes can communicate using either dedicated resources or a shared uplink cellular channel [53]. D2D communications are then triggered by the cellular network, and fall under continuous network management and control. For these reasons, they can also be employed for load balancing purposes [54]. Hence, D2D could become the ideal platform to develop data offloading in the future, because it may achieve higher resource utilization by reusing the spectrum of physically neighboring devices, while reliability may significantly increase thanks to a shorter link distance. Furthermore, D2D capabilities enable LTE to become a very interesting technology for public safety communications (PSC) [55]. Anyway, critical issues such as neighbor discovery, transmission scheduling, resource allocation and interference management, in particular in the case of multiple cell deployments, still need to be addressed in order to proceed to the effective integration in future cellular architectures. Related tutorials provide the reader with a broader overview on the existing research challenges and applications of D2D [56, 57].

Interference management and transmission coordination represent thorny problems that must not jeopardize the QoS of cellular users in the primary network. When two or more pairs of neighboring nodes are willing to communicate, they may use the same resources. In this case, interference is a major issue. The network could limit the maximum transmission power of D2D peers [51]. The optimization of radio resource allocation help decrease the

mutual interference between D2D communications and the primary cellular network [58]. Similarly, joint resource allocation and power control schemes can also be adopted [59]. Note that, for the intrinsic real time requirements of cellular networks, the computational complexity of resource allocation algorithm represents a tangible issue [60]. Resource allocation are treated extensively in [61, 62].

D2D communications as an underlay of cellular networks represent a significant leap forward towards the deployment of heterogeneous networks. D2D communications in this case share resources with cellular transmissions, therefore generating mutual interference. Consequently, resource allocation optimization, power control, and device discovery are key topics for the research community. However, the underlay approach does not exploit surplus bandwidth available through complementary technologies, but rather aims at taking advantage of parts of the LTE spectrum that may be under-utilized. Still, this could be the ideal technology to support the predicted data growth. Cellular operators can make profits on network-assisted D2D communications, supervising at the same time the resource consumption and the QoS of the network, which is difficult in out-of-band offloading techniques.

2.3 Taxonomy: Delayed offloading

In delayed offloading, content reception may be intentionally deferred up to a certain point in time, in order to reach more favorable delivery conditions. We include in this category the following types of traffic: (i) traffic with loose QoS guarantees on a per-content basis (meaning that individual packets can be delayed, but the entire content must reach the user within a given deadline) and (ii) truly delay-tolerant traffic (possibly without any delay guarantees). The relaxation in the delivery constraint allows also moving traffic opportunistically, which, by definition, can only guarantee a probabilistic delivery time. If data transfer does not end by the expected deadline, the cellular channel is employed as a fall-back means to complete the transfer, guaranteeing a minimal QoS. Despite the loss of the real-time support due to the added transmission delay, note that many mobile applications generate content intrinsically delay-tolerant. Enabling an alternate distribution method during peak-times (when the cellular network is overloaded or even in outage) becomes an interesting extension and represents a fundamental challenge for offloading solutions.

Most of the time, the offloading strategy relies on the cellular network to bootstrap the distribution process (to infect seed users in D2D-based offloading) or to ensure minimal QoE guarantees (fall-back transmissions when the deadline approaches). Before the deadline, the content is preferably delivered through the alternative technology. Unlike the approaches set forth in Section 2.2, delayed offloading directly exploits the mobility of nodes to create communication opportunities. As a side effect, performance heavily hinges upon the mobility pattern of users. A short digression on mobility characteristics is thus necessary to better catch the fundamental properties and inherent limits of delayed offloading.

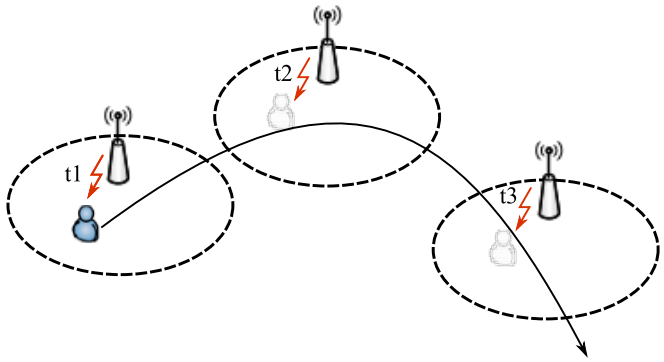


Figure 2.4: AP-based offloading. Mobility allows to receive delay-tolerant data from different APs at different times.

Since messages are forwarded only during contacts with users or APs, the statistical analysis of such encounters becomes particularly meaningful. First, the time until a new encounter occurs (the inter-contact time) gives an effective indication of the delivery capacity inside the opportunistic network. In addition, when contacts occur, knowing for how long they last (their contact time) would help us to foresee how many pending messages can be forwarded. The distribution of contact times also affects the total delivery capacity when multiple users compete for the same wireless channel, because contacts can be wasted due to contention and scheduling. These properties have been deeply investigated in trace-based studies [63, 64]. Common understanding is that inter-contact and contact times between mobile users often display a power law distribution with an exponential heavy tail. Analogous results hold also for contacts between users and fixed APs [12]. However, as pointed [65, 66], these results focus on aggregate inter-contact distributions, and are not representative of the network behavior, which instead depends on the properties of individual pairs. An interesting addition to the standard contact and inter-contact analysis considers an extended notion of contact relationships [67].

2.3.1 AP-based

AP-based strategies take advantage of a complementary networking backbone, often formed by fixed WiFi APs, to deliver data bypassing the cellular network. The complementary access network may be part of the cellular operator network, or may be completely separate. In the latter case, an agreement between operators should be envisioned. At first, this approach looks similar to the non-delayed case. The delay-tolerance of content is exploited here, with data exchange happening upon subsequent contacts between the user and different APs exploiting a sort of space-time diversity, as illustrated in Fig. 2.4. The movement of end-users creates contact opportunities with fixed APs defining the offloading capacity of the network.

Current research efforts aim at predicting the future offloading potential through past behaviors of users such as mobility, contacts with APs, and throughput. Using this prediction,

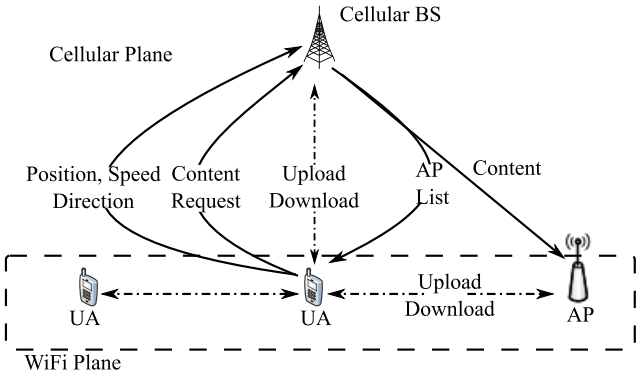


Figure 2.5: MADNet system architecture: when a mobile node wants to communicate, it makes a request to the cellular BS, which may replies directly forwarding the content through the cellular network or sending the content to a neighboring AP. The BS predicts the route of the nodes using the status information sent by the mobile node.

the offloading coordinator may decide which fraction of data to offload, when, and to whom. Possibly, downstream content is split in several pieces, which are then pro-actively sent to APs that nodes will (probably) encounter in the future. An alternative research area aims at identifying the optimal number of fixed APs and their geographical location, starting from a known user's mobility pattern. In the following sections, we present in detail these two approaches.

Prediction-Based Offloading

The prediction of node mobility combined with the knowledge of the geo-localization of fixed APs concur to enhance performance [68]. The predictor can inform the offloading coordinator of how many APs a mobile node will encounter during its route, when they will be encountered, and for how long the user will be in AP's range. The algorithm seeks to maximize the amount of delay-tolerant data to be offloaded to WiFi, ensuring also that data is transferred within its deadline. Similarly, the MobTorrent architecture exploits the hybrid infrastructure, data pre-fetching, and cache replication at fixed APs [69]. Download requests are issued through the cellular channel. Requested data is cached in advance to APs using location information and the mobility history of users.

A network-centered architecture named MADNet integrates cellular, WiFi APs, and mobile-to-mobile communications [5, 70]. MADNet employs the cellular network as a control channel. The system is explained in Fig. 2.5. When a mobile user asks for some content, the offloading coordinator replies with the list of the surrounding APs where it may pick up the requested data. The offloading coordinator predicts the neighboring APs by exploiting positioning information. Simulation results show that a few hundreds APs deployed citywide could offload half of the cellular traffic. Tradeoff between delay, QoS and energy efficiency, is discussed in [71]. The main contribution of the work is to explore the energy efficiency of delayed transmissions, because WiFi has, in general, better efficiency than cellular trans-

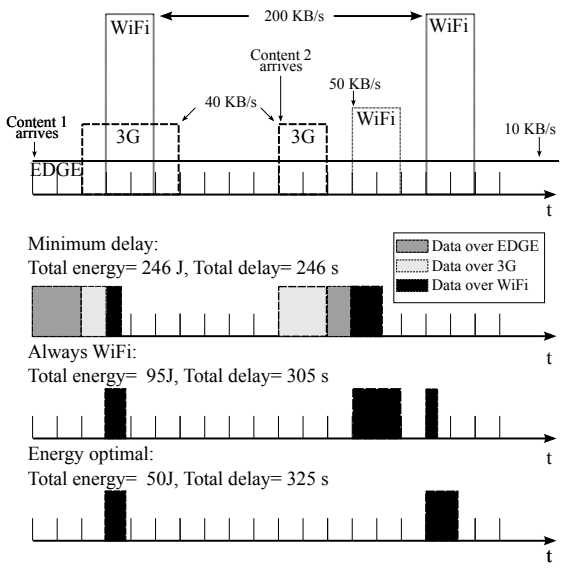


Figure 2.6: Three offloading strategies from the SALSA framework: The *Minimum delay* strategy minimizes the total delay, selecting always the channel with the fastest data rate; the *Always WiFi* strategy, uses only WiFi APs, regardless of their data rate; the *Energy optimal* strategy minimizes the energy consumption, using always the most energy efficient channel.

missions, as depicted in Fig. 2.6. The transmission decision relies on the prediction of the future available bandwidth for each possible access network, estimated as the average rate achieved over past transmissions, or as a function of the received RSSI. Go et al. suggest a heterogeneous city-scale mobile network that opportunistically offloads some cellular traffic to existing WiFi APs [72]. The core of the system relies on DTP (Delay Tolerant Protocol) to mask network disruptions from the application layer [73]. DTP binds the connection to a unique *flow ID* rather than to a tuple of physical IP addresses and ports, providing to applications the illusion of a continuous connection. The proposed system employs dedicated proxies located at the edge of the access network that hide user disconnections to application servers. Finally, Malandrino et al. relax the assumptions of an accurate prediction scheme by proposing a model that considers the uncertainty of mobility through a Gaussian noise process [74]. Each AP performs a joint pre-fetching and scheduling optimization through a linear programming problem, aimed at maximizing the aggregate data downloaded by users.

Focusing on user-centered policies instead, Wiffler is capable of exploiting the delay tolerance of content and the contacts with fixed APs [75]. Wiffler predicts future encounters with APs based on past contacts, deferring transmission only if this saves cellular traffic, employing an heuristic. By means of trace-based simulations, the authors show that with a prediction based only on the last four encounters, Similarly, Yetim et al. consider the decision of waiting for WiFi encounters rather than using the cellular connectivity as a scheduling problem [76]. Different sizes and deadlines are considered for content. Presuming that each content may be divisible in smaller scheduling units of MTU size, the scheduler exploits short windows

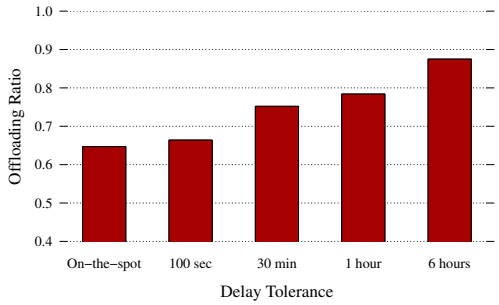


Figure 2.7: Offloading ratio of delayed AP transfers with various deadlines, 100% of traffic delayed, Seoul dataset. Increased delay-tolerance values result in an increased fraction of data offloaded.

of WiFi coverage to shift up to 23% of the total traffic away from the cellular network. Although delayed transfers may substantially improve the offloading performance of cellular networks, delaying all transfers up to their maximum delay tolerance is often an ineffective strategy. In case of absence of WiFi, each delayed transmission frustrates user experience. An ideal solution is to identify the optimal instant of time after which a user should stop deferring transmissions and start transferring data using the cellular interface, trading-off offloading efficiency and user satisfaction [77].

A key requirement to drive effective AP-based offloading is the ability to predict future capacity. The decision to wait for a possible upcoming offloading opportunity or to transmit data through the cellular channel (considered as a scarce and costly resource) is of utmost importance when dealing with delay-tolerant data. Distributed and centralized prediction methods have been developed based on the knowledge of prior encounters, mobility patterns, AP locations, and bandwidth availability. Future researches in this sense should also take into account the obvious trade-off between the overhead brought by context-awareness and the accuracy of prediction. Furthermore, most existing solutions for AP-based offloading rely on optimization frameworks, which are complex to solve and need heuristics. As a result, an interesting research topic might be to explore alternative self-adaptive approaches (e.g., based on machine learning techniques).

Optimal Deployment

Similarly to Section 2.2.1, we address here the feasibility and capacity of AP-based offloading, this time considering delay-tolerant content. Lee et al. demonstrate that increasing the delay-tolerance of content substantially improves the ratio of offloaded traffic, as depicted in Fig. 2.7 [12]. Additional findings suggest that the average completion time for delayed offloading is always much lower than the maximum deadline. Surprisingly, the authors discover that, with large content, delaying the transmission may result in faster completion times than not delaying it at all. This is motivated by the fact that WiFi usually offers higher data rates

than cellular networks, which translate into shorter aggregate completion times. Theoretical bounds for delayed data offloading with WiFi AP can be derived analytically as a function of the number of users and the availability of APs using queuing theory concepts [78]. Optimal placement of APs is discussed in [5, 79, 80]. Citywide coverage can be guaranteed by the integration of only a few hundreds of APs, offloading half of the cellular traffic. Simple heuristics for deployment suggest upgrading the network capacity in a limited number of locations. The underlying intuition is that most users pass by a limited number of hub locations during daily commutes. Thus, by upgrading only a tiny fraction of the network, providers may strategically support growing traffic with minimum investments.

In the context of vehicular networks, Abdrabou and Zhuang study the minimum number of APs to cover a road segment in order to guarantee a probabilistic connection time [81]. The maximum distance allowed between two neighboring APs is estimated analytically in [82]. Malandrino et al. model data downloading in a vehicular environment as an optimization problem, considering also the presence of fixed APs [83]. To counter the scarce availability of APs due to placement and maintenance costs, they take also into account parked vehicles, acting as additional APs, to assist in data distribution [84]. Astudillo et al. study the performance of broadcasting in a vehicular network using fixed APs to download data [85].

Similarly to the non-delayed offloading case, performance is tied to AP density. Nevertheless, the time dimension matters here, as increased delay-tolerance translates into an extended fraction of offloaded data. We may find a number of placement algorithms that exploit the delay tolerance of content by adding APs where people are most likely to transit. The problem has similarities with the optimal road-side unit placement strategies for ITS (Intelligent Transportation Systems) applications [86]. Cost based analysis proposed in the literature, help better understand the existing trade off between the cost of deploying more APs and the offloading benefit [5, 79, 83]; unfortunately, many solutions are not directly comparable due to differences in reference scenarios, use cases, and simulation parameters.

2.3.2 D2D

In delayed D2D offloading, content distribution is delegated to end users: in a broad sense, *users are the network*. They actively participate in the dissemination process by exploiting D2D communications. Mobility is an additional transport mechanism, creating opportunities for infected users to transfer data employing a delay-tolerant (DTN) approach [87].[†] DTNs allow content forwarding through store-carry-forward routing regardless of the existence of a connected path between senders and receivers, at the cost of additional reception delays. Golrezaei et al. analyze the theoretical performance bound for throughput [88]. Store-carry-forward routing coupled with simple caching policies at nodes could bring a linear throughput increase in the number of nodes. Apart from providing communication op-

[†] Delay-tolerant, disruption-tolerant, opportunistic, challenged, and intermittently-connected networks are used in the literature most of the time as synonyms, although sometimes they denote slightly different concepts. With respect to the offloading solutions, they can be considered as synonyms.

portunities when a fixed infrastructure is missing – such as in the case of tactical networks, or in least developed countries [89] – DTNs can be also coupled with a fixed infrastructure. Delayed D2D-based offloading is often seen as a quick and inexpensive way to increase mobile network capacity and to handle the predicted data tsunami [90]. Unlike AP-based approaches, the gain of this schema relies entirely on redundant traffic. However, this proves to be relevant for content access, as popularity follows Zipf-like distributions [91] – a small subset of content results extremely popular and is requested by a large number of co-located users, causing severe congestion and bandwidth shortage at RAN level. Moreover, the DTN approach supports conditions where standard multicast and broadcast approaches (also included in LTE [92]) cannot be used. For example, it supports all cases where popular content is requested by users during a given time window (short enough to guarantee that users are still physically co-located in the same region), but not necessarily at the exact same time. Note however, that D2D offloading is also beneficial when multicast in the cellular network can be used [93].

From its characteristics, it follows that the DTN approach can only address the diffusion of data with loose delivery constraints. Content is ideally supplied only to a small fraction of selected users among those who requested it. These *seeds* bootstrap the propagation by transferring content to users within their transmission range, as in Fig. 2.8. In this category, we also include strategies where the communication opportunities between nodes arise as a side effect of duty cycling of ad hoc interfaces. D2D interfaces are typically energy-hungry, and it is possible to apply energy saving policies to them, dynamically toggling between on and off states [94, 95].

A number of strategies can be used to disseminate the content among mobile nodes. In principle, any forwarding or data dissemination scheme proposed for opportunistic networks can be used. Hereafter, we just give a few examples. Interested readers can refer to [96, 97, 98] for dedicated surveys. From the seminal work of Vahdat and Becker that firstly proposed mobility-assisted epidemic forwarding [99], many routing protocols in the context of DTNs have been proposed. Notable works on forwarding strategies from Spyropoulos et al. [100], Lindgren et al. [101], and Burgess et al. [102] go beyond simple epidemics by tackling statistical and mobility characteristics of nodes, and targeting the case of separate subsets of users with different interests. Mathematical frameworks based on ODEs and Markovian models provide theoretical bounds on the performance of dissemination delay and the number of copies of the message in the network [103, 104]. Similarly, analytical bounds on dissemination delays are derived from the speed and density of nodes in [105, 106].

Most of the research efforts in this field focus on the design of efficient algorithms for the optimal selection of seed users, in order to minimize the number of users that receive the content through the cellular interface. On the other hand, a number of works deal with network architecture and protocol design. The former approach relies on social networking analysis or machine learning techniques to predict *which* users are the best gateways for content. The latter tackles the choice of *what* type of traffic to offload and *how*, defining communication

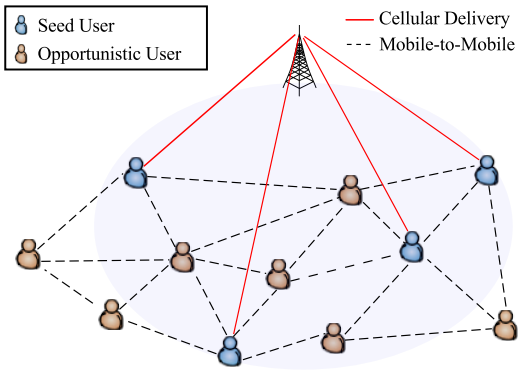


Figure 2.8: Data offloading through delay-tolerant networks. *Seed* users initially receive the content through the cellular network. Direct ad hoc transmissions are used to propagate the content in the network

protocols and network architectures. In the following paragraphs, we will detail better the two approaches.

Subset Selection

Ioannidis et al. propose pushing updates of dynamic content from the infrastructure to end-users [107]. They assume that the cellular infrastructure has a fixed aggregate bandwidth that needs to be allocated between end-users. Peers exchange opportunistically any stored content between them. A rate allocation optimization is proposed to maximize the average freshness of content among all end-users. Two centralized and distributed algorithms are presented. Similarly, Han et al. and Li et al. tackle the offloading problem employing a subset selection mechanism based on the user contact pattern [108, 90]. While in the first work Han et al. study how to choose a subset of dimension k to be initially infected [108], Li et al. consider the optimal subset selection as an utility maximization problem under multiple linear constraints such as traffic heterogeneity, user mobility, and available storage [90]. The subset selection problem is NP-hard, similarly to the case of the minimum AP set-selection problem presented in [79] and discussed in Section 2.3.1. Both works propose *greedy* selection algorithms to identify a sub-optimal target set. A point in common for all the subset selection strategies is that the network provider should be able to collect information about node contact rates in order to compute the best subset.

Using social networking arguments, Barbera et al. analyze the contact pattern between end-users, in order to select a subset of central VIP users that are important for the network in terms of *centrality* and *page-rank* [109]. The key idea is to transform these few central VIP users into data forwarders between standard nodes and the Internet. The authors exploit the repetitive and periodic mobility of humans to train the selection algorithm to build the networks' social graph over which the VIPs selection is made. An analogous approach is exploited by Chuang et al., which merge the subset selection problem with the concept of social relationship between end-users [110]. They propose a community-based algorithm that se-

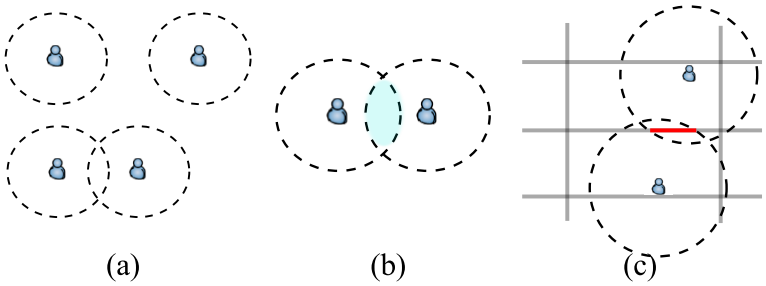


Figure 2.9: Coverage metrics in the TOMP framework: (a) the *Static Coverage* does not take into account any future movement, so nodes are considered in contact or not based on their present position; (b) the *Free Space Coverage* considers the possible movement of nodes in free space: the future meeting probability is the area of intersection of the two circles that represent the possible movements of the two nodes; (c) the *Graph-based Coverage* takes into account the underlying structure of road graph to limit the prediction to the road graph.

lects users belonging to disjoint social communities as initial seeds, in order to maximize the offloading efficiency. In effect, the selection of initial seeds based only on encounter probability proves to be insufficient, as users with high encounter probability might belong to the same community. The goal is to select the set of initial sources so that both cellular traffic load and delivery time are minimized. Also in this schema, mobile end-users are required to upload periodic information on the most frequent contacts, in order to let the centralized algorithm to choose the best subset of seed users.

Baier et al. approach the subset selection problem by predicting the movement of end-users in order to estimate future inter-device connectivity [111]. The system, named TOMP (*Traffic Offloading with Movement Predictions*), retrieves information about actual positioning and speed of mobile devices rather than connectivity patterns. The framework selects as seed users the nodes that have the best future connectivity likelihood with other nodes based on movement prediction. As explained in Fig. 2.9, TOMP proposes three coverage metrics to predict the future movements of nodes: static coverage, free-space coverage, and graph-based coverage.

Selecting high potential nodes as seeds of the dissemination process influences the performance of the offloading strategy. Wisely chosen seed users may infect a larger number of nodes, resulting in lesser late retransmissions. Subset selection algorithms commonly employ information on social interactions among users and their mobility patterns to figure out which nodes have the best features. Note that a control channel, binding the end-nodes to a central entity, is usually required in order to transfer context information. The performance of the offloading algorithm relies heavily on the understanding of the system dynamics. For this reason, it is essential to analyze how nodes meet creating communication opportunities in a fine-grained fashion, and characterize mobility at the microscopic level. Offloaded data vary from 30 to 50% for all the surveyed papers depending on the delay-tolerance and the dataset considered. However, apart one notable exception [110], only small scale and very specific datasets have been evaluated (typically around 100 users), providing a limited confi-

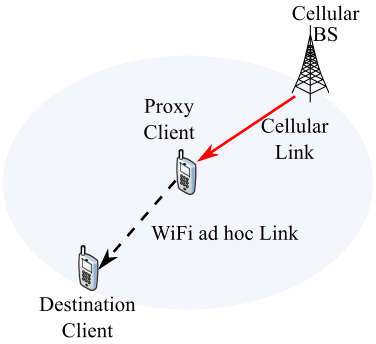


Figure 2.10: Network extension: destination client experiences bad cellular connectivity . After the discovery of a neighbor node with better channel conditions, data is routed through this "proxy client" in the cellular network.

dence in the generality of results.

Strategies

Luo et al. designed a new unified architecture for cellular and ad hoc networks, to leverage the advantages of each technology [112]. In this case, the goal is to increase the throughput experienced by mobile users by taking advantage of neighbors with better cellular connectivity, employed as a proxy. The working schema, as shown in Fig. 2.10, allows mobile users experiencing a low cellular downlink channel rate, to connect via ad hoc links to a neighbor with better cellular channel conditions. The proxy node then acts as a gateway for data traffic of its peers. Data is further relayed through IP tunneling via intermediate relay clients to the destination, using the ad hoc link. The paper proposes also two proxy discovery protocols (namely *on demand* and *greedy*), and analyzes the impact of the proxy relaying schema on the cellular scheduling.

Mayer et al. propose a routing scheme for the offloading of unicast message exchange between end-users [113]. The offloading schema is based on a simple assumption: the higher the probability that a message can be delivered through the infrastructure in case of failing opportunistic delivery, the longer DTN routing takes to deliver the message. In effect, the protocol initially attempts to deliver messages through opportunistic communications and switches to the infrastructure network only when the probability of delivering the message within the deadline becomes unlikely. This opportunistic/infrastructure routing decision is taken locally exploiting information exchanged with other nodes upon encounters. Key contextual information includes awareness for destination node and infrastructure capabilities. In this way, the system tries to offer a reliable message delivery, while saving cellular traffic at the same time.

Another solution, named RocNet, exploits the difference of traffic load among different locations [114]. Consider the distinct instantaneous traffic volume in a business district and a residential district during daytime. In case of localized RAN congestion, each delay-tolerant

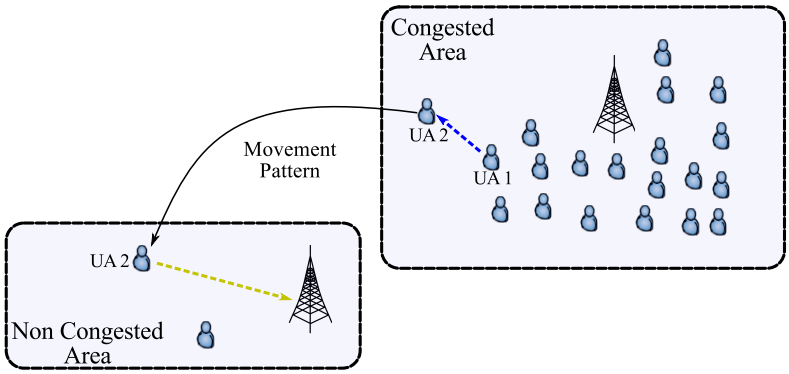


Figure 2.11: High level overview of RocNet. UA 1 is in a congested area. Upon discovering, UA 1 forwards data to UA 2 if it is more likely to move to a non congested area than UA 1.

data request originated in that area, instead of being transmitted to the overloaded cellular BS, is forwarded to a neighbor that is likely to head toward a less congested area, as shown in Fig. 2.11. A particle filter is employed to predict future movement pattern of neighbor users, starting from its movement history. When a terminal is in a congested area, a coefficient of variation is exchanged upon opportunistic meeting with neighbors, to decide which user is more likely to move to a low-congested area.

Finally, some architectures exploit the availability of hybrid delivery options (Cellular, APs, and opportunistic). Pitkanen et al. describe a system to extend the range of fixed WiFi APs through the DTN approach [115]. Delay-tolerant data is shifted from the cellular network to the closest WiFi AP, contributing to preserve cellular bandwidth for real-time and interactive applications. Similarly, Petz et al. introduce MADServer, an offloading-aware server that enables the distribution of web-based content through a multitude of access networks [116]. Both systems use contextual information from users, to predict where to cache data in advance, and are able to split the content into multiple pieces, independently delivered on different access networks. Small and time critical content is always transmitted over the cellular infrastructure, while large data, such as videos and pictures are offloaded only when it is beneficial and within deadline.

The definition of network architectures capable of exploiting different technologies to deliver content is a key milestone for the research community. The current trend is toward network-aided offloading schemes, where the cellular network guides its connected peers in the neighbor discovery and connectivity management phase. The routing scheme takes advantage of well-placed neighbors used as preferred gateways for data forwarding. The substantial use of context information harvested from end-users, or exchanged locally, is exploited to drive the routing decision through the optimal interface. Future challenges include the development of novel coordination mechanisms and inter-technology scheduling policies to control content retrieval between multiple access technologies and opportunistic networks. Cellular operators are particularly interested in the development of innovative ca-

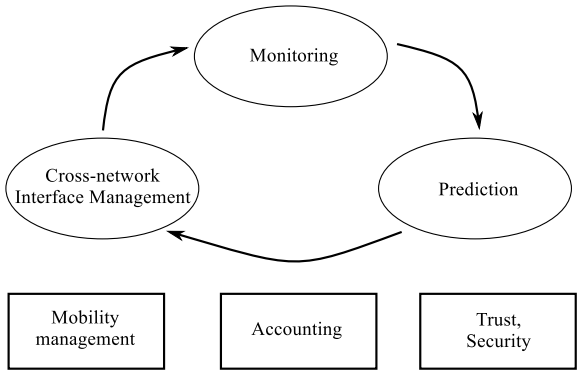


Figure 2.12: Offloading coordinator functional building blocks.

pacity models able to predict the additional gains provided by the activation of offloading, and to plan how much traffic they can divert from their core network.

2.4 Target offloading Architecture

The analysis conducted so far reveals that the various forms of offloading are quite different, in terms of both network infrastructures and delivery delay requirements. Despite this, it is still possible to identify from the specific solutions a number of common functionalities making up an advanced offloading scheme. The challenge is to go beyond what is done today, which is mainly a user-initiated offloading process. The opportunity for operators to drive the offloading process will provide them with better network management options.

In order to make this vision possible, we need to extract a number of generic high-level functionalities that make up the offloading system. This analysis is significant in view of the integration of offloading capabilities into future mobile networks architecture. Fig. 2.12 provides a high-level scheme to help us drive the discussion. Most of the works we surveyed consider an *offloading coordinator*, an entity specifically dedicated to the implementation of the actual offloading strategy. Its main task is to pilot the offloading operation depending on network conditions, users' requests, and operator offloading policy. While conceptually represented by a single entity, its physical location in the network may vary, and sometimes its implementation could be totally distributed. However, it is possible to identify, among all, three main interdependent functional blocks for the offloading coordinator: (i) monitoring, (ii) prediction, and (iii) cross-network interface management.

- Monitoring provides methods to track the actual data propagation spreading, user's requests, and to retrieve contextual information from nodes and the network. Retrieved information is necessary to evaluate and execute the offloading strategy. The monitoring block often requires the presence of a persistent control channel that allows end-users to interact with the offloading coordinator (e.g., the cellular channel is explicitly

employed with this purpose in [5, 117]). Harvested information is then passed to the prediction block to be processed.

- Prediction relates to the ability of the offloading coordinator to forecast how the network will evolve based on past observations. Typical prediction deals with mobility [68, 111], contact patterns of users [5], or expected throughput [41, 75]. Such predictions are then used to pilot the entire offloading process more efficiently. This is the block where typically the offloading intelligence resides. The complexity of the prediction should trade off its applicability, in order to guarantee the real-time operation of the offloading process. Predicted values are transmitted to the interface management block in order to drive the offloading process.

- Traditional approaches manage each interface independently. However, integrated management allows exploiting in parallel the benefit of each available interface. Cross-network interface management deals with deciding on which network the required content (or parts of it) will flow. Concepts such as load balancing, throughput maximization, congestion control, and user QoE (Quality of Experience) relates to this functional block. By exploiting this information, the network itself will be able to identify the current situation and optimize its performance. For instance, ANDSF and IFOM already use this capability [26, 29]. They are able to shift selected data on a given network interface, in order to obtain a benefit.

Additional transversal subjects emerge from the analysis of the literature. For instance, *mobility management*, *accounting*, and aspects related to *trust and security* are essential to support offloading strategies in mobile network architectures. Mobility management involves the seamless handover between different base stations due to the mobility of users. Accounting functionalities enable proper accounting and charging information for the offloaded traffic and users. This is a key component in order to design incentive mechanisms to stimulate the participation of mobile users in the offloading process. Finally, trust and security mechanisms guarantee the privacy and the integrity of both infrastructure and D2D communications. This block is essential since most offloading strategies transform the user into an active network element.

These can be regarded as the basic functional building blocks that mobile networks should provide to ensure offloading capabilities. Anyway, we stress that, depending on the specific implementation, the proposed functionalities may be present or not. For instance, mobility management modules are elemental in non-delayed offloading, in order to handle the handover between different APs, and to secure continuity of ongoing data session. On the other hand, the same block could be disregarded when dealing with delayed D2D transmissions.

2.5 Assessing Mobile Data Offloading

It is quite challenging to compare the performance of different offloading strategies only on the basis of the results reported in the literature, because the evaluated metrics often differ. In addition, we can assess the performance of offloading from the perspectives of both network operators and users, which have essentially divergent needs [118]. In this section, we will give hints on the metrics that we believe important for the evaluation of offloading strategies. In addition, we discuss simulators, mobility models, and testbeds, which play a significant role in performance evaluation.

Metrics

From a cellular operator's point of view, offloading should serve as a reserve of capacity, which may be added to the network in case of heavy congestion. For this reason, a significant challenge is to quantify the additional capacity brought by the use of offloading strategies. The most notable effect of offloading should be the reduction of traffic load and congestion in the primary network. Nevertheless, capacity improvements depend, among other things, on the number of mobile devices or wireless APs involved in the process, on the mobility of nodes, on the size and the delay-tolerance of the offloaded content. On the other hand, user satisfaction is often associated with Quality of Experience (QoE), so the received throughput and timely reception parameters are regarded as the most important parameters. Commonly employed metrics of interest today in the literature are the following:

- Offloading Ratio or Offloading Efficiency. It is the fundamental parameter to evaluate the effectiveness of any offloading strategy from an operator point of view. It is measured as the ratio of the total traffic offloaded (transferred through alternative channels) to the total traffic generated [12], or as the ratio of the total load of traffic that flows on the cellular channel after the offloading process to the traffic on the infrastructure in the absence of any offloading strategy [117].

- Offloading Overhead. The offloading overhead metric evaluates, in a broad sense, the amount of additional control data required by the offloading mechanism. For instance, as explained by Sankaran [28], in the IFOM scenario, the overhead is represented by the messages needed to exchange and discover IFOM capabilities between involved nodes. In the Push-and-Track scenario, the offloading overhead depends on the control traffic that flows into the infrastructure channel, intended to pilot the offloading process [117].

- Quality of Experience (QoE). From a user perspective, the most critical metric is the Quality of Experience (QoE), which is linked to its satisfaction. For any offloading class, the total achievable throughput is a common but important metric. The QoE indicator is then made up of several sub-metrics that depend on the application and the

type of offloading. For instance, video streaming QoE-metrics are the Peak Signal-to-Noise Ratio (PSNR) and the amount of packet loss. In delayed offloading, the delivery time is the most meaningful metric, representing the amount of time before content reception.

- Power Savings. In some works, the concept of offloading is associated with the power savings that may be attained by the nodes. This is possible because the WiFi interface is more efficient in terms of energy per bit than the cellular interface. Traffic offloading algorithms are interesting to achieve energy savings.

- Fairness. Fairness in terms of resource usage (in particular energy consumption) can be an important evaluation parameter. Fairer systems tend to distribute resources uniformly without relying too much on the same users. This aspect is critical in D2D offloading, where an unbalanced use of resources could lead to premature battery depletion. For instance, seed-based offloading strategies risk being unfair, because data is transmitted to a limited number of users that retransmit it on the secondary channel. Even if this strategy could reduce the overall energy consumption, it is unfair in terms of user's individual energy consumption.

It is important to note that evaluated metrics often depends on how performance is assessed. In particular, power saving is commonly evaluated in experimental works, while offloading efficiency is typically estimated through simulation. In general, simulation-based evaluations are likely to propose a system-wide approach, i.e., they consider the whole network, even with some approximation. On the other hand, evaluation based on real experiments, due to the inherent complexity of assembling large-scale scenarios, focus more on terminal-level parameters and small-scale experiments.

2.6 Cellular Networks and the Bandwidth Crunch: Alternative Solutions

As already mentioned, the intricate problem of mobile data explosion can be addressed in several ways. Hence, we briefly review alternative solutions to the capacity problem in cellular networks linked to data offloading. We identified five main categories related to data offloading, each one bringing advantages and disadvantages:

- Addition of small-size base station and/or femtocells.

- Multicasting/broadcasting data inside the cell.

- Integration of cognitive radio mechanisms.

- Proactive pushing of popular content on devices.

It is worth to note that many of these possibilities are orthogonal to each other, and can be deployed at the same time. In addition, the methods outlined in this section may also complement the strategies presented along the chapter.

Small-Sized and Femto-cell Deployment

The first solution adopted by the majority of cellular providers to face data growth is to scale the RAN by building more base stations with smaller cell size. Reducing the size of macro-cell increases the available bandwidth and cuts down the transmission power [119]. An obvious drawback is that operators have to build additional base stations. Equipment costs, site rental, backhaul, and power consumption, make this strategy very expensive in terms of CAPEX and OPEX. In addition, according to [120], only a small fraction of mobile users (around 3%) consume more than 40% of all mobile traffic. Consequently, the majority of users gets only a minimal benefit from this strategy, as heavy consumers will continue to grasp the bulk of the bandwidth.

Another possibility is to push the adoption of femtocells. The approach is analogous to AP-based offloading but makes use of the same access technology of the macro-cell. However, since femtocells work on the same frequency as the macro network, interference management becomes challenging [121]. Performance of femtocell-oriented offloading is investigated in [122, 123]; other works compare the gains brought by femtocells against AP-based offloading [15, 124]. Energy-related topics are presented in [125]. Interested readers should also refer to existing surveys on femtocells in the literature [126, 127]. The trend toward smaller cells is part of the so-called *HetNet* paradigm, in which cellular macro-cells coexist and overlay a myriad of smaller cells. This affects the design of the resource allocation scheme, and ongoing researches focus on the decision if a user should be served by the macro or by a closer small-cell. A flexible small-cell deployment helps in eliminating coverage holes, and increasing the network capacity in some regions inside a macro-cell [128].

Multicast/Broadcast

When many users in spatial proximity ask for the same data, multicast could emerge as a good alternative to data offloading for comparable use cases. Multicast employs a single radio link, shared among several users within the same radio cell.[‡] Logically there is no interaction, and users can only receive content. Multicast is a clever strategy to provide content to multiple users exploiting redundancy of requests, allowing in principle great resources saving.

Besides requiring modifications in the cellular architecture, multicast has intrinsic and still unresolved inefficiencies that limit its exploitation. Each user experiences different radio link conditions. This variability heavily reduces the effectiveness of multicast, since the base station must use a conservative modulation to ensure a successful to each user. Nodes that are closer to the base station are able to decode data at a higher rate, while others located near the edge of the cell have to reduce their data rate. Thus, the worst channel user dictates the performance, lowering the overall multicast throughput. This is the main reason why opportunistic offloading can be beneficial also in case of multicast, as we demonstrate in Chapter 4.

[‡]LTE proposes an optimized broadcast/multicast service through *enhanced Multimedia Broadcast/Multimedia Service* (eMBMS) [92].

Cognitive Radio Integration

The spectrum of frequencies available to mobile operators is already overcrowded, while other portions of the spectrum are relatively unused. The limited available bandwidth and the inefficiency in its use call for an opportunistic use of unoccupied frequencies [129]. Cognitive radios could dynamically detect unused spectrum and share it without harmful interference to other users, to shift data on it, enhancing the overall network capacity [130]. Cognitive radio can be employed to offload cellular networks [131], in cohabitation with the HetNet paradigm [132]. Cognitive technologies are thus capable of increasing spectrum efficiency and network capacity significantly.

Proactive Caching

Caching is a popular technique, commonly employed in web-based services in order to reduce traffic volume, the perceived delay, and the load on servers. Caching techniques work by storing popular data in a cache located at the edge of network. Some of these classical concepts can be re-utilized in mobile networks to tackle congestion at RAN. In order to avoid peak traffic load and limit congestion in mobile networks, techniques for predicting users' next requests and pre-fetching the corresponding content are available [133, 134, 135]. Data may be pro-actively cached directly at the user device, at cellular base station, or at IEEE 802.11 APs to improve the offloading process. The prediction is performed using statistical methods or machine learning techniques, and its accuracy is a key factor in performance. Note that some of these techniques may be (or are already) used in the delayed AP-based offloading schemes considered in Section 2.3.1.

2.7 Summary and Relationship with the Thesis

Mobile data offloading is a new and very hot topic, frequently identified as one of the enablers of next-generation mobile networks. In this chapter, we uncovered its key benefits, technological challenges, and current research directions. Offloading systems require a tighter integration within the cellular broadband infrastructures. Future cellular architectures should intelligently support an overall traffic increase of several orders of magnitude. Additional features still need to be developed to handle mobility of users, distributed trust, session continuity, and optimized scheduling policies. A very interesting research area concerns how to merge, in a fully integrated architecture, the different and often stand-alone offloading possibilities presented along this chapter. The contributions in this thesis try to answer to some of the questions raised here. In particular, after presenting a broad classification of current offloading strategies based on their requirements in terms of delivery guarantee, we presented the technical aspects and the state of the art for two main approaches. The former is more mature and proposes a tight integration between the cellular RAN and a complementary access network, allowing for real-time data offloading. The latter, still experimental, exploits

the delay tolerance of some types of data to optimize their delivery, and constitutes the core of this dissertation.

We identified some common functional blocks, proposing a high-level architecture valid for any mobile data offloading system. This architecture will serve as a reference point in the following chapter where we will deal with innovative offloading strategies. We further investigated open research and implementation challenges and alternatives to mitigate the cellular overload problem. User collaboration, especially in the opportunistic approach, is essential for any offloading strategy. In order to make offloading feasible, end-users must accept to share some resources (battery, storage space, etc.), and their wireless interface should be turned on. The central question here is how to motivate user participation. Mobile operators should propose a business concept for rewarding their customers, to make offloading attractive and fully functional at the same time with user participation. We attempt to clarify the relationship between the proposed incentives and the expected offloading benefit in chapter 5. Additional issues lie on the security and privacy plan of users employing mobile-to-mobile transmissions. Users rarely accept anyone stranger to access data stored on their devices. Further challenges include the development of an infrastructure to ensure distributed trust and security to terminals involved in the offloading process.

We all need people who will give us feedback. That's how we improve.

Bill Gates

3

DROiD: Adapting to Individual Mobility Pays Off in Mobile Data Offloading

Delivering large amounts of data over cellular networks remains a challenge because of the possible bottlenecks caused by the access to the radio channel. Moreover, this problem will be further exacerbated by the so called "data tsunami" foreseen for the coming years [1]. Such a complex scenario will load dramatically the existing cellular infrastructure; therefore, it is of paramount importance to find an alternative solution to cope with this problem, in order to save, whenever possible, cellular resources. While opportunistic networks offer additional capacity that can be leveraged to reduce congestion on the cellular network, timely delivery of content is an issue, due to the variability of human mobility and the resulting stochastic nature of forwarding events. When offloaded content must be delivered within given deadlines, we need offloading solutions that both meet these deadlines and reduce as much as possible the traffic carried by the cellular network.

Many seminal works on delayed offloading propose to pre-compute a set of optimal seeder nodes using the structure of the underlying opportunistic network or the contact patterns among users [90, 108]. Although this approach is undoubtedly elegant, it typically does not react well to changing conditions. The quest for optimality suffers also from two problems: (1) a limited scalability, due to the high complexity of finding the optimal solution (NP-hard problems), and (2) a strong dependence on the underlying mobility assumptions that are often unrealistic.

In order to better evaluate data offloading, we need to consider a more realistic scenario

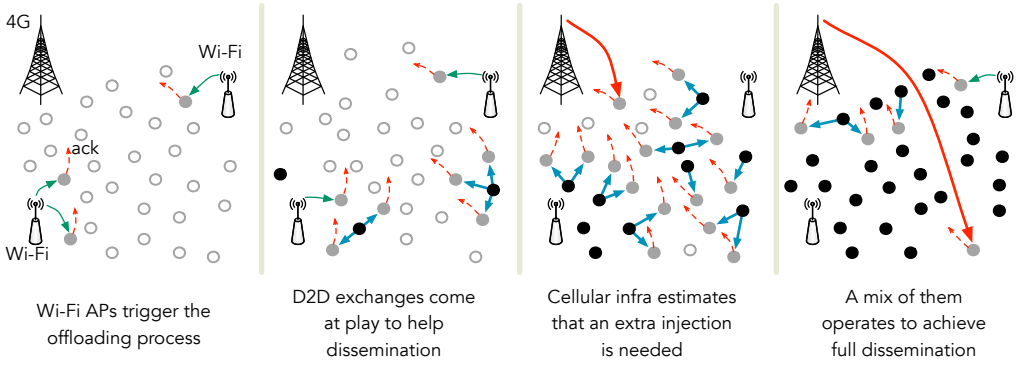


Figure 3.1: DROiD model: The dissemination process is kick started through cellular and/or AP transfers. Content is diffused among mobile devices through subsequent opportunistic contacts. Upon reception, users acknowledge the offloading agent using the feedback cellular channel. As acknowledgment messages are in general much smaller than data messages, we obtain at the end significant reduction of cellular traffic. The central system may decide at any time to re-inject copies through the cellular channel to boost the propagation. 100% delivery ratio is reached through fall-back re-injections.

with a large number of nodes and accurate mobility. We remove the optimality requirement, looking for a practical way of dealing with the problem. The availability of a cellular channel can be used to get feedback from users. The infrastructure can monitor and react to the evolution of data dissemination. In this chapter, we propose a novel offloading scheme to distribute popular data to a multitude of mobile users among different technologies (cellular, D2D, and AP-based).

3.1 Background

Our system, called DROiD (Derivative Re-injection to Offload Data), helps mobile operators relieve their access network by exploiting both the presence of WiFi access points and alternative transmission opportunities between users. The key feature of DROiD is to adapt to the heterogeneity of individual user mobility patterns. As a matter of fact, after running epidemic diffusions on mobility traces, we observed that the progression of the diffusion follows a characteristic stepwise evolution. Periods of fast progression alternate with periods where the dissemination stalls. We found the heterogeneity in user mobility to be the main responsible for this phenomenon (we detail this in Section 3.2).

We illustrate DROiD's operation in Fig. 3.1. The diffusion process is monitored through a persistent feedback channel that connects mobile users with an offloading coordinator. This monitoring control loop allows anticipate the correct re-injection decision. The system detects the formation of plateaux in the evolution of content diffusion. If needed, it also triggers *adaptive re-injection of additional copies in the system to finely control the pace at which the content is disseminated*. Because it makes no assumption on user mobility, DROiD leads

to better performance than previous strategies that are bounded to empirical functions that remains fixed for the entire dissemination process [117].

We evaluate the performance of DROiD through extensive simulations of a location-based service in a vehicular context using realistic mobility traces. In this service, content with traffic information or some infotainment announcement must be distributed to a multitude of users within a given maximum reception delay (in order to guarantee a minimal QoS on a per-content basis). We employ two realistic large-scale vehicular traces derived from multiple fine-grained traffic measurements in the city of Bologna and Koln (respectively 10,000 and 15,000 vehicles). We compare DROiD's performance under tight delays with other opportunistic and AP-based offloading solutions proposed in the literature, and with an oracle, taken as benchmarks. Our results highlight that DROiD substantially outperforms other offloading strategies that use an objective function for any considered delay tolerance value, reducing by more than half the infrastructure load.

As a summary, the contributions of this chapter are threefold:

- We turn the attention to the heterogeneity of contact patterns in opportunistic networks. We reveal that this heterogeneity is at the origin of the dynamic creation and dissolution of clusters. We harness this property to explain why epidemic diffusion presents a stepwise behavior.
- We propose DROiD, a low-complexity offloading framework that, thanks to a derivative-based re-injection strategy, better adapts to the contact patterns between nodes to offer enhanced offloading efficiency.
- We compare DROiD with other objective function-based strategies and show that it outperforms them even under tight delivery delay constraints. We evaluate our approach against common AP-based offloading strategies, considering also energy-saving schemes.

3.1.1 Principles behind DROiD

The system works by adopting a *Pub-Sub* paradigm, with users sending `subscribe` (`unsubscribe`) messages upon entering (leaving) the simulation area, in order to acknowledge their interest in a certain type of content. A central *offloading manager* represents the control plane of DROiD, and is dedicated to supervise data dissemination. The offloading manager decides when and to whom data is delivered on the cellular channel. This entity is conceptually represented by a single block, however, its physical mapping in the core/access network may be distributed (e.g., at E-UTRAN and EPC level in 3GPP LTE). We also considered the opportunity of having a citywide AP deployment that assists data distribution. Fixed WiFi APs are a conventional solution to offer high-speed Internet, reducing at the same time the network load, that can be deployed by operators or by municipalities [7].

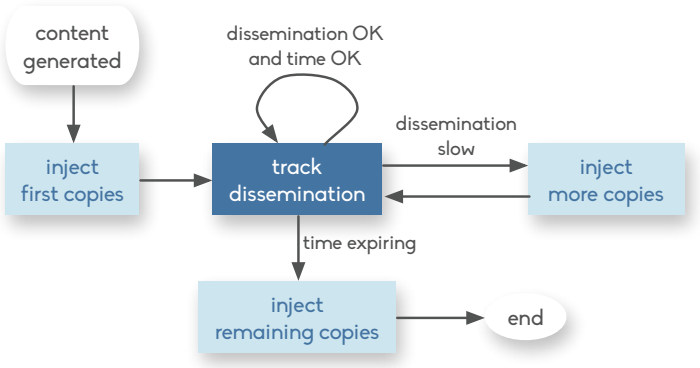
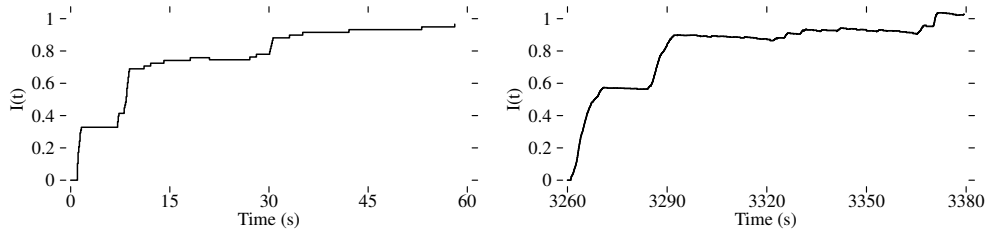


Figure 3.2: Flowchart of the high-level operation of the offloading manager in DROiD.

Fig. 3.2 provides a simplified scheme to illustrate the offloading manager operations. Depending on the considered scenario, a subset of subscribed users initially receives the content through the cellular or the AP infrastructure, and propagates it opportunistically employing D2D transmissions. Whenever a vehicle receives data from a neighboring user or AP, it acknowledges its reception to the manager using the cellular network, forming a feedback loop in the system. This simple mechanism allows DROiD to monitor in real time the evolution of the content dissemination process, and possibly to account for data usage. The manager continually estimates the infection ratio and may decide to re-inject additional copies of the content in order to boost the diffusion. The proposed system trades-off downlink data traffic for uplink control traffic, and since acknowledgments sent by mobile nodes on the infrastructure channel are relatively lightweight (compared to the size of the disseminated content), the system is expected to guarantee considerable reduction in the infrastructure load.

D2D and AP transmissions are truly opportunistic because they depend on the particular mobility of nodes. Transmission opportunities appear and disappear dynamically and abruptly. For these reasons, only probabilistic guarantees of successful content delivery and reception times can be given. To solve this issue, when the maximum delivery delay D approaches (i.e., the validity of content), and the time left is equal to the time required to send the message through the cellular infrastructure, denoted as P, the offloading manager enters a *panic zone* and pushes the content to all uninfected nodes through the infrastructure, guaranteeing full dissemination. Note that the feedback loop guarantees also a fall-back method to overcome various issues that may appear in the network, such as node failures or mobile users behaving selfishly – the occurrence of these events could heavily affect the opportunistic diffusion [136].



(a) Rollernet, 6 initial copies.

(b) Bologna, 10 initial copies

Figure 3.3: Epidemic diffusion of the content. The diffusion behavior alternates steep zones and flat zones that are the result of changing encounter probability among mobile nodes.

3.2 Stepwise epidemic diffusion and why adaptive offloading is needed

At the heart of DROiD is the idea of allowing the diffusion process to adapt to the idiosyncrasies of individual mobility patterns. To see what happens, let us take two examples using very different datasets: the small-scale Rollernet dataset, composed of only 62 nodes [137], and the Bologna dataset, composed of more than $10,000$ nodes. We plot in Fig. 3.3 the evolution of the content diffusion in the network. We start the diffusion by injecting a small number of initial copies (6, 5 and 10 respectively) to random nodes at t_0, and let the epidemic diffusion of the message progress with subsequent direct contacts. A node that has received the message is said *infected*, while a node that has not yet received the content is *sane*. The instantaneous infection ratio $I(t) \in [0, 1]$ follows a stepwise pattern, alternating plateaux (flat areas) to periods of heavy infection (steep areas) before reaching complete diffusion. We may find a similar dissemination evolution pattern for very different datasets such as the ones considered. This is a typical example of the way any given diffusion process progresses due to the randomness of contact patterns in opportunistic networks. In particular, the plateaux correspond to periods during which the dissemination does not make any progress, because no sane nodes come into range of the already infected nodes.

Let us now dig into the relationship between mobility patterns and progress of the epidemic diffusion. The first obvious point is that this phenomenon is intrinsically related to the heterogeneity of contact patterns, i.e., the fact that two different nodes do not meet on average the same number of other nodes. If the contact process of nodes is Poisson homogeneous and stationary, each pair of nodes meets with intensity λ. Assuming that each contact means an opportunity to transfer the content, neglecting the contact duration, the resulting epidemic diffusion follows the logistics equation [104] - the curve does not exhibit any plateau, which is in contrast to our observation. Since for homogeneous contact process each pair of nodes meet with the same probability, the longer the time that a copy has to propagate and the greatest benefit that copy brings. This would lead us to wrongly believe that the best offloading strategy is to inject the right amount of copies at t_0, forget about the infection

evolution, and let the opportunistic dissemination do the work for us.

To capture the heterogeneity of patterns, we adopt a Marked Poisson Process model of node contacts [138, 139]. In this model, the meeting times of any two nodes (i, j) follow a Poisson Process with rate $\lambda_{ij} = \lambda p_{ij}$. The inter-contact times T_{ij} are thus independent exponentials with parameter λ_{ij}, and the matrix $C = (p_{ij})$ captures the patterns of interactions between nodes. In the homogeneous case, C is the identity matrix, i.e., all nodes can see each other with the same probability. At any given time instant of the dissemination process, a set S of nodes is infected. We are interested in the random plateau duration T_S^p during which the dissemination does not progress. This corresponds to the random time during which this set of nodes does not meet any other nodes. Looking at the set of links between nodes in S and its complement, one can see that $T_S^p = \inf_{i \in S, j \notin S} T_{ij}$. By Poisson calculus, and noting the cut value $\partial S = \sum_{i \in S, j \notin S} p_{ij}$, T_S^p is an exponential random variable with parameter $\lambda \partial S$ [140]. The expected plateauing duration, once set S has been reached, is thus $1/\lambda \partial S$.

This simple argument shows that T_S^p is directly related to the structural properties of the contact matrix C, providing a natural connection between the community structure of the contact graph and the progression (or lack of progression) of the opportunistic dissemination process. In general, identifying structure in graphs consists in finding groups of nodes (called clusters) based on some similarity measure defined for the data elements [141]. Nodes in a community have high conductance (they are well knit to one another), and the ratio of the weight of inter-cluster edges to the total weight of all edges is low [142]. Applying these ideas to C (which represents the probability of two nodes to meet) means that a community S of users will spread the message quickly within the cluster (high conductance), but will reach a plateau once the nodes in the group all have the message, because the weight of inter-cluster edges and thus its cut value ∂S is low.

In practice, we observe strong dynamic clustering for the considered datasets. This observation provides the motivation of our further investigation of adaptive offloading strategies that are able to chase the individual mobility of nodes, re-injecting copies when the diffusion evolution runs into a plateau.

3.3 Shrinking the cellular load

Following the observations discussed in Section 3.2, we design an offloading solution capable of adapting to the varying content diffusion evolution.

The core of our approach resides on the intelligence of the offloading coordination. This latter is in charge of deciding *when* to re-inject additional copies depending on the evolution of the opportunistic dissemination process. Every re-injection decision is expected to bring benefit to the system; nevertheless, the gain depends on the re-injection time and the targeted node (to which copies are sent through the infrastructure). In this work, we do not focus on the choice of the nodes to be targeted, supported by the fact that previous work proved that a random node selection works better than many other more complicate selec-

tion schemes [117]. In fact, there is still a difficult trade-off to consider. On the one hand, if too many copies are injected in the beginning (in general, earlier injections have more time to diffuse), the system may be overestimated (as we do not know in advance how nodes will encounter). On the other hand, if the system injects too few copies in the beginning and waits for the panic zone to compensate for lags, many opportunistic encounters might be wasted because of the lack of enough copies in the network.

Re-injection is beneficial when the subsequent opportunistic transmissions save additional infrastructure pushes. Of course, the benefit can be null if the offloading coordination agent selects a node that would have received the message later from another node. Finding the good trade-off is difficult, as the offloading agent is essentially blind and the only information available is the list of currently subscribed users and the list of those who already received the content (inferred from acknowledgments).

The logic of our system is inspired by Push-and-Track [117]. However, DROiD achieves higher offloading efficiency by considering for the re-injection decision not only the actual dissemination level, but also the past infection trend. By noting that content diffusion has a stepwise pattern, DROiD anticipates and avoids the insurgence of long-lasting plateaux through its adaptive strategy. This is not the case for the static strategies proposed in Push-and-Track, in which the central coordinator consider only the distance between the instantaneous infection ratio and a fixed *a priori* target objective function [117]. Re-injection decisions, in this case, do not take into account the general evolution of the infection, but only the instantaneous value. Static strategies tend to react too late when the infection ratio is above the objective function but still not evolving, or to overreact when the infection evolves well but its instantaneous value still lies under the objective function. Late or too brutal re-injections result in a waste of messages pushed through the infrastructure. Another limitation of Push-and-Track is that it does not propose a single solution, but instead a multitude of objective functions; the problem is that different objective functions behave differently depending on the content lifetime and network status.

3.3.1 Derivative re-injection

We propose a low-complexity re-injection strategy, based only on the knowledge of the infection ratio evolution in the recent past. DROiD keeps in memory a short snippet of past infection ratio values. Each content has an associated tracker that stores the evolution of the infection ratio for a temporal sliding window of size W (i.e., at time t the values that will be considered are the ones between $[t - W, t]$). W is a design parameter that must be a multiple of the time step used for the evaluation of the infection ratio, τ, and smaller than D (recall, the validity of the content). The size of the sliding window trades off how far in time DROiD looks back, and dictates the reactivity to sudden changes in the infection ratio. It follows that the total memory footprint required by our system is fixed at W/τ for each content. In our experiments, we found that reasonable values of W fall in the range $[2\tau, 10\tau]$. In addition,

the coordinator should maintain the lists of subscribed and infected users, which are both linear in the number N of subscribed users.

As illustrated in Fig. 3.4, at evaluation time step t, the offloading coordinator performs a forward difference quotient on the instantaneous infection ratio $I(t)$ that approximates to a discrete derivative:

$$\Delta_I(t) = \begin{cases} \frac{I(t)-I(t-W)}{W}, & t - t_0 \geq W, \\ \frac{I(t)}{t-t_0}, & t - t_0 < W. \end{cases} \tag{3.1}$$

Note that $I(t)$ is not monotonically increasing, since nodes may exit the simulation area at any time. $\Delta_I(\cdot)$ approximates the slope of the infection ratio and is one of the parameters that influence the re-injection decision. DROiD re-injects additional copies of the content whenever the discrete derivative $\Delta_I(\cdot)$ is below a Δ_{lim} threshold computed on line as the ratio between the fraction of sane nodes and the time remaining before the panic zone. This because a steeper slope is needed when time gets closer to panic zone or the infection ratio lags (different from when we are at the beginning of the infection process). Formally speaking, we have:

$$\Delta_{\text{lim}}(t) = \frac{1 - I(t)}{(D - P) - (t - t_0)}. \tag{3.2}$$

As a final step, the injection rate $r_{\text{inj}}(t)$ is computed as a piecewise function, depending on the ratio of the current $\Delta_I(t)$ value and the Δ_{lim} threshold:

$$r_{\text{inj}}(t) = \begin{cases} c, & \Delta_I(t) \leq 0, \\ c\left[1 - \frac{\Delta_I(t)}{\Delta_{\text{lim}}(t)}\right], & 0 < \Delta_I(t) \leq \Delta_{\text{lim}}(t), \\ 0, & \Delta_I(t) > \Delta_{\text{lim}}(t), \end{cases} \tag{3.3}$$

where $c \in [0, 1]$ is a clipping value used to limit the overall amount of re-injected copies in the case of negative values of Δ_I.* Finally, $r_{\text{inj}}(t)$, which represents the percentage of uninfected nodes that need to be targeted, is multiplied to the number of uninfected nodes to find the number $\mathcal{R}(t)$ of copies to re-inject at t:

$$\mathcal{R}(t) = \lceil (1 - I(t)) \times |N(t)| \times r_{\text{inj}}(t) \rceil, \tag{3.4}$$

where $|N(t)|$ is the instantaneous number of nodes subscribed to the content update.

* Clipping value c represents the maximum fraction of content that can be pushed through the infrastructure at evaluation time. Negative values of $\Delta_I(\cdot)$ may happen in the case of infected nodes leaving the system. We address this issue in Section 3.5.

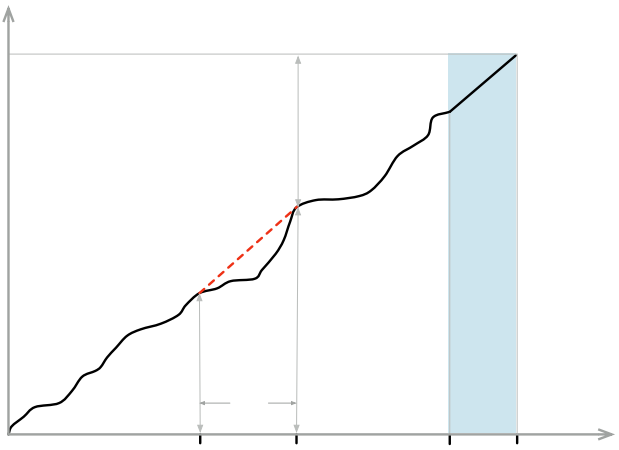


Figure 3.4: Discrete time slope detection performed by DROiD. For clarity we consider the content creation time $t_0 = 0$.

3.4 Datasets, scenarios, and simulation setup

We evaluate DROiD considering the problem of distributing popular content to a multitude of mobile nodes. We assume that nodes are equipped with two wireless interfaces (e.g., most smartphones or infotainment systems), so that they are able to communicate through two interfaces simultaneously. Possible combinations involve 3G and 4G to communicate with the cellular infrastructure and Bluetooth or WiFi ad hoc to communicate with neighboring devices.

3.4.1 Experimental datasets

Mobility traces. We employ two large-scale vehicular mobility traces representing the city of Bologna (Italy) and Koln (Germany). The Bologna dataset consists of $10,333$ nodes, covering a total of $20.6\,\text{km}^2$ and 191 km of roads, and was initially exploited to evaluate cooperative road traffic management strategies within the FP7 iTetris project [143]. The dataset is drawn from real traffic measurements acquired by 636 induction loops deployed citywide. The dataset captures real city traffic conditions, with speed of vehicles varying from 0 to around 50 km/h, depending on road congestion. We also consider the synthetic car traffic dataset of the metropolitan area of Koln, made public by the TAPASCologne project [144]. The dataset is larger than Bologna, simulating more than $15,000$ moving vehicles on an area of around $400\,\text{km}^2$.

The construction's methodology for both datasets is similar, based on statistical properties extracted from real world experiments and inserted into the SUMO mobility simulator to

Figure 3.5: Bologna map with fixed APs positions. Note the concentration of APs in the city center.

extract a microscopic-mobility model [145]. The simulated traffic mimics the everyday road activity in the two metropolitan area. From each mobility trace, we derive a contact trace that features contacts between nodes when the distance between them is below a given threshold (we consider in our analyses a range of 100 meters, in line with IEEE 802.11p specifications). The resulting trace for Bologna has a duration of about one hour; in average, 3, 500 nodes are present at the same time (because some nodes leave while others join during the observation period). For the Koln dataset, the trace lasts around two hours, from 6:00 am to 8:00 am; in average, more than 10, 000 nodes are present at the same time. The advantage of using these two large-scale traces is that, differently from other available datasets, we have a clear high turnover rate, due to vehicles entering and exiting the interest area, and no apparent social links. The distribution of contact durations is exponential for both traces. Most contacts are very short, confirming the highly dynamic nature of the two traces. Only few contacts last for more than a few minutes [117, 146].

For the evaluation section, we do not consider other existing datasets widely used by the research community, such as [137, 147, 148, 149, 150, 151]. Although extremely handy for the evaluation of opportunistic network strategies (as we did in Section 3.2 to illustrate the epidemic diffusion evolution), their limited sizes (below 100 nodes) and the particular settings where they have been collected (conferences, campus, etc.) make them unrealistic to properly evaluate offloading strategies.

AP spatial distribution. For the Bologna dataset, we extracted the location of existing WiFi Hotspot public deployment from [152]. The position of APs is shown in Fig. 3.5 along with a map of the city. We merged the location of APs with the vehicular mobility trace to extract a completely new dataset that includes vehicles mobility and AP positions. Employing the

same strategy as before, from the merged trace we derived the connectivity traces between vehicles and fixed APs. The outcome is a completely new time-variant graph with unidirectional edges connecting vehicles with APs and other vehicles.

To characterize this new dataset, we study the pairwise interactions among mobile nodes (vehicles) and fixed APs. Fig. 3.6 presents the distributions of contact and inter-contact times between vehicles and APs. Contact time represents the dwell time of a user inside the coverage of an AP. Inter-contact time accounts for the time duration between two subsequent contacts with any other AP. We observe that contact times are larger than inter-contact times. Unlike the vehicular-only trace, which is exponentially distributed [117], here the contact and inter-contact times are found to follow a log-normal distribution using MLE (Maximum Likelihood Estimation). This is not surprising, since log-normal distribution offers a more versatile model to catch the variability of contact/inter-contact times [65]. This hinges on the heterogeneity of the contact distributions among different AP-vehicle pairs.

Two anomalies make the study of the dataset interesting. First, a relevant number (around 20%) of inter-contact times are 0 s (not plotted in the figure), meaning that when a vehicle exits from the coverage area of an AP, it is already under the range of another. We may explain this aspect by noting from the map that in the central zone several APs are located very close together. Nevertheless, the sole contact distribution does not tell us the whole story, as very few APs are deployed in southern and western town districts, and many vehicles passing there enter and exit the system without falling into the coverage zone of any APs. We note that the observation that AP meetings occur in bursts, initially proposed in [75], proves to be true also for our dataset. We infer a strong correlation between the geographic position and the expected duration of contact and inter-contact times with APs. In addition, we note that around 80% of the contacts with APs last for more than 10 seconds. While this may be an acceptable duration for data transfers, short-lived contacts lasting less than that value could suffer from the duration of authentication and address granting procedures with an AP.

Fig 3.7 depicts the evolution of the number of vehicles in contact with at least an APs during a given time window (equivalent to the delay-tolerance of content reception in this case). The figure indicates the amount of vehicles that enter in the transmission range of least one AP during the considered distribution period. It results pretty intuitive that augmenting the delay tolerance increases the chance that a vehicle enters in the range of at least one AP. Still, the number of nodes benefiting from this transfer opportunity is limited, if measured against their total number in the system, lying always between 5 and 25% of present users. We extract this value even before considering if contact durations are large enough to transfer data, or if congestion permits communications. As we will show later, offloading strategies based exclusively on APs will have a very limited impact on the amount of saved data in this scenario.

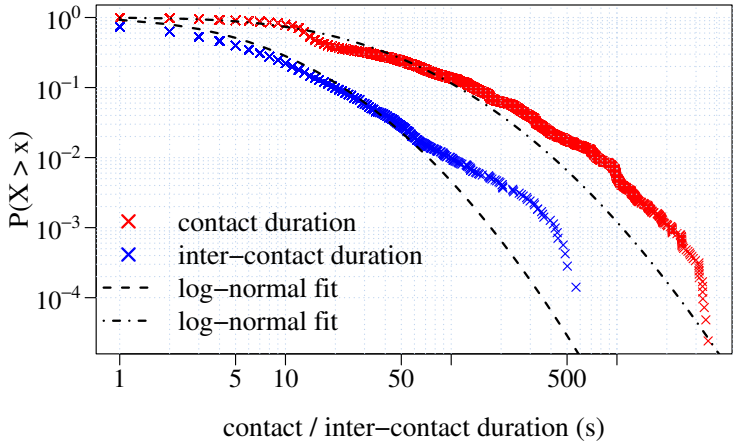


Figure 3.6: CCDF of contact and inter-contact times with fixed APs for the Bologna dataset. The distribution fits best with log-normal with $\mu = 3.098$ and $\sigma = 1.255$ for inter-contacts, and with $\mu = 1.644$ and $\sigma = 1.136$ for contacts.

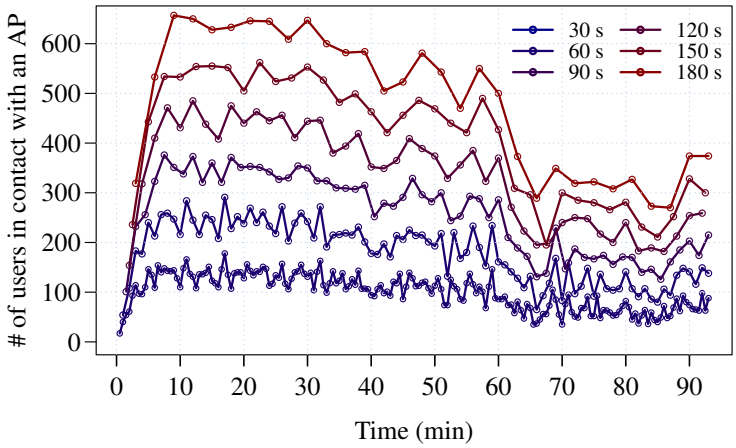


Figure 3.7: Time evolution of the total number of contact between vehicles and APs for different message lifetime.

3.4.2 Scenario

Without loss of generality, we consider a location-based traffic information service, where a centralized server issues a new content update every t_P seconds. DROiD must guarantee the delivery of each of the updates to all nodes interested in the content within a maximum delay D. Contents are issued periodically, with the previous one expiring when a new one is created (so $t_P = D$, and a single content is active in the system at a time). Possible contents of interest include popular geo-relevant data, such as localized traffic and roadwork alerts, generalized public utility information or geographic advertising; nevertheless, the proposed system also supports the efficient distribution of software updates for connected vehicles and mobile devices. The choice of which content to offload is made in advance, according to its delay-tolerant characteristics. Depending on the employed distribution strategy, contents may be delivered directly through the pervasive cellular network, through nearby fixed APs, or retrieved from a neighboring node in opportunistic fashion. Despite this work considers all users as interested in the content, the combined use of the *Pub-Sub* paradigm and `ack` messages makes the system easily extensible in the case of multiple contents and non uniform nodes' interest. Users may enter and leave the target area during the content lifetime, impacting the results, as we will see later.

3.4.3 Simulation setup

No network simulators among those publicly available today perform well in scenarios with several thousand nodes at the same time [153, 154]. Therefore, we built a streamlined event based simulator heavily inspired by the ONE simulator [153]. In our implementation, we consider a simple contact-based ad hoc MAC model, where a node may transmit only to a single neighbor at a time. Transmission times are deterministic since we do not take into account complex phenomena that occur in the wireless channel such as fading and shadowing. (we are not really interested here on the exact physical evolution of communications taking place during the offloading process). Communications consist of two different classes of messages (content and control). All transfers, including `ack` messages, may fail due to nodes moving out of each other's transmission range or exiting the simulation area. In addition, it is possible that the same message be concurrently received through the two interfaces. In that case, we consider the one that is processed first. The ad hoc routing protocol employed by nodes to disseminate the content is the epidemic forwarding.

Parameters in simulation are set to mimic the functioning of communication technologies currently available to consumers. In each simulation run, the downlink bit-rate for the infrastructure network is set to 100 KB/s, while uplink is fixed at 10 KB/s. These values are in line with the average bit-rate experienced by users of a typical HSPA network. The bit-rate for the ad hoc link is set to 1 MB/s, also in line with the advertised bit-rate of the IEEE 802.11p standard. The size of each content update is set at 100 KB. The size of the acknowledgement messages is 256 bytes, as it carries very little information (content and node identifiers). For

the other parameters, we use $\tau = 1$ seconds, $c = 0.05$, and $W = 5\tau$. The panic time duration P is fixed at 1 second.

3.4.4 A note on security

Security and privacy of data distribution is out of the technical scope of this work; nevertheless we are aware that this remains a relevant issue for disseminating content directly among end-users. For instance, in case such as software updates or road traffic information, extra security mechanisms must be in place to prevent malicious users to compromise trust or unauthorized settings to be installed in vehicular nodes. There is a large body of research work related to the security and privacy in VANET and DTNs, relying on public keys infrastructures (PKI) and digital certificates [155, 156]. The presence of a central coordinator and the existence of a persistent channel connecting each user to the coordinator, may simplify the adoption of these security mechanisms.

3.5 Results

3.5.1 Existing strategies

We investigate how our system performs under tight delivery constraints, when the maximum reception delay D lies in the range $[30, 180]$ seconds. This contrasts with what is done in the literature that consider long time scales for content reception (up to some hours). Instead, we are interested in very short maximum reception delays, in the order of minutes, as otherwise users would not realistically accept to trade-off reception delays for cellular capacity. State-of-the-art solutions, benefiting from more relaxed reception constraints, can take advantage of a sort of stochastic regularity in contact patterns of users [108, 157, 90]. Centralized optimization frameworks based on Monte Carlo sampling [108] or *temporal reachability graphs* [158] require the complete contact graph among users, and are known to have high computational complexity. Therefore, they are unable to evaluate the offloading strategy on large-scale datasets in real-time, as our framework does. Indeed, if the mobility patterns of subscribers change, the selected strategy might not be optimal anymore. In addition, none of the proposed strategies deal with nodes entering or leaving the system. These considerations make it difficult to compare existing approaches with our low-complexity approach, which targets the microscopic mobility of users and the unpredictable contact dynamics on small time intervals.

3.5.2 Reference strategies and evaluation

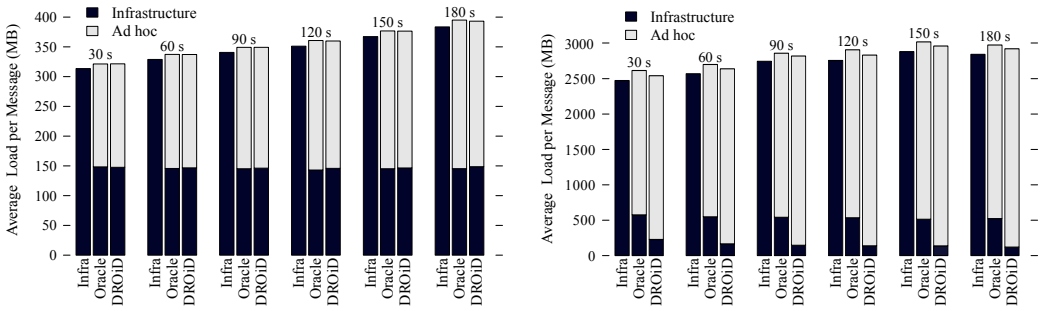
We employ two reference strategies for evaluation purposes: "infrastructure only" (Infra), and "con-nectivity-aware oracle" (Oracle). Infra represents the basic strategy, where there is no offloading at all, and the cellular infrastructure is the only means of distributing content. In the Oracle strategy, the coordinator has a real-time picture of the ad hoc connectivity of

the entire network (this is an unrealistic but useful assumption). The coordinator pushes the content to a random node within each existing connected component. The underlying idea is to push only one copy per connected component in order to get close to the minimal number of infrastructure copies. The coordinator has a perfect instantaneous view of the system; however, it does not account for transmission times and future movements of nodes. This strategy shows its limits in the Koln dataset, where the oracle tends to over-estimate the number of initial copies to push in the system, achieving less than optimal performance.

In addition to these baseline cases, we compare DROiD with Push-and-Track, which represents nowadays the offloading alternative that offers 100%-delivery ratio guarantees with tight delivery times. Since it offers primarily a methodology rather than a specific offloading strategy, it is difficult to state a priori which target objective function gives the best results [117]. To be as fair as possible, we compare DROiD with the objective function that gives, for each scenario, the best results, namely the *linear* and the *slow start* objective functions for the Bologna trace, and the *linear* and the *square root* strategies for the Koln trace.

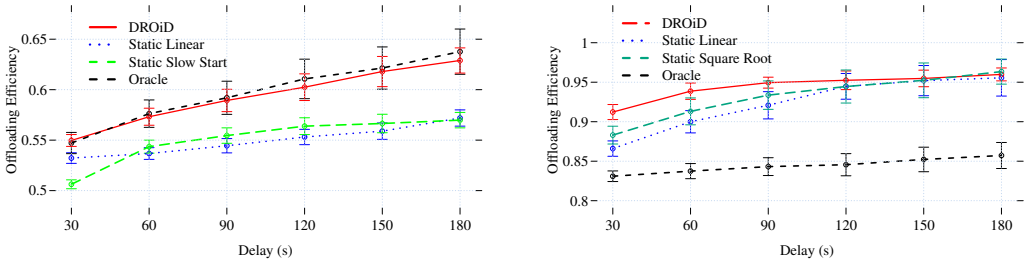
We also compare DROiD with other commonly employed offloading methods that takes advantage of the presence of fixed APs. To evaluate these strategies, we employ the new Bologna contact trace that includes the locations of APs and the mobility of vehicles (derived and analyzed in Section 3.4.1). In AP-based offloading, the coordinator pre-fetch each AP with the content to be distributed at time t_0. AP-based strategies do not make use of intermediate re-injections, whilst maintaining panic zone to guarantee a minimal QoS-level. In the first AP-based strategy, labeled *AP-only*, content distribution is only achieved upon direct contacts between vehicles and the designated AP. The other evaluated strategy, *AP + opportunistic*, builds on the latter, but proposes also the possibility of direct exchanges among nodes through ad hoc transmissions. We also propose and asses the performance of an extended version of DROiD that makes use of the presence of fixed APs. Finally, we evaluate the implementation of energy-saving strategies aimed at preserving the battery of mobile nodes. Each mobile node is allowed to opportunistically forward a given message only a limited number of times.

Methodology. All the results presented are averages over 10 simulation runs. Random seed is re-initialized at the beginning of each simulation run, and the starting time is shifted by $D/10$. We focus primarily on the aggregate load that flows through the cellular, wireless AP, and the ad hoc links. Load measurements take also into account `ack` messages as well as failed and aborted terminal-to-terminal transfers. `Ack`, `subscribe`, and `unsubscribe` messages constitute the infrastructure overhead. The offloading efficiency metric depends on the amount of traffic flowing on the cellular link when we use the offloading process, denoted with L, and the cellular traffic in the absence of any offloading strategy (i.e., Infra strategy), denoted with L_{ref}. Formally, the offloading efficiency is computed as $1 - L/L_{ref}$. We target the distribution of updates to *any* node that are part of the network for any period of time within $[t_0, t_0 + D]$. We denote with $N(t)$ the set of subscribed users at time $t \in [t_0, t_0 + D]$. Consequently, We define the set of all target nodes for the content as $K = \bigcup_t N(t), \forall t \in$



(a) Bologna trace **(b)** Koln trace

Figure 3.8: Infrastructure vs. ad hoc load per message sent using the Infra, the Oracle, and the DROiD strategies. Different maximum reception delays for messages are considered.



(a) Bologna trace **(b)** Koln trace

Figure 3.9: Offloading efficiency. Different maximum reception delays for messages are considered. 95% confidence intervals are plotted.

$[t_0, t_0 + D]$.

3.5.3 Comparison with static strategies

Cellular load. DROiD performs very well in terms of reduced infrastructure load, by delivering the majority of traffic through device-to-device communications even in the case of tight delays. Recall that all nodes entering the target area are expected to receive the content, regardless of their dwell time in the system. Therefore, the load in the Infra strategy increases with the message lifetime. Simulation results, plotted in Fig. 3.8 displays the average amount of traffic per message that flows through the infrastructure and ad hoc interfaces. In this picture, we compare DROiD to reference strategies only, to illustrate how DROiD consistently offload a relevant amount of data. We observe that DROiD approaches Oracle in the Bologna scenario, outperforming it in the Koln trace.

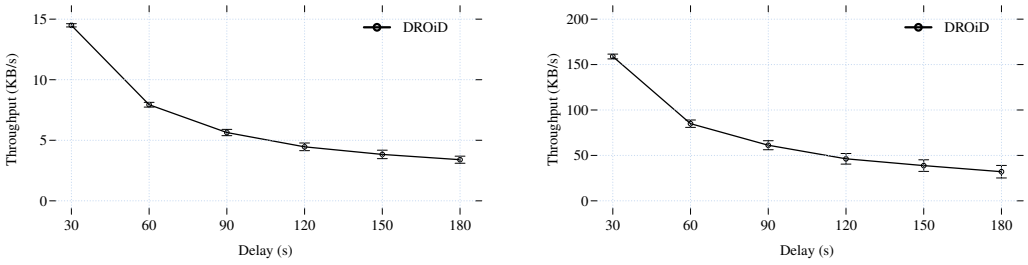
Table 3.1: Cellular overhead (%) for different strategies and reception delays.

Bologna	30s	60s	90s	120s	150s	180s
Oracle	0.31	0.36	0.39	0.43	0.45	0.49
DROiD	0.30	0.34	0.37	0.39	0.42	0.44
Linear	0.29	0.29	0.30	0.31	0.32	0.34
Slow Start	0.24	0.29	0.30	0.32	0.33	0.34
Koln						
Oracle	1.14	1.17	1.24	1.26	1.32	1.40
DROiD	2.47	3.62	4.5	4.87	5.29	5.71
Linear	1.89	2.23	3.08	3.97	5.01	6.32
Square root	1.70	2.57	3.30	3.87	4.45	5.10

Sudden variations in the infection ratio, due to nodes that dynamically enter and leave the simulation area, are well handled by the feedback mechanism of DROiD. While the load in the Infra strategy increases linearly in both scenarios (as a longer content lifetime implies a major number of nodes entering the system), the cellular load for Oracle and DROiD remain always nearly the same, translating in increased efficiency. If compared to Infra, the aggregate cellular and ad hoc load shows a limited overhead. This ad hoc overuse is not particularly worrisome, since direct transmissions have no monetary costs associated. The ad hoc overhead is dominated by failed and aborted transfers due to nodes moving out of each other's transmission range, or messages concurrently received on both interfaces. On the other hand, Table 3.1 compares the infrastructure overhead due to control messages. Note that our mechanism trade-off downlink traffic for uplink control traffic. Thanks to the small size of the `ack` messages, the feedback mechanism is never responsible for more than 6.4% of total cellular traffic. As expected, there is a linear relationship between the overhead and the number of contents received through ad hoc links (and the resulting `ack` messages). Note also that the ratio of control traffic increase for high efficiency. Figure 3.8 shows also a well-known phenomenon: an increase in the reception delay corresponds to an improved offload ratio. The ratio increase with larger delay-tolerance because the opportunistic dissemination has more time to propagate the content to the entire network and thus less copies need to be re-injected during the panic zone.

Offloading efficiency. Compared to static strategies, DROiD always leads to better results, as shown in Fig. 3.9. In the Bologna scenario, DROiD saves between 55% and 63% of traffic for different message delays. DROiD in this case always outperforms the static strategies in terms of offloading efficiency. Note that the results would have been even better if we had picked another objective function. In the Koln dataset, DROiD is upper bounded after only 90s by the 95% of efficiency.

Although DROiD and Oracle show more or less the same trend on the Bologna dataset,



(a) Bologna trace **(b)** Koln trace

Figure 3.10: Aggregate required throughput on the cellular channel for acks. 95% confidence intervals are plotted.

this result is achieved through two completely different strategies. On the one hand, Oracle, exploits additional knowledge on the connectivity status in the network, pushing the content to specific high potential nodes. On the other hand, DROiD has a much less complete, and slightly out of sync, view of the system, and employs its advanced derivative-based reinjection algorithm to guess when additional copies of the content are required. In the Koln scenario this additional knowledge does not help Oracle to obtain optimal offloading performance. Note that in this scenario, any feedback-based strategy does better than Oracle, which tends to overestimate the number of copies to inject in the beginning, being unaware of future movements of nodes. On the other hand, feedback-based strategies benefit from their infection-level awareness to decide when it is better to intervene with re-injections.

As a final remark, Oracle presents always larger confidence interval than DROiD (and static strategies in general). This is related to the mobility and turnover of nodes: the resulting connectivity changes in time influencing the Oracle prediction performance. A mobility and connectivity agnostic framework such as ours is less sensitive to this issue.

3.5.4 The impact of acks

DROiD needs a centralized offloading manager that takes the transmission decision and supervises data dissemination. Whenever a vehicle receives data from a nearby user, an `ack` message is sent on the cellular channel. This indicates that the system may overload the uplink cellular channel with too many simultaneous control messages. The occurrence of this situation is clearly unacceptable, as it is orthogonal to our goals.

In Table 3.1, we show the aggregate overhead due to acks. In the first place, thanks to the small size of `ack` messages, the feedback mechanism is never responsible for more than 6.4% of total cellular traffic (Koln scenario, 180s). As expected, there is a linear relationship between the overhead and the number of contents received through ad hoc links (and the resulting `ack` messages). As the number of acknowledgments increases, the cellular downlink

data decreases bringing up the overhead. This mechanism is at the base of the idea of trading off large downlink data content for uplink control traffic.

Fig. 3.10 presents the aggregate required uplink throughput to route acks. The total uplink traffic depends on the extension of the scenario and the number of involved nodes, explaining why the Koln scenario requires around 10 times more uplink bandwidth than Bologna. Both traces show the same decreasing trend. For shorter deadlines, the injection algorithm tends stimulate epidemic diffusion by injecting more content. The outcome is an increase in the number of simultaneous acknowledgments and in the required uplink bandwidth. However, using a random node selection helps distribute geographically the content, limiting the impact of acks on the same cellular base station.

3.5.5 Opportunistic or AP-based offloading?

How DROiD behaves when compared to most traditional AP-based offloading strategies? To answer this question, we run additional simulations to benchmark DROiD against other more conventional strategies based on the direct offloading from fixed hot-spots. AP-based offloading takes advantage of the presence of fixed infrastructure that can serve to offload the cellular network. Nevertheless relying on fixed deployment typically lacks of the flexibility of pervasive cellular networks, since transmission range and spatial density are limited for a physical reason.

We exploit the Bologna dataset, considering also the AP deployment, and maintaining the same simulation parameters as before. In Fig. 3.11 we may appreciate the efficiency of three alternative approaches to data offloading making use of fixed APs, simulated on the Bologna dataset and compared with DROiD. First of all, we note that *AP only* strategy gives extremely poor results. Even with larger tolerance to delays, this strategy is never capable of saving more than 20% of traffic. Thus, this strategy turns out to be unable to substantially relieve a large fraction of the cellular load. As already hinted in the dataset analysis phase, the main problem in this case is that APs are not ubiquitously available in each town district. Vehicles traveling in areas without WiFi coverage cannot download data from nearby APs, and will likely reach the panic zone without the content. This effect is only partially mitigated by an increase in delay-tolerance of the content. The analysis, carried out employing the real-world deployment in the city of Bologna, suggests that in order to offload a substantial part of traffic through fixed hot-spots, their deployment should be carefully planned, without black holes in spatial coverage.

As a second option, we consider our nodes to be able to communicate both with APs and other vehicles through direct ad hoc links, without considering any centralized coordinator to re-injects copies (the *AP + Opportunistic* strategy). This proves to be very beneficial for the overall offloading performance, increasing efficiency up to 50% with respect to *AP only*. The possibility to exchange data directly among users, together with mobility, allows to spread the infection in many areas that are not covered by fixed APs through *store-carry-forward* routing. Gains, as expected, rapidly improve as the reception delay increases. Fixed

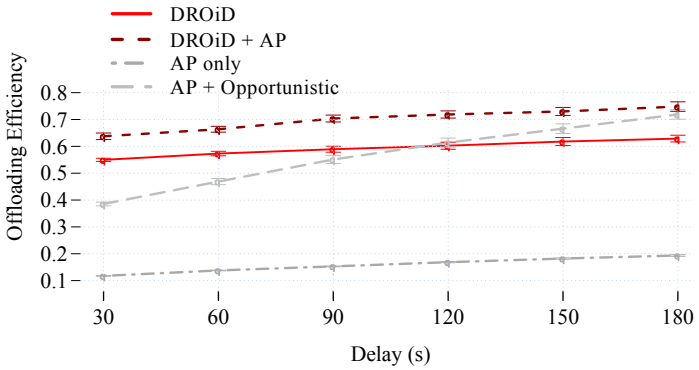


Figure 3.11: Bologna trace: offloading efficiency comparison between DROiD, AP-based and AP-based and AP + opportunistic distribution strategies. 95% confidence intervals are plotted.

APs, in this case, act as fixed (and free) infection source. Still, benefits of feedbacks and subsequent re-injections through the cellular link emerge for shorter reception delays (up to 120 s). DROiD here results always preferable than AP-based strategies. Once again, re-injections proves essential in the case of lagging infection evolution, and this is particularly true for short distribution intervals, where opportunistic contacts among vehicles may be scarce. For short distribution intervals, infected nodes hardly can carry the content far from AP location before its expiration. On the other hand, for longer delay tolerance values, the continuous use of the APs to infect neighbor nodes gives the *AP + opportunistic* strategy an edge. In order to take advantage of the fixed hot-spot infrastructure along with the feedback-based re-injection through derivative strategy, we evaluate DROiD coupled with APs. In this case, APs are pre-fetched with the content and initial distribution at t_0 is handed over to them. The re-injection algorithm intervenes only when the diffusion lags, so to overcome the difficulties encountered by users located away from APs range. This strategy emerges to be always the best, guaranteeing more than 65% of offloaded content. APs guarantee a steady infection rate to vehicles passing in their transmission range, letting the cellular infrastructure to target users in more isolated areas.

Further analysis of the average load flowing on each interface reveal interesting and unexpected information. In fact, from simulation logs it turns out that in the offloading strategies employing jointly opportunistic and AP-based communications (the *AP + opportunistic* and *DROiD + APs*), the fixed hot-spots are mainly used to bootstrap the dissemination, which is then carried over with subsequent direct communications between mobile nodes. The aggregate amount of data flowing through APs in these cases is roughly 10 times less than the case when hot-spots are the sole offloading options. Remarkably this little amount of data transferred through APs is very important to kick-start dissemination, offering an advantage compared to the non-AP based solutions (around 10% if we compare the two versions of DROiD).

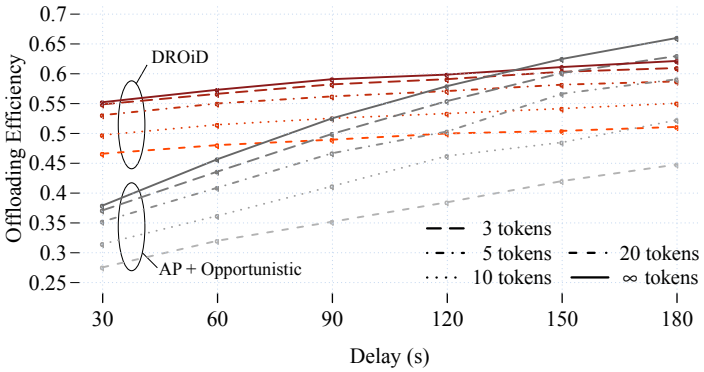


Figure 3.12: Offloading efficiency as a function of the number of transmission tokens for DROiD and *AP + Opportunistic*. Confidence intervals omitted for clarity.

3.5.6 Energy savings

A critical challenge to make mobile data offloading potentially attractive to end-users is to attenuate the impact of opportunistic communications on the battery of devices, concurring thus to increase their lifetime. For this reason, we analyze the impact that simple energy-saving methods have on offloading performance.

In our analysis, we compare DROiD with the *AP + opportunistic* strategy, fixing the maximum number of possible opportunistic transmission that a node can do for each message. To put energy saving strategies in practice, we offer to users only a fixed amount of *tokens* for each content, which is decreased each time the content is forwarded. When the token count is equal to 0, the node stops forwarding, and waits for the next content to appear. As we see from simulation results presented in Fig. 3.12, lowering the number of possible ad hoc transmissions has an impact on offloading performance, wasting possible contacts, lowering the network capacity. The impact of energy saving schemes is more pronounced for the *AP+opportunistic* schema, which sees performance highly lowered. The performance gap stretches as the delay-tolerance increases because nodes are more likely to run out of tokens. From the figure, we may appreciate that restricting the number of tokens to 20 does not bring a substantial performance hit for DROiD, while its influence is more pronounced in *AP+opportunistic*.

An energy saving scheme should trade off offloading efficiency for battery life, while ensuring that the overall energy cost is split equally between all the actor nodes. However, in opportunistic networks, contacts are typically imbalanced between nodes, and it is not uncommon to find nodes that during the message lifetime sustain an important number of data forwarding (these are typically highly central nodes that are targeted as preferred data carriers in other offloading strategies [157, 108]). To evaluate this aspect, we compare the fairness in the number of opportunistic transmission for the two schemes under evaluation with different token values and content reception delay. The traditional Jain's fairness index [159] is used to assess the number of transmissions made by a node for each content. We can see

	3	5	10	20	∞
30s	0.22	0.21	0.19	0.15	0.15
60s	0.27	0.25	0.23	0.2	0.17
90s	0.29	0.27	0.27	0.23	0.18
120s	0.3	0.33	0.3	0.25	0.21
150s	0.35	0.33	0.32	0.27	0.2
180s	0.39	0.37	0.34	0.29	0.21

(a) AP + Opportunistic

	3	5	10	20	∞
30s	0.43	0.4	0.34	0.29	0.28
60s	0.44	0.41	0.35	0.3	0.28
90s	0.46	0.42	0.36	0.31	0.29
120s	0.46	0.42	0.36	0.31	0.28
150s	0.47	0.43	0.37	0.31	0.28
180s	0.47	0.43	0.38	0.31	0.27

(b) DROiD

Figure 3.13: Jain's fairness index for different combinations of number of tokens and delay tolerance.

from Fig. 3.13 that DROiD always presents a better fairness than AP-based strategy. Analysis of nodes forwarding behaviors reveals that in the former strategy the number of users participating in content forwarding is sensibly lower than in DROiD. This depends on a mix of two factors: the efficiency is lower in general, so a greater number of nodes do not physically store the content to forward. Nevertheless, the effect relies also on the fact that AP-based strategies tend to concentrate data forwarders among those nodes that receive the content first. This is the reason why in the 180 s scenario with infinite tokens, DROiD shows better fairness but lower efficiency. Expectedly, fairness increases with an increasing delay-tolerance because nodes have more time to balance their number of transmissions. The advantage of a limited number of opportunistic transmissions is that it is a fair strategy for mobile users, at the cost of a lowered offloading efficiency. DROiD represents a good trade-off resulting in good transmission fairness, while offloading the most part of traffic.

3.6 DROiD in the wild: testbed implementation

Starting from the definition of the offloading system and protocols described in this chapter, we turned the proposed architecture into a real testbed by implementing DROiD. This testbed has been built in collaboration with the partners of the project FP7-MOTO [11]. Notably, a mobile operator (AVEA, the second largest cellular operator in Turkey) and a wireless hot-spots provider (FON, holder of one of the largest WiFi network in the world, with over 14 million hot-spots deployed worldwide) are involved in testbed development and testing. Below, we provide in detail the required components that allow data offloading in the testbed. The building blocks composing the architecture (illustrated in Fig. 3.14) interact with the infrastructure network of operators. The proposed offloading platform can be integrated either inside or outside the infrastructure.

To allow operators relieve the load on their network, the offloading architecture includes additional functionalities implemented by building blocks providing the Core MOTO Services. The testbed platform implements DROiD, and coordinates data dissemination. Any application can require the offloading platform to deliver a content while offloading the in-

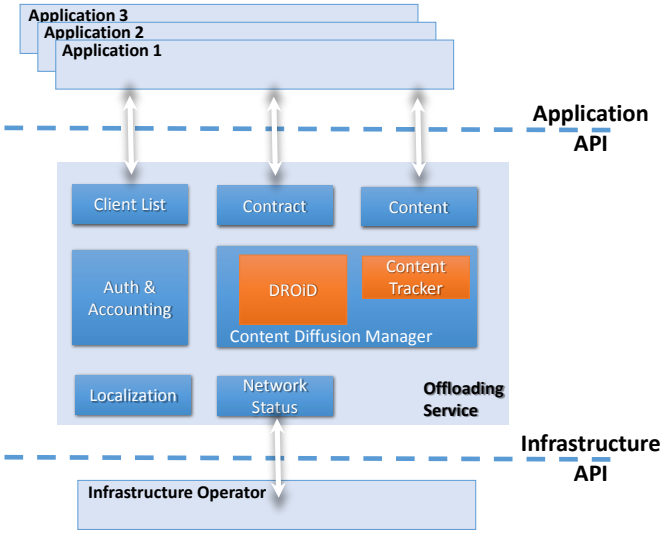


Figure 3.14: Offloading architecture built on top of DROiD in the context of the project FP7-MOTO.

frastructure network(s), using the Application API. Through this API, an application solicits the platform by giving the following inputs: the content to disseminate, the list of clients interested in this content (subscribers), and the service constraints to be met (SLA, e.g. the delay tolerance of the content). The testbed implements the following functional blocks:

- Client list, contract, content: to perform offloading, DROiD manages the content delivery to many clients according to the contract defining conditions for content delivery (e.g. maximum delivery delay). The list of customers, the SLA (Service Level Agreement), and the content to disseminate are given by any application that solicits the MOTO Services through the Application API. The content itself, or a link to the content hosted at the application level can be provided to MOTO.

- Localization: DROiD may need information on the localization of mobile clients to trigger offloading. Such knowledge can be gathered using the cell information provided either by operator networks (e.g., cell information in LTE, topological information on WiFi access points), or directly by the users (e.g., their GPS position if available).

- Content Diffusion Manager: this functional block implements DROiD strategy, and includes two sub-elements:

  - Dissemination strategy: this function is responsible for piloting the offloading process. It determines, from the list of clients that require the content and from localization and topological information gathered, the dissemination strategy to

be applied (e.g., deliver the content to clients A, B and C through the LTE network and ask them to relay the content in their neighborhood using D2D transmissions with specific routing policies). The dissemination process is monitored (cf. Content tracker) and the dissemination strategy can be updated during the dissemination process. After a given amount of time (panic zone), the traditional diffusion through the cellular network might be used to serve the content to all clients that did not receive it by other means.

  - Content tracker: this component receives acknowledgments sent by users when they receive the content. This element thus plays the role of monitoring the content dissemination process.

- Auth & Accounting: this module includes authentication, trust and credit management functionalities, as well as a database in which specific information on clients is maintained (trust and reputation indicators, credentials status).

- Network Status: this module allows maintaining information on the status of the available network infrastructures (e.g. remaining capacity for each network) that can be used to elaborate a dissemination scheme. Network status information can be requested by this component through the Infrastructure API, either on a regular basis or only when required.

The offloading architecture described above puts into practice the contributions in this chapter, demonstrating the efficient offloading of a primary network (e.g. cellular) by exploiting the direct ad hoc connectivity between clients or an alternative infrastructure (e.g. WiFi). The testbed has also been presented in two prominent conferences on mobile systems in the last year. [160, 161].

3.7 Conclusion

In this chapter, we first brought evidence of the stepwise behavior of the epidemic diffusion in opportunistic network, demonstrating that it depends on the dynamic clustering of nodes. We offered an analytical explanation of this behavior. To obtain efficient offloading in such a context, we proposed and evaluated DROiD, a low-complexity offloading strategy that adapts to the varying opportunistic dissemination evolution to improve the distribution of popular contents throughout a mobile hybrid network. Unlike most approaches in the literature, which seek to pre-compute the optimal subset of users to be reached through cellular communications, our system leverages on the availability of a control channel to track the evolution of content diffusion through user-sent acknowledgments. Thanks to a smart re-injection algorithm, DROiD guarantees better offloading performance than other state-of-the-art strategies. DROiD perceives when the evolution of the content diffusion stagnates,

and reacts in advance with respect to traditional strategies that considers only the actual infection rate.

Re-injection proves to be a successful strategy in the case of heavy mobility. As proved throughout past sections, even the Oracle, which is capable of knowing in advance the connectivity graph of nodes, struggles to offload an important amount of data, due to the inherent mobility of nodes that reduces the efficiency of predictions. The continuous analysis of the diffusion evolution proves to be a straightforward, but at the same time very effective method to react to the variability of contact patterns among nodes. The use of fixed WiFi APs could bring a further improvement, namely to kick start the distribution process and to deliver free copies of the content to users located inside their coverage range. Nevertheless, if we consider tight delivery times, the use of the pervasive cellular infrastructure is still required to target isolated node. Future work in this direction is manifold. First, we want to push the characterization of the epidemic diffusion further, especially in real scenarios. We also plan to investigate an analytical model that predicts the impact of intermittent connectivity on the dynamic formation and dissolution of clusters.

Finally, we have also defined the protocols involved in DROiD's process in order to experiment it in a real testbed. A real-world demonstrator, developed in collaboration with the partners of the FP7-MOTO project [11], has proved useful to carry on experimentations and test real world performance.

Remember the two benefits of failure. First, if you do fail, you learn what doesn't work; and second, the failure gives you the opportunity to try a new approach.

Roger Von Oech

4

Offloading LTE Multicast Data Dissemination through D2D communications

The reference scenario considered along this dissertation requires the distribution of the same piece of data to a community of interested users gathered within a limited geographical area (e.g., within a metropolitan area as discussed in Chapter 3.4). This is the case, for example, of software updates, on-demand videos, and road traffic information. In such situations, when data requests are spatially and temporally correlated, besides data offloading, another possible approach may address effectively the needs of operators: *LTE multicast*.

Traditionally, multicast helps save resources in the backbone. Multicast in the LTE standard improves the utilization of the last-hop radio link between the LTE base station (eNB) and user equipments (UEs). By exploiting the broadcast nature of the wireless channel, multicast benefits from a single unidirectional link, shared among several UEs inside the same radio cell. This permits, in principle, a more efficient use of network resources with respect to the case where each UE is reached through dedicated unicast transmissions. To ensure coexistence between multicast and unicast services, operators must reserve a fixed amount of resources for multicast transmissions.

Lately, field trials for video service during crowded sport events like the super-bowl have tested the effectiveness of multicast [162]. However, despite its attractive features, cellular multicast presents intrinsic and still unresolved issues that limit its exploitation: i) the rate

adaptation to the worst channel user, and ii) the lack of reliability. The reasons behinds of these inefficiencies will be investigated in detail in Section 3.2.

4.1 Background

In this chapter, *we explore the combination of opportunistic communications and multicasting in the LTE standard*. As we will see later, this strategy brings in significant reductions in the load at the base station providing noticeable offloading gains. The user with the worst radio conditions inherently limits the efficiency of standard LTE multicast because the transmission rate should match its capabilities. Hence, performance suffers and resources are wasted when the number of receivers increases. Moreover, users in good channel conditions unfairly receive lower rates due to their multicast group membership. By leveraging D2D communications, instead, we may obtain additional performance gains in terms of radio resources consumed at the eNB. Well-positioned users can participate in mitigating the inefficiencies of multicast, by handing over content to nodes with bad cellular channel signal through opportunistic communications. Despite the benefits of a hybrid distribution strategy are evident, in the design of the joint distribution strategy we face several challenges specific to the opportunistic and wireless domains:

- Performance of the opportunistic delivery hinges on the mobility pattern of users. In addition, opportunistic networks can only guarantee a probabilistic assurance of data reception.
- Understanding which fraction of users to reach through multicast and D2D transmissions is vital to offer a minimal QoS while guaranteeing resource savings.

Since a truly optimal solution is not conceivable without precise knowledge of future contact patterns, we attack the problem from a more practical point of view. We apply a Reinforcement Learning (RL) approach to decide which fraction of UEs should be reached through a multicast transmission and which should be served using opportunistic communications. A central controller installed at the eNB decides, for each packet of a content item to be disseminated, which fraction of users to reach with a multicast transmission. Each decision results in a certain use of the cellular network resources, which generates a reward associated to that choice. This reward is then used to guide (probabilistically) the future choices of the controller. Due to the many similarities in the formulation, we adopt the well-known *multi-armed bandit* RL technique to implement this algorithm [163].

To fully understand the performance of this joint multicast/D2D approach, it is necessary to evaluate the amount of radio resources consumed at the base station. This motivates us to introduce a finer model of radio resource consumption than previous works in the offloading literature. While this is well understood in the literature on physical aspects of cellular communications, existing proposals for opportunistic offloading do not consider heterogeneous channel conditions, assuming that delivering a given amount of data (i.e., a fixed size

packet) to different users has always the same cost for the operator [108]. For instance, in Chapter 3 we did not consider heterogeneous channel conditions, assuming that delivering a given amount of data (i.e., a packet) to different users has always the same cost for the operator. Such an assumption does not hold in reality, as resource consumption varies according to the channel condition experienced by each user. In other words, *transmitting the same piece of content to users with different channel conditions do lead to uneven costs at the base station*. To the best of our knowledge, we are the first to evaluate this aspect in the context of data offloading.

As a summary, the main contributions of this chapter are:

- Joint offloading strategy. Our strategy combines multicast with D2D opportunistic communications to improve cellular traffic offloading.

- RL-based multicast selection. The multicast emission is driven by a RL algorithm. Exploiting the knowledge of past rounds, the algorithm allocates transmission resources to the dissemination scenario, allowing substantial savings at the cellular base stations.

- Fine-grained resource consumption analysis. We evaluate resource consumption employing the smallest radio resource unit that can be assigned to users for data transmission. This analysis shows that existing macroscopic techniques fail to capture actual system behaviors.

- Performance evaluation. The RL strategy permits to save consistent amount of radio resources at the eNB (up to 89% with a 90 s deadline). Even in the worst case, the RL approach approximates an unfeasible strategy that picks the best fixed fraction of multicast users after exhaustive search.

4.1.1 Background on LTE and eMBMS rate allocation

LTE downlink transmission is based on OFDMA frames made of different frequency subcarriers having a spacing of 15 kHz. OFDMA frames are further divided in the time and frequency domain to form the Resource Blocks (RBs), which are the smallest radio resource unit that can be allocated by the packet scheduler. The eNB (cellular base station for LTE) supports different modulation and coding schemes (MCS) to adapt transmission to the variable channel characteristics of users. The MCS determines how much data is transmitted over each RB. Channel adaptation is driven by Channel Quality Indicator (CQI) feedbacks from the UEs. The reported CQI is a number between 0 (worst) and 15 (best) as listed in Table 4.1. The CQI indicates the most efficient MCS giving a Block Error Rate (BLER) of 10% or less. In unicast transmission, the eNB selects the MCS and the resources to allocate

Table 4.1: CQI / MCS Table for LTE [165].

CQI index	Modulation schema	code rate x 1024	Spectral Efficiency [bit/s/Hz]	Efficiency increase (w.r.t. CQI 1)
0		out of range	-	-
1	QPSK	78	0.1523	1×
2	QPSK	120	0.2344	1.54×
3	QPSK	193	0.3770	2.47×
4	QPSK	308	0.6016	3.95×
5	QPSK	449	0.8770	5.76×
6	QPSK	602	1.1758	7.72×
7	16-QAM	378	1.4766	9.69×
8	16-QAM	490	1.9141	12.57×
9	16-QAM	616	2.4063	15.80×
10	64-QAM	466	2.7305	17.93×
11	64-QAM	567	3.3223	21.81×
12	64-QAM	666	3.9023	25,62×
13	64-QAM	772	4.5234	29,70×
14	64-QAM	873	5.1152	33,59×
15	64-QAM	948	5.5547	36,47×

to each UEs, based on this feedback regarding the channel state. An analytical characterization of the channel adaptation mechanism can be found in [164]. A higher value of CQI allows the eNB to select an MCS such that it can transmit more information inside each RB. The number of RBs necessary to transmit a given amount of useful bits (or, equivalently, the amount of information transmitted per RB) is a typical measure of cost and thus efficiency of LTE transmissions.

Apart from unicast transmissions, new LTE releases propose also an optimized broadcast/multicast service through eMBMS (*enhanced Multicast Broadcast Multimedia Service*), a point-to-multipoint specification to transmit control/data information from the cellular base station (eNB) to a group of user entities (UEs) [92]. In eMBMS, all the users belonging to the same multicast group receive the same transmission. Channel heterogeneity (time varying and user-dependent) reduces the effectiveness of multicast because the eNB uses a single MCS for the entire multicast group downlink data. Usually, the selected MCS should be robust enough to ensure the successful reception and decoding of the data-frame for each UE in the multicast group. Thus, the worst channel among all the receivers dictates performance. It follows that an increase in the number of users in the multicast group boosts the probability that at least one user experiences bad channel conditions, degrading the overall throughput [166].

To exemplify the influence of poor quality users, we simulate a 500×500 m^2 single LTE cell with an increasing number of randomly located receivers using the ns-3 simulator [154]. Fig. 4.1 presents the minimum average channel quality in terms of CQI, reported at the eNB by users. In this configuration, users are static, and their location is uniformly distributed inside the eNB coverage area. From Fig. 4.1, we highlight two aspects. The first one is that,

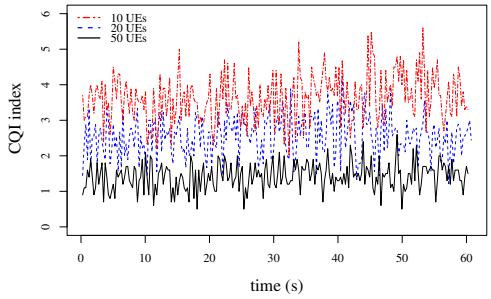


Figure 4.1: Minimum CQI for different multicast group sizes. 100 runs, confidence intervals are tight and not shown in figure.

when users are uniformly placed, there is a high chance of having at least one user experiencing a poor channel quality (e.g., even with only 10 users in the cell, we still have the worst CQIs not greater than about 4). The second point is that this behavior worsens as more users are added in the area. The result is that augmenting the number of multicast receivers clearly affects the attainable cell throughput. Table 4.1 shows also that an UE with the best CQI could theoretically receive 37 times the throughput of a user with the lowest index.

This greatly motivates us to investigate methods to offload multicast data using D2D connectivity to relieve the cellular infrastructure load, while reducing the influence of users experiencing poor radio conditions.

4.2 A reinforcement learning strategy for data dissemination

4.2.1 Content dissemination strategy and problem formulation

We address the dissemination of popular content to a set of N mobile users inside a single LTE cell. Each user is a multi-homed device that embeds both a LTE interface and a short-range technology that allows D2D communications. In simulation we consider IEEE 802.11g, however, the future integration of D2D capabilities within the LTE standard could be employed as well [167]. We want to transmit data with a guaranteed *maximum service delay* D, at the smallest cost for the cellular infrastructure (i.e., using the minimum number of RBs). We exploit the possibilities offered by D2D connectivity and store-and-carry forwarding. Specifically, instead of addressing all interested UEs with a single multicast transmission – that will likely result in a high cost in terms of used RBs – we address only a subset of the UEs (those in better channel quality), and exploit opportunistic D2D communications to reach the others. The challenging issue is that opportunistic dissemination is, by definition, unreliable, as it depends on many factors outside of the control of the cellular infrastructure (e.g., movement pattern of nodes, variable density of opportunistic neighbors, or interference on the D2D channel). The offloading strategy is essentially the same as DROiD. The

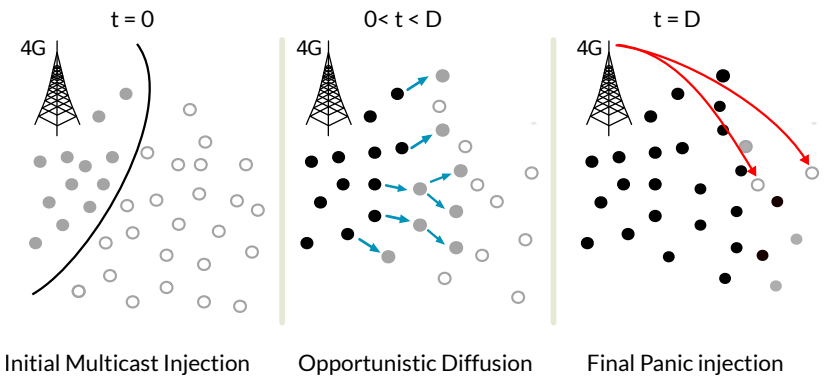


Figure 4.2: Users can decode data with a maximum modulation schema depending on their channel quality. The eNB may decide to multicast at higher rate. Users unable to decode data are reached through out-of-band D2D links and final panic retransmissions.

difference is that the initial injection is done via multicast transmissions, and no re-injections are made until the panic zone. Similarly, when the service delay reaches its maximum value D, the eNB pushes all the missing data to uninfected nodes using unicast transmissions. Of course, unicast transmissions represent the last opportunity to assure data reception [*] . In this scheme, the cost of disseminating content to interested UEs comes from i) the cost of the initial multicast transmission, and ii) the cost of the unicast transmissions in the panic zone.

Fig. 4.2 offers a representative example of the proposed dissemination strategy. To avoid the penalty due to the presence of UEs experiencing severe channel conditions, the eNB emits at a modulation that leaves them in outage. This is equivalent to restrict access to the multicast group only to the UEs in "relatively" good channel conditions. In the opportunistic dissemination phase, *outaged* UEs benefit from nearby nodes, fetching data through out-of-band D2D transmissions. This cooperative strategy is expected to be more efficient in terms of cellular resource consumption than multicast alone, given that the cellular rate increases and the D2D links typically exploit a much larger bandwidth than cellular communications. Finally, panic injections assure data reception to all users.

It is clear that such a scheme admits an optimal operating point. Reducing the set of UEs reached via the initial multicast transmission results in a lower cost for the multicast transmission. However, this may be paid with an additional cost for unicast transmissions in the panic zone, if the remaining UEs are not reached quickly enough through opportunistic dissemination. The challenge to identify this optimal operating point is that the cost of each possible configuration depends on future mobility of nodes, which is unknown at the time when the multicast transmission needs to be configured.

[*] The need for a guaranteed reception method is a common issue for multicast due to its shared nature. UEs have no assurance of reception due to the fact that the radio channel could suddenly degrade during data reception (e.g., due to fast fading or mobility). For this reason, mechanisms similar to panic retransmission are considered in several works in the literature [168, 169].

More precisely, the problem we address is the following: *how to select the initial set of seed users to be reached using multicast transmissions with the objective of minimizing the total number of physical resource blocks (RBs) needed for content dissemination*.

In our scheme, we use a single parameter I_0 to address this problem. Specifically, I_0 is the fraction of UEs that should be reached via the initial multicast transmission. Assuming that, when the multicast transmission is configured, UEs are ranked by decreasing value of CQI, this means that the eNB reaches only the best I_0 UEs in terms of channel quality. Optimally configuring I_0 is not trivial, because, while the cost of the multicast transmission is deterministic at the time when it is configured, the cost of the needed unicast transmissions in the panic zone is a stochastic variable, which depends on the pattern of mobility of UEs in the next D seconds.

We model this problem as a multi-armed bandit problem and we solve it through a Reinforcement Learning (RL) approach. As explained in detail in the rest of the section, our scheme is able to learn autonomously the best value of I_0, by observing the effect of different configurations on the resulting cost of disseminating a given content. Assuming that multiple content items need to be disseminated over time to a given set of UEs, our scheme actually learns the best *probability distribution* over the possible values of I_0, that results in the minimum cost in terms of RBs for a given stochastic mobility pattern of UEs. Without prior knowledge on the mobility patterns, and given that mobility is stochastic, learning the best distribution of I_0 is the only practical choice for a learning framework.

4.2.2 Background on multi-armed bandit algorithm

Let us now briefly introduce the general formulation of a multi-armed bandit problem (bandit for short). In the simplest case, in a bandit problem there is a set of K unknown probability distributions $\langle F_{D_1}, ..., F_{D_k} \rangle$ with associated expected values $\langle \mu_1, ..., \mu_k \rangle$ and variances $\langle \sigma_1^2, ..., \sigma_k^2 \rangle$.

For the sake of illustration, let us assume that F_{D_i} describes the distribution of the outcomes of the i^{th} arm on a slot machine (the bandit); the player is viewed as a gambler whose goal is to collect as much money as possible by pulling these arms over many turns. Initially, the distributions F_{D_i} are completely unknown to the player. At each turn, $t = 1, 2, ...$, the player selects an arm, with index $j(t)$, and obtains a reward $r(t) \sim D_{j(t)}$. Since the player does not know in advance the distribution F_{D_i}, it has to explicitly test the i^{th} action with a trial-and-error search. Therefore, the player has two conflicting objectives: on the one hand, finding out which distribution has the highest expected value (or explore the distribution space); on the other hand, gaining as much rewards as possible while playing (or exploit its knowledge). Reinforcement Learning algorithms specify a probabilistic strategy by which the player should choose an arm $j(t)$ at each turn. Clearly, the effectiveness of the solution depends on how the gambler handles the exploration/exploitation dilemma when testing the different arms iteratively. Exploitation maximizes its reward at present time; at the same time, exploration may lead to a greater total reward in the future.

4.2.3 Learning algorithm

The general multi-armed bandit formulation can be specialized as follows. First of all, in our problem each arm of the bandit corresponds to a different I_0 threshold. Thus, K is the number of different thresholds chosen for multicast emission. It follows that F_{D_i} is the distribution of the amount of RBs that are used during the entire dissemination process when I_0 is used as threshold. More precisely, $D_i = m_i + X_i$, where m_i is the fixed and known number of RBs that are used for a multicast transmission at the MCS of needed to reach the I_0 best UEs in terms of channel quality, and X_i is the random variable that models the total number of RBs used for the unicast transmissions during the panic zone. Note that X_i depends on many factors, including the set of seeds that are activated, the network topology and node mobility, as well as the dissemination strategy. In our case, each turn corresponds to the dissemination of a content composed of a multitude of packets that are transmitted independently. After the content deadline the reward for each threshold is updated. Assuming that $I_0 = i$ was used for the n^{th} multicast transmission, the obtained reward is computed as:

$$\mu_i(n) = \frac{1}{m_i + x_i(n)}, \tag{4.1}$$

where $x_i(n)$ is the number of RBs that are used for the unicast transmissions in the n^{th} panic zone. Note that the higher the number of used RBs and the lower the reward. To dynamically estimate the *average* reward $\mu_i(n)$ for each value of I_0 we use a classical exponential moving average with rate α:

$$\mu_i(n) = \alpha\mu_i(n-1) + (1-\alpha)\mu_i(n). \tag{4.2}$$

Now, we must define the policy to select at time $n+1$ the next I_0 value given the knowledge of the average rewards estimated at time n. Different learning methods have been proposed in the literature for the armed bandit problems.

The simplest one is the *ϵ-greedy* algorithm that selects with probability $(1-\epsilon)$ the I_0 value with the maximum accumulated reward (greedy action), while it selects with probability ϵ one of the remaining I_0 values at random (with uniform probability) independently of the reward estimates (exploration action). More formally, let $\pi_{i(n)}$ be the probability to set $I_0 = i$ for the transmission of the n^{th} packet, and $i^*(n) = \arg\max_i \mu_i(n-1)$. Then, in the ϵ-greedy algorithm it holds that $\pi_{i^*(n)} = 1 - \epsilon$.

Another class of learning algorithms is known as *pursuit* methods, in which the π probabilities are selected to strengthen the last greedy selection. Specifically, let $i^*(n)$ be the greedy value of I_0 defined above. Then, just prior to selecting the CQI for the transmission of the n^{th} packet, the greedy probability is reinforced as follows

$$\pi_{i^*(n)}(n) = \pi_{i^*(n)}(n-1) + \beta[\pi_{MAX} - \pi_{i^*(n)}(n-1)], \tag{4.3}$$

while all the non-greedy probabilities are updated as follows

$$\pi_{i(n)}(n) = \pi_{i(n)}(n-1) + \beta[\pi_{MIN} - \pi_{i(n)}(n-1)], i \neq i^*. \quad (4.4)$$

Here π_{MAX}, π_{MIN} are respectively the upper and the lower bound that the probability $\pi_{i(n)}(n)$ can take $\forall i, n$. In equations B.2 and B.3 the greedy choice is increased, but never more than π_{MAX}, and each non-greedy choice is reduced, but no less than π_{MIN}. This guarantees that the pursuit method is able to cope with the possible non-stationarity of the problem we are considering, i.e. the distribution of rewards can change over time due to the underlying mobility. Compared to the pursuit method, the ϵ-greedy strategy presents a threshold effect by which the choice that has the maximum accumulated reward immediately gets the highest probability, while in pursuit the likelihood of the same option gradually increases by a factor β proportional to the distance to the maximum bound π_{MAX} (similar remarks hold for the non greedy choices). So in pursuit the evolution of the distribution over the possible choices is less drastic and more gradual.

4.2.4 Wrap-up of the dissemination strategy

As a summary, the key principles behind the joint multicast/D2D approach are: i) at initial time, the eNB sends data to the best I_0 CQI-ranked UEs through a single multicast emission. A RL algorithm is employed to learn the experimental distribution ($\pi_{i^*}(n)$) for the I_0 parameter; ii) the UEs that have received the data through the multicast emission start disseminating it in a D2D (epidemic) fashion; iii) before the maximum *deadline* D, we define a time interval, a *panic zone* where all the nodes that have not yet retrieved the content (either with the initial multicast emission or in D2D fashion) receive it through unicast cellular retransmissions. The proposed scheme allows all UEs to receive data by the deadline (as long as the panic zone is sufficiently large). It adapts to different *deadlines* – the larger ones allowing for more D2D dissemination. Its performance relies on the RL algorithm that permits the cellular base-station to learn by experience the best transmission rate for each multicast emission.

4.3 Performance Evaluation

4.3.1 Methodology and parameters

We consider location aware content distribution in a pedestrian scenario such as a shopping mall or a crowded touristic landmark. Possible contents of interest include location based broadcasting with advertisement, geo-relevant data, alerts, and public utility information; nevertheless, the proposed system also supports the efficient distribution of over-the-air software updates.

We simulate UDP constant bit-rate downlink flows, with packet size $s_k = 2,048$ bytes and a total content size of 8 MB. Each packet is distributed independently using the multi-

Table 4.2: ns-3 simulation parameters.

Parameter	Value
Cellular layout	Isolated cell, 1-sector
LTE downlink bandwidth	5 MHz (25 RBs)
Frequency band	1865 MHz (Band 3)
CQI scheme	Full Bandwidth
eNb TX-power	51 dBm
Pathloss	Cost 231
Fast fading	Extended Pedestrian A (EPA) model
Cell dimension	200×200 m^2
eNb position	100, 100 m
eNb antenna height	30 m
User antenna height	1.5 m
Multicast group size N	10, 25, 50 UEs
Service delay D	30, 60, 90 s

armed bandit algorithm. The considered traffic has a loose QoS guarantees on a per-content basis (meaning that individual packets can be delayed, but the entire content must reach the user within the given deadline D).

The synthetic mobility of UEs is implemented according to a Random-Waypoint model on a 200×200 sq.m. area. Nodes move in this space with a speed falling between 1 and 2.5 m/s (pedestrian speed).[†] The synthetic mobility trace is the input of a packet-level simulator. Indeed, we implemented the multi-armed bandit algorithm in the ns-3 network simulator, which emulates the full LTE and Wi-Fi stack, allowing very realistic simulations [171, 172]. The network is composed of an eNB placed in the center of the interest area, a remote server that provide the content, and multiple mobile devices. All the UEs connects to the same eNB during the experiments. Since ns-3 does not natively support cellular multicast, we implemented an additional module that interacts with the packet scheduler to emulate single-cell multicast. The multicast module decides, upon each transmission, the fraction of UEs to be reached directly, i.e. the I_0 parameter, based on the multi-armed bandit algorithm. It receives the CQI from the standard LTE modules, and sets the MCS of the multicast transmission to reach the intended UEs. We fix the bandwidth allocated for the multicast service at 5 MHz. 3GPP standard recommends not to reserve more than 60% of RBs to multicast [92], so a 5 MHz value could represent respectively 50% or 25% of RBs in a typical 10 or 20 MHz deployment. The other simulation parameters for the LTE cell are listed in Table 4.2.

Additionally, we implemented DTN store-carry-forward routing mechanism at UEs to support D2D opportunistic communications. This is an implementation of the conventional epidemic forwarding mechanism [99]. Regardless of its reception method, an unexpired packet can be forwarded on the Wi-Fi interface upon meeting with neighbors. Neighbor discovery is implemented through a beaconing protocol triggered each 250 ms. UEs peri-

[†] While the realism of the random waypoint model is questionable in general, it has been shown that it realistically reproduces movement patterns of groups of users moving in a confined physical area [170]. Therefore, we consider it appropriate for simulation in our target scenario.

odically broadcast beacon messages containing their identifier and the list of buffered packets. Upon beacon reception UEs update their vicinity information and can transmit packets opportunistically.

All the simulation results are averages over 10 independent runs. Unless otherwise stated, confidence intervals are not shown, as they are very tight (usually in the order of 5% of the average value). We assess the performance for different values of N (the number of users inside the cell) and D (the maximum reception delay)so to evaluate performance under different loads.

Implementation assumption: In simulation we make the following simplifications:

- HARQ-level retransmissions and RLC-level feedback are disabled in multicast. This is a reasonable assumption: otherwise the eNB should merge the *ack/nack* messages received from all the UEs, and decide which is the best retransmission strategy. We guarantee the maximum content delivery time D with *panic zone* retransmissions.
- The PUCCH channel is employed to acknowledge data reception towards the eNB. Panic zone retransmissions are then triggered looking at the list of received acknowledgments.
- The multi-armed bandit algorithm acts as a packet scheduler. It employs a cross-layer design at the eNB. By exploiting signaling from physical layer (i.e., the amount of RBs consumed and the CQI for each UEs are used to evaluate the reward), the algorithm decides the MCS of each multicast transmission.

We are aware that our simulation-based evaluation has some limits. First, we consider a simplified version of the eMBMS standard. The proposed approach requires deeper integration with the eNB scheduler. For now, we leave out the discussion on incentives that are vital to convince users to agree to spend their battery and storage resources to relay data to someone else. This is an orthogonal problem that is addressed in the opportunistic networking literature through appropriate mechanisms and that we will treat in the next chapter.

4.3.2 Reference strategies

We compare our proposal with four different strategies for content delivery. The main performance indexes we consider are (i) the number of RBs used by the eNB to deliver the content by the stated deadline, and (ii) the offloading ratio, i.e. the fraction of UEs that are served via D2D opportunistic communications with respect to the case where only multicast is used. The considered strategies are the following ones:

- Multicast-only is the basic strategy, where UEs have no other means than the cellular network to receive data.

UEs	D	Pursuit	Epsilon	Fixed–Best
10	30s	69	65	55
	60s	66	64	51
	90s	59	61	66
25	30s	52	11	69
	60s	75	74	71
	90s	73	77	83
50	30s	25	–34	55
	60s	30	–19	72
	90s	76	80	85

(a)

UEs	D	Pursuit	Epsilon	Fixed–Best
10	30s	73	63	63
	60s	73	70	54
	90s	75	73	73
25	30s	69	12	70
	60s	80	80	71
	90s	83	84	83
50	30s	54	–28	55
	60s	55	–31	72
	90s	88	87	89

(b)

Figure 4.3: Levels of cellular offloading for the considered scenarios. (a) Aggregate. (b) Steady-state. Savings are referred to the multicast-only scenario (%).

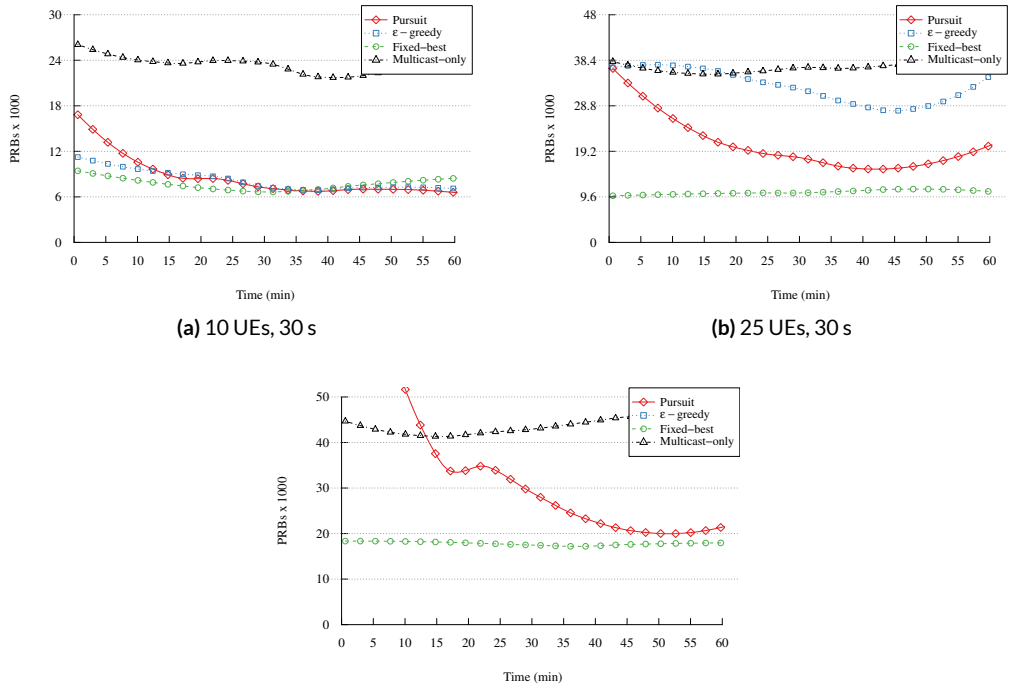
- FIXED-BEST minimizes the number of RBs maintaining a static allocation of multicast users (I_0 is fixed during all the simulation duration). Since the optimal size of the multicast group is unknown, we ran extensive simulations to find experimentally the I_0 value that minimizes the aggregate RB usage. This strategy represents the experimental benchmark for all the RL-methods.

- ϵ-GREEDY estimates the reward using the exponential moving average presented in Eq. 4.2. This simple algorithm selects the greedy value of I_0 with probability $1 - \epsilon$. In our implementation, we selected $\epsilon = 0.05$ and $\alpha = 0.5$. We motivate this choice as a trade-off between different requirements. We need to maintain the exploration phase active in order to cope with the non-stationarity of the underlying process. However, transmitting with a wrong CQI can lead to significant efficiency loss.

- PURSUIT selects the I_0 transmission probability following Eqs. B.2 and B.3. In this case, the transmission probability pursues the greedy action by adapting the likelihood of emission to the temporal evolution of the system. In simulation we fixed $\beta = 0.3$, $\pi_{MIN} = 0.01$, $\pi_{MIN} = 0.95$. We will give more explanations on the choice of these parameters later on.

4.3.3 EVALUATION

Fig. 4.3 provide a summary of the resource savings (aggregate over 1 hour, and after the transient state) for the two considered algorithms (ϵ-greedy and pursuit), related to the basic *Multicast-only* approach. The RL solution to the joint multicast-D2D problem is an effective method to save resources at the eNB, approaching and even surpassing *Fixed-best* in more than one occasion.



(a) 10 UEs, 30 s

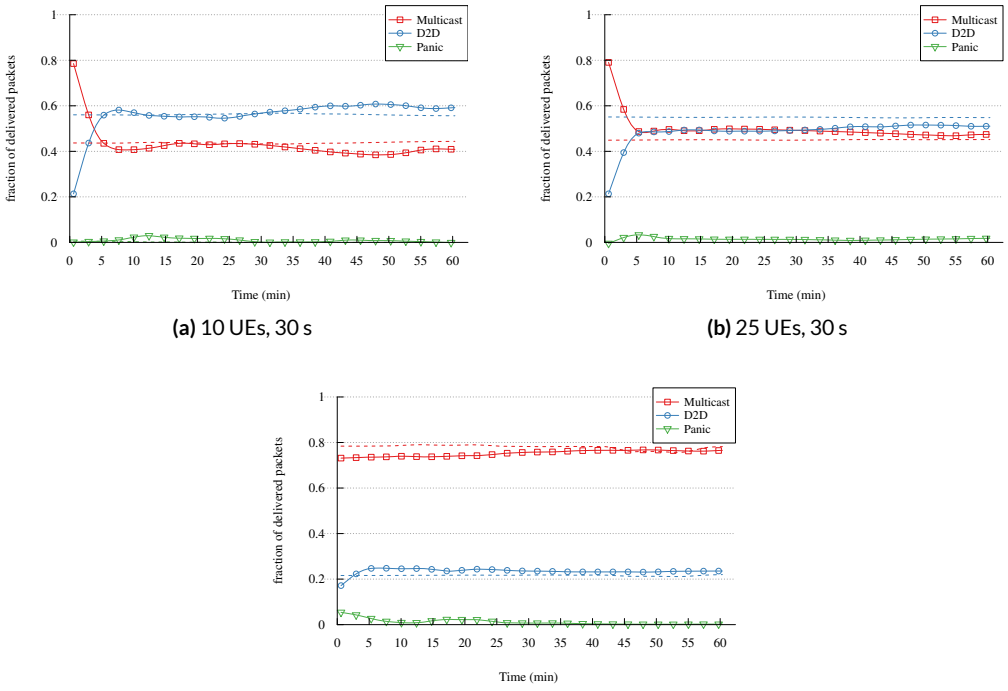
(b) 25 UEs, 30 s

(c) 50 UEs, 30 s

Figure 4.4: RBs usage for *Multicast-only* (black), *ϵ-greedy* (blue), *Fixed-best* (green), and *pursuit method* (red). Content is divided into 4000 packets of 2048 bytes. Plots are averaged over 10 runs, 95 % confidence intervals are not plotted but are knit.

Focusing on Fig. 4.3(b), which depicts the RBs savings after the training phase, our system allows saving up to 88% of RBs for the 90s scenario if compared to *Multicast-only*. This result confirms that the right synergy in the utilization of multicast and D2D resources allows for significant resource savings. Even with shorter deadlines, the pursuit method performs very well, saving at least 54% of RBs. The main difference between RL-based methods and the benchmark represented by *Fixed-best* is that in the latter the value of I_0 is fixed and pre-computed in advance. Its performance is stable over all the dissemination periods, but this optimal value is the outcome of an extensive trial and error simulation phase. Conversely, a learning strategy is able to predict the distribution for selecting the values of I_0 that results in saving comparable with those of *fixed best*, without relying on any prior knowledge and adapting to the network state.

We can observe in Fig. 4.4 this effect for the tightest deadline considered (30 s). The behavior of the RL methods (*pursuit* and *ϵ-greedy*) and *Fixed-best* are significantly different. *Pursuit* and *ϵ-greedy* are based on a learning algorithm; they therefore need time to learn the most appropriate distribution for I_0. Once trained, their performance is often on par or even better than the best fixed-value strategy represented by *Fixed-best*. Another advantage is that



(a) 10 UEs, 30 s

(b) 25 UEs, 30 s

(c) 50 UEs, 30 s

Figure 4.5: Pursuit method, reception method. Dashed lines are the objective ratio for *Fixed-best*. Content is divided into 4000 packets of 2048 bytes. Plots are averaged over 10 runs, 95 % confidence intervals are not plotted but are knit.

even when those strategies are trained, they continue to explore the solution space, being able to cope with the possible non-stationarity of the contact process that rules the opportunistic diffusion. Conversely, *fixed-best* is locked to a static value of the parameter I_0 and insensitive to variations in the mobility of UEs. One of the key advantages of the RL strategies are that they can autonomously find the trade-off between multicast and D2D transmissions in a reasonable time - without extensively search all the entire parameter space.

4.3.4 Comparison between ϵ-greedy and pursuit method

Given the heavy request for mobile data today, operators are mainly concerned about radio resource usage. To examine the impact on RB consumption, we fix the deadline at 30 s and vary the number of multicast UEs in the cell from 10 to 50. Intuitively, more UEs asking for the same content should require more infrastructure resources. On the other hand, the number of contact opportunities increases, offering more possibilities to offload the network.

Fig. 4.4 gives hints on the actual amount of RBs devoted to distribute data in the considered scenarios. Unlike many other works in the literature, the use of the ns-3 simulator allows us to evaluate precisely the amount of radio resources consumed at the eNB. In gen-

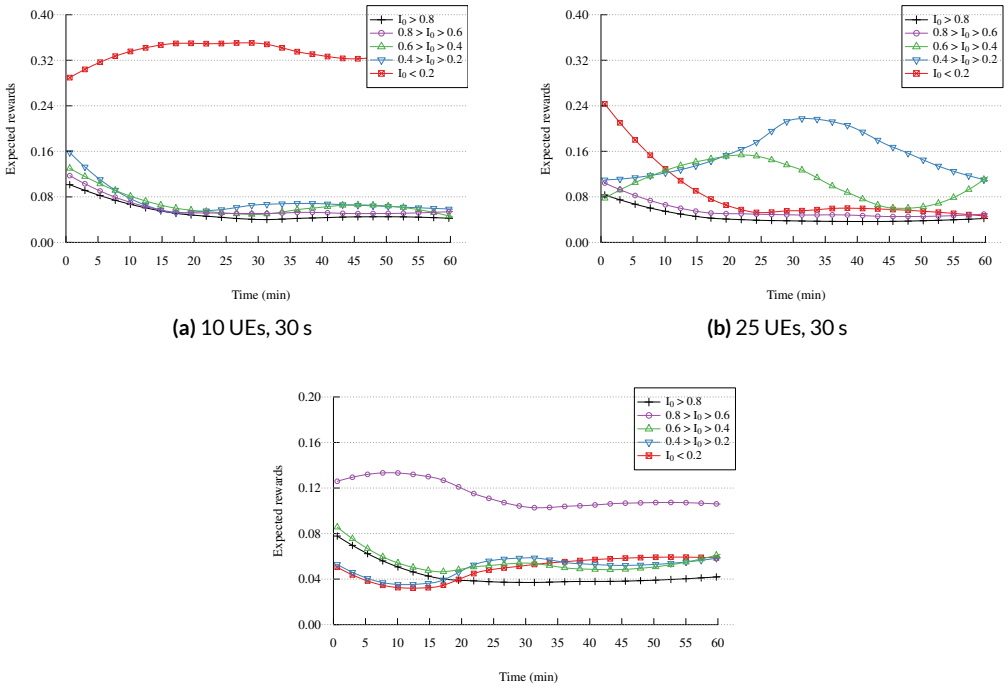
eral, we note that when it converges, the *ϵ-greedy* method is faster than the *pursuit* method (e.g., Fig. 4.4(a)). In many cases, even when *ϵ-greedy* fails to converge, *pursuit* approaches the behavior of the *fixed-best* strategy. This depends on the fact that *ϵ-greedy* always selects the value of I_0 that maximizes the expected rewards. Instead, *pursuit* has an indirect selection method that better adapts to the temporal evolution of the system. The added complexity is however beneficial in most cases, as it results into an improved performance (Fig. 4.4(b) and Fig. 4.4(c)). The reinforcement given by Eq. B.2, allows smoothing out the inherent variations in epidemic diffusion that prevent the proper prediction in the *ϵ-greedy* method. The effect appears when the number of targeted UEs increases while keeping a tight deadline (i.e., 30 s). In those scenarios, the variability in performance of the opportunistic diffusion prevents the *ϵ-greedy* method to learn properly the best distribution for selecting I_0. An example of this effect is illustrated in Fig. 4.4(c). In that case, the *pursuit method* succeeds in matching the *fixed-best* strategy. The *ϵ-greedy* method instead diverges nearly instantaneously, failing to learn an appropriate policy.

We draw the lesson that the *ϵ-greedy*, method owing to its simplicity, does not fit well scenarios with significant variability. For those cases, the *pursuit* method is a better match. On the other hand, in scenarios where the variability of the opportunistic process is low – i.e., when the deadline is large – the *ϵ-greedy* approach allows for a quicker convergence time.

4.3.5 Detailed evaluation of the pursuit method

We plot in Fig. 4.5 the fraction of packets partitioned by their reception method. Considering the same deadline and increasing the number of UEs has the effect of reducing the share of D2D transmissions. While a larger number of UEs should multiply the contact opportunities, many of them are not adequately exploited because UEs can transmit only to one neighbor at a time. The result is that fixing the deadline, the share of UEs addressed through D2D transmissions is upper bounded. In this case, we analyze the detailed evolution of our strategy considering the shorter deadline (30 s). A similar analysis can be done with the longer deadlines. Note, however, that for a larger number of UEs, even though the fraction of UEs addressed through D2D transmission is limited (e.g., 20% in the case of 50 UEs), the resulting advantage in terms of RB saving is much higher (around 55% for that case, from results in Figure 4). In this case, the RL algorithm understands that it is better to see the 20% of the UEs that are experiencing bad cellular quality through D2D. Serving them through multicast is expected to result in a too high cost in terms of RBs, due to the need of reducing too much the MCS. On the other hand, decreasing too much the share of UEs served by the multicast transmission brings the opposite effect, with a considerable amount of RBs spent for unicast transmissions in the panic zone.

We also take note of a peculiarity. Looking at the reception methods in Fig. 4.5, the convergence time looks like always less than 10 minutes. Comparing this value to Fig. 4.4, however, we realize that the actual convergence (in terms of RBs employed at the eNB) happens much later in time (around 40 min). This anomaly is justified by the fact that even a small amount of



(a) 10 UEs, 30 s

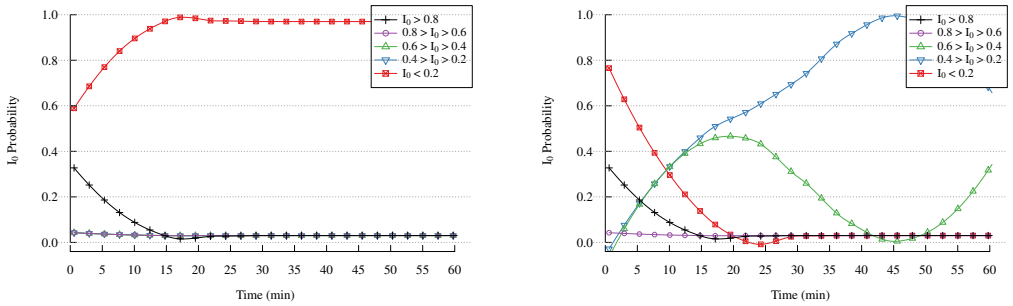
(b) 25 UEs, 30 s

(c) 50 UEs, 30 s

Figure 4.6: Pursuit method, average reward values for I_0. Content is divided into 4000 packets of 2048 bytes. Plots are averaged over 10 runs, 95 % confidence intervals are not plotted but are knit.

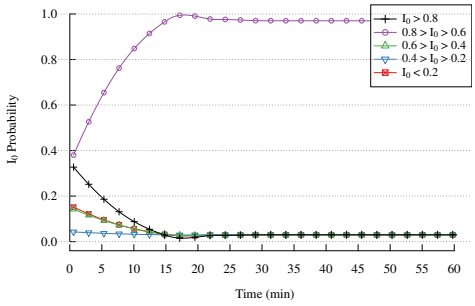
unicast retransmissions in the panic zone consumes many more resources than the multicast emission. The fine-tuning required to reach an optimal RBs usage level is thus responsible for this longer convergence time. When the number of UEs is large, the choice of the appropriate distribution for I_0 becomes fundamental in order to avoid congesting the cell with too many panic retransmissions.

Considering the detailed mechanisms of the *pursuit* method described in Sec. 4.2.3, Fig. 4.6 and 4.7 compare the rewards and I_0 probabilities respectively. In the figures, we quantized the values of I_0 to form five levels. In two cases out of three (namely for 10 and 50 UEs), there is a set of values for I_0 that performs clearly better than the others do. In the 25 UEs scenario instead, distribution of I_0 is more spread out. There is a clear tendency to prefer higher values of I_0 though, in the range between 0.2 and 0.6. The best I_0 value is the one that is not affected too much by the loss in spectral efficiency due to the reduced multicast rate, but at the same time can guarantee a low penalty due to unicast panic re-injections. In the 10 UEs scenario, emitting with a multicast rate that targets one or two users ($I_0 \leq 0.2$) is sufficient to achieve high efficiency. On the other hand, we note that increasing the multicast group size, the best distribution of I_0 shifts towards higher values. Intuitively, the penalty due to panic re-injections is extremely severe in these cases and the pursuit algorithm tends to



(a) 10 UEs, 30 s

(b) 25 UEs, 30 s



(c) 50 UEs, 30 s

Figure 4.7: Pursuit method, average emission probabilities for I_0. Content is divided into 4000 packets of 2048 bytes. Plots are averaged over 10 runs, 95 % confidence intervals are not plotted but are knit.

allocate more seeders in the opportunistic domain.

Finally, the emission probability follows the pattern of the rewards with the exception that the greedy probability is always reinforced until the value π_{MAX} and the non-greedy probabilities are reduced until π_{MIN}.

4.4 Conclusion

In this chapter, we have presented a hybrid distribution strategy, jointly leveraging LTE multicast and opportunistic D2D communications to distribute popular content with guaranteed delays. Multicast is an advantageous option to distribute popular data into a cellular network. However, performance is determined by the UE with the worst channel quality inside the multicast group. We proposed a framework to counter the inefficiencies of cellular multicast and offload part of the traffic using D2D communications.

The proper balance of multicast and D2D transmissions is achieved using a multi-armed bandit learning strategy. We proposed and evaluated two different algorithms under variable multicast group size and reception delays. Simulation results prove that D2D communications permit to configure multicast transmission in a more efficient way, saving resources

and improving the overall cell throughput. On the other hand, we have also shown that the used learning algorithms are able to obtain performance comparable (and in several cases even superior) to the best possible strategy that uses a fixed split between multicast and D2D communications, which can only be identified *after* exhaustive search, and is thus practically unfeasible. On the other hand, the proposed learning algorithms are able to dynamically learn the best balance between multicast and D2D transmissions, and, to do so, need a reasonable learning time.

I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

Sir Arthur Conan Doyle

5 Incentives for D2D Offloading: Discussion and Modeling

5.1 Introduction

The effectiveness of D2D offloading strategies revolves around the participation of users as opportunistic forwarders. Incentives are then needed to stimulate participation and to reward users acting as data relays. Existing proposals assume that all seeders are, by default, also forwarders in the D2D domain. Such an assumption may lead to suboptimal results when forwarders must be rewarded for transmitting content on behalf of the infrastructure. In those scenarios, uncontrolled D2D communications may generate additional costs without necessarily bringing gains to data dissemination. *we introduce a clear separation between seeders and forwarders*, as shown in Fig. 5.1. Seeders receive content via the cellular infrastructure, but only nodes promoted as forwarders are allowed to transmit content opportunistically. The separation between seeders and forwarders provides operators with an additional degree of freedom. The balance between instantaneous cost and future benefits of *seeding* and *forwarding* decisions is strategic to data dissemination, given that available resources (bandwidth and rewards) are limited.

We investigate the following problem: *which fraction of seeders should be promoted as forwarders and when should this happen*? We answer this question using a mathematical support that leads to an optimal solution. Content diffusion in opportunistic networks is comparable to the spreading of a disease in a population. We model the dissemination process using a variant of the classic Susceptible-Infected-Recovered (SIR) epidemic model from Kermack-

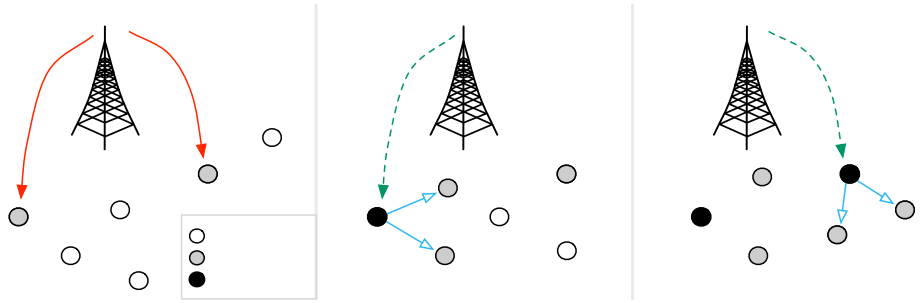


Figure 5.1: Offloading process: the infrastructure selects two nodes as content seeders (Fig. 5.1(a)), deciding that one of the seeders should be promoted as forwarder (Fig. 5.1(b)). Later on, the infrastructure estimates that it is worth promoting another node because the D2D transmissions are not enough to guarantee sufficient dissemination (Fig. 5.1(c)).

McKendrick [173]. Since operators strive to optimize the distribution cost, we translate the possible decisions they can take (injection and forwarding) into a cost function. We apply Pontryagin's Maximum Principle to minimize the cost function subject to the state-equations that govern the network. As far as we know, no existing works contemplate the difference between seeders and forwarders, failing thus to quantify the trade-off that exists between performance and cost in a more realistic system. How to select the best forwarders has been considered in the literature from the spatial re-use, throughput, and interference points of view, rather than from an incentive/economic consideration. Performance evaluation of truly opportunistic forwarding has been treated extensively, for example by using ordinary differential equations (ODEs) [104] or Markov chains [103]. Instead, our work proposes an extended model that couples opportunistic dissemination and infrastructure connectivity. Indeed, a central offloading coordinator controls the cellular injections and the promotion of users to the forwarding state to reach optimal data dissemination. In summary, the main contributions in this chapter are:

- Seeder-forwarder model. We propose a model for opportunistic offloading, where the network operator controls content injections and the amount of users that participate in data forwarding.

- Optimal control. We formulate opportunistic data offloading as an optimal control problem to minimize the cost of data dissemination in a hybrid scenario. The cost function trades off monetary and network resources consumed to reach a certain dissemination level and the user satisfaction.

- Cost functions. We prove how to solve the opportunistic offloading problem for different classes of cost functions. We show that under plausible cost functions, the

control of the injection is continuous, while promotions have an on-off behavior.

- Evaluation. We show the sensitivity of the optimal controls to different values of the contact rate and delay tolerance. We evaluate the optimality of our strategy against other heuristics. Finally, we confirm the benefit of the proposed seeder-forwarder model compared to the simplified two-state model currently employed in literature.

The remainder of the chapter is structured as follows. We present an overview of the seeder-forwarder model in Section 5.2. Cost-related issues in opportunistic offloading are discussed in Section 5.3. The optimal control problem is formulated and solved analytically in Section 5.4. Numerical results are presented in Section 5.5. Finally, Section 5.6 draws the conclusion and presents the future work.

5.2 System Overview

We formulate opportunistic offloading as an optimal control problem, modeling the evolution of the content dissemination using a variant of the classic SIR model. Content of interest may include software updates, geo-located information, traffic updates and targeted advertising. Public safety applications may also benefit from D2D communications.

In this model, some users request data and are referred to as *interested*. We consider that, initially, all the nodes are in the *interested* state. At this stage, the operator can only use cellular transmissions to reach a subset of the *interested* users. Nodes that receive the content enter the *seeder* state, still not playing any active role in data distribution. At this point, the coordinator can promote a fraction of them to the *forwarder* state to diffuse the content[*].

Cost

Incentives to reward user participation in data dissemination can be offered by using virtual credit schemes or discounts. Promotion to the forwarder state does not represent a cost in itself for operators, but enables users to be rewarded for the D2D content distribution. For now, we do not consider any additional cost related to overhead, signalization and maintenance of forwarder users, which is left as a future work. Injections performed using the cellular infrastructure consume resources and have a direct cost that depends on the resource availability in the network. Beside the cost for injection and rewards, the model considers indirect cost as well, because the fraction of uninfected (or unsatisfied) nodes at the end of the content lifetime depends on the strategy that has been adopted. We strive to devise a coherent injection/promotion strategy to minimize the aggregate cost of offloading.

[*] We adopted here a slight variation of the traditional nomenclature in the SIR model: "interested" users in the original SIR model are analogous to *interested* of our model. Similarly, "infective" and "seeder" nodes are named *forwarders* and *seeders*, respectively. We do not use the same names, as their roles are a bit different.

Table 5.1: List of parameters.

Parameter	Definition
$n_I(t)$	fraction of interested nodes at t
$n_S(t)$	fraction of seeder nodes at t
$n_F(t)$	fraction of forwarder nodes at t
$\lambda(t)$	contact rate at t
$u_I(t)$	direct injection rate at t
$u_S(t)$	promotion rate at t
$I_{max}(t)$	maximum injection rate at t
T	content lifetime
$\Phi(.)$	final payoff function for interested nodes at T
$f(.)$	instantaneous cost function for injection
$g(.)$	instantaneous cost function for promotion

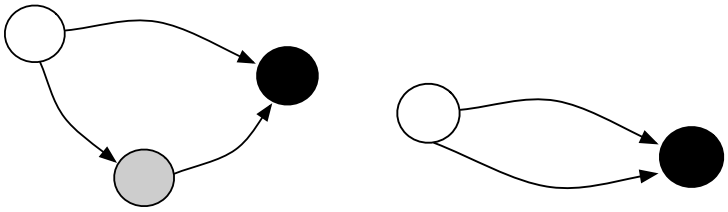
Network model

The system consists of N mobile nodes and a single content to be distributed by the mobile infrastructure to all the nodes within the lifetime T. Intermediary nodes can be used as opportunistic relay. Following the notation introduced above, nodes can be in the *interested*, *seeder*, or *forwarder* states. Their respective fractions during T are $n_I(t)$, $n_S(t)$, $n_F(t)$. At steady state, we have $n_I(t) + n_S(t) + n_F(t) = 1 \ \forall t \in [0, T]$, and we can always represent the system using only two of the above state. An interested node can receive the content whenever in contact with a forwarder neighbor, but not with a seeder. Table 5.1 provides a summary of the parameters used along the chapter (some of them are explained in the following sections).

Encounters and communication opportunities

In the real world, the system under observation can be described with discrete values (e.g., the number of present users, the number of cellular transmissions performed). For ease of modeling, we consider instead continuous values for the state and the controls. We assume that N is large and that encounters are homogeneous, i.e., nodes are equally likely to meet each other. Consistently with the literature, we use a *mean field* model that is accurate for a large population number. Opportunistic dissemination between mobile users can be regarded as the spread of infective disease – not surprisingly *epidemic* routing is a conventional forwarding strategy in opportunistic networks. As with a disease contagion in a population, content spreads from *forwarder* to *interested* nodes when such a pair enters in physical proximity.[†] The evolution of the states can be described by a system of ODEs along with a set of initial and terminal constraints. We consider the contact rate $\lambda(t)$ that rules the encounter

[†] Our system shares also similarities with peer-to-peer (P2P) networks, which features seeders forwarding traffic to client nodes.



(a) Seeders-forwarders **(b)** Two-state

Figure 5.2: State transition rates for the seeder-forwarder and two-state model. In the latter case, all the users that receive the content from the cellular channel are considered forwarders.

of any two nodes at time t. At each instant, we have $n_F(t)$ forwarders capable of meeting $n_I(t)$ interested nodes. As shown in Fig. 5.2a, interested nodes become forwarders with rate $\lambda(t)n_I(t)n_F(t)$.

Injections and Promotions

We consider a central offloading coordinator that manages the cellular injections and the promotion of seeders to the forwarder state. In our model, cellular injections increase the rate at which nodes leave the interested state for the seeder state. The intensity at which injections are performed at t is denoted by $u_I(t)$, which is a bounded Lebesgue integrable function with $0 \leq u_I(t) \leq 1 \, \forall t \in [0, T]$. Consequently, as shown in Fig. 5.2a, $u_I(t)n_I(t) \leq I_{max}(t)$ describes the rate of injected copies. The injection rate is bounded by $I_{max}(t)$, which is a measure of the instantaneous available load on the cellular network.

Seeder nodes carry the content but need to be promoted in order to contribute to data dissemination. Operators can promote only the necessary fraction of seeders. This is done via a control channel that binds users with the central coordinator. As a result, nodes shift to the infective state at intensity $u_S(t)$, a bounded Lebesgue integrable function with $0 \leq u_S(t) \leq 1 \, \forall t \in [0, T]$. This increases the fraction of nodes in the forwarder state by a rate $u_S(t)n_S(t)$.

Therefore, the following system of ODEs controls the evolution of the interested, seeders and forwarders in the system:

$$\frac{dn_I(t)}{dt} = -\lambda(t)n_I(t)n_F(t) - u_I(t)n_I(t), \tag{5.1a}$$

$$\frac{dn_S(t)}{dt} = u_I(t)n_I(t) - u_S(t)n_S(t), \tag{5.1b}$$

$$\frac{dn_F(t)}{dt} = \lambda(t)n_I(t)n_F(t) + u_S(t)n_S(t), \tag{5.1c}$$

with initial states $n_I(0) = i_0$, $n_S(0) = s_0$, and $n_F(0) = 1 - i_0 - s_0$. For the offloading problem under consideration, $i_0 = 1$, and $s_0 = 0$, since we consider all users to be in the interested state at the beginning of content diffusion.

The set of equations above are called state dynamics. They describe how the current states n_I, n_S, n_F change at time t as a reaction to the control signals u_I, u_S. Notice that $\frac{\partial n_I}{\partial t} + \frac{\partial n_S}{\partial t} + \frac{\partial n_F}{\partial t} = 0$, therefore the model can always be expressed using only two of the above equations.

5.3 The cost of offloading

The optimal offloading strategy consists in minimizing the number of interested nodes at time T, while implementing a cost-savvy injection/promotion campaign. If operators had no capacity or monetary limitations, then the optimal strategy would be to inject the maximum amount of data via the cellular channel. When capacity is limited, operators must improve their strategy from both the operational and budgetary points of view. In Eq. 5.2, we consider a cost function J that is general enough to grasp various types of cost incurred by operators:

$$J(T) = \underbrace{\Phi[n_I(T)]}_{\text{payoff}} + \int_0^T \underbrace{f[u_I(t)\, n_I(t)]}_{\text{injection}} + \underbrace{g[\lambda\, n_I(t)\, n_F(t)]}_{\text{reward}}\, dt. \tag{5.2}$$

where $\Phi[n_I(T)]$ is the final payoff, representing the cost incurred by the operator for not having satisfied the fraction $n_I(T)$ of users by the deadline. This can lead to a loss of earnings due to missed deliveries, or to extra costs in terms of final injections [117, 174]. $f[u_I(t)\, n_I(t)]$ captures the instantaneous cost in terms of network resources for the injections over the cellular channel. Finally, forwarders are rewarded with $g[\lambda\, n_I(t)\, n_F(t)]$, which represents reductions or virtual credits accorded to users each time they make an opportunistic transmission. The integral portrays the growing cost over time of these two latter terms. Note that the promotion control u_S does not appear inside the cost function. Promoting a node to the forwarder state does not directly generate a cost. However, the node will be able to transmit data opportunistically, possibly increasing the rewarding cost for the operator. For physical reasons, $\Phi(.)$, $f(.)$, and $g(.)$ should be any monotonically increasing and piecewise differentiable function, with $\Phi(0) = f(0) = g(0) = 0$ (the cost for doing nothing should always be zero).

At any time t, the offloading controller decides the value of control signals u_I and u_S. The decision is taken by assessing the fraction of nodes in each compartment (n_I, n_S and n_F), the remaining time before the deadline, and the contact pattern between nodes. The applied controls lead to two consequences: (i) a direct effect, which generates the instantaneous costs $f(.)$ and $g(.)$ for the operator, and (ii) an indirect effect, represented by the future change in states formalized by Eq. 5.1. The optimal offloading strategy requires the coordinator to plan

its injection and promotion strategies by *minimizing the cost for the operator while maximizing the rate of change of state variables*.

5.4 Optimal offloading formulation

We formulate the optimal control problem considering only two state variables n_I and n_S. This is possible because $n_F(t) = 1 - n_I(t) - n_S(t)$. The system can be controlled by the tuple $< u_I, u_S >$ belonging to the set of all the admissible controls $U = \{u_I, u_S\}$, where u_I, u_S are Lebesgue integrable with $u_I, u_S \in [0, 1]$. The idea is to minimize the cost function J subject to the state evolution constraints in Eq. 5.1:

$$\min_{u_I(t), u_S(t) \in U} J, \tag{5.3a}$$

subject to:

$$\frac{dn_I}{dt} = -\lambda(t) n_I(t)(1 - n_I(t) - n_S(t)) - u_I(t) n_I(t), \tag{5.3b}$$

$$\frac{dn_S}{dt} = u_I(t) n_I(t) - u_S(t) n_S(t), \tag{5.3c}$$

$$n_F(t) \geq 0, n_I(t) \geq 0, n_S(t) \geq 0,$$

$$n_I(t) + n_F(t) + n_S(t) = 1,$$

$$n_I(0) = i_0, n_S(0) = s_0, n_f(0) = 1 - i_0 - s_0. \tag{5.3d}$$

5.4.1 General solution

The existence of an optimal solution can be proved by applying the Filippov-Cesari theorems [175]. For instance, if the functions inside the integral in Eq. 5.2 are continuous, bounded, and convex in controls, with bounded derivatives, the control signals $u_I(t)$ and $u_S(t)$ take values in a closed set. Also, Eqs. B.6b and B.6c are linear in the controls. This guarantees the existence of an optimal solution. We apply Pontryagin's Maximum Principle [176] to solve the above problem and derive the optimal control (Theorem 3.4 in [177]). The conditions of Pontryagin's maximum principle reduce the computation of an optimal strategy to the solution of a boundary value problem for a system of differential equations.

Let the tuple $(n_I^*(.), n_S^*(.), u_I^*(.), u_S^*(.))$ be an optimal solution to the problem formalized in Eq. 5.3.[‡] There exist continuous and piecewise continuously differentiable adjoint functions $p_i(t)$ and $p_s(t)$ that maximize the present-value *Hamiltonian* function H.

[‡]Throughout the chapter, variables with the star superscript (e.g., $u_I^*(t)$) represent the value at the optimum.

For the sake of ease of mathematical manipulation, we transform the problem into a *maximization problem* by multiplying the Hamiltonian by -1. We also remove the dependence from time whenever possible, in order to make reading also easier:

$$\begin{aligned} H(n_{I,S}, u_{I,S}, p_{i,s}, t) = & - f[u_I\, n_I] \\ & - g[\lambda\, n_I\, (1 - n_I - n_S)] \\ & + p_i[-\lambda n_I(1 - n_I - n_S) - u_I n_I] \\ & + p_s[u_I n_I - u_S n_S]. \end{aligned} \tag{5.4}$$

The Hamiltonian function, in analogy with a corresponding concept occurring in traditional mechanics, balances the rate of change of states and the cost incurred by operators. Indeed, the Hamiltonian is a generalized profit rate that includes both direct and indirect effects, and has to be maximized at each instant. The *weights* for the state variables are given by the adjoint functions p_i and p_s, which represent the marginal increase of H due to an increment in the state. Consequently, the adjoint equations p_i and p_s evaluated at the optimum are:

$$p_i^*(t) = -\frac{\partial H(.)}{\partial n_I}\bigg|_{n^*_{I,S}, u^*_{I,S}, p^*_{i,s}} = \tag{5.5a}$$

$$= \frac{\partial f(.)}{\partial n_I} + \frac{\partial g(.)}{\partial n_I} - p_i\,[\lambda\,(2n_I - 1 + n_S) - u_I] - p_s\,u_I,$$

$$p_s^*(t) = -\frac{\partial H(.)}{\partial n_S}\bigg|_{n^*_{I,S}, u^*_{I,S}, p^*_{i,s}} = \tag{5.5b}$$

$$= \frac{\partial g(.)}{\partial n_S} - p_i \lambda n_I + p_s u_S.$$

With transversality conditions

$$p_i(T) = \frac{\partial \Phi(n_I^*(T), T)}{\partial n_I}, \tag{5.6a}$$

$$p_s(T) = \frac{\partial \Phi(n_I^*(T), T)}{\partial n_S} = 0. \tag{5.6b}$$

According to the maximum principle (Theorem 3.4 in [177]), there exist optimal controls, a tuple $< u_I^*, u_S^* > \in U$ of continuous and piecewise continuously differentiable functions, and their corresponding solutions n_I^*, n_S^* that maximize the Hamiltonian H satisfying

Eqs. 5.5 and 5.6:

$$u^*_{I,S}(t) \in \underset{u_{I,S} \in U}{\arg\max}\, H(n_{I,S}, u_{I,S}, p_{i,s}, t). \tag{5.7}$$

This canonical system composed of four coupled ODEs and the transversality conditions determines a boundary value problem (BVP) that can be solved numerically.

5.4.2 Example of application

We give an example of the benefit of the seeder-forwarder model by solving the optimization problem for a class of cost functions $\Phi(.)$, $f(.)$, and $g(.)$. Thanks to the flexibility of the model, the cost functions can be replaced at will, to take into account the specificities of certain types of network and their operating costs.

We consider an exponential function for the final payoff $\Phi(x) = e^x - 1$, a power-law function for the direct injections $f(x) = bx^\alpha$, with $\alpha \geq 2$, and a linear function $g(x) = cx$ to reward forwarders. Recall just that Φ, f and g should be monotonically increasing functions that start at zero. The motivation for the choice of these functions is straightforward and follows the discussion in Section 5.3. The final payoff function $\Phi(x)$ starts at zero and then rapidly increases to model the cost for missing the delivery of the content by the deadline T. $f(x)$ represents the cost for injecting data on the cellular channel during the content lifetime. The power-law accounts for the fact that the more simultaneous cellular data transmissions, the less efficient they are in terms of radio resources at the cellular base station. The power-law coefficient α depends on the considered network and on its overall congestion conditions. The cost function $g(x)$ is linear, since the reward offered to forwarders for each opportunistic transmission they perform is fixed.

By substituting the cost functions $\Phi(.)$, $f(.)$, and $g(.)$ in Eqs. 5.4, 5.5, and 5.6, the Hamiltonian, the adjoint functions, and the transversality conditions become:

$$\begin{aligned}
H &= -b(n_I u_I)^\alpha - c(\lambda\, n_I\, (1 - n_I - n_S)) + \\
&\quad + p_i[-\lambda n_I(1 - n_I - n_S) - u_I n_I] + \\
&\quad + p_s[u_I n_I - u_S n_S], && \text{(5.8a)} \\
p^*_i(t) &= b\,\alpha\, n_I^{\alpha-1} u_I^\alpha + c\,\lambda\,(1 - 2n_I - n_S) \\
&\quad - p_i\,[2\lambda n_I - \lambda + \lambda n_S - u_I] - p_s\, u_I, && \text{(5.8b)} \\
p^*_s(t) &= -c\,\lambda n_I - p_i \lambda n_I + p_s u_S, && \text{(5.8c)} \\
p_i(T) &= e^{n^*_I(T)}, && \text{(5.8d)} \\
p_s(T) &= 0. && \text{(5.8e)}
\end{aligned}$$

Injections

Given that $f(x)$ is strictly convex, we can extract $u_I(t)$ using the Hamiltonian maximization condition ($\frac{\partial H}{\partial u_I} = 0$ evaluated at the optimum), along with the restriction on the maximum injection rate ($n_I(t)u_I(t) \leq I_{max}(t) \forall t$):

$$u_I^*(t) = \begin{cases} 0, & \text{if } \psi(t) < 0, \\ \frac{\psi(t)}{n_I(t)}, & \text{if } 0 \leq \psi(t) \leq I_{max}(t), \\ \frac{I_{max}(t)}{n_I(t)}, & \text{if } \psi(t) \geq I_{max}(t). \end{cases} \tag{5.9}$$

Equivalently, we have that $u_I^*(t) = \frac{\min[\max[\psi(t),0],I_{max}]}{n_I(t)}$, where $\psi(t) = \sqrt[\alpha-1]{\frac{p_i^* - p_s^*}{-\alpha b}}$.

Promotions

In the case of $u_S(t)$, since the Hamiltonian is linear in the control variable u_S, the maximization condition $\frac{\partial H}{\partial u_S} = 0$ is trivially satisfied and independent of u_S. The control in this case is called *singular* (Definition 3.40 in [177]) with a *bang-bang* solution, i.e., a control that switches discontinuously between one extreme to the other.

The Hamiltonian maximization condition cannot help us determine the optimal control. It follows that we have to use another method to find when the switching points happen. To do so, we can rewrite the Hamiltonian in Eq. 5.8a as:

$$\begin{aligned} H = &- b(n_I u_I)^\alpha - c(\lambda\, n_I\, (1 - n_I - n_S)) + \\ &+ p_i[-\lambda n_I(1 - n_I - n_S) - u_I n_I] + \\ &+ p_s[u_I n_I] - u_S[p_s\, n_S]. \end{aligned} \tag{5.10}$$

We define the switching function $\sigma = (p_s n_S)$. From Eq. 5.7, it is clear that, to maximize the Hamiltonian, u_S should take its maximum value when $\sigma < 0$ (and maximum value when $\sigma > 0$). Since, we have by construction $u_S \in [0, 1]$, it follows that:

$$u_S^*(t) = \begin{cases} 0, & \text{if } \sigma > 0, \\ 1, & \text{if } \sigma < 0. \end{cases} \tag{5.11}$$

In order to be able to retrieve the evolution of the state and adjoint variables, we have to solve a system of coupled differential equations (respectively Eqs. B.6b, B.6c, 5.8b, 5.8c, 5.8d, and 5.8e), with a mix of initial and final conditions (boundary value). We solved it numerically by using the shooting method from the R package *bvpSolve* to compute the evolution of the state and adjoint variables as well as the optimal control (see next section) [178].

5.5 Numerical results

We cover the problem of identifying the best injection and rewarding strategy that an operator should put in place to reach the optimal offloading performance. For this purpose, we include in this section numerical data for the simulations that we performed using R. One of the main strengths of the proposed model is that every parameter can be easily tuned.

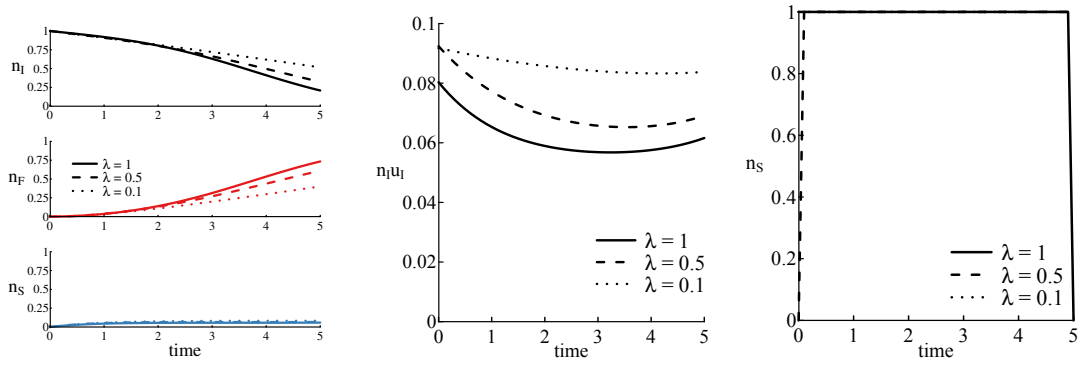
First, we conduct a sensitivity analysis of the value of the key parameters to understand their implications in the offloading strategy. We notice that performance strongly depends on both the content deadline and the contact rate of nodes. Then, we explore the scenarios where the seeder-forwarder model brings advantages over a more classic two-state model. Finally, we address some implementation issues. We compare the optimal strategy with several heuristic strategies, investigating under which conditions and limits they can be adopted.

5.5.1 Symmetric injections, on-off promotions

We investigate "when" the coordinator should inject copies of the content and promote seeders as forwarders. The goal is to understand theoretically the implications of the various parameters of our model. Figs. 5.3 and 5.4 display the time evolution of states and control variables. We look at what happens for different deadlines and contact rates.

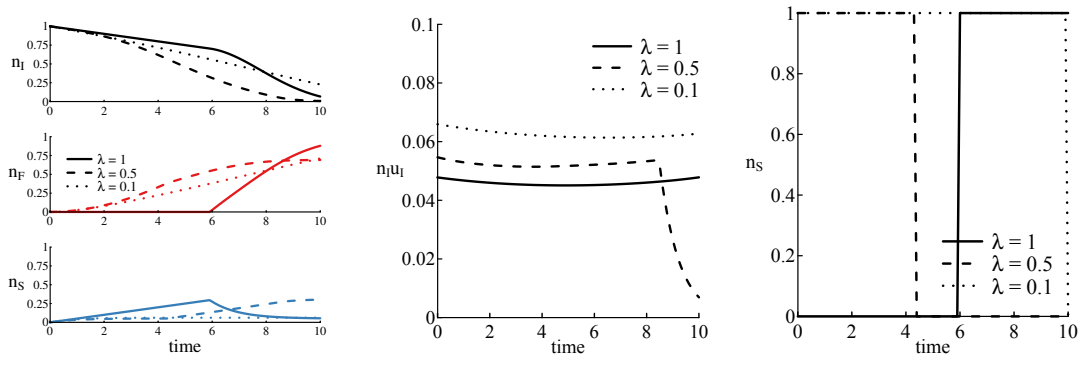
By comparing Figs. 5.3 and 5.4, we discover that the impact of promotion is stronger for shorter deadlines. From this, we conclude that the deadline strongly influences the rewarding strategy. Short deadlines (e.g., $T = 5$) are not sufficient to yield complete dissemination under the provided cost-function. Instead, for $T = 10$ we obtain complete data delivery (at least for the best contact rates). Intuitively, when content dissemination is incomplete by the deadline, the final payoff $\Phi(n_I(T))$ takes a large part of the cost functional J. This behavior confirms a well-known phenomenon in the opportunistic literature: an increased delivery delay improves the fraction of nodes that receive the content, since opportunistic forwarders have more time to meet and exchange data. The effect is particularly evident if we compare the curves with the same contact rate. The added dissemination time allows less effort to control injections and promotions, thus lower costs for cellular operators. On the other hand, increased delivery times could negatively affect the user satisfaction.

Besides distribution delay, lower contact rates also entail stronger injection and promotion efforts. From Figs. 5.3b and 5.4b, we observe that the injection control is stronger at the beginning and at the end of the dissemination period, following a nearly symmetric path. The symmetric pattern depends on the evolution of interested users in Eq. 5.1a, which relies on the interactions between n_I and n_F. Therefore, the injection rate is higher when a few nodes are in the forwarder or the interested states – respectively at the beginning and at the end of the dissemination period. Injections help fix the slow start and the slow convergence time of opportunistic dissemination. Wang et al. previously pointed out the symmetric trend of the injection control, although for a simplified model [179].



(a) state evolution **(b)** injections **(c)** promotions

Figure 5.3: Optimal offloading for different contact rates λ. $T = 5s$. Other parameters: $I_{max} = 0.1, \alpha = 2$, $b = 10, c = 1, i_0 = 1, s_0 = 0$.



(a) state evolution **(b)** injections **(c)** promotions

Figure 5.4: Optimal offloading for different contact rates λ. $T = 10s$. Other parameters: $I_{max} = 0.1, \alpha = 2$, $b = 10, c = 1, i_0 = 1, s_0 = 0$.

Promotions (visible in Fig. 5.3c and 5.4c) follow an interesting pattern. When $T = 5$, the control is always at its maximum. In this case, the shorter deadline is the main responsible for the poor content dissemination. In effect, even with an extreme promotion strategy (always promote), dissemination levels never reach 100% at the deadline. Longer deadlines, instead, allow devising better promotion strategies. For instance, when $T = 10$, promotions show three distinct patterns that depend on the contact rates. For $\lambda = 0.1$, the control is always at its maximum for all the dissemination duration to counter the low contact rate. $\lambda = 0.5$ presents an on-off behavior, with promotions that stop when the amount of forwarders reaches significant levels (in order to self-sustain without costing too much to operators). Finally, for $\lambda = 1$, the promotion control is activated only after half of the dissemination period. This collides with the desire to attain the widest possible dissemination of the content. Although at first sight this might seem counter-intuitive, we must remember

that, in the model, operators pay a small fee for each opportunistic transmission performed by users. Under higher contact rates, opportunistic dissemination has to be limited in order to save monetary resources.

As anticipated, control signal u_S takes the form of a *bang-bang* control with exactly one on-off switch. Recall that injections performed when $u_S(t) = 0$ serve only to satisfy the fraction of users that will likely not receive the content by the deadline, without further improving the dissemination (because these nodes do not become forwarders). Moreover, we point out that the optimal strategy does consider moments where no additional forwarders are needed ($u_S(t) = 0$). This strengthens the idea that separating forwarders and seeders is beneficial from a cost-benefit point of view.

5.5.2 Rewarding forwarders: When is it worth?

We investigate in which cases it is worth separating seeders and forwarders from an operator's point of view. Including the seeder state in the picture is motivated by the fact that not all users carrying the content may be required to forward data. This may depend on a mix of factors, such as the contact pattern of users or the delay tolerance of the content. Eventually, controlling the forwarder state of nodes becomes essential when the operator wants to reward user participation. Allowing an additional state can also be useful to understand the implications of a particular reward strategy aimed to involve users in cooperation for data dissemination. As many works suggest, offering some kind of incentive, i.e., discounts or virtual credits, motivates user participation in opportunistic offloading [180]. However, current models in the literature consider all nodes storing the data as potential forwarders as in the model depicted in Fig. 5.2b.

Separating seeder and forwarder nodes is advantageous for operators if compared to the classic two-state model. We plot in Fig. 5.5 the evolution of the cost function J divided by its three main components $\Phi(T)$, $F(T)$, and $G(T)$ for the optimal strategy. From Eq. 5.2, we have:

$$F(T) = \int_0^T f[u_I(t)\, n_I(t)]dt, \tag{5.12}$$

$$G(T) = \int_0^T g[\lambda\, n_I(t)\, n_F(t)]dt. \tag{5.13}$$

$\Phi(T)$ is the final payoff value, due to nodes that have not received the content by T, $F(T)$ is the total cost due to injection, and $G(T)$ is the total cost due to rewarding forwarder users.

From Fig. 5.5, we can appreciate that as the deadline increases, the seeder-forwarder model improves its performance compared to the two-state model. In our evaluations, for $T > 5$, the number of uninfected nodes at the deadline decreases steadily, reducing the relative weight

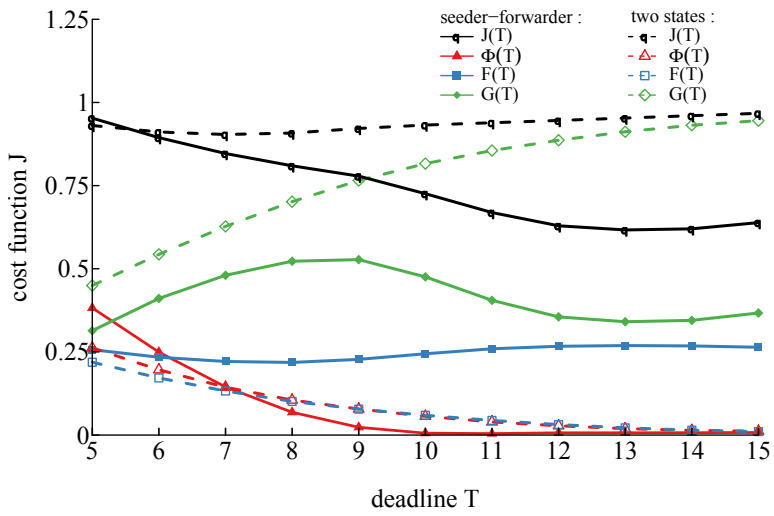


Figure 5.5: Cost functional J and its main components for the optimal strategy using two offloading models (seeder-forwarder and two-state), varying the deadline T. Other parameters: $\lambda = 0.5\, I_{max} = 0.1, \alpha = 2, b = 10, c = 1$, $i_0 = 1, s_0 = 0$. Note that for $T \leq 5$ (not plotted) the cost functional is dominated by the final payoff $\Phi(T)$ due to missed data deliveries.

of $\Phi(T)$ on the overall cost. In this context, the rewarding cost $G(T)$ takes the larger part of J, accounting for the cost to reward forwarders. The cost for rewarding users increases linearly for the two-state model as the deadline stretches, making up nearly the entirety of the cost functional. This confirms that an uncontrolled number of forwarders can interfere with the will of operators to cut operational expenses. Instead, *a separation between seeders and forwarders offers improved flexibility in the control of the offloading evolution, bringing clear benefits in terms of distribution costs*. Note that for short deadlines (for $T \leq 5$ in the example), all connection opportunities should be used as captured by the two-state model. Indeed, the cost-functional is dominated by the final payoff $\Phi(T)$, whose value depends on missed data deliveries. The seeder-forwarder model introduces an additional transition delay from seeder to forwarding state (the ODE formulation requires a non-null time to transit to a state) so its benefits come into play in the non-trivial situations of content with larger deadline requirements.

5.5.3 Implementation considerations

We first investigate how intuitive heuristics not requiring any optimization framework perform compared to the optimal strategy, and what lessons can be learned having the knowledge of the optimal injection and promotion controls. Finally, we consider the impact of realistic values of the contact rate.

We compare the optimal strategy against three other heuristics. The first heuristic, named *initial control*, mimics an operator wanting to rely only on an initial subset of forwarders. These forwarders are the only way to distribute content until the deadline is reached. This

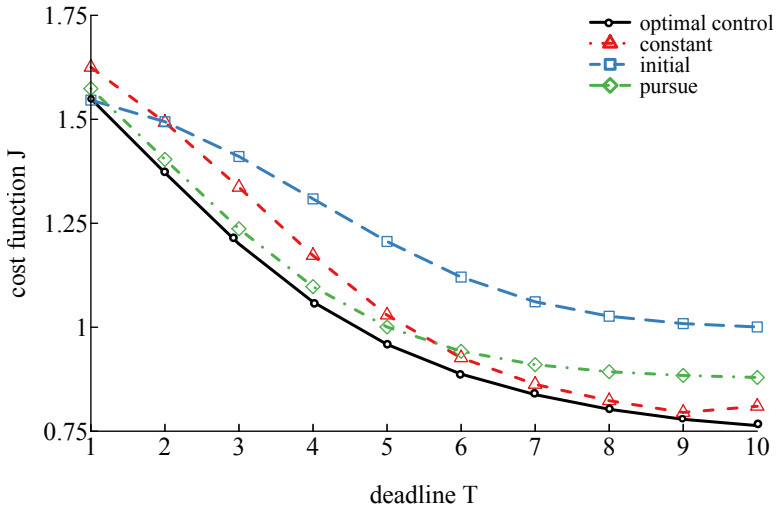


Figure 5.6: Cost functional J for different control strategies varying the deadline T. Other parameters: $\lambda = 0.5$ $I_{max} = 0.05, \alpha = 2, b = 10, c = 1, i_0 = 1, s_0 = 0$.

strategy relies on an initial injection at the rate $I_{\max}$, without performing any further injections. The second strategy, named *constant control*, steadily injects at a fixed rate of $\frac{I_{\max}}{2}$. These two strategies are static and do not require any knowledge of how the dissemination evolves. In both cases, the promotion control $u_R(t)$ is fixed at 1 for all the dissemination delay. Finally, we consider a more dynamic strategy, named *pursue control*, where both injection and promotion controls follow the evolution of the interested nodes $n_I(t)$. In this case, the control is strong at the beginning of the dissemination, gradually descending as the time goes by, following $n_I(t)$. The rationale behind this choice is the fact that copies with a large dissemination time are more effective in content dissemination. We compare these strategies in terms of the cost function introduced in Eq. 5.2, varying the deadline T and the contact rate λ.

FFig. 5.6 displays the cost functional J by varying the deadline. As expected, the functional J for the optimal control is always smaller than all the other heuristic strategies. However, for shorter deadlines ($T < 6$) the *pursue* strategy results very close to the optimal. This is the first lesson we can draw: with shorter deadlines, a control that follows the rate of interested nodes comes close to the optimal. As the deadline increases, the efficiency of the *pursue* strategy decreases. On the other hand, we note that the *constant* strategy approaches the optimum for larger deadlines. Indeed, the steady injection profile at rate $\frac{I_{max}}{2}$ is very similar to the one in the optimal strategy (depicted in Fig. 5.4b). Promotions are not adapted, concurring to increase the overall cost. Lastly, relying only on an initial set of forwarders, without any additional injection during the dissemination duration (such as in the *initial* strategy), brings considerable efficiency drops. This happens because the set of initial seeders (which is inherently limited by $I_{\max}$) cannot cover the entire network in T. Fig. 5.7 shows the trend of the cost functional J varying the contact rate λ. In general, the relative performance of the

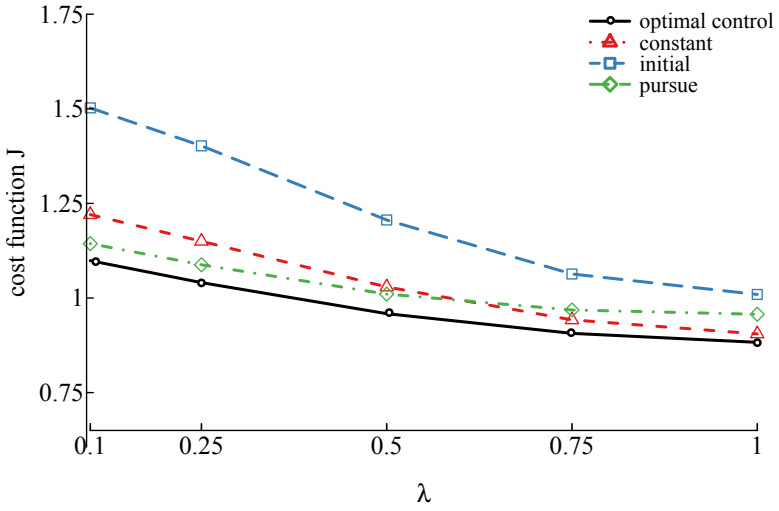


Figure 5.7: Cost functional J for different control strategies varying the contact rate λ. Other parameters: $T = 5s$, $I_{max} = 0.05, \alpha = 2, b = 10, c = 1, i_0 = 1, s_0 = 0$.

heuristics is the same as in the previous case.

Fig. 5.7 outlines the importance of the contact rate in the performance of the offloading strategy. λ is at the base of the opportunistic diffusion between mobile users. However, the contact rate depends on the mobility pattern of users, and can vary in time. λ may also include the uncertainties induced by the wireless channel and the movement of nodes. Operators should estimate the value of λ in order to adapt the optimal solution to current network conditions. In this context, offloading architectures that employ a feedback mechanism can prove very useful [117, 174].

5.6 Conclusion and outlook

In this chapter, we proposed a novel analytical framework for opportunistic offloading that captures the differences between seeders and forwarders. With our approach, mobile operators are able to finely control the dissemination evolution through external controls such as infrastructure injections and promotion of seeders as forwarders. Optimal strategies trade off the cost in terms of network and monetary resources with the total dissemination rate. We applied then the Pontryagin's Maximum Principle to devise an optimal offloading strategy that minimizes the distribution costs for the operator. One of the main strengths of the model is that every parameter can be easily tuned. We demonstrated its sensitivity to different values of the contact rate and delay-tolerance, and evaluated the advantages of the proposed model over the simple interested-forwarder model. In particular, we showed that when we have enough time flexibility, introducing the separation between seeders and forwarders is strongly beneficial for the cellular operator.

Besides evaluating the optimal strategy in case of time-varying parameters, we believe future developments can leverage the techniques presented in this chapter to handle a more general case of stochastic diffusion processes following a Markov decision model. In this case, the evolution of the diffusion evolves following stochastic values, and the applied control depends on the observation of the system. The epidemic diffusion model can be extended taking into account forwarders that stop sharing content due to battery or storage constraints. Promotions to the forwarder state can bring additional costs for state overhead and maintenance. Finally, a birth-death process can be included to represent the arrival and departure of the users in the interest area.

6

Conclusion & Perspectives

Global mobile data usage witnessed a spectacular increase in recent years driven by the boom in the smart mobile devices market. Future forecast expects the growth to continue at the same pace, calling for efficient strategies to cope with this surge. As of today, operators are already under heavy pressure, attempting to accommodate such an unprecedented amount of mobile traffic on their networks. Therefore, they must intervene with major investments to scale their access networks. Nevertheless, expenses to buy more licensed band or build more base stations are very high. Unfortunately, the increase in network capacity brought by these methods will hardly keep up with traffic growth.

Mobile data offloading represents an attractive solution to relieve the load on the mobile network infrastructure at small cost, benefiting from the existence complementary technologies. Offloading has the potential to solve the longstanding RAN overloading problem, by shifting data over alternative and less congested networks. The discussion provided in this thesis strongly advocates the use of alternative mobile access networks for offloading purposes. In particular, we focused on the "newest" type of data offloading, based on direct data exchange among users using device-to-device (D2D) communications. The idea is to benefit from the increased density of mobile users and the delay tolerance of a number of content types to shift a portion of the traffic from the primary (cellular) channel to an alternative D2D channel.

6.1 Summary of Contributions

6.1.1 Survey on Data Offloading Techniques in Cellular Networks

The literature lacked a comprehensive survey on mobile data offloading. We classified current offloading approaches based on their requirements in terms of delivery guarantee, presenting the technical aspects and the state of the art for two main approaches. The former is more mature and proposes a tight integration between the cellular RAN and a complementary access network, allowing for real-time data offloading. The latter, still experimental, exploits the delay tolerance of some types of data to optimize their delivery. As the outcome of this literature survey, we identified some common functional blocks, proposing a general high-level architecture valid for any mobile data offloading system. We investigated open research and implementation challenges and the existing alternatives to mitigate the cellular overload problem. We believe that our survey work highlights current research directions and trends in the data offloading field, and can help trigger further research activities in the area.

6.1.2 Large Scale Mobile Opportunistic Data Offloading

We highlighted the stepwise behavior of the epidemic diffusion in opportunistic network, demonstrating that it depends on the dynamic clustering of nodes. To obtain efficient offloading in large scale scenarios, we proposed DROiD, a re-injection based offloading scheme that adapts to the dissemination evolution improving the distribution of popular contents. DROiD tracks the evolution of content diffusion through user-sent acknowledgments on the cellular channel. DROiD offers better offloading performance than other state of the art strategies. DROiD perceives when the evolution of the content diffusion lags, reacting in advance with respect to traditional strategies that consider only the actual infection rate. Simulations were carried out to confirm that the proposed strategy consistently improves existing offloading systems, performing sometimes better than an oracle that has the real-time full picture of the ad hoc connectivity of the entire network. We also evaluated our system against more conventional AP-based systems, in addition to the impact that the implementation of energy-saving strategies at terminal-level has on the overall efficiency and fairness. The easy-to-implement heuristic for content re-injection proposed by DROiD helps operators save large amount of data traffic, allowing them to better manage their network resources.

6.1.3 Offloading LTE Multicast with D2D Communications

In this chapter, we first explored the shortcomings of existing multicast implementations in cellular networks. Multicast alone targets similar use-cases as offloading, and in line of principle is an advantageous strategy to distribute popular data into a cellular network. However, the user with the worst channel quality inside the multicast group determines performance. To counter the inefficiencies of cellular multicast, we proposed a framework that exploits D2D capabilities at UEs. The proper balance of multicast and D2D transmissions is achieved

by using a multi-armed bandit learning strategy. We proposed and evaluated two different algorithms under variable multicast group size and reception deadline. Simulation results prove that D2D communications allow increasing the multicast transmission rate, saving resources and improving the overall cell throughput. At the same time, the analysis demonstrates that both algorithms have a reasonable convergence time. D2D communications are an effective means of offloading multicast traffic in LTE networks, improving resource utilization at the eNB and increasing the available cell throughput.

6.1.4 Incentives for D2D Offloading

Incentives are a fundamental issue in D2D offloading. We proposed a novel analytical framework for D2D offloading that captures the differences between seeders and forwarders. This approach allows mobile operators to finely control the dissemination evolution through external controls such as infrastructure injections and promotion of seeders as forwarders. Optimal strategies trade off the cost in terms of network and monetary resources with the total dissemination rate. After formalizing the diffusion and the cost model, we applied the Pontryagin's Maximum Principle to devise an optimal offloading strategy that minimizes the distribution costs for the operator. One of the main strengths of the model is that every parameter can be tuned easily. The solution is evaluated numerically for a sample cost-function. We demonstrated its sensitivity to different values of the contact rate and delay-tolerance, and evaluated the advantages of the proposed model over the simple interested-forwarder model. In particular, when we have enough time flexibility, introducing the separation between seeders and forwarders help cellular operators optimize the cost related to the offloading process.

6.2 Open Challenges and Perspectives

Mobile data offloading remains a new and very hot topic, frequently identified as one of the enablers of next-generation mobile networks. Future research directions are manifold. Effective offloading systems require a tighter integration within the 3GPP and the wireless broadband infrastructures. Additional features still need to be developed to handle mobility of users, distributed trust, session continuity, and optimized scheduling policies. Offloading strategies may take advantage both of the AP connectivity and D2D communication opportunities. An unified architecture requires reconsidering existing wireless network paradigms. Therefore, future cellular architectures should intelligently support the distribution of heterogeneous classes of services, including real-time and delay-tolerant flows, to cope with an overall traffic increase of several orders of magnitude. A fine comprehension of data traffic and mobility patterns of nodes is required. It is critical to understand which types of traffic can be safely diverted on complementary data channels and which cannot, based on their delivery requirements. Additionally, fundamental research should focus on how nodes move and meet, creating communication opportunities in a fine-grained fashion.

Besides research challenges, the implementation of offloading strategies results in a variety of practical challenges. Academia and industry must tackle such challenges in order to make offloading a viable answer to the mobile data overload problem. To date, both technical and adoption-related challenges complicate the widespread introduction of offloading. The foremost technical challenge is related to the lack of a widely accepted mechanism to handle transparently several flows in parallel on different interfaces (nor protocols resilient to link failures, communication disruptions, and capable of handling substantial reception delays). As pointed out throughout the dissertation, various mechanisms have been proposed, but there is not yet a consensus on a de-facto standard. From the user perspective, a major concern comes from the dramatic battery drain of multiple wireless interfaces simultaneously turned on, even in idle mode. As of today, this combined use will seriously reduce the battery life of mobile devices. Possible solutions may be the design of low-powered network interfaces or the implementation of energy saving policies (a sort of duty cycle to switch on and off network interfaces), although privacy concerns prevent network operators to force a device to turn on and off a network interface.

Regarding user adoption, we should not forget that user collaboration, especially in the opportunistic approach, is essential for any offloading strategy. In order to make offloading feasible, end-users must accept to share some resources (battery, storage space, etc.), and their wireless interface should be turned on. The central question here is how to motivate users to participate. Mobile operators should propose a business concept for rewarding their customers, to make offloading attractive and fully functional at the same time with user participation. Additional issues lie on the security and privacy plan of users employing mobile-to-mobile transmissions. Users rarely accept anyone stranger to access data stored on their devices. Further challenges include the development of an infrastructure to ensure distributed trust and security to terminals involved in the offloading process.

A key question concerns the role of the network provider in the offloading process. In our work, we proposed network-centered offloading strategies, where the cellular network controls offloading. However, the debate is still open. Should the operator drive carefully the offloading process, or are end-nodes sufficiently autonomous to decide for themselves the best offloading strategy? In other words, future implementations should clarify how much the offloading process will be user-driven or operator-driven. Both strategies present advantages and disadvantages. An operator may have a better view of the overall network, while a user may have only a local and obviously partial view. On the other hand, an operator-driven offloading strategy may tend to give priority to a certain type of traffic or class of users, while a distributed offloading strategy may result more fair. The debate on these issues is still very open, and more research is needed along the lines introduced in this thesis.

A

List of Publications

Under review

- F. Rebecchi, M. Dias de Amorim, and V. Conan, "Circumventing Plateaux in Cellular Data Offloading using Adaptive Content Reinjection," under major revision *IEEE Transactions on Network and Service Management*.
- F. Rebecchi, L. Valerio, R. Bruno, V. Conan, M. Dias de Amorim, and A. Passarella, "A Multi-Armed Bandit Resource Allocation Scheme for D2D-aided Cellular Multicast," submitted to *Elsevier Computer Communications*.
- F. Rebecchi, M. Dias de Amorim, and V. Conan, "Seeders vs. Forwarders: Optimal Control of D2D Offloading under Rewarding Conditions," submitted to *IEEE MASS 2015*.

Published

- F. Rebecchi, M. Dias de Amorim, and V. Conan, "The Cost of Being Altruistic: Optimal D2D Offloading under Rewarding Conditions," In *Algotel 2015*, Beaune, France, June 2015.
- F. Benbadis, F. Rebecchi, F. Cosnier, M. Sammarco, M. Dias de Amorim, and V. Conan, "Demo: D2D Rescue of Overloaded Cellular Channels," In *ACM MobiSys*, Florence, Italy, May 2015.
- F. Rebecchi, M. Dias de Amorim, V. Conan, A. Passarella, R. Bruno and M. Conti, "Data Offloading Techniques in Cellular Networks: A Survey," In *IEEE Communications Surveys & Tutorials*, 2015.

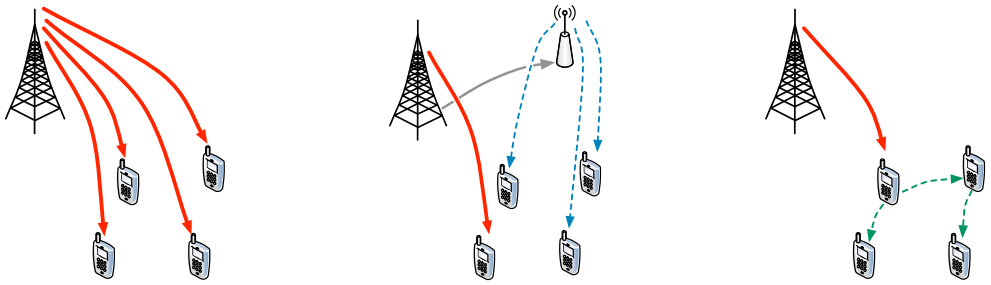
- F. Rebecchi, M. Dias de Amorim, and V. Conan, "Flooding Data in a Cell: Is Cellular Multicast Better than Device-to-Device Communications?," In *Ninth ACM MobiCom Workshop on Challenged Networks (CHANTS)*, Maui, HI, September 2014.
- F. Benbadis, F. Rebecchi, F. Cosnier, M. Sammarco, M. Dias de Amorim, and V. Conan, "Demo: Opportunistic Communications to Alleviate Cellular Infrastructures: the FP7-MOTO Approach," In *Ninth ACM MobiCom Workshop on Challenged Networks (CHANTS)*, Maui, HI, September 2014.
- F. Rebecchi, M. Dias de Amorim, and V. Conan, "DROiD: Adapting to Individual Mobility Pays Off in Mobile Data Offloading," In *IFIP Networking*, Trondheim, Norway, June 2014.
- F. Rebecchi, M. Dias de Amorim, and V. Conan, 'Adaptive Mobile Traffic Offloading," In *Algotel 2014*, Le-Bois-Plage-en-Ré, France.

B

Résumé de la thèse en français

B.1 Introduction

Les réseaux mobiles sont une partie intégrante de notre vie quotidienne. Leurs capacités permettent des applications qui auraient été inconcevables il y a seulement dix ans. Poussé par la popularité croissante des appareils mobiles intelligents et l'introduction de plans de données abordables par les opérateurs cellulaires, le trafic mobile mondial est en plein essor. Des applications mobiles riches en données, telles que les flux audio et vidéo, les réseaux sociaux, ou les services basés sur le cloud, sont de plus en plus populaires parmi les utilisateurs. Durant le laps de temps qu'aura pris la rédaction de cette thèse (2012 - 2015), le trafic de données mobile a été multiplié par trois, et devrait augmenter de près de dix fois d'ici à 2019. Il est également prévu que les deux tiers de ce trafic soient liés à la vidéo (avec ou sans exigences de temps réel) en 2017. Les réseaux cellulaires subissent une forte pression en essayant de faire face à cette surcharge sans précédent de données. Accueillir cette croissance nécessite d'importants investissements à la fois dans le réseau d'accès radio (RAN) et dans le cœur du réseau. Les utilisateurs doivent partager les mêmes ressources limitées sans fil, ce qui pose des problèmes de capacité. Des progrès considérables sont constamment faits à la couche physique pour augmenter le débit, mais cela n'est ni suffisant, ni rentable pour faire face à l'augmentation de la demande. Ce problème concerne principalement les opérateurs mobiles, car ils doivent veiller à ce que le client soit satisfait, ainsi qu'à la rentabilité des entreprises, compte tenu de la tendance vers des modèles d'affaires à forfait. En d'autres termes, l'augmentation exponentielle du trafic circulant dans leur RAN ne génère pas suffisamment de revenus supplémentaires à allouer à d'autres améliorations de RAN. Cela crée ce que Molleryd et d'autres. appellent l'*écart de revenus*.



(a) Infrastructure only.

(b) AP-based offloading.

(c) Terminal-to-terminal offloading.

Figure B.1: Les deux principales approches de délestage comparés à (i) un système d'infrastructure traditionnelle, (ii) délestage à travers d'un point d'accès sans fil, (iii) délestage grâce aux transmissions de dispositif à dispositif (D2D).

B.1.1 Le délestage: une solution possible

Ces circonstances ont favorisé l'intérêt des méthodes alternatives pour atténuer l'affluence sur le réseau cellulaire. Dans cette thèse, nous portons notre attention vers l'une de ces solutions, qui a récemment suscité l'intérêt de la communauté de la recherche : le délestage de données mobiles. Une approche intuitive qui consiste à tirer parti de la bande passante inutilisée à travers différentes technologies sans fil. Dans ces conditions, nous définirons le délestage comme *l'utilisation d'une technologie sans fil pour transférer des données ciblées, à l'origine, pour circuler à travers un réseau cellulaire, afin d'améliorer certains indicateurs clés de performance*. Outre l'avantage évident de soulager la charge du réseau d'infrastructures, déplacer les données sur une technologie sans fil complémentaire conduit à un certain nombre d'autres améliorations, y compris l'augmentation du débit, la réduction du temps de réception du contenu, l'extension de la couverture du réseau, l'augmentation de la disponibilité du réseau et une meilleure efficacité énergétique. Ces améliorations concernent, à la fois, les opérateurs mobiles et leurs utilisateurs. Par conséquent, le délestage est souvent décrit dans la littérature comme une stratégie gagnant-gagnant. Malheureusement, tout cela a un coût, et un certain nombre de défis doivent être relevés, principalement ceux liés à la coordination de l'infrastructure, à la mobilité des utilisateurs, ainsi qu'à la continuité de service, la tarification, et les modèles d'affaires.

Pour la commodité du lecteur, nous montrons dans la Fig. B.1 les deux principales approches de délestage dans les réseaux cellulaires en comparaison avec le mode traditionnel qu'utilise seulement l'infrastructure (Fig. B.1a). Détourner le trafic via des points d'accès Wi-Fi fixes, comme dans la Fig. B.1b, représente une solution classique pour réduire le trafic

sur les réseaux cellulaires. Les destinataires situés à l'intérieur d'une zone de couverture Wi-Fi (généralement beaucoup plus petite que celle d'une macro cellule) pourraient l'utiliser comme une alternative intéressante au réseau cellulaire quand ils ont besoin d'échanger des données. De plus, la popularité croissante des smartphones proposant plusieurs options de communication permet de déployer un réseau de dispositif à dispositif (D2D) qui repose sur la communication à courte portée directe entre les utilisateurs mobiles, sans aucune nécessité d'un'infrastructure de support (Fig. B.1c). Le délestage D2D représente un sujet de recherche majeur au cœur de cette thèse. Bénéficiant d'intérêts partagés entre des utilisateurs co-localisés, un fournisseur de services cellulaires peut décider d'envoyer un contenu populaire uniquement à destination d'un sous-ensemble d'utilisateurs via le réseau cellulaire, en leur laissant la mission de propager l'information grâce à des communications D2D.

B.2 Taxonomie et architecture

Au-delà de la distinction évidente entre les approches basées sur l'utilisation de points d'accès Wi-Fi et l'utilisation de transmissions D2D déjà mentionnées dans l'introduction, un autre aspect joue un rôle majeur dans la catégorisation du délestage. Plus particulièrement, nous tenons compte des exigences en termes de garanties de livraison des applications générant le trafic. Pour cette raison, nous prenons en considération une dimension temporelle dans notre classement, selon le délai que les données peuvent tolérer lors de la livraison. Cela se traduit en deux catégories supplémentaires : (i) le délestage non-retardé et (ii) le délestage retardé.

Nous considérons donc ces deux dimensions orthogonales (garanties de retard de livraison et approche technique), qui correspondent à quatre combinaisons possibles. La plus grande différence entre ces deux mécanismes (retardés et non-retardés) résulte de la façon dont la réception du contenu est manipulée.

Les diverses formes de délestage sont très différentes, tant en termes d'infrastructure réseau, qu'au niveau des exigences de retard de livraison. Malgré cela, il est possible identifier, à partir des solutions spécifiques, un certain nombre de fonctionnalités génériques qui composent le système de délestage. Le défi est d'aller au-delà de ce qui est fait aujourd'hui, qui est essentiellement un processus initié par l'utilisateur. Cette analyse est importante en vue de l'intégration des capacités de délestage pour une future architecture des réseaux mobiles. La plupart de la littérature que nous avons analysée envisage un coordinateur de délestage, qui est une entité spécifiquement dédiée à la mise en œuvre de la stratégie de délestage. Sa tâche principale est de piloter les opérations de délestage en fonction des conditions du réseau, des demandes faites par les utilisateurs, ainsi que de la politique opérateur. Alors que le coordinateur est conceptuellement représenté par une seule entité, son emplacement physique dans le réseau peut varier, et parfois sa mise en œuvre peut être totalement distribuée. Cependant, il est possible d'identifier, parmi tous, trois blocs fonctionnels interdépendants pour le coordinateur de délestage : (i) le suivi, (ii) la prévision, et (iii) la gestion des interfaces inter-réseau.

- Le *suivi* fournit des méthodes pour suivre la propagation des données en temps réel et les demandes de l'utilisateur, et pour récupérer des informations contextuelles à partir de nœuds et du réseau. Ces informations sont nécessaires pour évaluer et exécuter la stratégie de délestage. Le bloc de suivi nécessite souvent la présence d'un canal de contrôle persistant qui permet aux utilisateurs finaux d'interagir avec le coordinateur du délestage.

- La *prévision* permet au coordinateur d'anticiper la façon dont le réseau évoluera sur la base des observations passées. Les prédictions typiques traitent de la mobilité, des modèles de contact, ou encore du débit attendu. Ces prédictions sont ensuite utilisées pour piloter le processus de délestage plus efficacement.

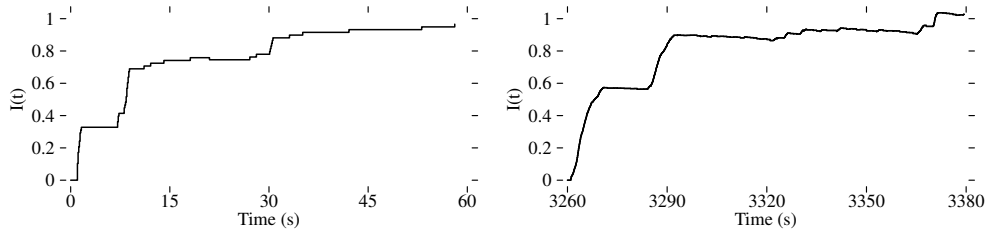
- Une *gestion intégrée des interfaces de communications* permet d'exploiter en parallèle le bénéfice de chaque interface disponible. Des concepts tels que l'équilibrage de charge, la maximisation du débit, le contrôle de congestion, ou bien la QoE (*Quality of Experience*) se rapportent à ce bloc fonctionnel. En exploitant ces informations, le réseau lui-même sera en mesure d'identifier la situation actuelle et d'optimiser ses performances.

Mais il y a bien d'autres sujets transversaux qui ressortent de l'analyse de la littérature. Par exemple, la gestion de la mobilité, de la comptabilité, et des aspects liés à la confiance, mais aussi la sécurité, sont essentiels pour permettre le délestage dans les réseaux mobiles. Ceux-ci peuvent être considérés comme les blocs fonctionnels de base que les réseaux mobiles devraient offrir afin de pouvoir fournir une réelle capacité de délestage. Quoi qu'il en soit, nous soulignons que, en fonction de la mise en œuvre spécifique, les fonctionnalités proposées peuvent être absentes.

B.3 L'adaptation à la mobilité individuelle aide le délestage

Nous proposons un système, appelé DROiD (Re-injection dérivative pour le délestage de données), qui aide les opérateurs mobiles à soulager leur réseau d'accès en exploitant à la fois la présence de points d'accès Wi-Fi et les possibilités de transmission directe entre utilisateurs (D2D). La caractéristique principale de DROiD est de s'adapter à l'hétérogénéité de la mobilité des utilisateurs. Après avoir exécuté des diffusions épidémiques sur des traces de mobilité réelles, nous avons observé que la progression de la diffusion suit une évolution caractéristique par paliers, comme en Fig. B.2. Des périodes de progression rapide s'alternent avec des périodes où la diffusion stagne; en particulier, les plateaux correspondent aux périodes pendant lesquelles la diffusion ne fait pas de progrès, car aucun des nœuds sains ne rentre en communication avec des nœuds déjà infectés. Nos recherches nous ont permis de déterminer que le principal responsable de ce phénomène n'est autre que l'hétérogénéité de la mobilité des utilisateurs.

Pour modéliser cette hétérogénéité, nous adoptons un modèle de processus de Poisson marquée des contacts entre les nœuds. Dans ce modèle, les temps de contact de deux nœuds



(a) Rollernet, 6 copies initiales. **(b)** Bologna, 10 copies initiales.

Figure B.2: Diffusion épidémique du contenu. La diffusion alterne des zones escarpées et des zones plates qui sont le résultat de l'évolution des probabilités de rencontre entre les nœuds mobiles.

(i, j) suivent un processus de Poisson de taux $\lambda_{ij} = \lambda p_{ij}$. Le temps d'inter-contact T_{ij} est donc exponentiel indépendant avec paramètre λ_{ij}, et la matrice $C = (p_{ij})$ capte les modes d'interaction entre les nœuds. Dans le cas homogène, C est la matrice d'identité, à savoir, tous les nœuds peuvent se voir l'un l'autre avec la même probabilité. À chaque instant du processus de diffusion, un ensemble S de nœuds est infecté. Nous sommes intéressés par la durée du plateau aléatoire T_s^p au cours duquel la diffusion ne progresse pas. Ceci correspond au temps au cours duquel cet ensemble de nœuds infectés ne contactent aucun autre nœud. En regardant les liens entre les nœuds dans l'ensemble S et son complément, nous remarquons que $T_S^p = \inf_{i \in S, j \notin S} T_{ij}$. Par calcul de Poisson, et notant la valeur de seuil $\partial S = \sum_{i \in S, j \notin S} p_{ij}$, nous voyons que T_s^p est une variable aléatoire exponentielle de paramètre $\lambda \partial S$. La durée de plafonnement prévue, une fois que l'ensemble S a été atteint, est donc $1/\lambda \partial S$.

À partir de cette connaissance théorique, nous illustrons le fonctionnement de DROiD dans la Fig. B.3. Le processus de diffusion est contrôlé par un canal de retour persistant qui relie les utilisateurs mobiles avec un coordinateur de délestage. Cette boucle de commande permet d'anticiper la décision de réinjection. Le système détecte la formation de plateaux dans l'évolution de la diffusion du contenu. Si nécessaire, il déclenche également la réinjection de copies supplémentaires dans le système pour contrôler finement la vitesse à laquelle le contenu est diffusé. DROiD conduit à de meilleures performances que les stratégies précédentes dans la littérature, qui sont bornées à des fonctions empiriques qui restent fixe pour l'ensemble du processus de diffusion.

B.3.1 Évaluation

Nous étudions la façon dont notre système fonctionne sous la contrainte d'une livraison serrée, lorsque le retard maximum de réception D est compris entre $[30, 180]$ secondes. Cela contraste avec ce qui se fait dans la littérature, qui ne tient qu' exclusivement compte de longues échelles de temps pour la réception du contenu (jusqu'à quelques heures).

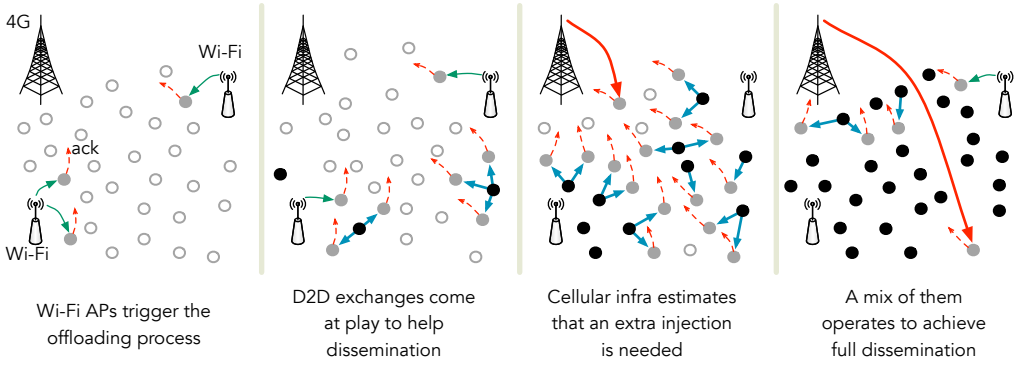
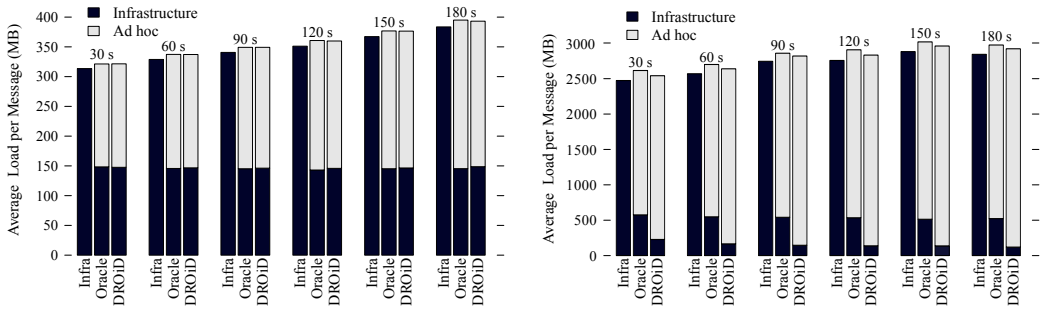


Figure B.3: Modèle de DROiD : Le processus de diffusion commence par des transferts depuis les station de base cellulaires et/ou des points d'accès Wi-Fi. Le contenu est diffusé parmi les appareils mobiles, grâce à des contacts opportunistes. Lors de la réception du contenu, les utilisateurs informent le coordinateur en utilisant le canal cellulaire de rétroaction. Comme les messages d'accusé de réception sont en général beaucoup plus petits que les messages de données, on obtient une réduction significative de trafic cellulaire. Le coordinateur central peut décider à tout moment de réinjecter des copies à travers le canal cellulaire pour stimuler la propagation. Les 100% du taux de livraison dans le délai sont atteints grâce aux réinjections finales.

Nous considérons un service d'informations de trafic basé sur la localisation ; les contenus sont publiés périodiquement, l'un perd son utilité quand un nouveau est créé (un seul contenu est actif dans le système). Le contenu qu'on vise comprend des données géo-pertinents telles que la circulation et les travaux routiers, des alertes localisées, des informations d'utilité publique ou bien de publicité ; néanmoins, le système proposé prend également en charge la distribution des mises à jour logicielles pour les véhicules connectés et les appareils mobiles. Selon la stratégie de distribution utilisée, les contenus peuvent être livrés soit directement à travers le réseau cellulaire (omniprésent), soit à travers des points d'accès Wi-Fi à proximité, soit récupérés par le nœuds voisins de façon opportuniste. Malgré le fait que nous considérons tous les utilisateurs comme intéressés par le contenu, l'utilisation combinée du paradigme *Publish-Subscribe* et des messages d'accusé de réception rend le système facilement extensible dans le cas de multiples contenus et d'intérêts non uniformes. Les utilisateurs peuvent aussi entrer et sortir de la zone d'intérêt à tout moment, en ayant un impact sur les résultats, comme nous le verrons plus tard.

DROiD fonctionne très bien en termes de réduction de charge sur l'infrastructure, en fournissant la majorité du trafic grâce à des communications D2D, même dans le cas de délais de livraison très serrés. Nous rappelons que tous les nœuds qui entrent dans la zone d'intérêt sont ciblés pour recevoir le contenu, indépendamment de leur temps de séjour dans le système. Par conséquent, la charge de la stratégie Infra, qui considéré seulement des transmissions sur le réseau cellulaire, augmente avec la durée de vie du message. Les résultats de simulation, tracées dans la Fig. B.4 affichent le trafic moyen par message qui circule à travers l'infrastructure et les interfaces D2D. Dans cette image, nous comparons DROiD et les stratégies de référence pour illustrer comment DROiD décharge constamment, une quantité im-



(a) Bologna **(b)** Koln

Figure B.4: Charge sur l'infrastructure et le D2D par message envoyé en fonction des strategies Infra, Oracle, et DROiD. Différents retards de réception maximale pour les messages sont considérés.

portante de données. Dans la simulation de Bologne, DROiD s'approche d'un oracle qui connait à tout moment la connectivité opportuniste des nœuds. Cependant, dans le scénario de Koln, nous remarquons que DROiD surpasse l'oracle en termes de trafic déplacé sur les interfaces D2D des nœuds.

Les brusques variations du taux d'infection, dues aux nœuds entrant et quittant la zone de simulation, sont bien gérées par le mécanisme de rétroaction de DROiD. Alors que la charge de la stratégie Infra augmente de manière constante dans les deux scénarios (comme une durée de vie plus longue implique un nombre plus important de nœuds qui entrent dans le système), la charge cellulaire pour Oracle et DROiD restera toujours à peu près la même.

B.3.2 Discussion

Dans cette contribution, nous avons, dans un premier temps, apporté la preuve que la diffusion épidémique s'effectue par paliers ; cela dépend fondamentalement de la dynamique de regroupement des nœuds. Nous avons offert une explication analytique de ce comportement. Pour obtenir un délestage efficace dans un tel contexte, nous avons proposé et évalué DROiD, une stratégie de déchargement à faible complexité qui s'adapte à l'évolution de la diffusion opportuniste pour améliorer la distribution du contenus populaires à travers un réseau hybride mobile.

B.4 Délestage par des communications multicast et D2D

Traditionnellement, le multicast permet d'économiser les ressources dans le *backbone* du réseau. Le multicast cellulaire améliore l'utilisation du lien radio entre la station de base et les équipements d'utilisateurs. En exploitant la nature *broadcast* du canal sans fil, le multicast utilise une seule liaison unidirectionnelle, partagée entre plusieurs utilisateurs au sein d'une même cellule radio. En dehors des transmissions unicast, le LTE propose un service de diffusion optimisée grâce à l'eMBMS (*enhanced Multicast Broadcast Multimedia Service*), une spécifica-

tion point-à-multipoint pour transmettre des données de la station de base cellulaire (eNB) à un groupe d'entités utilisatrices (UEs). Dans l'eMBMS, toutes les UEs qui appartiennent au même groupe reçoivent la même transmission. Ceci permet, en principe, une utilisation plus efficace des ressources du réseau par rapport au cas où chaque UE est atteint à travers des transmissions unicast dédiées. Cependant, en dépit de ses caractéristiques attrayantes, le multicast dans le LTE présente des problèmes intrinsèques, et encore non résolus, qui limitent son exploitation, notamment (i) l'adaptation à l'utilisateur avec le pire canal, et (ii) le manque de fiabilité.

L'hétérogénéité du canal (variant dans le temps et dépendant de l'utilisateur) réduit l'efficacité du multicast car l'eNB utilise une seule transmission pour l'ensemble des participants. La modulation et le codage sélectionnés doivent être suffisamment robustes pour assurer la réception et le décodage des données pour chaque UE dans le groupe de multicast. Ainsi, le pire canal parmi tous les récepteurs dicte la performance. Il en résulte que l'augmentation du nombre d'utilisateurs dans le groupe de multicast augmente la probabilité qu'au moins un utilisateur connaisse de mauvaises conditions de canal, dégradant le débit global. En outre, des utilisateurs dans des bonnes conditions de canal reçoivent un débit inférieur à leurs capacités, en raison de leur appartenance au groupe de multicast. En misant sur des communications D2D, nous pouvons obtenir des gains de performance en termes de ressources radio (blocs de ressources, RBs) consommées à l'eNB. Des utilisateurs bien positionnés peuvent atténuer les inefficacités du multicast, en relayant le contenu aux nœuds en mauvais condition de canal, grâce aux communications opportunistes. Malgré le fait que les avantages d'une stratégie de distribution hybride (multicast et D2D) soient évidents, sa conception fait face à plusieurs défis spécifiques aux domaines opportuniste et sans fil :

- Le bénéfice de la livraison opportuniste dépend de la configuration de la mobilité des utilisateurs. En outre, les réseaux opportunistes ne peuvent donner qu'une assurance partielle de réception.

- Pour offrir le minimum de qualité de service requis quant aux utilisateurs, tout en garantissant des économies de ressources, il est essentiel de diviser, de façon optimale, les utilisateurs entre la réception multicast et D2D.

Étant donné qu'une solution optimale n'est pas concevable sans une connaissance précise des motifs de contacts futurs, nous nous attaquons au problème d'un point de vue pratique. Nous adoptons une approche d'apprentissage par renforcement pour décider quelle fraction d'UE devra être atteinte à travers la transmission multicast et quelle fraction devra être attente en D2D. Un contrôleur central installé à l'eNB est en charge, pour chaque paquet à diffuser, de cette décision. Chaque décision se traduit par une certaine utilisation de ressources du réseau cellulaire (RBs), ce qui génère une récompense associée à ce choix. Cette récompense est ensuite utilisée pour guider (de façon probabiliste) les choix futurs du contrôleur. En raison des nombreuses similitudes dans la formulation, nous adoptons la tech-

nique d'apprentissage par renforcement appelé *bandit manchot multi bras* pour mettre en œuvre cet algorithme.

B.4.1 Strategie de distribution

Nous nous adressons à la diffusion du contenu pour un ensemble d'utilisateurs mobiles dans une cellule LTE. Comme dans notre contribution précédente (DROiD), chaque utilisateur intègre à la fois une interface cellulaire (LTE) et une technologie à courte portée qui permet des communications D2D. Dans les simulations menées, nous considérons le standard IEEE 802.11g. En revanche, l'intégration future des capacités de D2D dans la norme LTE pourrait aussi être utilisée. Nous voulons transmettre des données avec une garantie de délai maximum de réception D, et ce au moindre coût pour l'infrastructure cellulaire.

Au lieu d'aborder toutes les UEs intéressées par une seule transmission multicast – qui se traduirait probablement par un coût élevé en termes de RBs utilisés – nous adressons seulement un sous-ensemble des UEs (ceux avec la meilleure qualité de canal), en exploitant des communications opportunistes pour atteindre les autres. La diffusion opportuniste est, par définition, peu fiable, car elle dépend de plusieurs facteurs hors du contrôle de l'infrastructure cellulaire (par exemple, le modèle de mouvement des nœuds, la variabilité des voisins opportunistes, ou encore les interférences sur le canal D2D). La stratégie de délestage qui nous proposons ici est essentiellement le même que DROiD. La différence est que l'injection initiale est effectuée par l'intermédiaire de transmissions multicast, et les réinjections sont réalisées seulement dans la zone de panique. Lorsque le retard de service atteint sa valeur maximale D, l'eNB pousse toutes les données manquantes vers les nœuds non infectés en utilisant des transmissions unicast. Bien sûr, les transmissions unicast représentent la dernière opportunité pour assurer la réception des données. Dans ce schéma, le coût de diffusion aux UEs intéressées provient de (i) le coût de la transmission initiale de multicast, et (ii) le coût des transmissions unicast dans la zone de panique. La Fig. B.5 offre un exemple représentatif de la stratégie de diffusion proposée. Pour éviter le nivellement vers le bas en raison de la présence d'UEs avec des mauvaises conditions de canal, l'eNB émet à une modulation qui ne leur permet pas de recevoir. Dans la phase de diffusion opportuniste, les UEs bénéficient des noeuds à proximité afin de récupérer les données par le biais de transmissions D2D.

Il est évident qu'un tel système admet un point de fonctionnement optimal. Réduire l'ensemble des UEs atteint par le multicast réduit le cout d'envoi. Toutefois, cela peut être payé avec des coûts supplémentaires à cause des transmissions unicast dans la zone de panique, si les UEs initialement non ciblées ne sont pas atteints assez rapidement par les transmissions D2D. Identifier ce point de fonctionnement optimal est compliqué, parce que le coût de chaque configuration possible dépend de la mobilité future des nœuds (inconnue au moment de la transmission multicast). Plus précisément, le problème que nous abordons est le suivant : *comment sélectionner la configuration initiale des utilisateurs destinés à être atteints en utilisant des transmissions multicast, avec l'objectif de minimiser les ressources nécessaires pour la diffusion du contenu.*

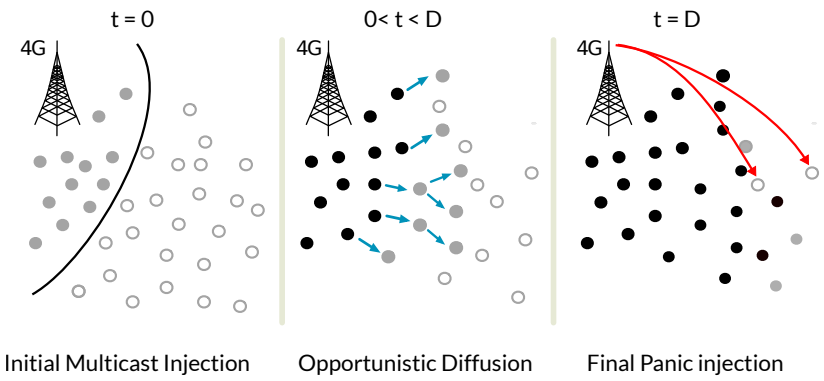


Figure B.5: Les utilisateurs peuvent décoder les données avec un schéma de modulation donné en fonction de leur qualité de canal. L'eNB peut décider de transmettre en multicast avec un débit plus élevé. Les utilisateurs incapables de décoder les données sont atteints par des liens D2D et les retransmissions de panique finales.

Dans notre système, nous utilisons un seul paramètre I_0 qui représente la fraction d'UEs qui reçoivent la transmission initiale en multicast. Cela signifie que l'eNB, dans un premier temps, atteint seulement les meilleurs I_0 UEs en termes de qualité de canal. Trouver la façon optimale de configurer I_0 n'est pas banal. Alors que le coût de la transmission multicast est déterministe, le coût des transmissions unicast nécessaires dans la zone de panique est une variable stochastique qui dépend de la mobilité des UEs pendant le cycle de vie du contenu. Nous modélisons ce problème à l'aide d'une approche d'apprentissage par renforcement (RL) que nous résolvons avec la stratégie du bandit manchot multi bras. Notre système est, en effet, capable d'apprendre de manière autonome la distribution des valeurs de I_0, en observant l'effet des différentes configurations sur le coût résultant de la diffusion d'un contenu. Sans connaissance préalable sur la mobilité des nœuds, trouver la meilleure répartition des valeurs de I_0 est le seul choix pratique pour une stratégie d'apprentissage.

B.4.2 Algorithme d'apprentissage

La formulation de l'algorithme bandit manchot multi bras peut être spécialisée comme suit. Tout d'abord, dans notre problème chaque bras du bandit correspond à un seuil I_0 différent. Il en résulte que la distribution F_{d_i} représente la quantité de RBs qui sont utilisés pendant le processus de diffusion lorsque I_0 est utilisé comme seuil. Plus précisément, $d_i = m_i + x_i$, où m_i est le numéro fixe et connu de RBs utilisés pour une transmission multicast au MCS nécessaire pour atteindre les I_0 meilleurs UEs en termes de qualité de canal, et x_i est la variable aléatoire qui modélise le nombre total de ressources utilisées pour les transmissions unicast au cours de la zone de panique. Dans notre cas, chaque tour correspond à la diffusion d'un contenu composé d'une multitude de paquets qui sont transmis de façon indépendante. Après le délai maximum de réception du contenu, la récompense pour chaque seuil est mise à jour. En supposant que $I_0 = i$ ait été utilisé pour l'énième transmission, la récompense

obtenue est calculée comme suit :

$$\mu_i(n) = \frac{1}{m_i + x_i(n)}. \tag{B.1}$$

Pour estimer dynamiquement le *récompense moyenne* $\mu_i(n)$ nous utilisons une moyenne mobile exponentielle classique avec taux α.

La politique pour choisir la prochaine valeur de I_0 se base sur des méthodes d'apprentissage qui ont été proposées dans la littérature pour les problèmes de bandit manchot multi bras. La stratégie plus simple est l'algorithme *ϵ-greedy* qui sélectionne la valeur de I_0 avec la récompense cumulée maximale avec une probabilité $(1 - \epsilon)$. Une autre classe d'algorithmes d'apprentissage est connue comme *méthode de la poursuite* (pursuit-method), dans lequel les probabilités sont sélectionnées pour renforcer la dernière sélection gourmande. Plus précisément, si $i^*(n)$ est la valeur de I_0 avec la récompense maximale, alors juste avant de sélectionner le MCS pour la transmission du énième paquet, la probabilité est renforcée comme suit :

$$\pi_{i^*(n)}(n) = \pi_{i^*(n)}(n-1) + \beta[\pi_{MAX} - \pi_{i^*(n)}(n-1)], \tag{B.2}$$

tandis que toutes les probabilités non gourmandes sont mises à jour comme suit :

$$\pi_{i(n)}(n) = \pi_{i(n)}(n-1) + \beta[\pi_{MIN} - \pi_{i(n)}(n-1)], i \neq i^*, \tag{B.3}$$

ou π_{MAX}, π_{MIN} sont respectivement la limite supérieure et inférieure que la probabilité $\pi_{i(n)}(n)$ prenne $\forall i, n$. Dans les équations B.2 et B.3, le choix gourmand est augmentée, mais jamais au delà de π_{MAX}, et le non-gourmand est réduit, mais jamais en dessous de π_{MIN}. Cela garantit que la méthode de poursuite est en mesure de faire face à la possible non-stationnarité du problème que nous étudions, à savoir le fait que la distribution des récompenses puisse changer au fil du temps en raison de la mobilité sous-jacente.

B.4.3 Évaluation

Nous considérons comme cas d'usage la distribution de contenus dans un scénario piéton comme dans un centre commercial ou un point de repère touristique bondé. Nous simulons des flux constant UDP, avec des paquets de taille $s_k = 2048$ octets et une taille du contenu total de 8 MO. Chaque paquet est distribué indépendamment des autres en utilisant l'algorithme de bandit manchot multi bras. La mobilité synthétique des UEs est mise en œuvre selon un modèle Random-Waypoint sur un zone de 200×200 m^2. Les nœuds se déplacent dans cet espace avec une vitesse se situant entre 1 et $2,5$ m/s (vitesse piétonne). Le réseau est composé d'un eNB placé au centre de la zone d'intérêt, un serveur distant qui fournit les contenus, et des multiples dispositifs mobiles. Étant donné que ns-3 ne supporte pas nativement le multicast LTE, nous avons mis en place un module supplémentaire qui interagit avec l'ordonnanceur de paquets pour émuler les effets du multicast en une seule cel-

lule. Le module de multicast reçoit les informations sur la qualité de canal (CQIs) de chaque UE, et décide la fraction d'UEs à atteindre directement, c'est-à-dire le paramètre I_0. La bande passante allouée pour le service de multicast est fixé à 5 MHz. En outre, nous avons mis en œuvre un mécanisme de routage DTN épidémique aux nœuds. La découverte de voisins est mise en œuvre par le biais d'un protocole de balisage déclenché toutes les 250 ms.

Nous comparons notre proposition avec quatre stratégies différentes pour la livraison des contenus. Les principaux indices de performance que nous considérons dans l'évaluation sont (i) le nombre de RBs utilisés par l'eNB pour délivrer le contenu et (ii) le taux de délestage. Les stratégies de diffusion envisagées sont les suivantes :

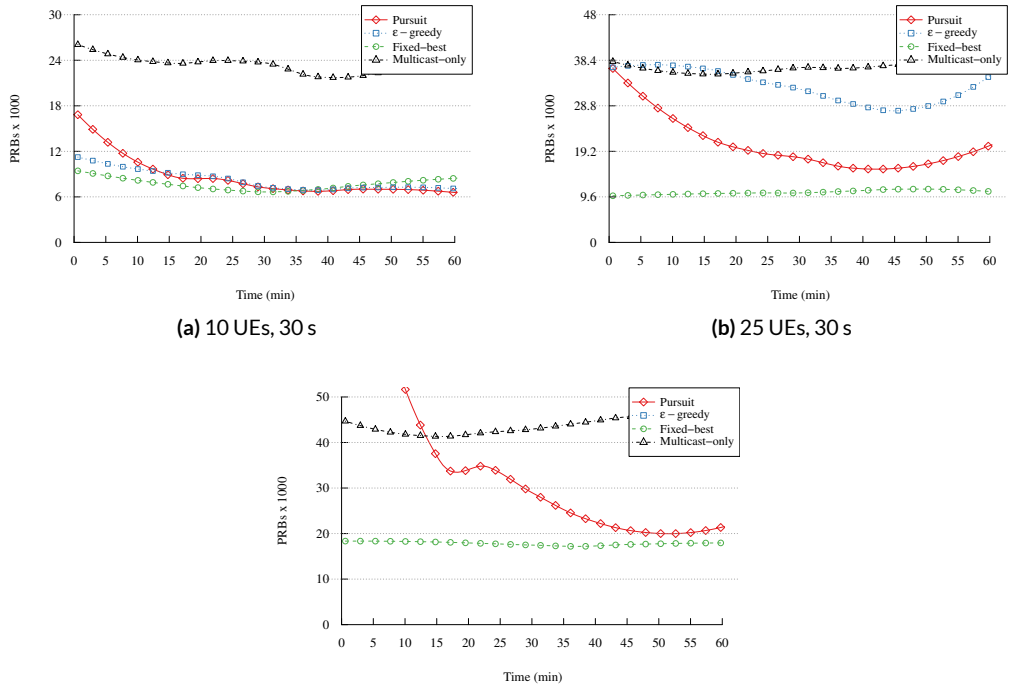
- MULTICAST-ONLY est la stratégie de base, où les UE n'ont aucun autre moyen que le réseau cellulaire pour recevoir des données.

- FIXED-BEST minimise le nombre de RBs avec une allocation statique des utilisateurs de multicast (I_0 reste fixe pendant toute la durée de la simulation). Étant donné que la taille optimale du groupe de multicast est inconnue, nous avons effectué des simulations pour trouver expérimentalement la valeur de I_0 qui minimise l'utilisation de RBs. Cette stratégie représente la référence expérimentale pour les méthodes d'apprentissage.

- ϵ-GREEDY estime la récompense en utilisant la moyenne mobile exponentielle. Ce simple algorithme sélectionne la valeur avec la meilleure récompense de I_0 avec une probabilité de $1 - \epsilon$. Dans notre implémentation, nous avons sélectionné $\epsilon = 0.05$ et $\alpha = 0.5$.

- PURSUIT sélectionne I_0 suivant la méthode présenté dans les équations B.2 et B.3. Dans ce cas, la probabilité d'émission poursuit l'action gourmande en s'adaptant à l'évolution temporelle du système. Dans la simulation, nous avons fixé $\beta = 0.3$, $\pi_{MIN} = 0.01$ et $\pi_{MIN} = 0.95$.

Les techniques d'apprentissage permettent d'économiser jusqu'à 88% de RBs pour un scénario avec délai de livraison fixé à 90 s par rapport à la stratégie *Multicast-only*. En général, la solution proposée se rapproche et même dépasse *Fixed-best* en plusieurs occasions. Ces résultats confirment qu'une synergie dans l'utilisation des transmissions multicast et D2D permet d'importantes économies de ressources à l'eNB. Les stratégies de RL peuvent trouver de manière autonome le meilleur compromis entre multicast et D2D en un délai raisonnable (toujours inférieure à 1 heure) – sans chercher intensivement tout l'espace de paramètres. Nous pouvons observer dans la Fig. B.6 le processus d'apprentissage pour le délai le plus serré qu'on considère (30 s). Les stratégies *pursuit* et *ϵ-greedy* ont besoin de temps pour apprendre la distribution la plus appropriée pour I_0. Une fois formés, leur performance est souvent à la hauteur ou même mieux que la stratégie représentée par *Fixed-best*, où la valeur de I_0 est fixe



(a) 10 UEs, 30 s

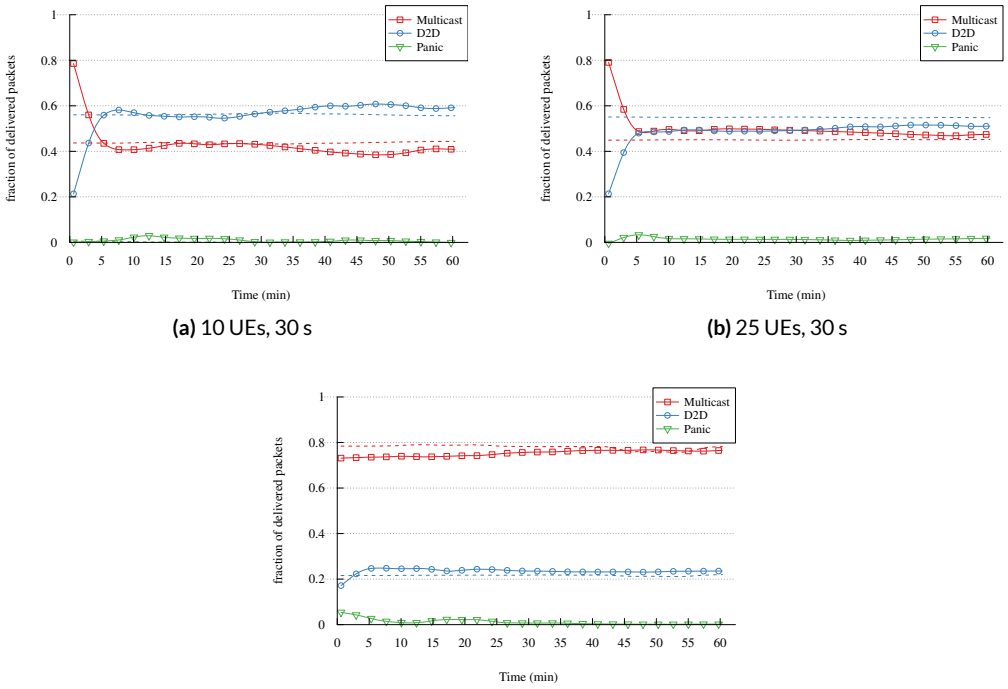
(b) 25 UEs, 30 s

(c) 50 UEs, 30 s

Figure B.6: Utilisation de RB pour les stratégies *Multicast-only* (noir), *ϵ-greedy* (blue), *Fixed-best* (vert), et *pursuit method* (rouge).

et pré-calculée. Dans cette dernière stratégie, la performance est stable sur toute la période de diffusion, mais ce chiffre est le résultat d'un processus étendu de *trial and error*. Un autre avantage des techniques d'apprentissage est que, bien que déjà formées, elles continuent à explorer l'espace des solutions, étant capables de faire face à la non-stationnarité du processus de contact qui régit la diffusion opportuniste.

En revanche, nous nous rendons compte que la méthode *ϵ-greedy*, en raison de sa simplicité, ne correspond pas bien aux scénarios qui présentent une variabilité importante de la diffusion opportuniste. Dans ces cas, la méthode *pursuit* est mieux adaptée. D'autre part, dans les scénarios où la variabilité du processus opportuniste est faible – comme dans le cas où le délai de livraison est long – l'approche *ϵ-greedy* permet des temps de convergence plus rapides. La Fig. B.6 donne des indications sur le montant réel de RBs consacrées à distribuer des données dans les scénarios envisagés. Contrairement à beaucoup d'autre approches dans la littérature, l'utilisation du simulateur ns-3 nous permet d'évaluer précisément la quantité de ressources radio consommées au eNB pour la distribution de chaque paquet. La Fig. B.7 se concentre sur la méthode de la poursuite et sépare les paquets par leur méthode de réception. Bien que dans un scénario avec un plus grand nombre d'UEs les occasions de contact opportunistes se multiplient, beaucoup d'entre elles ne sont pas exploitées de manière adéquate,



(a) 10 UEs, 30 s

(b) 25 UEs, 30 s

(c) 50 UEs, 30 s

Figure B.7: Méthode de réception dans *Pursuit*. Les lignes pointillées se réfèrent à la stratégie *Fixed-best*. Le contenu est divisé en 4000 paquets de 2048 octets.

puisque les UEs peuvent transmettre à un seul voisin à la fois. Le résultat est que la part des UEs adressée par les transmissions D2D est limitée. Cependant, même si la fraction d'UEs adressés par les transmissions D2D est limitée (par exemple, 20 % dans le scénario avec 50 UEs), l'avantage résultant en termes d'économie de RBs est beaucoup plus élevé (au moins 55 %).

B.4.4 Discussion

Nous avons présenté une stratégie de distribution hybride, en misant conjointement sur le multicast LTE et les communications D2D opportunistes, enfin de distribuer des contenus populaires avec des retards de livraison garantis. Le multicast est une option avantageuse pour distribuer des données dans un réseau cellulaire. Cependant, la performance dans une cellule est déterminée par l'utilisateur avec la pire qualité de canal. Nous avons donc proposée une solution basé sur des techniques d'apprentissage pour lutter contre les inefficacités du multicast cellulaire et distribuer une partie du trafic en utilisant des communications D2D entre les terminaux. Les résultats de simulation montrent que la stratégie proposée permet

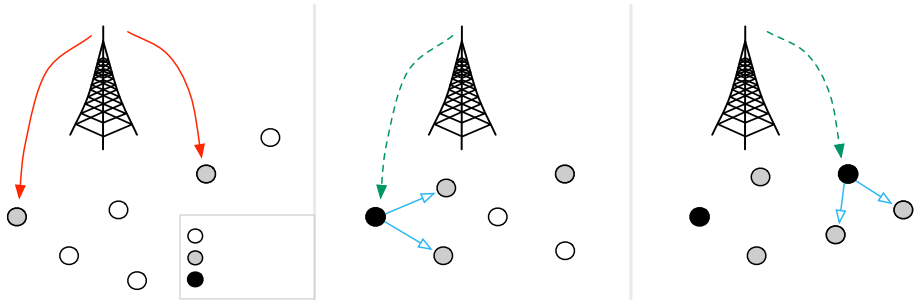


Figure B.8: Processus de délestage: l'infrastructure sélectionne deux nœuds comme des porteurs du contenu (fig. B.8 (a)), en décidant que l'un des porteurs devraient être promus comme relayeur (figure B.8 (b)). Plus tard, l'infrastructure estime qu'il vaut la peine de promouvoir un autre nœud, car les transmissions D2D ne suffisent pas à garantir une diffusion suffisante (Fig. B.8(c)).

de configurer la transmission multicast d'une manière plus efficace, améliorant l'économie des ressources et le débit global de la cellule.

B.5 Le délestage D2D et la question des récompenses

L'efficacité des stratégies de délestage D2D, telles que celles introduites précédemment, dépend de la participation des utilisateurs en tant que relais opportunistes. Des récompenses sont alors nécessaires pour stimuler la participation et motiver les utilisateurs qui agissent comme relais de données. Les propositions existantes supposent, en effet, que tous les porteurs du contenu (*seeds* en anglais) soient, par défaut, aussi des relayeurs dans le domaine D2D. Une telle hypothèse peut conduire à des résultats non optimaux lorsque les relayeurs doivent être récompensés pour la transmission du contenu au nom de l'infrastructure. Dans un scénario de ce type, des communications D2D incontrôlées peuvent générer des coûts supplémentaires sans nécessairement apporter des gains à la diffusion des données. *Nous proposons d'introduire une séparation claire entre les porteurs et les relayeurs*, comme le montre la Fig. B.8. Les porteurs reçoivent du contenu à travers des injections via l'infrastructure cellulaire, mais seulement les nœuds promus comme relayeurs sont ensuite autorisés à transmettre le contenu de façon opportuniste (en D2D). Avec cette séparation entre porteurs et relayeurs, les opérateurs disposent d'un degré de liberté supplémentaire. L'équilibre entre coût instantané et avantages futurs portés par les décisions d'injection et promotion est stratégique pour la diffusion des données, étant donné que les ressources disponibles (bande passante et récompenses) sont limitées.

Nous formulons le délestage opportuniste comme un problème de contrôle optimal, modélisant l'évolution de la diffusion d'un contenu en utilisant une variante du modèle classique *SIR*. Le contenu d'intérêt peut inclure des mises à jour logicielles, des informations

géo-localisées, des mises à jour sur l'état de la circulation, mais aussi de la publicité ciblée. Des applications de sécurité publique peuvent également bénéficier des communications D2D. Dans le modèle proposé, certains utilisateurs demandent des données, c'est-à-dire qu'ils sont intéressés par un certain contenu. Pour cette raison, nous considérons, en premier lieu, que tous les nœuds se trouvent dans l'état intéressé. À ce stade, l'opérateur ne peut utiliser que des transmissions cellulaires pour atteindre un sous-ensemble d'utilisateurs intéressés (qu'on appellera des injections). Les nœuds intéressés qui reçoivent le contenu passent à l'état porteurs, toujours sans jouer aucun rôle dans la distribution des données. À ce stade, le coordinateur peut promouvoir une fraction d'entre eux vers l'état relayeur pour diffuser le contenu.

Une incitation pour récompenser la participation des utilisateurs comme relayeurs peut être offerte en utilisant des systèmes de crédits virtuels ou de remises. La promotion vers l'état de relayeur ne représente pas un coût en soi, mais permet aux utilisateurs d'être récompensés pour la distribution du contenu en D2D. À côté des coûts pour les injections et les récompenses, le modèle qu'on propose considère ainsi des coûts indirects, parce que la fraction de nœuds non infectés (ou insatisfaits) à la fin de la durée de vie du contenu dépend de la stratégie de délestage adoptée. Le système se compose de N nœuds mobiles et d'un contenu unique qui devra être distribué par l'infrastructure avant la fin de sa durée de vie T. Suite à la notation introduite ci-dessus, les nœuds peuvent être dans les états *intéressé*, *porteur*, ou *relayeur*. Leurs fractions respective sont $n_I(t)$, $n_S(t)$, $n_F(t)$.

Dans le monde réel, le système sous observation peut être décrit avec des valeurs discrètes (par exemple, le nombre d'utilisateurs présents, le nombre de transmissions cellulaires effectuées). Au contraire, pour faciliter la modélisation, nous prenons en compte des valeurs continues pour les états et les contrôles. Nous supposons que la valeur de N soit grande, et que les rencontres soient homogènes, c'est-à-dire, les nœuds ont la même probabilité de se rencontrer. Conformément à la littérature, nous utilisons un modèle qui est exact pour une grande population. La diffusion opportuniste entre utilisateurs mobiles est considérée comme la propagation des maladies infectieuses. De la même manière qu'une infection dans une population, le contenu se propage à partir des *relayeurs* aux *intéressées* quand une telle paire entre en proximité physique. L'évolution des états est donc décrite par un système d'équations différentielles avec un ensemble de contraintes initiales et finales.

Nous considérons un coordinateur de délestage central qui gère les injections (intéressé $\rightarrow$ porteur) et les promotions (porteur $\rightarrow$ relayeur). Dans le modèle proposé, les injections cellulaires augmentent la vitesse à laquelle les nœuds quittent l'état intéressé vers l'état porteur. L'intensité des injections est représentée par $u_I(t)$, ce qui est une fonction délimitée et intégrable au sens de Lebesgue avec $0 \leq u_I(t) \leq 1 \; \forall t \in [0, T]$. Par conséquent, $u_I(t)n_I(t) \leq I_{max}(t)$ décrit le taux des copies injectés. La vitesse d'injection est limitée à tout instant par $I_{max}(t)$, qui est une mesure de la charge instantanée disponible sur le réseau cellulaire. Les nœuds porteurs stockent le contenu, mais ils doivent être promus afin de contribuer à la diffusion des données. Cela se fait par l'intermédiaire d'un canal de contrôle qui relie les utilisateurs et le coordinateur central. Par conséquent, les nœuds abandonnent l'état

porteur avec intensité $u_S(t)$, qui est une fonction délimitée et intégrable au sens de Lebesgue avec $0 \leq u_S(t) \leq 1 \, \forall t \in [0, T]$. Cela augmente la fraction de nœuds dans l'état relayeurs d'un taux $u_S(t)n_S(t)$. Par conséquent, le système d'équations qui suit contrôle l'évolution des états intéressés, porteurs et relayeurs dans le système est :

$$\frac{dn_I(t)}{dt} = -\lambda(t)n_I(t)n_F(t) - u_I(t)n_I(t), \tag{B.4a}$$

$$\frac{dn_S(t)}{dt} = u_I(t)n_I(t) - u_S(t)n_S(t), \tag{B.4b}$$

$$\frac{dn_F(t)}{dt} = \lambda(t)n_I(t)n_F(t) + u_S(t)n_S(t), \tag{B.4c}$$

La stratégie de délestage optimale consiste à minimiser le nombre de nœuds encore dans l'état intéressé au temps T, tout en mettant en œuvre une campagne d'injection et promotion avisée. Si les opérateurs n'avaient pas de limites de capacité ou monétaires, alors la stratégie optimale serait d'injecter la quantité maximale de données via le canal cellulaire. Lorsque la capacité est limitée, les opérateurs doivent améliorer leur stratégie d'un point de vue opérationnel et budgétaire. Dans l'équation B.5, nous considérons une fonction de coût J qui est assez générale pour saisir différents types de coûts supportés par les opérateurs:

$$J(T) = \underbrace{\Phi[n_I(T)]}_{\text{récompense finale}} + \int_0^T \underbrace{f[u_I(t)\, n_I(t)]}_{\text{injection}} + \underbrace{g[\lambda\, n_I(t)\, n_F(t)]}_{\text{recompense D2D}} dt. \tag{B.5}$$

où $\Phi[n_I(T)]$ est la récompense finale, représentant le coût encouru par l'opérateur pour n'avoir pas satisfait la fraction d'utilisateurs $n_I(T)$ avant la fin de la validité du contenu. $f[u_I(t)\, n_I(t)]$ tient compte du coût en termes de ressources réseau pour les injections sur le canal cellulaire. Enfin, les nœuds relayeurs sont récompensés avec $g[\lambda\, n_I(t)\, n_F(t)]$ chaque fois qu'ils font une transmission opportuniste. L'intégrale dépeint le coût croissant au fil du temps de ces deux derniers termes. Le contrôle de la promotion u_S n'apparaît pas à l'intérieur de la fonction de coût. En effet, promouvoir un nœud vers l'état relayeur ne génère pas directement un coût pour l'opérateur. Cependant, une fois dans l'état relayeur, les nœuds sont en mesure de transmettre des données de façon opportuniste et être éventuellement récompensés par l'opérateur.

Le système peut être contrôlé par le tuple $< u_I, u_S >$ appartenant à l'ensemble de tous les contrôles admissibles. L'idée est de minimiser la fonction de coût J, sous réserve des contraintes dans l'évolution des états identifiés dans l'équation. B.4 :

$$\min_{u_I(t),u_S(t)\in U} J, \tag{B.6a}$$

subject to:

$$\frac{dn_I}{dt} = -\lambda(t)n_I(t)(1 - n_I(t) - n_S(t)) - u_I(t)n_I(t), \tag{B.6b}$$

$$\frac{dn_S}{dt} = u_I(t)n_I(t) - u_S(t)n_S(t), \tag{B.6c}$$

$$n_F(t) \geq 0, n_I(t) \geq 0, n_S(t) \geq 0,$$
$$n_I(t) + n_F(t) + n_S(t) = 1,$$
$$n_I(0) = i_0, n_S(0) = s_0, n_f(0) = 1 - i_0 - s_0. \tag{B.6d}$$

L'existence d'une solution optimale est prouvée utilisant le théorème Filippov-Cesari. Nous appliquons le principe maximale de Pontryagin pour résoudre le problème ci-dessus et trouver le contrôle optimal. Considérez le solution optimale de l'Eq. B.6 $< n_I^*(.), n_S^*(.), u_I^*(.), u_S^*(.) >$.[*] Alors, des fonctions adjointes continues et continûment différentiable à morceaux $p_i^*(t)$ et $p_s^*(t)$ que maximisent la fonction hamiltonien existent :

$$\begin{aligned} H(n_{I,S}, u_{I,S}, p_{i,s}, t) = & - f[u_I\, n_I] \\ & - g[\lambda\, n_I\, (1 - n_I - n_S)] \\ & + p_i[-\lambda n_I(1 - n_I - n_S) - u_I n_I] \\ & + p_s[u_I n_I - u_S n_S]. \end{aligned} \tag{B.7}$$

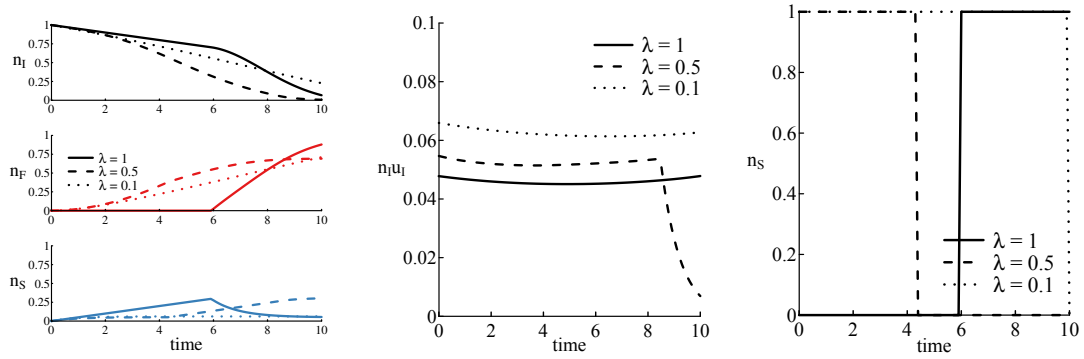
Les équations adjointes optimales sont :

$$\begin{aligned} p_i^*(t) = & -\frac{\partial H(.)}{\partial n_I}\bigg|_{n_{I,S}^*, u_{I,S}^*, p_{i,s}^*} = \\ & = \frac{\partial f(.)}{\partial n_I} + \frac{\partial g(.)}{\partial n_I} - p_i\,[\lambda\,(2n_I - 1 + n_S) - u_I] - p_s\,u_I, \end{aligned} \tag{B.8a}$$

$$\begin{aligned} p_s^*(t) = & -\frac{\partial H(.)}{\partial n_S}\bigg|_{n_{I,S}^*, u_{I,S}^*, p_{i,s}^*} = \\ & = \frac{\partial g(.)}{\partial n_S} - p_i\lambda n_I + p_s u_S. \end{aligned} \tag{B.8b}$$

Dans la suite, nous considérons une fonction exponentielle pour la récompense finale $\Phi(x) = e^x - 1$, une fonction en loi de puissance pour les injections directes $f(x) = bx^\alpha$ ($\alpha \geq 2$), et

[*] Les variables avec l'exposant étoiles (par exemple, $u_I^*(t)$) représentera la valeur à l'optimum.



(a) state evolution **(b)** injections **(c)** promotions

Figure B.9: Optimal offloading for different contact rates λ. $T = 10s$. Other parameters: $I_{max} = 0.1, \alpha = 2$, $b = 10, c = 1, i_0 = 1, s_0 = 0$.

une fonction linéaire $g(x) = cx$ pour récompenser les relayeurs. En raison de l'espace limité disponible, nous sautons la dérivation de la solution, qui suit une tractation standard, et qui nous fournissons directement.

Étant donné que $f(x)$ est strictement convexe, nous pouvons extraire $u_I^*(t)$ en utilisant la condition hamiltonien de maximisation ($\frac{\partial H}{\partial u_I} = 0$ évaluée à l'optimum), ainsi que la limitation de la vitesse d'injection maximale:

$$u_I^*(t) = \frac{\min[\max[\psi(t), 0], I_{max}]}{n_I(t)}, \qquad \psi(t) = \sqrt[\alpha-1]{\frac{p_i(t)^* - p_s(t)^*}{-\alpha\ b}}. \tag{B.9}$$

Comme l'équation B.7 est linéaire dans les variables de contrôle u_S, la condition de maximisation est trivialement satisfaite et indépendante de u_S. Le contrôle dans ce cas est appelé singulier avec une solution bang-bang, à savoir, une commande qui bascule de manière discontinue entre un extrême à l'autre. Nous définissons la fonction de commutation $\sigma = (p_s n_S)$. Par construction $u_S \in [0, 1]$, donc il suit que $u_S^*(t) = 1(-\sigma)$.

Pour résoudre le système d'équations différentielles couplées, nous adoptons le *shooting method* à partir du package R *bvpSolve* pour calculer l'évolution des variables d'état ainsi que le contrôle optimal. la Fig. B.9 propose un exemple des variables d'état et de contrôle pour différentes valeurs du taux de contact λ. En général, les injections sont plus fortes au début et à la fin de la période de diffusion (Fig. B.9b).

Les promotions (Fig. B.9c) affichent trois motifs différents. Pour $\lambda = 0, 1$, le contrôle est toujours à son maximum. Dans des scénarios avec un faible taux de contact, considérer un état porteur additionnel n'apporte aucune amélioration. Dans les deux autres cas, la stratégie optimale ne prévoit pas une transition indiscriminée vers l'état relayeur. Par exemple, $\lambda = 0.5$ présente un comportement on-off, avec des promotions qui se terminent lorsque le montant des relayeurs atteint des niveaux significatifs. Enfin, quand $\lambda = 1$, la promotion est déclenchée seulement après la moitié de la période de diffusion. Bien qu'à

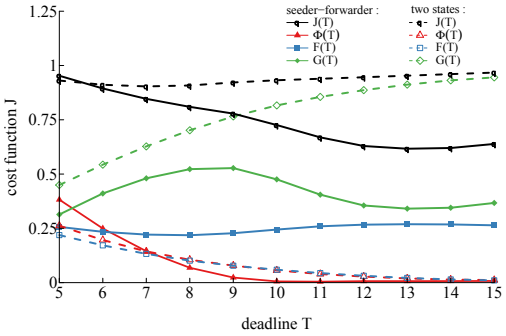


Figure B.10: Fonction de coût J pour la stratégie optimale en utilisant deux modèles de délestage (porteur-relayeur et deux-état), variant le delay de délestage $T, \lambda = 0.5, I_{max} = 0.1, \alpha = 2, b = 10, c = 1$.

première vue cela peut sembler contre-intuitif, nous ne devons pas oublier que dans le modèle proposé les opérateurs doivent récompenser chaque transmission de D2D effectuée par les utilisateurs. Nous tirons la leçon que dans un régime de taux de contact élevé, la diffusion opportuniste doit être limitée afin d'économiser les ressources monétaires.

Nous étudions aussi quand il est utile d'envisager une séparation entre les porteurs et les relayeurs. Nous comparons notre modèle à un modèle à deux états classique, où tous les porteurs sont également relayeurs. Fig. B.10 montre l'évolution de la fonction de coût J divisée par ses trois composantes principales $\Phi(T)$, $F(T)$ et $G(T)$. Avec des délais courts, lorsque les nœuds ont peu de possibilités de contact, le modelé à deux états a un léger avantage en terme de coût, vu que J est dominée par la récompense finale $\Phi(T)$.

Inversement, pour des délais plus longs, le modèle à trois états améliore considérablement la fonction de coût J. Pour $T > 5$, le nombre de nœuds non infectés à la date limite diminue, réduisant le poids de $\Phi(T)$ sur le coût global. La plus grande partie de J est due à la récompense des relayeurs opportunistes (décrit par $G(T)$). Le coût pour récompenser les utilisateurs augmente linéairement pour le modèle à deux états avec le délai de délestage. Un nombre incontrôlé des relayeurs interfère avec la volonté des opérateurs de réduire les coûts opérationnels. Une séparation entre les porteurs et les relayeurs offre une meilleure flexibilité dans le contrôle de l'évolution de délestage, permettant la mise en œuvre de stratégies moins coûteuses.

B.5.1 Discussion

Nous avons proposé un cadre analytique pour le délestage opportuniste qui capte les différences entre les porteurs et les relayeurs. Avec notre approche, les opérateurs mobiles sont en mesure de contrôler finement l'évolution de la diffusion à l'aide des contrôles externes tels que les injections à travers l'infrastructure et les promotions des porteurs en relayeurs. Nous avons ensuite appliqué le principe maximal de Pontryagin pour concevoir une stratégie de délestage optimal qui minimise les coûts de distribution pour l'opérateur. L'un des principaux avantages du modèle proposé est que chaque paramètre peut être facilement réglé. Nous

avons démontré sa sensibilité à des différentes valeurs du taux de contact et de tolérance aux délais de livraison. Nous avons également évalué les avantages du modèle proposé en le comparant à un modelé classique à deux états. En particulier, nous avons démontré que, lorsque nous avons des délais de livraisons suffisamment longs, l'introduction d'une séparation entre les nœuds semoirs et les nœuds relayeurs est fortement bénéfique pour un opérateur cellulaire.

B.6 Conclusion générale

L'utilisation globale des données mobiles a connu une augmentation spectaculaire au cours des dernières années, engendrée par le boom du marché des dispositifs mobiles. Les prévisions, qui atteignent une croissance égale à aujourd'hui pour les prochaines années, appellent à des stratégies efficaces pour faire face à cette montée. À compter d'aujourd'hui, les opérateurs sont déjà sous forte pression, en essayant de recevoir un quantité sans précédent de trafic mobile sur leurs réseaux. Par conséquent, ils doivent intervenir avec d'importants investissements à l'échelle de leurs réseaux d'accès. Néanmoins, les dépenses pour acheter plus de bandes, ou pour construire des stations de base, sont très élevées. Malheureusement, l'augmentation de la capacité du réseau offerte par ces méthodes ne pourra pas suivre la croissance du trafic.

Le délestage de données mobiles représente une solution bon marché pour soulager la charge sur l'infrastructure du réseau mobile, tout en bénéficiant des technologies complémentaires existantes, en déplaçant les données sur des réseaux alternatifs moins encombrés. Le débat suscité par cette thèse préconise fortement l'utilisation de réseaux alternatifs d'accès mobile à des fins de délestage. Plus précisément, nous nous sommes concentrés sur un type de délestage, basé sur l'échange direct de données entre utilisateurs, utilisant des communications de dispositif à dispositif (D2D). L'idée étant de profiter de l'augmentation de la densité d'utilisateurs mobiles, ainsi que de la tolérance aux retards d'un certain nombre de types du contenu pour déplacer une partie du trafic du canal primaire (cellulaire) vers un canal de D2D alternatif.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update (2014 – 2019)," 2015.

[2] P. Taylor, "Data overload threatens mobile networks," accessed: 2013-08-21. [Online]. Available: http://www.ft.com/intl/cms/s/0/caeb0766-9635-11e1-a6a0-00144feab49a.html

[3] B. G. Mölleryd, J. Markendahl, J. Werding, and O. Mäkitalo, "Decoupling of revenues and traffic - is there a revenue gap for mobile broadband ?" in *Conference on Telecommunications Internet and Media Techno Economics (CTTE)*, Ghent, Belgium, Jun. 2010, pp. 1–7.

[4] S. Curtis, "Can you survive on 4g alone?" accessed: 2013-11-06. [Online]. Available: http://www.telegraph.co.uk/technology/internet/10272292/Can-you-survive-on-4G-alone.html

[5] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li, "Cellular traffic offloading through WiFi networks," in *IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, Valencia, Spain, Oct. 2011.

[6] L. Korowajczuk, *LTE, WiMAX and WLAN Network Design, Optimization and Performance Analysis*. John Wiley & Sons, 2011.

[7] "Data Offload – Connecting Intelligently," White Paper, Juniper Research, 2013.

[8] iPass, "iPass application." [Online]. Available: http://www.ipass.com/

[9] Guglielmo, "BabelTen application," accessed: 2013-08-21. [Online]. Available: http://www.guglielmo.biz/Servizi.aspx?lan=eng

[10] W. Mohr and W. Konhauser, "Access network evolution beyond third generation mobile communications," *IEEE Communications Magazine*, vol. 38, no. 12, pp. 122–133, Dec. 2000.

[11] "MOTO FP7 Project http://www.fp7-moto.eu/."

[12] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 536–550, Apr. 2013.

[13] P. Fuxjager, I. Gojmerac, H. R. Fischer, and P. Reichl, "Measurement-based small-cell coverage analysis for urban macro-offload scenarios," in *Vehicular Technology Conference (VTC Spring)*, Yokohama, Japan, May 2011, pp. 1–5.

[14] S. Liu and A. Striegel, "Casting Doubts on the Viability of WiFi Offloading," in *ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, Helsinki, Finland, 2012, pp. 25–30.

[15] L. Hu, C. Coletti, N. Huan, I. Z. Kovács, B. Vejlgaard, R. Irmer, and N. Scully, "Realistic indoor wi-fi and femto deployment study as the offloading solution to lte macro networks," in *IEEE Vehicular Technology Conference (VTC Fall)*, Quebec City, QC, Sep. 2012, pp. 1–6.

[16] N. Ristanovic, J.-Y. Le Boudec, A. Chaintreau, and V. Erramilli, "Energy efficient offloading of 3G networks," in *IEEE International Conference on Mobile Ad-Hoc and Sensor Systems (MASS)*, Valencia, Spain, Oct. 2011, pp. 202–211.

[17] E. Bulut and B. K. Szymanski, "Wifi access point deployment for efficient mobile data offloading," in *ACM international workshop on Practical issues and applications in next generation wireless networks*, Istanbul, Turkey, 2012, pp. 45–50.

[18] E. Oliveira and A. Viana, "From routine to network deployment for data offloading in metropolitan areas," in *Sensing, Communication, and Networking (SECON), 2014 Eleventh Annual IEEE International Conference on*, June 2014, pp. 126–134.

[19] F. Mehmeti and T. Spyropoulos, "Performance analysis of on-the-spot mobile data offloading," in *IEEE GLOBECOM*, Atlanta, GA, Dec 2013, pp. 1577–1583.

[20] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[21] S. Wietholter, M. Emmelmann, R. Andersson, and A. Wolisz, "Performance evaluation of selection schemes for offloading traffic to IEEE 802.11 hotspots," in *IEEE International Conference on Communications (ICC)*, Ottawa, Canada, jun 2012, pp. 5423–5428.

[22] 3GPP, "3GPP TS 24.312: Access Network Discovery and Selection Function (ANDSF) Management Object (MO) (Rel. 10)," 2011.

[23] ——, "3GPP TR 23.829: Local IP Access and Selected IP Traffic Offload (LIPA-SIPTO) (Rel. 10)," 2011.

[24] ——, "3GPP TS 23.261: IP flow mobility and seamless Wireless Local Area Network (WLAN) offload (Rel. 10)," 2011.

[25] D. Hagos and R. Kapitza, "Study on performance-centric offload strategies for lte networks," in *6th Joint IFIP Wireless and Mobile Networking Conference (WMNC)*, Dubai, 2013, pp. 1–10.

[26] Y. M. Kwon, J. S. Kim, J. Gu, and M. Y. Chung, "ANDSF-based congestion control procedure in heterogeneous networks," in *IEEE International Conference on Information Networking (ICOIN)*, Bangkok, 2013, pp. 547–550.

[27] K. Samdanis, T. Taleb, and S. Schmid, "Traffic Offload Enhancements for eUTRAN," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 3, pp. 884–896, 2012.

[28] C. B. Sankaran, "Data Offloading Techniques in 3GPP Rel-10 Networks: A Tutorial," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 46–53, 2012.

[29] A. de la Oliva, C. Bernardos, M. Calderon, T. Melia, and J. Zuniga, "Ip flow mobility: Smart traffic offload for future wireless networks," *IEEE Communications Magazine*, vol. 49, no. 10, pp. 124–132, Oct. 2011.

[30] R.Stewart, "Stream control transmission protocol," *RFC 4960*, 2007.

[31] X. Hou, P. Deshpande, and S. R. Da, "Moving bits from 3g to metro-scale WiFi for vehicular network access: An integrated transport layer solution," in *IEEE International Conference on Network Protocols (ICNP)*, Vancouver, Canada, Oct. 2011, pp. 353–362.

[32] M. A. P. Gonzalez, T. Higashino, and M. Okada, "Radio access considerations for data offloading with multipath tcp in cellular / wifi networks," in *International Conference on Information Networking (ICOIN)*, Bangkok, Jan. 2013, pp. 680–685.

[33] "MultiPath TCP -Linux Kernel Implementation," accessed: 2014-06-17. [Online]. Available: http://multipath-tcp.org/pmwiki.php/Users/Android

[34] I. van Beijnum, "Multipath TCP lets Siri seamlessly switch between Wi-Fi and 3G/LTE," accessed: 2014-06-17. [Online]. Available: http://arstechnica.com/apple/2013/09/multipath-tcp-lets-siri-seamlessly-switch-between-wi-fi-and-3glte/

[35] 3GPP, "3GPP TSG SA: Feasibility Study for Proximity Services (ProSe) (Rel. 12)," 2012.

[36] M. Dohler, D.-E. Meddour, S. M. Senouci, and A. Saadani, "Cooperation in 4g - hype or ripe?" *IEEE Technology and Society Magazine*, vol. 27, no. 1, pp. 13–17, Spring 2008.

[37] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, Mar 2000.

[38] S. Weber, J. Andrews, and N. Jindal, "An overview of the transmission capacity of wireless networks," *IEEE Transactions on Communications*, vol. 58, no. 12, pp. 3593–3604, December 2010.

[39] A. Ozgur, O. Leveque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3549–3572, Oct 2007.

[40] L. Al-Kanj, Z. Dawy, and E. Yaacoub, "Energy-aware cooperative content distribution over wireless networks: Design alternatives and implementation aspects," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 1736–1760, Fourth 2013.

[41] M. Stiemerling and S. Kiesel, "Cooperative p2p video streaming for mobile peers," in *IEEE Computer Communications and Networks (ICCCN)*, Zurich, Switzerland, Aug 2010, pp. 1–7.

[42] P. Karunakaran, H. Bagheri, and M. Katz, "Energy efficient multicast data delivery using cooperative mobile clouds," in *18th European Wireless ConferenceEuropean Wireless*, April 2012, pp. 1–5.

[43] S.-S. Kang and M. W. Mutka, "A mobile peer-to-peer approach for multimedia content sharing using 3g/wlan dual mode channels," *Wireless Communications and Mobile Computing*, vol. 5, no. 6, pp. 633–645, 2005.

[44] M.-F. Leung and S.-H. Chan, "Broadcast-based peer-to-peer collaborative video streaming among mobiles," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 350–361, March 2007.

[45] D. Zhu and M. Mutka, "Cooperation among peers in an ad hoc network to support an energy efficient IM service," *Pervasive and Mobile Computing*, vol. 4, no. 3, pp. 335 – 359, 2008.

[46] S. Hua, Y. Guo, Y. Liu, H. Liu, and S. Panwar, "Scalable video multicast in hybrid 3g/ad-hoc networks," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 402–413, April 2011.

[47] M. Ramadan, L. E. Zein, and Z. Dawy, "Implementation and evaluation of cooperative video streaming for mobile devices," in *IEEE Personal, Indoor and Mobile Radio Communications, (PIMRC)*, Cannes, France, Sept 2008, pp. 1–5.

[48] L. Keller, A. Le, B. Cici, H. Seferoglu, C. Fragouli, and A. Markopoulou, "Microcast: Cooperative video streaming on smartphones," in *ACM MobiSys*, 2012, pp. 57–70.

[49] S. Sharafeddine, K. Jahed, N. Abbas, E. Yaacoub, and Z. Dawy, "Exploiting multiple wireless interfaces in smartphones for traffic offloading," in *Black Sea Conference on Communications and Networking (BlackSeaCom)*, July 2013, pp. 142–146.

[50] S. Andreev, A. Pyattaev, K. Johnsson, O. Galinina, and Y. Koucheryavy, "Cellular traffic offloading onto network-assisted device-to-device connections," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 20–31, April 2014.

[51] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to lte-advanced networks," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42–49, 2009.

[52] B. Raghothaman, E. Deng, R. Pragada, G. Sternberg, T. Deng, and K. Vanganuru, "Architecture and protocols for LTE-based device to device communication," in *International Conference on Computing, Networking and Communications (ICNC)*, San Diego, CA, Jan. 2013, pp. 895 – 899.

[53] M. J. Yang, S. Y. Lim, H. J. Park, and N. H. Park, "Solving the data overload: Device-to-device bearer control architecture for cellular data offloading," *Vehicular Technology Magazine, IEEE*, vol. 8, no. 1, pp. 31–39, Mar. 2013.

[54] J. Liu, Y. Kawamoto, H. Nishiyama, N. Kato, and N. Kadowaki, "Device-to-device communications achieve efficient load balancing in LTE-advanced networks," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 57–65, April 2014.

[55] T. Doumi, M. Dolan, S. Tatesh, A. Casati, G. Tsirtsis, K. Anchan, and D. Flore, "LTE for public safety networks," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 106–112, 2013.

[56] D. Feng, L. Lu, Y. Yuan-Wu, G. Ye Li, S. Li, and G. Feng, "Device-to-device communications in cellular networks," *IEEE Communications Magazine*, vol. 52, no. 4, pp. 49–55, 2014.

[57] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Enable device-to-device communications underlaying cellular networks: challenges and research aspects," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 90–96, 2014.

[58] M. Zulhasnine, C. Huang, and A. Srinivasan, "Efficient resource allocation for device-to-device communication underlaying lte network," in *IEEE 6th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Niagara Falls, ON, Oct. 2010, pp. 368–375.

[59] C.-H. Yu, K. Doppler, C. B. Ribeiro, and O. Tirkkonen, "Resource sharing optimization for device-to-device communication underlaying cellular networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2752–2763, 2011.

[60] F. Malandrino, C. Casetti, and C.-F. Chiasserini, "A fix-and-relax model for heterogeneous lte-based networks," in *IEEE 21st International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS)*, San Francisco, CA, Aug. 2013, pp. 308–312.

[61] M. Hasan, E. Hossain, and D. Kim, "Resource allocation under channel uncertainties for relay-aided device-to-device communication underlaying lte-a cellular networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 2322 – 2338, Apr. 2014.

[62] Y. Li, T. Wu, P. Hui, D. Jin, and S. Chen, "Social-aware d2d communications: qualitative insights and quantitative analysis," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 150–158, 2014.

[63] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, June 2007.

[64] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnović, "Power law and exponential decay of intercontact times between mobile devices," *IEEE Transactions on Mobile Computing*, vol. 9, no. 10, pp. 1377–1390, Oct 2010.

[65] V. Conan, J. Leguay, and T. Friedman, "Characterizing pairwise inter-contact patterns in delay tolerant networks," in *ACM Autonomics*, Rome, Italy, 2007, pp. 19:1–19:9.

[66] A. Passarella and M. Conti, "Analysis of individual pair and aggregate intercontact times in heterogeneous opportunistic networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 12, pp. 2483–2495, Dec 2013.

[67] A. Tatar, T. Phe-Neau, M. D. de Amorim, V. Conan, and S. Fdida, "Beyond contact predictions in mobile opportunistic networks," in *IEEE Wireless On-demand Network Systems and Services (WONS)*, April 2014, pp. 65–72.

[68] V. Siris and D. Kalyvas, "Enhancing mobile data offloading with mobility prediction and prefetching," in *ACM international workshop on Mobility in the evolving internet architecture (MobiArch)*, Istanbul, Turkey, 2012, pp. 17–22.

[69] B. B. Chen and M. C. Chan, "Mobtorrent: A framework for mobile internet access from vehicles," in *IEEE INFOCOM*, 2009, pp. 1404–1412.

[70] A. Y. Ding, B. Han, Y. Xiao, P. Hui, A. Srinivasan, M. Kojo, and S. Tarkoma, "Enabling energy-aware collaborative mobile data offloading for smartphones," in *Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2013 10th Annual IEEE Communications Society Conference on*. IEEE, 2013, pp. 487–495.

[71] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger, and M. J. Neely, "Energy-delay tradeoffs in smartphone applications," in *International conference on Mobile systems, applications, and services - MobiSys*, San Francisco, CA, 2010, pp. 255 – 270.

[72] Y. Go, Y. G. Moon, and K. S. Park, "Enabling DTN-based data offloading in urban mobile network environments," in *International Conference on Future Internet Technologies*, Seoul, Korea, 2012, p. 48.

[73] Y. Go, Y. G. Moon, G. Nam, and K. S. Park, "A disruption-tolerant transmission protocol for practical mobile data offloading," in *ACM international workshop on Mobile Opportunistic Networks*, Zurich, Switzerland, 2012, pp. 61–68.

[74] F. Malandrino, C. Casetti, C. Chiasserini, and M. Fiore, "Offloading cellular networks through ITS content download," in *IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, Seoul, South Korea, Jun. 2012, pp. 263–271.

[75] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *ACM Mobisys*, San Francisco, CA, Jun. 2010.

[76] O. B. Yetim and M. Martonosi, "Adaptive usage of cellular and WiFi bandwidth: An optimal scheduling formulation," in *ACM CHANTS*, Istanbul, Turkey, Aug. 2012.

[77] D. Zhang and C. Yeo, "Optimal handing-back point in mobile data offloading," in *IEEE Vehicular Networking Conference (VNC)*, Nov. 2012, pp. 219–225.

[78] F. Mehmeti and T. Spyropoulos, "Is it worth to be patient? analysis and optimization of delayed mobile data offloading," in *IEEE INFOCOM*, Toronto, ON, Apr. 2014, pp. 2364 – 2372.

[79] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci, "Taming the mobile data deluge with drop zones," *IEEE/ACM Transactions on Networking*, vol. 20, no. 4, pp. 1010–1023, Aug. 2012.

[80] C. Lochert, B. Scheuermann, C. Wewetzer, A. Luebke, and M. Mauve, "Data aggregation and roadside unit placement for a vanet traffic information system," in *ACM international workshop on VehiculAr Inter-NETworking*, San Francisco, CA, 2008, pp. 58–65.

[81] A. Abdrabou and W. Zhuang, "Probabilistic delay control and road side unit placement for vehicular ad hoc networks with disrupted connectivity," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 1, pp. 129–139, 2011.

[82] Y. Wang, J. Zheng, and N. Mitton, "Delivery delay analysis for roadside unit deployment in intermittently connected vanets," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2014, pp. 155–161.

[83] F. Malandrino, C. Casetti, C. Chiasserini, C, and M. Fiore, "Optimal content downloading in vehicular networks," *IEEE Transactions onMobile Computing*, vol. 12, no. 7, pp. 1377–1391, 2013.

[84] F. Malandrino, C. Casetti, C. Chiasserini, C. Sommer, and F. Dressler, "Content downloading in vehicular networks: Bringing parked cars into the picture," in *IEEE Personal Indoor and Mobile Radio Communications (PIMRC)*, 2012, pp. 1534–1539.

[85] D. Astudillo, E. Chaput, and A.-L. Beylot, "Bulk data transfer through vanet infrastructure," in *Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th*, June 2013, pp. 1–5.

[86] E. Hossain, G. Chow, V. C. Leung, R. D. McLeod, J. Mišić, V. W. Wong, and O. Yang, "Vehicular telematics over heterogeneous wireless networks: A survey," *Computer Communications*, vol. 33, no. 7, pp. 775 – 793, 2010.

[87] K. Fall and S. Farrell, "DTN: an architectural retrospective," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 5, pp. 828–836, Jun. 2008.

[88] N. Golrezaei, A. Dimakis, and A. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4286 – 4298, Jul. 2014.

[89] N. Anciaux, L. Bouganim, T. Delot, S. Ilarri, L. Kloul, N. Mitton, and P. Pucheral, "Opportunistic data services in least developed countries: Benefits, challenges and feasibility issues," *SIGMOD Rec.*, vol. 43, no. 1, pp. 52–63, May 2014.

[90] Y. Li, G. Su, P. Hui, D. Jin, L. Su, and L. Zeng, "Multiple mobile data offloading through delay tolerant networks," in *ACM CHANTS*, Las Vegas, NV, Sep. 2011.

[91] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *ACM SIGCOMM conference on Internet measurement*, San Diego, CA, 2007, pp. 1–14.

[92] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and rel-11 enhancements," *IEEE Communication Magazine*, vol. 50, no. 11, pp. 68–74, Nov. 2012.

[93] F. Rebecchi, M. D. de Amorim, and V. Conan, "Flooding data in a cell: Is cellular multicast better than device-to-device communications?" in *ACM CHANTS*, Maui, HI, Sep. 2014.

[94] E. Biondi, C. Boldrini, A. Passarella, and M. Conti, "Optimal duty cycling in mobile opportunistic networks with end-to-end delay guarantees," in *European Wireless*, Barcelona, Spain, May 2014.

[95] E. Biondi, C. Boldrini, M. Conti, and A. Passarella, "Duty cycling in opportunistic networks: the effect on intercontact times," in *The 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (ACM MSWiM 2014)*, Montreal, Canada, Sep. 2014.

[96] L. Pelusi, A. Passarella, and M. Conti, "Opportunistic networking: data forwarding in disconnected mobile ad hoc networks," *IEEE Communications Magazine*, vol. 44, no. 11, pp. 134–141, 2006.

[97] C. Boldrini and A. Passarella, "Data dissemination in opportunistic networks," *Mobile Ad Hoc Networking: Cutting Edge Directions,*, pp. 453–490, 2013.

[98] F. Mezghani, R. Dhaou, M. Nogueira, and A.-L. Beylot, "Content dissemination in vehicular social networks: taxonomy and user satisfaction," *Communications Magazine, IEEE*, vol. 52, no. 12, pp. 34–40, December 2014.

[99] A. Vahdat and D. Becker, "Epidemic routing for partially connected ad hoc networks," Technical Report CS-200006, Duke University, Tech. Rep., 2000.

[100] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and wait: An efficient routing scheme for intermittently connected mobile networks," in *ACM SIGCOMM Workshop on Delay-tolerant Networking*, Philadelphia, PA, 2005, pp. 252–259.

[101] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic routing in intermittently connected networks," *ACM SIGMOBILE mobile computing and communications review*, vol. 7, no. 3, pp. 19–20, 2003.

[102] J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine, "Maxprop: Routing for vehicle-based disruption-tolerant networks." in *IEEE INFOCOM*, vol. 6, 2006, pp. 1–11.

[103] R. Groenevelt, P. Nain, and G. Koole, "The message delay in mobile ad hoc networks," *Performance Evaluation*, vol. 62, no. 1–4, pp. 210 – 228, 2005.

[104] X. Zhang, G. Neglia, J. Kurose, and D. Towsley, "Performance modeling of epidemic routing," *Computer Networks*, vol. 51, no. 10, pp. 2867 – 2891, 2007.

[105] P. Jacquet, B. Mans, and G. Rodolakis, "Information propagation speed in mobile and delay tolerant networks," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5001–5015, Oct 2010.

[106] E. Baccelli, P. Jacquet, B. Mans, and G. Rodolakis, "Highway vehicular delay tolerant networks: Information propagation speed properties," *IEEE Transactions on Information Theory,*, vol. 58, no. 3, pp. 1743–1756, 2012.

[107] S. Ioannidis, A. Chaintreau, and L. Massoulié, "Optimal and scalable distribution of content updates over a mobile social network," in *IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009.

[108] B. Han, P. Hui, V. S. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan, "Mobile data offloading through opportunistic communications and social participation," *IEEE Transactions on Mobile Computing*, vol. 11, no. 5, pp. 821–834, May 2012.

[109] M. V. Barbera, A. C. Viana, M. D. de Amorim, and J. Stefa, "Data offloading in social mobile networks through VIP delegation," *Ad Hoc Networks*, vol. 19, pp. 92–110, 2014.

[110] Y. Chuang and K.-J. Lin, "Cellular traffic offloading through community-based opportunistic dissemination," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Shanghai, China, Apr. 2012, pp. 3188–3193.

[111] P. Baier and K. Rothermel, "TOMP: Opportunistic traffic offloading using movement predictions," in *IEEE Conference on Local Computer Networks (LCN)*, Oct. 2012.

[112] H. Luo, X. Meng, R. Ramjee, P. Sinha, and L. Li, "The Design and Evaluation of Unified Cellular and Ad-Hoc Networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 9, pp. 1060–1074, Sep. 2007.

[113] C. Mayer and O. Waldhorst, "Offloading infrastructure using delay tolerant networks and assurance of delivery," in *IFIP Wireless Days (WD)*, Niagara Falls, ON, Oct. 2011, pp. 1–7.

[114] H. Izumikawa and J. Katto, "RoCNet: Spatial mobile data offload with user-behavior prediction through delay tolerant networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Shanghai, China, 2013, pp. 2196–2201.

[115] M. Pitkanen, T. Karkkainen, and J. Ott, "Opportunistic web access via wlan hotspots," in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Mannheim, Germany, Apr. 2010, pp. 20–30.

[116] A. Petz, A. Lindgren, P. Hui, and C. Julien, "MADServer: A server architecture for mobile advanced delivery," in *ACM CHANTS*, Istanbul, Turkey, 2012, pp. 17–22.

[117] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, and M. D. de Amorim, "Push-and-track: Saving infrastructure bandwidth through opportunistic forwarding," *Pervasive and Mobile Computing*, vol. 8, no. 5, pp. 682–697, Oct. 2012.

[118] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: technical and business perspectives," *IEEE Wireless Communications*, vol. 20, no. April, pp. 104–112, 2013.

[119] J. M. Chapin and W. H. Lehr, "Mobile broadband growth, spectrum scarcity, and sustainable competition," in *TPRC*, 2011, pp. 1–36.

[120] "Mobile data offload for 3G networks," White Paper, IntelliNet Technologies, 2011.

[121] M. Yavuz, F. Meshkati, S. Nanda, A. Pokhariyal, N. Johnson, B. Raghothaman, and A. Richardson, "Interference Management and Performance Analysis of UMTS/HSPA+ Femtocells," *IEEE Communications Magazine*, vol. 47, no. September, pp. 102–109, 2009.

[122] D. Calin, H. Claussen, and H. Uzunalioglu, "On femto deployment architectures and macrocell offloading benefits in joint macro-femto deployments," *IEEE Communications Magazine*, vol. 48, no. 1, pp. 26–32, Jan. 2010.

[123] J. Gora and T. E. Kolding, "Deployment aspects of 3g femtocells," in *International Symposium on Personal, Indoor and Mobile Radio Communications*, Tokyo, Japan, Sep. 2009, pp. 1507–1511.

[124] L. Hu, C. Coletti, N. Huan, P. Mogensen, and J. Elling, "How much can wi-fi offload? a large-scale dense-urban indoor deployment study," in *IEEE Vehicular Technology Conference (VTC Spring)*, Yokohama, Japan, May 2012, pp. 1–6.

[125] D. Karvounas, A. Georgakopoulos, D. Panagiotou, V. Stavroulaki, K. Tsagkaris, and P. Demestichas, "Opportunistic exploitation of resources for improving the energy-efficiency of wireless networks," *IEEE International Conference on Communications (ICC)*, pp. 5746–5750, Jun. 2012.

[126] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Communications Magazine*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[127] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, Present, and Future," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[128] J. Hoadley and P. Maveddat, "Enabling small cell deployment with HetNet," *IEEE Wireless Communications*, vol. 19, no. 2, pp. 4–5, Apr. 2012.

[129] I. F. Akyildiz, W. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer Networks*, vol. 50, no. 13, pp. 2127 – 2159, 2006.

[130] K. Berg and M. Katsigiannis, "Optimal cost-based strategies in mobile network offloading," in *International Conference on Cognitive Radio Oriented Wireless Networks*, Stockholm, Sweden, Jun. 2012.

[131] P. Grønsund, O. Grøndalen, and M. Lähteenoja, "Business Case Evaluations for LTE Network Offloading with Cognitive Femtocells," *Elsevier Telecommunications Policy*, vol. 37, no. 2–3, 2013.

[132] H. ElSawy, E. Hossain, and D. I. Kim, "Hetnets with cognitive small cells: user offloading and distributed channel access techniques," *IEEE Communications Magazine*, vol. 51, no. 6, 2013.

[133] A. J. Mashhadi and P. Hui, "Proactive Caching for Hybrid Urban Mobile Networks," University College London, Tech. Rep., 2010.

[134] F. Malandrino, M. Kurant, A. Markopoulou, C. Westphal, and U. Kozat, "Proactive seeding for information cascades in cellular networks," in *IEEE INFOCOM*, Orlando, FL, Mar. 2012, pp. 1719–1727.

[135] M. Fiore, C. Casetti, and C. Chiasserini, "Caching strategies based on information density estimation in wireless ad hoc networks," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 5, pp. 2194–2208, Jun 2011.

[136] M. Karaliopoulos, "Assessing the vulnerability of DTN data relaying schemes to node selfishness," *IEEE Communications Letters*, vol. 13, no. 12, pp. 923–925, Dec. 2009.

[137] P.-U. Tournoux, J. Leguay, F. Benbadis, J. Whitbeck, V. Conan, and M. D. de Amorim, "Density-aware routing in highly dynamic DTNs: The RollerNet case," *IEEE Transactions on Mobile Computing*, vol. 10, no. 12, pp. 1755–1768, Dec. 2011.

[138] I. Carreras, D. Miorandi, and I. Chlamtac, "A framework for opportunistic forwarding in disconnected networks," in *ICST International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*, San Jose, CA, USA, Jul. 2006.

[139] A. Picu, T. Spyropoulos, and T. Hossmann, "An analysis of the information spreading delay in heterogeneous mobility DTNs," in *IEEE WoWMoM*, San Francisco, CA, USA, Jun. 2012.

[140] P. Bremaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer, 1999.

[141] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, Aug. 2007.

[142] G. W. Flake, R. E. Tarjan, and K. Tsioutsiouliklis, "Graph clustering and minimum cut trees," *Internet Mathematics*, vol. 1, no. 4, pp. 385–408, 2004.

[143] "FP7 European project iTETRIS: The integrates wireless and traffic platform for real-time road traffic management solutions," http://www.ict-itetris.eu.

[144] S. Uppoor and M. Fiore, "Large-scale urban vehicular mobility for networking research," in *IEEE Vehicular Networking Conference (VNC)*, Amsterdam, Netherlands, 2011, pp. 62–69.

[145] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "SUMO – Simulation of Urban MObility: An overview," in *International Conference on Advances in System Simulation*, Barcelona, Spain, Oct. 2011.

[146] S. Uppoor and M. Fiore, "Insights on metropolitan-scale vehicular mobility from a networking perspective," in *Proceedings of the 4th ACM International Workshop on Hot Topics in Planet-scale Measurement*, ser. HotPlanet '12, Low Wood Bay, Lake District, UK, 2012, pp. 39–44.

[147] CRAWDAD, "Community resource for archiving wireless data at dartmouth," http://crawdad.cs.dartmouth.edu.

[148] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott, "Impact of human mobility on opportunistic forwarding algorithms," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 606–620, Jun. 2007.

[149] N. Eagle and A. Pentland, "CRAWDAD data set mit/reality (v. 2005-07-01)," 2005, http://crawdad.cs.dartmouth.edu/mit/reality.

[150] J. Burgess, B. N. Levine, R. Mahajan, J. Zahorjan, A. Balasubramanian, A. Venkataramani, Y. Zhou, B. Croft, N. Banerjee, M. Corner, and D. Towsley, "CRAWDAD dataset umass/diesel (v. 2008-09-14)," 2008, http://crawdad.cs.dartmouth.edu/umass/diesel.

[151] M. Piorkowski, N. Sarafijanovoc-Djukic, and M. Grossglauser, "A parsimonious model of mobile partitioned networks with clustering," in *International Conference on Communication Systems and NetworkS (COMSNETS)*, Bangalore, India, Jan. 2009.

[152] http://www.comune.bologna.it/wireless, accessed on 24-03-2014.

[153] A. Keränen, J. Ott, and T. Kärkkäinen, "The ONE simulator for DTN protocol evaluation," in *ICST International Conference on Simulation Tools and Techniques (SIMUTools)*, Rome, Italy, Mar. 2011.

[154] NS-3, "Network simulator," http://www.nsnam.org.

[155] Y. Z. J. Sun, C. Zhang and Y. M. Fang, "An identity-based security system for user privacy in vehicular ad hoc networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 9, pp. 1227–1239, Sep. 2010.

[156] E. Johnson, H. Cruickshank, and Z. Sun, "Providing authentication in delay/disruption tolerant networking (dtn) environment," in *Personal Satellite Services*. Springer, 2013, pp. 189–196.

[157] M. V. Barbera, J. Stefa, A. C. Viana, M. D. de Amorim, and M. Boc, "VIP delegation: Enabling vips to offload data in wireless social mobile networks," in *IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)*, Barcelona, Spain, Mar. 2011.

[158] J. Whitbeck, M. D. de Amorim, V. Conan, and J.-L. Guillaume, "Temporal reachability graphs," in *ACM Mobicom*, Istanbul, Turkey, Aug. 2012, pp. 377–388.

[159] R. Jain, D.-M. Chiu, and W. R. Hawe, *A quantitative measure of fairness and discrimination for resource allocation in shared computer system*. Eastern Research Laboratory, Digital Equipment Corporation, 1984.

[160] F. Benbadis, F. Rebecchi, F. Cosnier, M. Sammarco, M. Dias de Amorim, and V. Conan, "Demo: opportunistic communications to alleviate cellular infrastructures: the fp7-moto approach," in *ACM CHANTS*, Maui, HI, 2014, pp. 85–88.

[161] ——, "Demo: D2d rescue of overloaded cellular channels," in *ACM MobiSys*, Florence, Italy, May 2015.

[162] Z. Honig, "Verizon demonstrating LTE Multicast during Super Bowl XLVIII (hands-on video)," http://www.engadget.com/2014/01/29/verizon-lte-multicast/.

[163] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. MIT Press, 1998.

[164] B. Sas, E. Bernal-Mor, K. Spaey, V. Pla, C. Blondia, and J. Martinez-Bauset, "Modelling the time-varying cell capacity in lte networks," *Telecommunication Systems*, vol. 55, no. 2, pp. 299–313, 2014.

[165] 3GPP, " TS 36.213 V11.2.0 Rel.11: Evolved universal terrestrial radio access (e-utra); physical layer procedures," 2013.

[166] R. Bhatia, L. Li, L. Haiyun, and R. Ramjee, "ICAM: integrated cellular and ad hoc multicast," *IEEE Trans. on Mobile Computing*, vol. 5, no. 8, pp. 1004–1015, Aug 2006.

[167] K. Doppler, M. Rinne, C. Wijting, C. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to lte-advanced networks," *IEEE Comm. Mag.*, vol. 47, no. 12, pp. 42–49, Dec 2009.

[168] G. Tan, S. Ma, D. Jiang, Y. Li, and L. Zhang, "Towards optimum hybrid arq with rateless codes for real-time wireless multicast," in *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*. IEEE, 2012, pp. 1953–1957.

[169] M. Rahman, H. Cheng-Hsin, A. Hasib, and M. Hefeeda, "Hybrid multicast-unicast streaming over mobile networks," in *IFIP Networking*, Trondheim, Norway, June 2014, pp. 1–9.

[170] C. Boldrini and A. Passarella, "Hcmm: Modelling spatial and temporal properties of human mobility driven by users' social relationships," *Computer Communications*, vol. 33, no. 9, pp. 1056–1074, 2010.

[171] N. Baldo, "The ns-3 lte module by the lena project."

[172] D. Azzarelli, E. Pierattelli, A. Marchetto, L. D'Orazio, F. Rebecchi, A. Passarella, R. Bruno, and G. Mainetto, "Description and development of moto simulation tool environment – release b," 2015, http://cordis.europa.eu/docs/projects/cnect/9/317959/080/deliverables/001-D512v2Ares2015964509.pdf.

[173] W. Kermack and A. McKendrick, "Contributions to the mathematical theory of epidemics – I," *Bulletin of mathematical biology*, vol. 53, no. 1, pp. 33–55, 1991.

[174] L. Valerio, R. Bruno, and A. Passarella, "Adaptive data offloading in opportunistic networks through an actor-critic learning method," in *ACM CHANTS*, Maui, HI, Sep. 2014.

[175] W. Fleming and R. Rishel, *Deterministic and stochastic optimal control*. Springer, 1975.

[176] V. G. Boltyanskii, R. V. Gamkrelidze, and L. S. Pontryagin, "The theory of optimal processes. I. The maximum principle," DTIC Document, Tech. Rep., 1960.

[177] D. Grass, J. P. Caulkins, G. Feichtinger, G. Tragler, and D. A. Behrens, *Optimal control of nonlinear processes: with applications in drugs, corruption, and terror*. Springer, 2008.

[178] K. Soetaert, J. Cash, and F. Mazzia, "Package bvpSolve, solving boundary value problems in R," *Journal of Statistical Software*, vol. 33, no. 9, pp. 1–25, 2010.

[179] X. Wang, M. Chen, Z. Han, T. T. Kwon, and Y. Choi, "Content dissemination by pushing and sharing in mobile cellular networks: An analytical study," in *Mobile Ad-hoc and Sensor Systems (MASS), 2012 IEEE 9th International Conference on*. IEEE, 2012, pp. 353–361.

[180] X. Zhuo, W. Gao, G. Cao, and S. Hua, "An incentive framework for cellular traffic offloading." *IEEE Transactions on Mobile Computing*, vol. 13, no. 3, pp. 541–555, 2014.