Communication centrée sur les utilisateurs et les contenus dans les réseaux sans fil

Zheng Chen

NNT : 2016SACLC092

Thèse de doctorat de l'Université Paris-Saclay préparée à CentraleSupélec

Ecole doctorale n°580
Sciences et technologies de l'information et de la communication
Spécialité de doctorat : Réseaux, information et communications

par

Mme Zheng Chen

Communication Centrée sur les Utilisateurs et les Contenus dans les Réseaux sans Fil

Thèse présentée et soutenue à Gif-sur-Yvette, le 16 décembre 2016.

Composition du Jury :

M.	Hikmet Sari	Professeur Université Paris-Saclay	(Président du jury)
M.	Jean-Marie Gorce	Professeur INSA Lyon	(Rapporteur)
M.	Wan Choi	Professeur Associé KAIST	(Rapporteur)
M.	Giuseppe Caire	Professeur TU Berlin	(Examinateur)
M.	Philippe Ciblat	Professeur Telecom ParisTech	(Examinateur)
M.	Thrasyvoulos Spyropoulos	Professeur Associé EURECOM	(Examinateur)
Mme.	Mari Kobayashi	Professeur CentraleSupélec	(Directrice de thèse)
M.	Marios Kountouris	Chercheur Principal Huawei Technologies	(Encadrant de thèse)

Acknowledgments

This three-year PhD is full of adventures and challenges at the same time. Everything went differently from what I was expecting, yet nothing ended as what I was fearing for. From my perspective, PhD study is not just about learning new things and applying them in our research fields. We also learn to deal with all the unexpected changes and frustration that might occur at any moment in our research and in other aspects of our lives. During these three years, I grew up as a researcher and as a person, with a lot of ups and downs, embracing numerous magic moments and cruel reality. When this journey finally comes to the end, I feel extremely grateful for all the chain reactions one after another that have brought me where I belong. For that, I would like to give my deepest gratitude to all the people who have accompanied me through this part of my life and offered me their kindest support.

First of all, I would like to thank my PhD supervisor Marios Kountouris, for being a very wise, cool and supportive supervisor, who gives inspiration and guide instead of giving instructions. I really enjoyed all the brainstorming discussions, from which I learned thinking about research problems from different perspectives and in a broader picture. I would also like to thank my PhD director Mari Kobayashi, for her support and help during these years. My sincere thanks to all the jury members for serving in my PhD committee, especially Hikmet Sari, for his kindness and support in the Telecom department.

I would like to give my special thanks to Tony Quek, for inviting me to visit his group in SUTD, which turns out to be a life-changing decision in many ways. And my other collaborators, Jemin Lee and Vangelis Angelakis, with whom I had very nice collaboration experience and learned many things about research. Beaucoup de remerciement à Cathérine, José, Huu-Hung pour leur gentillesse et leur aide à résoudre tous les problèmes que j'avais renconté à l'école. Many thanks to my officemates, Matha, Salah Eddine, Bakarime, and all my friends, previous and current colleagues, Chao, Chien-Chun, German, Meryem, Andrés, Victor, Asma, Maialen, Kenza, Fei, Azary..., for making every workday full of laughter and interesting discussions.

I would like to thank my parents for their unconditional love and support. I wouldn't have made it this far without their encouragement and sacrifice. Thanks to the beautiful sea and sky in Crete, my mind remained peaceful during the thesis writing, which was supposed to be a tough period for most of the last-year PhD students. At last, I would like to thank Nikos, my love and the other half of my soul, for every moment we have lived together.

Contents

Acronyms

AC	access control. 51
ASE	area spectral efficiency. xiii, xiv, 17–19, 29, 35, 40, 41, 44, 45, 47–50, 53–55
BPP	binomial point process. 103
BS	base station. xiii, 18, 21, 27, 28, 35–37, 43, 50–52, 55, 101, 130
CCDF	complementary cumulative distribution function. 27
CoMP	coordinated multi-point. iii, 26, 100, 101
CS/CB	coordinated scheduling and coordinated beamforming. 101
CSI	channel state information. 22, 38, 39
CSMA	carrier sense multiple access. 21, 28
D2D	device-to-device. i, iii, xiii, xiv, xvii, 18–29, 35–42, 44–55, 57, 68, 78, 81–92, 94, 95, 100, 124, 129, 130
DoF	degree of freedom. 132
DTMC	discrete time markov chain. 63, 64
FDMA	frequency division multiple access. 104

GZ guard zone. 51

HCP hard-core process. 28

IRM independent reference model. 24, 25

JT joint transmission. 101, 104, 106–109, 111, 114, 118

LCD largest content diversity. xv, 103–106, 108, 111, 113–116, 119–122, 130

LFU least frequently used. 23, 26

LRU least recently used. 23

M2M machine-to-machine. 68

MAC medium access control. 21

MBS macro base station. 17, 22

MIMO multiple-input multiple-output. 131, 132

MPC most popular content. xv, 84–86, 103–105, 108, 111–116, 118–122, 130

MPR mutipacket reception. 29, 57, 58, 60

ODSA optimal dual-solution searching. 85

PCP Poisson cluster process. 28, 132

PDF probability density function. 88, 97

PGFL probability generating functional. 61, 144

PHP Poisson hole process. 22, 28, 38

PHY physical layer. iii, 23, 26, 27, 101, 124, 130

PMF probability mass function. 102

PPP Poisson point process. xiii, xv, 27, 29, 30, 36, 38–40, 42–45, 50, 58, 60–62, 83, 84, 87, 88, 102, 109, 117, 144

PT parallel transmission. 101, 104, 106, 107, 111, 114, 117, 118

PT-OS parallel transmission with orthogonal spectrum assignment. 106–108, 110, 111

PT-SS parallel transmission with successive decoding based spectrum sharing. 106–111

QoE quality-of-experience. 23

QoS quality-of-service. 30, 65, 130

RRM radio resource management. 28, 124, 132

RSS received signal strength. 38

SBS small base station. iii, xv, 17–19, 24, 26–28, 30, 82–92, 94, 95, 100–121, 124, 130–132

SC small cell. 82–90, 92, 94

SCDP successful content delivery probability. 101, 106–111, 117, 118

SCN small cell network. iii, xv, 17, 19, 24, 27, 100–103, 106, 112, 113, 115, 116, 121, 124, 130

SIC successive interference cancellation. 106, 107, 109, 111

SINR signal-to-interference-plus-noise ratio. 23, 27, 55, 58, 60, 63, 87, 95, 107

SIR signal-to-interference ratio. xiii, 12, 22, 28, 29, 35, 37, 49, 50, 55, 78, 87, 100, 108–110, 118–120, 129, 132

SNM shot noise model. 25

TDMA time division multiple access. 104

VoD video-on-demand. 18

List of Figures

List of Tables

Resumé en Français

Contexte et Motivation

La prolifération des appareils mobiles et le développement des services de téléphonie mobile mettent une pression énorme sur la capacité et l'efficacité des réseaux sans fil actuels. De la plus récente rapport annuel publié par Cisco, le trafic de données mobiles a augmenté de 4000 fois pendant les 10 dernières années, et il est très susceptible d'augmenter de près de huit fois entre 2015 et 2020, où 75 % des données trafic sera vidéo [1]. Les méthodes traditionnelles pour augmenter la capacité de réseau tels que plus de spectre et plus d'efficacité spectrale ne seront suffisantes par rapport à leur gains potentiels. Un moyen efficace d'augmenter la capacité du réseau est lié à l'efficacité spatiale du spectre, qui se réalise par réduire la distance émetteur-récepteur et permettre les communications simultanées à courte distance pour augmenter l'efficacité spatiale du spectre.

Le déploiement d'équipements hétérogène, les stations pico et femto par exemple, a été introduite comme une extension des réseaux cellulaire actuels, afin d'améliorer le débit du réseau et d'augmenter la réutilisation spatiale des ressources de communication [2, 3]. Les petites stations de base peuvent non seulement réduire le trafic des données sur les stations de base macro, mais aussi servir les utilisateurs mobiles dans de petites cellules avec moins de délai. Bien que les réseaux aux petites cellules ont attiré une énorme attention pour le potentiel d'augmenter la capacité du réseau à faible consommation d'énergie, il y a beaucoup de défis sur le déploiement, y compris l'auto-organisation, le backhauling, le handover et la gestion des interférences [4].

A part du déploiement des réseaux à petites cellules, la communication centrée sur les utilisateurs a également apparue comme une technique promettant pour décharger le trafic de données cellulaire. La communication *device-to-device* (D2D) a été identifié comme une des technologies de rupture dans la conception de la future cinquième génération (5G) des réseaux cellulaires [5]. L'idée d'intégrer la communication D2D dans les réseaux cellulaires est de permettre la communication directe entre deux utilisateurs à proximité sans passer par les stations de base ou le cœur du réseau [6]. Cette conception a le potentiel de gérer les communications locales plus efficace et de décharger le trafic des données sur les réseaux cellulaires. La communication à courte

distance permet également l'existence de multiples liaisons D2D simultanées sans appliquer techniques d'accès multiple orthogonal. Par conséquent, intégrer D2D dans les réseaux cellulaires améliore la réutilisation des ressources spatiales. Il a de nombreux autres avantages tels que la couverture améliorée et le délai réduit, ainsi l'opportunité de permettre des nouveaux services basés sur la communication locale. Une vue d'ensemble des services de proximité D2D est donnée dans [7], qui présente des activités de normalisation 3GPP et des challenges principales dans la conception des réseaux cellulaires avec D2D intégrée. En parallèle avec les progrès de la standardisation de D2D, les recherches théoriques en cours sur D2D intégrée dans les réseaux cellulaires couvre de nombreux aspects différents de la conception du réseau et l'analyse des performances en utilisant des approches diverses [8]. Cependant, à cause du développement des applications informatiques, par exemple, la vidéo à demande (VoD), les réseaux sociaux et le partage de contenu, les demandes de vidéo deviennent la majorité du trafic de données à travers du réseau sans fil. Il est important d'être conscient des propriétés suivantes sur les services liées aux contenus vidéos;

- Le haut degré de réutilisation asynchrone de contenu, c'est-à-dire, le même contenu peut être consulté par les différents utilisateurs d'une manière asynchrone. En général, le délai entre les demandes est trop grand pour appliquer la méthode multi-diffusion. Par conséquent, les demandes seront traitées individuellement.

- La corrélation spatiale et sociale sur les demandes de vidéo. Les préférences des utilisateurs par rapport aux contenus demandés sont souvent spatialement corrélées, par exemple, les étudiants dans le même campus sont très susceptibles de demander les vidéos de catégorie similaire. En raison de leurs liens sociaux, à savoir, leur connectivité sur les réseaux sociaux, un contenu vidéo vu par un utilisateur influent peut facilement se propager dans une petite communauté d'utilisateurs connectés sur les réseaux sociaux qui sont aussi physiquement à proximité.

En conséquence de ces propriétés, si on traite la demande de chaque utilisateur indépendamment, il y aura la transmission répétée du même fichier vidéo à plusieurs utilisateurs à proximité, qui entraîne énorme gaspillage d'énergie et de ressources spectrales. Par conséquent, caching aux périphéries du réseau, par exemple aux petites stations de base et aux dispositifs des utilisateurs, est apparue comme une solution potentielle pour distribuer du contenu vidéo sans fil d'une façon plus efficace.

Les algorithmes de remplacement pour la mise en cache de document Web sont parfois réutilisables dans les systèmes de mise en cache sans fil. Cependant, nous devons être conscients des différences fondamentales entre les réseaux de caching filaires et sans fil. Récemment, ils ont apparu un grand nombre de recherche qui envisagent à étudier le caching proactif au bord de réseau basé sur la prédiction de demandes de contenu par les utilisateur. Les contenus qui

sont très susceptibles d'être demandés sont préchargées à partir du réseau de base pendant les heures creuses et mis en cache aux petites stations de base ou aux appareils avant d'être demandé par les utilisateurs [9]. Bien que l'idée du caching proactif semble simple et clair, il reste de nombreuses questions fondamentales auxquelles il faut répondre afin de comprendre la conception du réseau de caching sans fil avant de procéder à la mise en œuvre des algorithmes de mise en cache.

Dans cette thèse, notre intérêt principal est d'étudier les méthodes de déchargement du trafic cellulaires par la communication centré sur les utilisateurs et les contenus dans les réseaux sans fil. Dans la suite de ce chapitre introductif, nous donnons un aperçu général sur les direction de recherche dans cette thèse, à savoir, communication D2D dans les réseaux cellulaires et la mise en cache proactive au bord de réseau. Premièrement, nous présentons brièvement l'état de l'art de ces directions de recherche et les motivations pour les thèmes de recherche présentés dans cette thèse. Nous donnons aussi une introduction concise sur l'outil mathématique de base utilisée dans cette thèse pour l'analyse de la modélisation et de la performance topologie du réseau - *la géométrie aléatoire*. Deuxièmement, nous présentons le plan de cette thèse et les publications au cours de cette thèse.

Communication D2D dans les Réseaux Cellulaires

Comme mentionné au début de ce chapitre, l'intégration de la communication D2D aux réseaux cellulaires présente de nombreux avantages tels que l'amélioration de l'efficacité spatiale du spectre, la latence réduite et le déchargement du trafic cellulaire, etc. Malgré les avantages, il y a encore des challenges à faire face pour avoir l'implémentation réussie, y compris la gestion des interférences, la possibilité d'auto-organisation, la découverte des appareils du réseau, et l'allocation des ressources.

L'idée principale de la communication D2D dans les réseaux cellulaires est de permettre la communication directe entre deux utilisateurs mobiles à proximité pour échanger des informations sans passer par les points d'accès cellulaires ou le cœur du réseau. D'abord, D2D ressemble beaucoup aux réseaux ad hoc. Toutefois, l'assistance de l'infrastructure de réseau cellulaire rend D2D dans les réseaux cellulaires différent que les réseaux typiques ad hoc. Lorsque les appareils sont couverts par le réseau, la communication D2D dépends des infrastructures de réseaux cellulaires pour les fonctions de contrôle telles que la sélection du mode, l'allocation de puissance, etc. Lorsque les appareils sont hors de la couverture cellulaire, la communication entre les appareils peut se réaliser de manière ad hoc ou basé sur les clusterhead [7].

Overlay vs. Underlay

Vue par l'utilisation du spectre, D2D peuvent utiliser le spectre radiofréquence autorisé (*intra bande*) ou le spectre non autorisé (*hors bande*) [8]. La plupart des travaux existants sur les communications D2D sont basé sur *intra-bande* D2D et surtout sur D2D *undelaid* dans les réseaux cellulaires où les utilisateurs D2D réutilisent le spectre radiofréquence autorisé d'une manière opportuniste [10–14]. Dans ce type de réseaux, les liaisons cellulaires reçoivent l'interférence inter-couche des transmissions D2D concurrentes, alors que les liaisons D2D reçoivent l'interférence de la couche D2D et de la couche cellulaire. Par conséquent, il nécessite d'avoir une gestion appropriée de l'accès D2D afin d'atteindre une meilleure efficacité spatiale du spectre sans nuire considérablement la qualité des liaisons cellulaires.

Une autre option pour *intra bande* D2D est D2D *overlaid* des réseaux cellulaires, ce qui signifie que les utilisateurs D2D et les utilisateurs cellulaires utilisent les ressources du spectre orthogonales, ainsi ne ne créent pas d'interférence entre eux. Dans ce type de réseau, le gain de l'allocation des ressources joue le rôle le plus important afin d'attribuer de façon optimale les ressources radio entre les utilisateurs cellulaire et les utilisateurs D2D [15, 16].

Gestion d'Interférence

L'intégration de D2D dans les réseaux cellulaires nécessite de nouveaux protocoles pour gérer le partage des ressources entre les communications D2D et cellulaires. En particulier, dans les réseaux cellulaires avec D2D underlaid, les méthodes de la gestion d'interférence sont très importantes pour atteindre le gain de réutilisation du spectre sans nuire à l'expérience utilisateur cellulaire, qui est l'une des matières de base dans cette thèse.

La majorité des études sur la gestion d'interférence existantes réside dans les aspects suivants:

- **Contrôle de puissance.** Le contrôle de puissance est une technique efficace et largement utilisé dans le système sans fil pour l'atténuation des interférences. En ajustant la puissance de transmission de chaque utilisateur D2D en fonction de la densité des nœuds du réseau et l'état du canal, nous pouvons non seulement contrôler le niveau d'interférence dans l'ensemble du réseau, mais aussi l'interférence causée par chaque utilisateur D2D à ses récepteurs voisins co-canal [17–21]. Plusieurs stratégies de contrôle de puissance pour les réseaux D2D ont été développées et évaluées en utilisant le modèle de déploiement de réseau déterministe pour l'optimisation de différents métriques de performance [22–27].

- **Contrôle d'accès distribué**. Le mécanisme d'accès au canal au couche MAC, connue pour le protocole d'accès multiple, permet de multiples nœuds d'accéder aux mêmes ressources physiques sans la nécessité de

contrôle global. Les protocoles de contrôle d'accès tels que Carrier Sense Multiple Access (CSMA) et ALOHA sont des techniques souvent utilisées dans les réseaux de capteurs sans fil et les réseaux ad hoc pour atténuer les interférences entre les transmissions simultanées à proximité [28, 29]. Le concept de la zone de garde par CSMA inspire également des études récentes sur D2D intégrée dans les réseaux cellulaires à considérer l'atténuation des interférences en mettant des zones d'exclusion autour des utilisateurs actifs [30–33]. Avec l'évaluation sur la qualité du canal, un système ALOHA avancé appelé *opportuniste ALOHA* a été proposé dans [34] et peut être facilement appliqué dans les réseaux D2D. Des politiques similaires sur l'ordonnancement distribué et opportunistes sont également applicables dans les réseaux cellulaires avec D2D intégré [35, 36] .

- **Sélection de mode.** Pour les utilisateurs qui peuvent soit fonctionner en mode D2D soit en mode cellulaire, de décider quel mode pour sélectionner en fonction de la charge du réseau instantanée, l'état de canal et le niveau d'interférence est essentielle pour le débit du réseau [37, 38]. La sélection de mode peut être soit semi-statique ou dynamique, selon que le critère de sélection est basée sur le l'évaluation long-terme ou l'état du réseau instantané. La sélection de mode dynamique est souvent accouplée avec le contrôle de puissance de telle sorte que les liens cellulaires et les lien D2D peuvent partager les ressources d'une manière efficace[39, 40].

Contrôle d'Accès D2D Distribué

Dans un cellule macro avec une grande quantité de nœuds D2D, si la channel station information (CSI) globale est disponible aux stations de base, on peut appliquer les méthodes de contrôle centralisée de puissance sur les liaisons cellulaires et D2D, ce qui donne la performance optimale [17]. Cependant, le partage des CSI globale pourrait introduire l'entête de signalisation trop lourde dans un réseau dense avec une grande quantité de dispositifs D2D.

Sans la connaissance de CSI globale, un système de contrôle d'accès simple mais efficace pour les nœuds D2D est le contrôle d'accès opportuniste, c'est à dire, un émetteur D2D est autorisé à être actif s'il satisfait à un critère prédéfini. L'avantage de cette type de contrôle d'accès est que, la décision de chaque nœud sur son accès au réseau peut être effectué de façon indépendante sur la connaissance de sa propre condition, à savoir, le gain du canal, l'interférence reçue en provenance des nœud intervenants les plus proches ou la combinaison des deux. Un modèle de réseau aléatoire pour les réseaux cellulaires avec D2D underlaid est proposé dans [17], et le contrôle de d'accès D2D sur la qualité du canal a été étudié. Les techniques de contrôle d'accès et d'activation opportunistes pour les réseaux femto à deux couches sont proposés dans [41], là, contrairement à [17], la connaissance de rapport signal

sur interférence (SIR) est exploitée pour augmenter encore le débit total. L'ordonnancement basé sur l'interférence et le canal pour les réseaux ad hoc est examiné dans [42], et un ordonnanceur distribué pour les réseaux d'égal à égal ad hoc est proposé dans [43]. Pourtant, aucun de ces ouvrages existants sur la planification à seuil tient compte du de l'interférence totale que chaque liaison reçoit de toutes les transmissions simultanées. En outre, l'analyse des performances et l'optimisation de l'accès opportuniste basé sur l'information SIR n'a pas été pris en considération.

Zones de Garde Cellulaire

Les zones de garde (régions d'exclusion) autour de récepteurs cellulaires ont été considérés dans les réseaux cellulaires avec D2D underlaid comme un moyen d'augmenter le débit de D2D sans dégrader significativement la qualité de service du réseau cellulaire [44–47]. Les protocoles basés sur la zone de garde cause la dépendance entre les émetteurs de différentes couches, à savoir, les stations de base appartenant à différents niveaux présentent répulsion. Les résultats récents sur *Poisson Hole Process* (PHP) peut être utilisée pour calculer les interférences et la probabilité de couverture dans les réseaux à deux niveaux avec les zones de garde [48]. Un modèle de réseau cognitif où aucun utilisateur secondaire peuvent se situer dans les zones de garde des utilisateurs primaires est considérée dans [49], où des bornes sur la probabilité de panne sont fournis. Une analyse similaire pour les réseaux cellulaires hétérogènes avec les dépendances intra-couche et inter-couche peut également être trouvée dans [50]. Néanmoins, aucun de ces travaux ont considéré le contrôle d'accès D2D décentralisée et l'activation de la liaison combinée avec les zones de garde cellulaires afin d'offrir la garantie minimum sur la qualité de la liaison cellulaire tout en optimisant la performance du couche D2D.

Protocole Cognitive Sensibles au Délai

Dans la majorité des études existantes sur les réseaux cellulaires avec la communication D2D underlaid ou les réseaux cognitifs avec le partage de canal, la mesure de la qualité des liaisons de communication est basée sur l'hypothèse que les émetteurs ont toujours un paquet à transmettre. Avec cette hypothèse, la probabilité de couverture basé sur le rapport signal sur interférence plus bruit est souvent analysée comme le paramètre principal pour évaluer la performance du réseau.

En plus de la probabilité de couverture, le délai subi par les utilisateurs cellulaires est un autre critère important pour évaluer la qualité d'expérience des utilisateurs, en particulier le délai de file d'attente lorsque les nœuds ont les arrivées de paquets en rafale. En utilisant des techniques de la théorie des files d'attente, nous pouvons caractériser le délai de file d'attente dans certains cas [51]. En proposant un protocole d'accès partagé entre les nœuds primaire et secondaire selon la file d'attente dans le nœud primaire (cellulaire), nous

pouvons optimiser la performance du réseau sous les contraintes de délai sur l'utilisateur primaire. Au meilleur de notre connaissance, l'optimisation du débit dans les réseaux D2D étant conscient du délai de l'utilisateur cellulaire n'a jamais été étudiée dans la littérature.

Caching Proactif au Bord du Réseau

Caching n'est pas une idée nouvelle en informatique, mais récemment, les recherches sur les réseaux sans fil a commencé à envisager d'introduire des capacités de caching aux périphéries du réseau. Ce qui rend le caching sans fil différent du caching web est que le succès de la livraison de contenu dans les réseaux sans fil dépend fortement du gain de canal de transmission de contenu. Il existe deux grandes catégories de techniques de caching dans les deux réseaux filaires et sans fil, à savoir *caching réactif* et *caching proactif*. La plupart des algorithmes existants sur le remplacement de cache, comme LFU et LRU, appartiennent à la catégorie du caching réactif où le contenu mis en cache sont mises à jour au moment où une donnée a été demandée à partir du cache. Le concept de caching proactif dans les réseaux sans fil, en particulier aux périphéries du réseau, est apparue récemment et a reçu beaucoup d'attention pour son grand potentiel dans la réduction du trafic de données sans fil et l'amélioration l'efficacité énergétique. L'idée clé de la mise en cache proactif sans fil est de précharger le contenu pendant les heures creuses avant d'être demandé localement par les utilisateurs [9, 52–55]. Caching dans les réseaux sans fil, comme un moyen d'exploiter le degré élevé de réutilisation de contenu asynchrone causé par des applications centrées sur l'information, présente de nombreux avantages prometteurs tels que le déchargement du trafic de données cellulaires et de la consommation d'énergie réduite. En stockant de façon proactive le contenu à proximité du réseau bord où les utilisateurs locaux partagent les préférences de contenu similaires, les transmissions répétées du même contenu à partir du réseau de base aux utilisateurs locaux, sont évités. En particulier, l'introduction de l'unité de caching dans les réseaux à petites cellules réduit également la charge de trafic de backhaul, offre ainsi un meilleur service aux utilisateurs de petites cellules sous les contraintes de capacité de backhaul [56]. Lorsqu'un utilisateur demande un fichier déjà stocké dans les stations de base couvrantes, la latence de service est largement réduite car il n'a pas besoin de passer par le backhaul pour récupérer le contenu à partir de serveurs distants. En outre de caching dans les petites cellules, l'espace de stockage sur les appareils des utilisateurs peut également être exploitée pour stocker les contenus. Dans [57] les auteurs donnent un aperçu éclairant sur les idées fausses et les barrières commerciales relatives au caching dans les futurs réseaux sans fil.

Prédiction des Demandes

La prédiction des demandes des utilisateurs est principalement basé sur la popularité de contenu en ligne, par exemple, des vidéos YouTube les plus populaires sont très susceptibles d'être demandés à nouveau. La modélisation de la popularité de contenu globale/locale est très important pour la performance des systèmes de mise en cache sans fil. Dans la littérature du réseau centré sur les information, une hypothèse largement utilisé est le *independent reference model* (IRM), ce qui suppose que la probabilité d'une demande d'un certain contenu est constant et indépendant du passé [58, 59] Dans la littérature, une hypothèse intensivement utilisée est que nous avons une parfaite connaissance de la popularité du contenu, qui suit la distribution Zipf, à savoir, la popularité du i-ème fichier le plus populaire est $p_i = \frac{\Omega}{i^\gamma}$, où $\Omega = \left(\sum_{j=1}^{N} j^{-\gamma} \right)^{-1}$ est le facteur de normalisation et γ est le paramètre de forme de la distribution Zipf, qui définit le niveau de concentration sur les demandes des utilisateurs [60]. Lorsque γ est élevé, cela signifie que la plupart des demandes sont générées à partir d'un certain nombre de fichiers les plus populaires.

En plus de la popularité du contenu global, certains travaux existants considèrent les préférences des utilisateurs spécifiques et l'impact sur les décisions de mise en cache optimales [61, 62]. Sachant que les demandes des utilisateurs sont souvent spatialement et temporellement corrélés, nous pourrions avoir les observations suivantes concernant la production et la distribution des demandes de contenu:

- La distribution de popularité globale peut être tout à fait différente de la popularité de contenu locale lorsque les utilisateurs dans la même communauté ont souvent des préférences de contenu similaires.

- La popularité d'un contenu est variable dans le temps. Le modèle IRM utilisé pour l'étude des algorithmes de mise en cache qui ne considèrent pas la localité temporelle pourrait conduire à des résultats trop simplifiées.

La corrélation spatiale des demandes des utilisateurs peut être estimée par des méthodes d'apprentissage afin de prédire avec plus de précision les demandes des utilisateurs. Dans [63] et [64] un modèle Shot Noise (SNM) est considéré pour modéliser la génération des demandes des utilisateurs avec la localité temporelle et géographique sur la popularité du contenu. Tenant compte des liens sociaux entre les utilisateurs à proximité, le contenu de diffusion parmi les utilisateurs D2D connectés sur les réseaux sociaux est étudié dans [65], avec un algorithme proposé pour maximiser le déchargement du trafic cellulaire offert par la communication D2D .

Placement de Contenu

Caching dans les réseaux sans fil a été discuté dans de nombreux scénarios différents concerne par rapport à où mettre les contenus, par exemple, aux petites stations de base cellulaire [52, 66, 67], aux dispositifs d'utilisateur [68, 69], et la combinaison des deux [54, 70]. Dans [53] la distribution optimale du contenu dans les cas de caching codé et non codé sont étudiés dans un réseau femtocaching. Dans [71] le compromis entre le débit et la probabilité d'interruption est discuté dans les réseaux de caching aux appareils avec la communication D2D activée. Sous le même modèle de réseau, dans [55] la distance de collaboration optimale a été étudiée en fonction des paramètres du modèle. Dans [72] une stratégie de caching aléatoire est considéré pour les appareils aléatoirement distribués qui peuvent servir les utilisateurs à proximité. Un algorithme a été proposé pour trouver les probabilités de caching optimales. La mobilité des dispositifs des utilisateurs permet l'échange et la diffusion des contenus stockés, mais aussi affecte la transmission réussie du contenu [73]. Dans [54] la collaboration entre femtocaching et la communication D2D est présentée pour la mise en cache distribuée dans les femtocells avec une capacité de backhaul faible. Dans un réseau de caching sans fil où les capacités de stockage sont autorisées dans les appareils des utilisateurs, les helpers et les petites stations de base, les probabilités optimales de caching ont été étudiées dans [74]. Dans [70] un problème d'optimisation conjointe dans les réseaux de caching aux petites cellules et aux appareils est étudiée, qui vise à déterminer les politiques optimales de caching et transmission qui minimisent une fonction de coût.

Dans une topologie de réseau statique sans tenir compte du canal à évanouissement aléatoire lors de la phase de transmission de contenu, le placement de contenu optimal peut être déterminé par résoudre les problèmes d'optimisation combinatoire avec certaines contraintes [75–77]. Dans les réseaux stochastiques, les stratégies de mise en cache proactive couramment utilisés se concentrent principalement dans les cas suivants:

- **Cache les contenus les plus populaires partout**. Il est similaire à la politique *least frequently used* (LFU) dans le cache web, donne une performance optimale avec les stations de base qui ont les régions de couverture isolées.

- **placement de cache probabilistes**. Lorsque les petites stations de base ont des zones de couverture qui se chevauchent, les utilisateurs peuvent être servis par plusieurs petites stations de base. Le ratio de réussite du cache peut être améliorée par l'adoption d'un *placement de cache probabiliste* qui aide à améliorer la diversité des contenus dans les caches locales[78]. Cette stratégie est également adapté pour le cas de caching D2D où des appareils sont très susceptibles d'être au sein de la distance de communication de plus q'un des appareils en voisinage [74, 79]. L'idée de base du caching probabiliste est que chaque nœud décide de stocker les contenus avec une certaine distribution de probabilité par rapport à

sa capacité maximale du cache. Le détail de cette stratégie est décrite dans [78], dénommé comme *optimal geographical caching*. Récemment, la probabilité de la transmission réussie et la densité de réception réussie ont été considérés comme des objectifs d'optimisation alternatives, qui donne un point de vue alternative sur le caching probabiliste optimale [80–82].

- **Caching coopérative**. Lorsque les utilisateurs sont dans la zone de couverture de plus qu'une petites stations de base, si la coopération entre les petites stations de base est autorisé, les espaces de cache parmi les petites stations de base peuvent être considéré comme en entité. On peut permettre plusieurs petites stations de base à coopérer dans l'espace de cache dans la couche PHY pour la diffusion de contenu. Par conséquent, la mise en cache doit être appliquée d'une manière qui ne soit pas " homogène "pour tous les petites stations de base, mais de façon que le gain de la coopération peut être exploitée. Des études récentes de la mise en cache sans fil avec les techniques de multipoint coordonnée (CoMP) offrent de nouvelles perspectives sur les avantages de la mise en cache pour atteindre le gain de coopération au couche PHY. Un schéma appelé le PHY caching a été proposé dans [83], qui donne les lois d'échelle asymptotique dans les réseau ad hoc. Considérant la transmission coopérative par les helpers de caching, [66] étudie la mise en cache optimale comme un moyen d'équilibrer la diversité de caching et le gain de coopération. En plus de la transmission coopérative, la coopération au niveau du cache dans les réseaux aux petites cellules peut être réalisée en tenant compte des capacités de caching de multiples petites stations de base comme une entité. Cependant, la probabilité de caching n'est plus identique pour tous les petites stations de base, ce qui nécessite un contrôle centralisé pour les décisions de placement de contenu. L'idée de la coopération au niveau du cache a été discuté dans la littérature dans différents scénarios. Dans [84], la coopération aux petites cellules avec la méthode de mise en cache avec les différents seuils est proposé afin de combiner les avantages de la mise en cache distribuée et la transmission coopérative au couche PHY. Une stratégie de caching sensitive au backhaul pour un groupe de stations de station est étudié dans [56], qui résout un problème d'optimisation pour réduire au minimum le délai de téléchargement. Néanmoins, aucun des ouvrages existants fournit des solutions efficaces pour l'utilisation du cache dans les réseaux coopérative aux petites cellules sans les algorithmes itératifs.

Géométrie Aléatoire

Les processus ponctuels spatiaux ont été intensivement utilisés dans l'étude des réseaux D2D pour leur importance dans la modélisation de la distribution des nœuds dans les réseaux à grande échelle. Les modèles de la géométrie aléatoire ont été introduits dans les réseaux sans fil dans les années 1960. Plus

tard, il y avait l'utilisation extensive de modèles aléatoires spatiaux dans les réseaux ad hoc, les réseaux de capteurs et les réseaux cognitifs. Les modèles de processus ponctuels permet la caractérisation et l'évaluation de toutes les configurations possibles de l'ensemble du réseau sans fil au lieu d'une configuration spécifique. Avec la moyenne spatiale, on peut obtenir la valeur moyenne pour certains métriques de performance, tels que la probabilité de transmission réussite et l'efficacité spatiale du spectre [85]. Des études antérieures dans [86–88] permet une analyse résoluble sur la couverture du réseau et de la connectivité basée sur la métrique de performance fondamentale - le rapport signal sur interférence plus bruit (SINR), qui ont été étendue cas avec de nombre arbitraire des stations de base [89]. Basé sur ces résultats, en utilisant la modélisation et les approches analytiques similaire , on peut caractériser la performance moyenne de réseaux cellulaires avec D2D underlaid. Les processus ponctuels sont appropriés pour modéliser la distribution des nœuds D2D dans les réseaux cellulaires pour les raisons suivantes:

- Les appareils des utilisateurs sont souvent mobiles, qui peut être facilement capté par le caractère aléatoire de points générés à partir de processus ponctuels stochastique.
- Lorsque nous considérons le contrôle d'accès D2D distribué et la gestion des interférences pour un grand nombre des appareils D2D dans les réseaux cellulaires, la performance d'une spécifique liaison D2D a moins de valeur générale que la performance moyenné spatialement sur nœuds répartis de façon aléatoire.

Le modèle spatial le plus souvent utilisé est le *Poisson point process* (PPP) homogène, qui est souvent utilisé pour modéliser la distribution de utilisateurs D2D sans corrélation spatiale. Il existe de nombreux autres processus ponctuels qui peuvent être utilisés pour modéliser le réseau stochastique avec la dépendance entre les nœuds. Par exemple, dans un réseau cognitive basée sur la zone de protection où les nœuds secondaires ne peuvent pas se trouver dans la zones de garde de nœuds primaires, la corrélation spatiale peuvent être bien capturé par un *Poisson Hole Process* (PHP) [48, 50]. Dans un réseau ad hoc avec les zones de garde entre les nœuds actifs qui impose une distance minimale entre deux nœuds à proximité, la distribution des nœuds peuvent être modélisées par un *Hardcore Process* (HCP) [90, 91]. Lorsqu'on considère l'effet de regroupement des utilisateurs, *Poisson Cluster Process* (HCP) peuvent être appliqués pour modéliser la distribution des nœuds dans ce type de réseau [92, 93]. Cependant, la complexité analytique des processus ponctuels augmente souvent avec la corrélation spatiale des nœuds. Dans certains cas, les modèles spatiaux stochastiques pourraient conduire à une analyse intraitable, nécessite donc des approximations et des hypothèses pour augmenter la traçabilité du modèle de réseau considéré.

Il est important d'être conscient que les processus ponctuels ne sont pas toujours applicable. Les propriétés fondamentales de ces modèles spatiaux sont le caractère aléatoire de la distribution des nœuds et la moyenne spatiale. Dans

un système sans fil avec l'allocation de ressources et l'ordonnancement avancé basé sur les conditions de canal instantanées des nœuds du réseau, l'analyse de la performance du réseau en utilisant les processus ponctuels devient souvent intraitable. En outre, la distribution de stations de base dans le monde réel est pas exactement aléatoire. Avec la planification soigneuses sur les locations des stations de base, la performance du réseau sera certainement mieux que la valeur attendue avec l'hypothèse que le stations de base sont distribués au hasard. Cependant, l'analyse théorique sur les modèles spatiaux stochastiques nous fournit quelques indications de base sur l'impact des paramètres du réseau sur la performance moyenne des réseaux sans fil à grande échelle. Les valeurs théoriques sur la performance du réseau peut également être considérée comme une limite inférieure de la performance qu'on pourra attendre sur le déploiement du réseau actuel.

Plan de Thèse

Cette thèse se compose de deux grandes parties. Dans la partie I, intitulé comme **Device-to-Device (D2D) Underlaid Cellular Networks**, nous nous focalisons sur la gestion des interférences et l'optimisation de l'accès D2D dans les réseaux cellulaires avec D2D underlaid. Plus explicitement,

Dans **Chapitre 2 – Distributed SIR-aware Opportunistic Access Control**, nous proposons un système de contrôle d'accès décentralisé pour la gestion des interférences dans les réseaux cellulaires avec D2D underlaid. Notre méthode combine le contrôle d'accès opportuniste basé sur le rapport signal sur interférence et les régions d'exclusion cellulaires. Les expressions analytiques et des approximations pour la probabilité de couverture sur les liens cellulaires et sur les liens D2D sont dérivés. Nous caractérisons l'impact du rayon de la zone de protection et le seuil du rapport signal sur interférence pour les liens D2D sur l'efficacité spatiale du spectre et la probabilité de couverture cellulaire, d'où nous dérivons les valeurs optimales de ces paramètres pour atteindre le débit D2D maximale en assurant une couverture minimale pour les utilisateurs cellulaires. Les simulations valident l'exactitude de nos résultats analytiques et montrent le gain de notre conception proposé par rapport aux solutions existantes.

Dans **Chapitre 3 – Priority-based Shared Access with Delay Constraints**, nous analysons un réseau d'accès partagé avec un nœud primaire (cellulaire) et les nœuds secondaire (D2D) distribués de façon aléatoire dont la distribution suit une PPP homogène. Les nœuds secondaires utilisent un protocole d'accès aléatoire qui leur permet d'accéder au canal, avec des probabilités qui dépendent de la file d'attente du nœud primaire. En supposant un système avec les arrivées de paquets aléatoire au nœud primaire et le file d'attente saturé aux secondaires, nous étudions le débit du réseau secondaire et le délai moyen primaire, ainsi que l'impact de la probabilité d'accès et la puissance de transmission pour les nœuds secondaires. Nous formulons un

problème d'optimisation pour maximiser le débit du réseau secondaire sous contraintes de délai pour le nœud primaire. Dans le cas où aucun contrôle de congestion est effectué, on donne la solution sur la probabilité optimal pour un nœud secondaire d'accéder au réseau. Nos résultats numériques montrent l'impact des paramètres du réseau sur la performance du protocole d'accès partagé proposé avec les différentes priorités.

Dans la deuxième partie de la thèse, intitulée comme **Proactive Caching at the Network Edge**, nous étudions le rôle du caching aux petites cellules et aux appareils des utilisateurs. Plus explicitement,

Dans **Chapitre 4 – Stochastic Wireless Caching Networks**, nous étudions le caching probabiliste dans les réseaux sans fil et nous visons à répondre aux deux questions suivantes: 1) où mettre les contenus dans un réseau sans fil? 2) quel objectif à considérer pour le placement de contenu optimal? Pour répondre à la première question, nous modélisons un réseau cellulaire sans fil en utilisant la géométrie aléatoire et analyser la performance de deux types de réseaux, à savoir, caching aux appareils des utilisateurs et caching aux petites cellules. Nous obtentions des expressions analytiques sur la performance de chacun scénario. Nos résultats montrent que la performance du caching par rapport à où stocker les contenus dépend fortement de la densité d'utilisateur et la distribution de la popularité des contenus.

Afin de répondre à la seconde question, nous considérons le caching probabiliste dans les réseaux stochastiques D2D. Notre objectif est de comparer les performances des probabilités de caching optimales obtenues avec deux objectifs différents: maximiser la probabilité de réussite du cache et maximiser la densité des demandes traitées avec succès par des caches locaux. Les résultats de simulation montrent que les probabilités de caching optimales obtenues par l'optimisation du débit montre un gain notable en termes de densité des demandes des utilisateurs servis avec succès, en particulier dans l'environnement avec les utilisateurs dense.

Dans **Chapitre 5 – Small Cell Cooperative Caching**, nous considérons les réseaux aux petites cellules où les petites stations de base sont regroupés en clusters disjoints. Nous proposons un schéma de caching coopérative où une partie de l'espace disponible est réservé pour stocker les contenus les plus populaires dans tous les petites stations de base, tandis que le reste est utilisé pour stocker en collaboration les différentes partitions des contenus moins populaires dans différentes petites stations de base, comme un moyen d'accroître la diversité de contenu local. Selon la disponibilité et le placement du contenu demandé, multipoint coordonnée avec soit une transmission conjointe ou la transmission parallèle est utilisé pour fournir le contenu à l'utilisateur qu'il le demande. Notre analyse montre un compromis entre la diversité de transmission et la diversité de contenu dans notre conception de caching-transmission collaborative. Nous étudions également l'utilisation optimale du cache pour deux objectifs: maximiser la performance du service de cache et maximiser l'efficacité énergétique. Nos résultats de simulation montrent que le schéma

proposé permet d'obtenir un gain combiné de la coopération au niveau du cache et au niveau du signal.

Dans la troisième partie, qui est la dernière partie de cette thèse, nous donnons les conclusions, les perspectives et les annexes.

Publications

- **Articles de Revues Internationales à Comité de Lecture**

  [J6] **Z. Chen**, N. Pappas, M. Kountouris, V. Angelakis, "Energy Harvesting in D2D Underlaid Cellular Networks with Bursty Traffic", à soumettre à *IEEE Transactions on Green Communications and Networking*, Dec. 2016.

  [J5] N. Pappas, **Z. Chen**, "On the Effect of Bursty Traffic in Caching Helpers", soumis à *IEEE Communication Letters*, Sep. 2016.

  [J4] * **Z. Chen**, M. Kountouris, "Distributed SIR-aware access control for D2D underlaid cellular networks with guard zones", soumis à *IEEE Transactions on Wireless Communications*, Jul. 2016.

  [J3] * **Z. Chen**, N. Pappas, M. Kountouris, "Probabilistic Caching in Wireless D2D Networks: Hit Optimal vs. Throughput Optimal", en révision, *IEEE Communication Letters*, Oct. 2016.

  [J2] * **Z. Chen**, N. Pappas, M. Kountouris, V. Angelakis, "On the performance of delay aware shared access with priorities", en révision, *IEEE Transactions on Mobile Computing*, Oct. 2016.

  [J1] * **Z. Chen**, J. Lee, T. Q. S. Quek, M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks", en révision, *IEEE Transactions on Wireless Communications*, Jul. 2016.

- **Articles de Conférences Internationales à Comité de Lecture**

  [C6] * **Z. Chen**, M. Kountouris, "D2D caching vs. small cell caching: where to cache content in a wireless network?", in *IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Edinburgh, United Kingdom, Jul. 2016, pp. 1-6.

  [C5] **Z. Chen**, N. Pappas, M. Kountouris, V. Angelakis, "Throughput analysis of smart objects with delay constraints", *IEEE 17th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Coimbra, Portugal, Jul. 2016, pp. 1-6.

[C4] **Z. Chen**, J. Lee, T. Q. S. Quek and M. Kountouris, "Cluster-centric cache utilization design in cooperative small cell networks", in *IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, Jun. 2016, pp. 1-6.

[C3] **Z. Chen**, M. Kountouris, "Cache-enabled small cell networks with local user interest correlation", in *IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Stockholm, Sweden, Jul. 2015, pp. 680-684.

[C2] **Z. Chen**, M. Kountouris, "Guard zone based D2D underlaid cellular networks with two-tier dependence", in *IEEE International Conference on Communication Workshop (ICCW)*, London, UK, Jun. 2015, pp. 222-227.

[C1] **Z. Chen**, M. Kountouris, "Distributed SIR-aware opportunistic access control for D2D underlaid cellular networks", in *Proc. IEEE Global Communication Conference (GLOBECOM)*, Austin, TX, Dec. 2014, pp. 1540-1545.

Chapter 1

Introduction

1.1 Background and Motivation

Mobile device proliferation and the development of mobile network services are creating tremendous pressure on the capacity and efficiency of current wireless networks. From the recent annual visual network index (VNI) reports released by Cisco, mobile data traffic has grown 4,000-fold over the past 10 years and almost 400-million-fold over the past 15 years. Mobile traffic is highly likely to have an eight-fold increase between 2015 and 2020, and 75% of the data traffic is expected to be video [1]. Traditional methods for boosting the network capacity such as seeking broader spectrum and increasing per-link spectral efficiency are either cost ineffective or with limited potential gains and scalability virtues. An effective way to increase the network capacity is by spatial reuse of spectrum resources, that is, reducing the transmitter-receiver distance and allow concurrent short-distance communication links to increase the area spectral efficiency (ASE).

Network densification with heterogeneous equipment deployment (e.g. pico and femto base stations) has been introduced as an expansion to existing macrocell networks to improve network throughput and spatial reuse of communication resources [2, 3]. Such dense deployment of self-organizing, low-power and short-range SBSs can reduce the traffic load on the macro base stations (MBSs), and serve mobile users in small cells with lower latency. Although small cell networks (SCNs) have attracted enormous attention for its great potential in increasing the network capacity with lower energy consumption, there are many technical and deployment challenges that need to be addressed, including self-organization, backhauling, handover and interference management [4]. For instance, the small cell access points require high-speed backhaul to be connected to the core network. In dense SCNs where the density of small access points will be comparable to the density of users, backhaul availability and capacity may become the performance and cost bottleneck.

Another promising technology to offload cellular traffic and increase re-

source utilization is device-centric communication, which has been identified as one of the disruptive technologies in the design of future fifth generation (5G) cellular networks [5]. The key idea of enabling device-to-device (D2D) communication in cellular networks is to allow direct communication between two mobile users in proximity bypassing the BSs or core network, while cellular infrastructure plays an active role in assisting D2D by the control functions such as synchronization, mode selection, resource allocation and scheduling [6]. D2D integrated cellular networks has the potential to handle local communication more efficiently. The short-distance transmission also allows multiple concurrent D2D links without the need of orthogonal multiple access techniques. Therefore, compared to the conventional BS-centric network structure, D2D in cellular networks brings opportunity for improved spatial resource reuse, i.e., higher ASE. Other advantages of D2D includes the enhanced link coverage and reduce end-to-end latency, while enabling new location-based services and reliable public safety communications. An overview of D2D proximity services in 3rd Generation Partnership Project (3GPP) standardization activities and of the main challenges in designing D2D-enhanced cellular standards is given in [7]. In parallel with the standardization progress of D2D, ongoing theoretical research on D2D in cellular networks covers many different aspects of network design and performance analysis using various approaches [8].

In additional to the attempt of offloading cellular traffic by handling user requests locally with shortened communication distances, as with small cells and D2D communication, another perspective is to investigate the source of wireless data traffic. Due to the development of information-centric applications, such as video-on-demand (VoD), social networks and content sharing, wireless demand for video content has become the dominant source of the exponentially growing wireless data traffic, and it has the following properties:

- High degree of asynchronous content reuse, which means that the same content can be viewed by different users in an asynchronous way. In general, the time differences among the user requests are large enough that the multicast system is not an available option to serve the asynchronous requests.
- Spatial and social correlation of content requests. Users preferences regarding the requested contents are often spatially correlated, e.g., students in the same campus are very likely to request similar categories of videos. Because of the social ties among mobile users, for instance, their connectivity on social networks, a video content viewed by an influential user can easily spread over a small community of connected users on social networks who are also physically in proximity.

Based on these facts, treating each user request independently will result in repeated transmission of the same video file to the users in proximity, which causes huge waste of communication resources. Inspired by the web caching, enabling caching capabilities at the network edge, such as SBSs and user devices, has emerged as a potential solution to exploit the content correlation

of user requests and serve local users with lower latency and energy consumption. Caching in SCNs is also a promising method to turn cache capacity into backhaul capacity so that the backhaul load can be largely reduced.

Existing web cache replacement algorithms in computer science are sometimes reusable in wireless caching systems. However, we must be aware of the fundamental differences between wired and wireless caching networks that result in different design and implementation of caching systems. Recently, there has been a considerable large amount of studies considering proactive caching at the network edge based on user request prediction. The contents that are very likely to be requested are prefetched from the core network during off-peak hours and cached at SBSs or user devices before being requested by the users [9]. The idea of proactive caching seems simple, yet there are many fundamental questions to be answered in order to bring light to the wireless caching network design before proceeding to the implementation of caching algorithms.

In this thesis, our main interest is to study cellular traffic offloading methods by considering user-centric and content-aware communications in wireless networks. In the remainder of this introductory chapter, we give a general overview of the major research directions of this thesis, namely D2D underlaid cellular networks and proactive caching at the network edge. First, we introduce briefly the state-of-art of these research directions and the motivations behind the research topics presented in this thesis. We give also a concise introduction on the basic mathematical tool utilized in this thesis for the network topology modeling and performance analysis – *Stochastic Geometry*. Second, we present the outline of this thesis and the related publications during the course of this PhD.

1.2 D2D Underlaid Cellular Networks

Enabling D2D communication in cellular networks has been proposed to exploit the physical proximity of mobile users in order to increase the network capacity by improving the spatial reuse of spectrum resources. This new paradigm is expected to bring many advantages such as the improved ASE, reduced latency and cellular data traffic offloading. However, the implementation of D2D in current cellular networks introduce conceptional and technical complications in the network design and operation, including the interference management, protocol design, node discovery and privacy issues.

The key idea of D2D communication in cellular networks is to allow direct link between two mobile users in proximity to exchange information without passing through the cellular access points or the core network. Motivated by the proximity-based services and applications, D2D technologies have been standardized by the 3GPP Release 12, mainly to support the need of public safety and commercial mobile networks. In the meanwhile, various research

in academia has been carried out both on the theoretical analysis of fundamental problems and the practical issues of supporting D2D in current cellular networks.

From the first sight, D2D shares many common features with ad hoc networks. However, the role of cellular network infrastructure makes D2D in cellular networks different from a typical ad hoc network. When the device is within network coverage, D2D communication relies on cellular infrastructure for the control functions such as mode selection, power control etc. When the device is outside the cellular coverage, the communication between devices may take place in a ad hoc manner or clusterhead-based manner [7].

1.2.1 Overlay vs. Underlay

Existing studies on D2D integrated cellular networks can be divided into two major categories, depending on the spectrum bands in which D2D communication occurs. The first is *inband* if D2D communication occurs on licensed cellular spectrum, while the second is *outband* if D2D links exploit the unlicensed spectrum [8]. Most existing works focus on *inband* D2D and especially on D2D *underlaid* cellular networks, where D2D users opportunistically access the licensed spectrum utilized by cellular users [10–14]. In these network deployments, due to spectrum sharing between D2D and cellular communications, concurrent transmissions cause both intra-tier interference among the D2D pairs and cross-tier interference between D2D and cellular users. Therefore, new interference management solutions must be proposed in order to deal with the interference issues in D2D underlaid cellular networks.

Another option for *inband* D2D is D2D *overlaid* cellular networks, which means that the D2D and cellular users are assigned with orthogonal spectrum, thus will not create interference between each other. In such network the resource allocation plays the most important role in order to optimally assign radio resources between cellular and D2D users [15, 16].

1.2.2 Interference Management

The integration of D2D in cellular networks requires new protocol designs and functionalities to manage the resource sharing between D2D and cellular communications. Particularly, in D2D underlaid cellular networks, the interference management solutions are of great importance to achieve the spectrum reuse gain without harming cellular user experience, which is one of the core subjects in this thesis.

The majority of the existing interference management studies lie in the following aspects:

- **Power control.** Power control is an effective and widely used technique in wireless system for the interference mitigation. By adjusting the trans-

mission power of each D2D user depending on the network node density, user location and the channel condition, we can not only control the overall interference level, but also the interference caused by each D2D user to its nearby co-channel receivers [17–21]. Several D2D power control strategies are developed and evaluated using the deterministic network deployment model for optimizing different performance metrics [22–27].

- **Distributed medium access control**. The channel access mechanism at medium access control (MAC) layer, known as multiple access protocol, allows multiples nodes to access the same physical channel/spectrum resources without the need of global scheduling or control. MAC protocols such as carrier sense multiple access (CSMA) and slotted ALOHA are widely used techniques in wireless sensor networks and ad hoc networks to mitigate interference among nearby concurrent transmissions [28, 29]. The concept of CSMA guard zone also inspires recent studies in D2D underlaid cellular networks to consider interference mitigation by setting interference limited areas around active cellular or D2D users [30–33]. When the BSs are equipped with multiple antennas, setting interference-limited areas would reduce the multi-user diversity, but still shows significant gain over the traditional interference management schemes. With channel quality-based evaluation and selection, an enhanced spatial ALOHA scheme called opportunistic ALOHA was proposed in [34] and can be easily extended to the case with D2D users. Similar distributed opportunistic scheduling policies, such as threshold scheduling [35] and opportunistic channel probing [36], are also applicable in D2D underlaid cellular networks.

- **Mode selection.** For users that can either operate in D2D mode or cellular mode, deciding which mode to select according to the instantaneous network load, channel condition and interference level is critical to both single-link data rate and the overall network throughput [37, 38]. The mode selection can be either semi-static or dynamic, depending on whether the selection criteria is based on long-term or instantaneous network condition and channel gains. Dynamic mode selection are often coupled with power/resource allocation so that D2D and cellular links can share the time/spectrum resources in a efficient way [39, 40].

In this thesis, we mainly consider the distributed opportunistic access control and the guard zone technique to schedule the set of D2D users that are allowed to access cellular spectrum as a means to harness the intra-tier and cross-tier interference in D2D underlaid cellular networks.

Distributed Opportunistic D2D Access Control

Different from cognitive radio network, the spectrum sharing in D2D underlaid cellular networks is managed with cellular network assistance. In a macrocell

with D2D nodes underlaying in the network, if the global channel state information (CSI) is available at the MBS, one can apply centralized power control on the cellular and D2D links based on the CSI of each link, which leads to the optimal performance and also gives an upper bound of the distributed power control and opportunistic medium access control methods. However, the global CSI sharing might introduce high signaling overhead in a dense network with large amount of D2D devices.

Without the knowledge of global CSI, a simple yet efficient access control/scheduling scheme for the D2D nodes is the opportunistic access control, i.e., a D2D transmitter is allowed to be active if it satisfies a predefined criteria. With such scheme, the access decision of each node can be made independently based on the knowledge of each transmission link itself, i.e., the channel gain, the received interference from the nearest interfering node or the combination of both. In [17], a random network model for D2D underlaid cellular networks is proposed and channel-aware power control and link activation algorithms are developed. Different opportunistic access control and link activation techniques for two-tier femtocell networks are proposed in [41]. Therein, contrary to [17], SIR knowledge is exploited to further increase the aggregate throughput. Interference-channel aware scheduling for large-scale ad hoc networks is investigated in [42], and in [43] a synchronous distributed scheduler for peer-to-peer ad hoc networks is proposed. None of these existing works on threshold-based scheduling takes into account the aggregate interference that a potential link receives from all concurrent transmissions. Moreover, performance analysis and optimization of SIR-aware opportunistic access has not considered.

Cellular Guard Zones

Guard zones (exclusion regions) around cellular receivers have been considered in D2D underlaid cellular networks as a means to boost the D2D throughput without significantly degrading the quality of cellular links [44–47]. With stochastic spatial models for the network topology modeling, guard zone based protocols lead to inter-tier dependence among transmitters, i.e. BSs belonging to different tiers exhibit repulsion. Recent results on Poisson hole process (PHP) can be used to calculate interference and coverage probability in guard-zone based two-tier networks [48]. A cognitive radio network model where no secondary users can lie within the guard zones of primary users is considered in [49], where bounds on the outage probability are provided. Similar analysis for heterogeneous cellular networks with intra-tier and inter-tier dependence can also be found in [50]. Nevertheless, none of these works have considered decentralized D2D access and link activation combined with cellular exclusion regions in order to offer minimum guarantee on the cellular link quality while optimizing the performance of D2D tier.

Delay-Aware Shared Access Protocol

In the vast majority of existing studies on D2D underlaid cellular networks or other spectrum sharing cognitive networks, the measurement of link quality is based on the assumption that the nodes are backlogged, i.e., the transmitters always have a packet to transmit. Under this assumption, the coverage probability which depends on the target signal-to-interference-plus-noise ratio (SINR) is often analyzed as a baseline metric to characterize the network performance. In additional to the coverage probability, the delay experienced by cellular users is another important criterion for the user quality-of-experience (QoE) evaluation, especially the queueing delay when the nodes have bursty packet arrivals, which cannot be evaluated directly with SINR-related metrics. By using techniques from queueing theory, we are able to characterize the queueing delay at a network node with bursty packet arrivals in some cases [51]. In a shared access network with delay-sensitive primary user, by properly designing a delay-aware access protocol for the lower-priority (secondary) users depending on congestion level of the queue in the primary node, we can optimize the network performance with respects to the primary user delay constraints. To the best of our knowledge, delay-aware throughput optimization in D2D underlaid cellular networks (or shared access network with priorities) has never been studied in the literature.

1.3 Proactive Caching at the Network Edge

Caching is not a novel idea in computer science, but recently the research on future wireless networks has started to consider introducing caching capabilities at the network edge. What makes wireless caching different from the traditional web caching is that the success of wireless content transmission depends heavily on the physical layer (PHY) content transmission. There are two major categories of caching techniques in both wired and wireless networks, namely *reactive caching* and *proactive caching*. Most of the existing cache replacement algorithms, such as least frequently used (LFU) and least recently used (LRU), belong to the reactive caching category where the cached content is updated at the moment when a data is requested from the cache. In additional to the reactive cache replacement algorithms that can be reapplied in a similar way, the concept of proactive caching in wireless networks, especially at the network edge, has emerged recently and received enormous attention. The key idea of wireless proactive caching is to prefetch content during off-peak hours before being requested locally by the end users [9, 52–55]. The implementation of proactive caching contains at least two major stages:

1. Predict user requests based on known information of the popular contents and the profiles of users such as their locations, content preferences and social ties etc.
2. Apply content placement according to a certain caching strategy and the

predicted user request pattern.

Wireless caching, as a means to exploit the high degree of asynchronous content reuse caused by information-centric applications, has many promising advantages such as cellular data traffic offloading and reduced energy consumption. By proactively storing content close to the network edge where local users might share similar content preferences, the repeated transmissions of the same content from the core network to local users, are avoided. Particularly, introducing caching unit in SCNs also reduces the backhaul traffic load, thus offers improved service to the small cell users under the backhaul capacity constraints [56]. When a local user requests for a file already cached in its covering SBS, the service latency is largely reduced since it does not have to pass through the backhaul to retrieve the content from remote servers.

Besides small cell caching, storage space on user devices can also be exploited for smart caching at the mobile side, fulfilling the demands of other devices in proximity through D2D communication. The mobility of users helps the content dissemination among a group of users who might share similar content preferences. The social ties among users who are connected on social networks and who have regular daily encounters also create opportunity for proactive social-aware caching in user devices. Despite the promising aspects of wireless caching, there is no clear conclusion yet about the design and implementation of this idea in current network architecture. In [57] the authors give an enlightening overview on the technical misconceptions and the business barriers regarding the implementation of caching techniques in future wireless networks.

1.3.1 User Request Prediction

The prediction of network-wide user requests is mainly based on the popularity of content online, e.g., the most popular YouTube videos are very likely to be requested again. In the literature of information-centric network, a widely used assumption is the independent reference model (IRM), which assumes that the probability of a request for a certain content is constant and independent of the past [58,59]. Based on the IRM, one commonly used distribution for the video content popularity is the Zipf distribution, i.e., the popularity of the i-th most popular file is $p_i = \frac{\Omega}{i^\gamma}$, where $\Omega = \left(\sum_{j=1}^{N} j^{-\gamma} \right)^{-1}$ is the normalization factor and γ is the shape parameter of Zipf distribution, which defines the concentration level of user requests [60]. When γ is high, it means that most of the requests are generated from a few most popular files.

Knowing that the user requests are often spatially and temporally correlated, we might have the following observations regarding the generation and distribution of content requests:

- The global popularity distribution can be quite different from the lo-

cal content popularity, when users in the same community share similar content preferences.

- The popularity of a content is time-varying. The classical IRM which does not consider the temporal locality might lead to over-simplified results.

In additional to the global content popularity which is often used as reference information for user request prediction, some existing works consider specific user preferences and the impact on the optimal caching decisions [61, 62]. The spatial correlation of user requests in local areas can be estimated by machine learning methods in order to predict user requests more accurately. In [63] and [64] a shot noise model (SNM) is considered to model the user request generation which captures the temporal and geographical locality of content popularity. Taking into account the social ties among users in proximity, content dissemination among D2D users connected on social networks is studied in [65], with an algorithm proposed for maximizing the cellular traffic offloading offered by D2D communication.

1.3.2 Content Placement

Caching in wireless networks has been discussed in many different scenarios regarding where to cache content, such as at small cell base stations or femto access points [52, 66, 67], at user devices [68, 69], and the combination of both [54, 70]. In [53] the optimal content assignment in the coded and uncoded cases are studied in a femtocaching network. In [71] the throughput-outage tradeoff is discussed in wireless D2D caching networks using a protocol model for the spatial scheduling of coexisting D2D links. Under the same network model, in [55] the closed-form expression is given for the optimal collaboration distance as a function of the model parameters. The impact of mobility of user devices on cache hit performance in cached D2D networks is investigated in [73]. In [54] the collaboration between femtocaching and D2D communication is presented for distributed content caching in femtocells with low backhaul capacity combined with user terminals acting as caching helpers. In a wireless caching network where caching capabilities are enabled in user devices and distributed helpers, the cache hit probability optimization in such two-tier caching system is investigated in [74]. In [70] a joint optimization problem in small cell and D2D caching network is studied, which aims at determining the optimal transmission and caching policies which minimize a generic cost function.

In a static network topology without consideration of the random fading channel during the content transmission phase, the optimal content placement can be determined by solving combinatorial optimization problems with certain constraints [53, 75–77]. In stochastic wireless networks where the caching helpers and user devices are randomly distributed, when considering the content popularity as the main reference for user request prediction, the commonly

used proactive caching strategies mainly lie in the following cases:

- **Cache the most popular content everywhere**. It is similar to the LFU replacement policy in Internet caching, gives optimal performance with non-overlapping SBS coverage or with isolated caches.

- **Probabilistic caching placement**. When SBSs have overlapping coverage areas, users have more than one potential serving SBSs. The cache hit ratio can be improved by adopting an optimal *probabilistic caching placement* policy to increase content diversity in the caches of potential serving SBSs [78]. This strategy is also suitable for the case with D2D caching where user devices are very likely to be within the communication distance of more than one neighbor devices [72, 74, 79]. The core idea of probabilistic caching is that a node caches each file in the content library with a certain probability with respect to its maximum cache capacity. The detail of this caching strategy is described in [78], referred as *optimal geographic caching*. Apart from the cache hit probability optimization, recently the transmission success probability and the density of successful reception have been considered as alternative optimization objectives, which shift the optimization study of the probabilistic caching placement from simple hit-optimal perspective to the case with more consideration on the content delivery phase [80–82].

- **Cooperative caching**. When users in the coverage area of more than one SBS, if cooperation among multiple SBSs is allowed, the cache space of the cooperative SBSs can be considered as en entity. One can allow multiple SBSs to cooperate both in cache space (i.e., cache-level cooperation), and in the physical layer for content delivery (i.e., signal-level cooperation). Therefore, the caching placement should be applied in a way that is not "homogeneous" for all the SBSs, but in a way that the base station cooperation gain can be exploited. Recent studies in wireless caching with coordinated multi-point (CoMP) techniques provide new perspectives on the benefits of caching to achieve physical layer (PHY) cooperation gain. A PHY caching scheme called cache-induced dual-layer coordinated multi-point (CoMP) was proposed in [83], providing asymptotic scaling laws of wireless ad hoc network with such scheme. Considering cooperative transmission via caching helpers, [66] investigates the optimal caching placement as a means to balance diversity and cooperation gain. In addition to signal-level cooperative transmission, cache-level cooperation in SCN can be realized by considering the caching capabilities of multiple SBSs as an entity and selectively cache more diverse contents in different SBSs in improve the cache hit probability. However, the cache probability of a certain file is no longer identical for all the SBSs, requiring local centralized control for cache placement decisions. The idea of cache-level cooperation has been discussed in the literature in different scenarios. In [84], small cell cooperation with threshold-based caching method is proposed to combine the advantages of distributed caching

and PHY layer cooperative transmission. Backhaul-aware caching placement strategy for a group of cooperative BSs is studied in [56] by solving an optimization problem to minimize the average download delay. Nevertheless, none of the existing works provide efficient solutions for the cache utilization policy in cooperative SCNs without relying on iterative algorithms.

1.4 Stochastic Geometry

Stochastic spatial models from stochastic geometry was first introduced in wireless networks in early 1960s. Later on, spatial models from stochastic geometry have been extensively used in mobile ad hoc networks, sensor networks and cognitive networks. Point process models from stochastic geometry allows the characterization and evaluation of network performance over all possible configurations of the entire wireless network instead of just one specific configuration. With spatial averaging, one can obtain the average/expected value of some basic performance metrics such as the transmission success probability, based on the complementary cumulative distribution function (CCDF) of the SINR [85]. Earlier studies in [86–88] provides tractable analysis on the wireless network coverage and connectivity, which were further extended in the case with arbitrary number of tiers of BSs [89]. With the help of the existing results, using similar modeling and analytical approaches, we can characterize the spatially averaged performance of D2D underlaid cellular networks when using stochastic point processes to model the distribution of D2D users.

Stochastic point processes are suitable to model the node distribution in D2D underlaid cellular networks and in SCNs for the following reasons:

- User devices are usually with high mobility, which can be easily captured by the randomness of points generated from stochastic point processes, without the need of using mobility models for the user movement prediction.

- Irregular network deployments of the SBSs and small cell access points is more likely to be modeled by a stochastic spatial model than the traditional hexagonal grid model.

- When considering distributed access and interference management for massive number of D2D pairs in cellular networks, the performance of one specific D2D link is of less general value than studying the spatially averaged performance of randomly distributed users.

The most commonly used spatial model is homogeneous/stationary PPP where the intensity of points, i.e., the average number of points existing in any bounded region, is constant. Homogeneous PPP are often used to model the distribution of D2D users without spatial correlation or dependence. Some

other point processes can be used to model stochastic network with dependence. For instance, in a guard zone-based cognitive network where the secondary nodes cannot lie in the guard zones of primary nodes, the spatial correlation can be well captured by a PHP [48,50]. In a pure D2D or ad hoc network, with CSMA guard zones among active nodes which imposes a minimum distance between two nearby active node, the node distribution can be modeled by a hard-core process (HCP) [90, 91]. When considering the clustering effect of users, Poisson cluster process (PCP) models, such as Matérn cluster process and Thomas cluster process, can be applied to model the node distribution in such network [92, 93]. However, the analytical complexity of spatial points processes often increases with the spatial correlation of nodes. Some stochastic spatial models might lead to intractable performance analysis, thus requires approximations and assumptions in order to analyze the performance of the considered network model.

In this thesis, stochastic point processes are intensively used to model the distribution of user devices and SBSs. However, it is important to be aware that stochastic geometry is not a one-fits-all tool. The fundamental properties of these spatial models are the randomness of node distribution and the spatial averaging. In a wireless system with advanced radio resource management (RRM) based on the instantaneous channel conditions of network nodes, the analytical tractability of network performance with stochastic point processes will no longer hold. Moreover, the distribution of BSs in real world is not exactly random. With careful BS planning, the network performance will be definitely better than the expected value obtained with the assumption that the BSs are randomly distributed. However, the theoretical analysis with stochastic spatial models provides us some basic insights on the impact of network parameters on the average performance of large wireless networks. The obtained theoretical values of performance metrics can be also seen as a lower bound of the network-wide performance we can expect from the real-world network deployment.

1.5 Thesis Outline

This thesis consists of two major parts and is organized as follows. In Part I, entitled as **Device-to-Device (D2D) Underlaid Cellular Networks**, we focus on the interference management and the optimization of distributed D2D access schemes in D2D underlaid cellular networks.

In **Chapter 2 – Distributed SIR-aware Opportunistic Access Control**, we propose a decentralized access control scheme for interference management in multi-cell D2D underlaid cellular networks. Our method combines SIR-aware link activation with cellular exclusion regions in a case where D2D links opportunistically access the licensed cellular uplink spectrum. Analytical expressions and tight approximations for the coverage probabilities of cellular and D2D links are derived. We characterize the impact of the guard zone radius

and the SIR threshold on the D2D ASE and cellular coverage. A tractable approach was proposed in order to find the SIR threshold and guard zone radius, which maximize the ASE of the D2D communication while ensuring sufficient coverage probability for cellular uplink users. Simulations validate the accuracy of our analytical results and show the performance gain of our proposed scheme compared to existing state-of-the-art solutions.

In **Chapter 3 – Priority-based Shared Access with Delay Constraints**, we analyze a shared access network with a fixed primary (cellular) node and randomly distributed secondary (D2D) nodes whose distribution follows a homogeneous PPP. The secondaries use a random access protocol allowing them to access the channel with probabilities that depend on the queue size of the primary. Assuming a system with mutipacket reception (MPR) receivers having bursty packet arrivals at the primary and saturation at the secondaries, our protocol can be tuned to alleviate congestion at the primary. We study the throughput of the secondary network and the primary average delay, as well as the impact of the secondary node access probability and transmit power. We formulate an optimization problem to maximize the throughput of the secondary network under delay constraints for the primary node, which in the case that no congestion control is performed has a closed form expression providing the optimal access probability. Our numerical results illustrate the impact of network operating parameters on the performance of the proposed priority-based shared access protocol.

In Part II of the thesis, entitled as **Proactive Caching at the Network Edge**, we investigate the role of proactive caching at the network edge such as small cells and user devices.

In **Chapter 4 – Stochastic Wireless Caching Networks**, we study the probabilistic caching placement in wireless stochastic networks and we aim at answering the following two questions: 1) where to cache content in a wireless network? 2) which objective to consider for the optimal content placement? To answer the first question, we model a wireless cellular network using stochastic geometry and analyze the performance of two network architectures, namely caching at the mobile device allowing device-to-device (D2D) connectivity and local caching at the radio access the network edge (small cells). We provide analytical expressions for key performance metrics, including the cache hit probability, the density of cache-served requests and the average power consumption. Our results reveal that the performance of cache-enabled networks with either D2D caching or small cell caching heavily depends on the user density and the content popularity distribution.

In order to answer the second question, we consider the probabilistic caching placement in stochastic wireless D2D caching networks. The goal is to compare the performance of the optimal caching probabilities obtained with two different objectives: maximizing the cache hit probability and maximizing the density of successfully handled requests by local caches. Simulation results show that the optimal caching probabilities obtained by throughput optimiza-

tion shows notable gain in terms the density of successfully served user requests, particularly in dense user environment.

In **Chapter 5 – Small Cell Cooperative Caching**, we consider a cluster-centric SCN with combined design of cooperative caching and transmission policy. SBSs are grouped into disjoint clusters, in which in-cluster cache space is utilized as an entity. We propose a combined caching scheme where part of the available cache space is reserved for caching the most popular content in every SBS, while the remaining is used for cooperatively caching different partitions of the less popular content in different SBSs, as a means to increase local content diversity. Depending on the availability and placement of the requested content, either joint transmission (JT) or parallel transmission (PT) is used to deliver content to the served user. Using PPP for the SBS location distribution and a hexagonal grid model for the clusters, we provide analytical results on the successful content delivery probability of both transmission schemes for a user located at the cluster center. Our analysis shows an inherent tradeoff between transmission diversity and content diversity in our cooperation design. We also study optimal cache space assignment for two objective functions: maximization of the cache service performance and the energy efficiency. Simulation results show that the proposed scheme achieves performance gain by leveraging cache-level and signal-level cooperation and adapting to the network environment and user quality-of-service (QoS) requirements.

In Part III, which is the last part of this thesis, we give the conclusions, future perspectives and the appendices.

1.6 Publications

List of publications during the course of this PhD where those marked with $*$ are presented entirely or partially in this manuscript.

- **Journal Articles**

  [J6] **Z. Chen**, N. Pappas, M. Kountouris, V. Angelakis, "Energy Harvesting in D2D Underlaid Cellular Networks with Bursty Traffic", to be submitted to *IEEE Transactions on Green Communications and Networking*, Dec. 2016.

  [J5] N. Pappas, **Z. Chen**, "On the Effect of Bursty Traffic in Caching Helpers", submitted to *IEEE Communication Letters*, Sep. 2016.

  [J4] * **Z. Chen**, M. Kountouris, "Distributed SIR-aware access control for D2D underlaid cellular networks with guard zones", submitted to *IEEE Transactions on Wireless Communications*, Jul. 2016.

  [J3] * **Z. Chen**, N. Pappas, M. Kountouris, "Probabilistic Caching in Wireless D2D Networks: Hit Optimal vs. Throughput Optimal", under revision in *IEEE Communication Letters*, Oct. 2016.

[J2] * **Z. Chen**, N. Pappas, M. Kountouris, V. Angelakis, "On the performance of delay aware shared access with priorities", under revision in *IEEE Transactions on Mobile Computing*, Oct. 2016.

[J1] * **Z. Chen**, J. Lee, T. Q. S. Quek, M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks", under revision in *IEEE Transactions on Wireless Communications*, Jul. 2016.

- **Conference Papers**

[C6] * **Z. Chen**, M. Kountouris, "D2D caching vs. small cell caching: where to cache content in a wireless network?", in *IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Edinburgh, United Kingdom, Jul. 2016, pp. 1-6.

[C5] **Z. Chen**, N. Pappas, M. Kountouris, V. Angelakis, "Throughput analysis of smart objects with delay constraints", *IEEE 17th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Coimbra, Portugal, Jul. 2016, pp. 1-6.

[C4] **Z. Chen**, J. Lee, T. Q. S. Quek and M. Kountouris, "Cluster-centric cache utilization design in cooperative small cell networks", in *IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, Jun. 2016, pp. 1-6.

[C3] **Z. Chen**, M. Kountouris, "Cache-enabled small cell networks with local user interest correlation", in *IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Stockholm, Sweden, Jul. 2015, pp. 680-684.

[C2] **Z. Chen**, M. Kountouris, "Guard zone based D2D underlaid cellular networks with two-tier dependence", in *IEEE International Conference on Communication Workshop (ICCW)*, London, UK, Jun. 2015, pp. 222-227.

[C1] **Z. Chen**, M. Kountouris, "Distributed SIR-aware opportunistic access control for D2D underlaid cellular networks", in *Proc. IEEE Global Communication Conference (GLOBECOM)*, Austin, TX, Dec. 2014, pp. 1540-1545.

Part I

Device-to-Device (D2D) Underlaid Cellular Networks

Chapter 2

Distributed SIR-aware Opportunistic Access Control

In this chapter we consider a multi-cell device-to-device (D2D) underlaid cellular network, in which uplink cellular users intend to communicate with their nearest base station (BS) while multiple D2D links coexist in the same spectrum. We propose a decentralized opportunistic access scheme for the D2D users which builds on distributed signal-to-interference ratio (SIR)-based threshold scheduling and cellular exclusion regions. The main idea of our scheme is that a potential D2D link is allowed to access the cellular spectrum if the D2D transmitter is located outside the guard zones around cellular BSs (receivers) and whenever its SIR exceeds a predefined threshold. We provide analytical expressions on the probability of successful transmission in the D2D tier and on the coverage probability in the cellular tier. Based on these analytical expressions and tight approximations, we analyze the effect of the exclusion zone radius and of the SIR threshold on network-wide key performance metrics.

Furthermore, we consider the optimization problem of maximizing the area spectral efficiency (ASE) of D2D communications while keeping the cellular coverage probability above a certain level. We propose a tractable approach to solve the aforementioned optimization problem and derive in closed form the approximate optimal access probability and optimal SIR threshold. Simulations validate the accuracy of our analytical results and show the performance gain of our proposed scheme compared to existing state-of-the-art solutions.

This chapter is organized as follows. In Section 2.1, we present the D2D underlaid cellular network model. The proposed distributed access control scheme is presented in Section 2.2 and its performance is analyzed in Section 2.3. The proposed scheme is optimized in Section 2.4. Simulation results are provided in Section 2.5 the concluding remarks are given in Section 2.6.

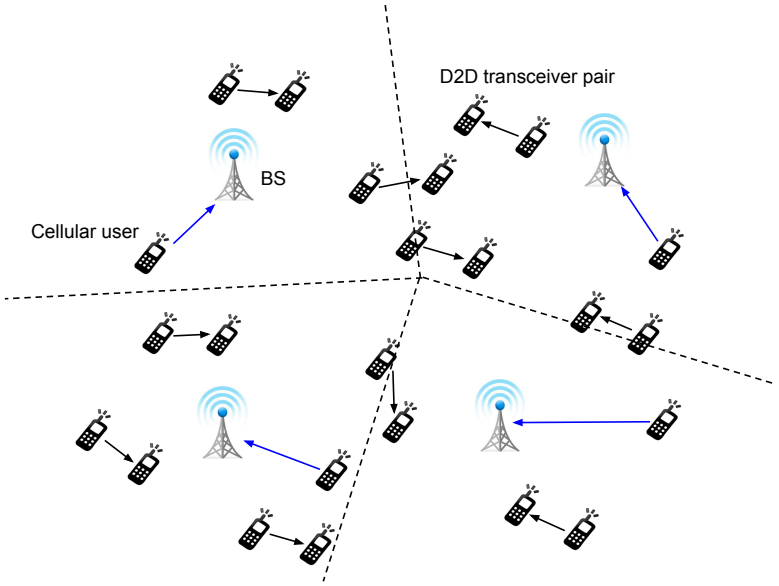


Figure 2.1: Multi-cell D2D underlaid cellular network model. Uplink cellular users transmit to their associated BSs, while multiple D2D links access the same spectrum.

2.1 Network Model

The locations of cellular BSs are modeled as a homogeneous Poisson point process (PPP) $\Phi_M = \{y_i : y_i \in \mathbb{R}^2\}$ in the two-dimensional Euclidean plane $\mathbb{R}^2$ with intensity λ_M, where y_i denotes the location of the i-th BS. Cellular users are placed according to some independent stationary point process and are associated to the closest base station. The coverage area of a BS is represented by a Poisson-Voronoi tessellation (PVT) on the plane. We assume that in each Voronoi cell there is always one active cellular user scheduled.[1] Denote by $\Phi_U = \{u_i : u_i \in \mathbb{R}^2\}$ the set of active cellular uplink transmitters, since each point u_i is randomly dropped in the Voronoi cell covered by y_i in Φ_M, the density of Φ_U will also be equal to λ_M.

The distribution of potential D2D transmitters follows a marked PPP $\Phi_D = \{(x_i, e_i) : x_i \in \mathbb{R}^2, e_i \in \{0, 1\}\}$ with intensity λ_D, where x_i denotes the location of the i-th D2D transmitter and e_i denotes its transmission mode: $e_i = 1$ means that the transmitter is active, otherwise $e_i = 0$. Potential D2D receivers are distributed at random isotropic directions around their respective transmitters and at a fixed distance d. All base stations, uplink users and D2D nodes are equipped with a single antenna.

Due to resource sharing among the cellular uplink users and D2D pairs, the success of cellular and D2D transmissions depends on both intra-tier and cross-tier interferences. Without loss of generality, conditioning on having a D2D receiver at the origin and its associated D2D transmitter at x_i with fixed

[1]This implies that the user density is much higher than λ_M, so that there is always at least one users to be served in the coverage region of each BS.

distance d from the receiver, its received signal-to-interference-plus-noise ratio (SINR) is given by

$$\mathrm{SINR}_i^{\mathrm{D}} = \frac{P_d |h_{i,i}|^2 d^{-\alpha}}{\sum\limits_{x_j \in \Phi_D \setminus \{x_i\}} e_j P_d |h_{j,i}|^2 d_{j,i}^{-\alpha} + \sum\limits_{u_k \in \Phi_U} P_c |h_{k,i}|^2 d_{k,i}^{-\alpha} + \sigma^2}, \tag{2.1}$$

where P_d and P_c denote the transmit powers of the D2D and cellular user, respectively; $h_{j,i}$ denotes the small-scale channel fading from transmitter j to the i-th D2D receiver, with $|h_{j,i}|^2 \sim \exp(1)$ (Rayleigh fading); $d_{j,i}$ denotes the distance from transmitter j to i-th D2D receiver. We consider a distance-dependent pathloss attenuation, which follows a standard power law, i.e. $d^{-\alpha}$ where $\alpha > 2$ is the pathloss exponent; σ^2 denotes the background thermal noise variance.

Similarly, for the cellular uplink communication, conditioning on having a BS at the origin and its associated cellular user at u_i, the received SINR at the i-th BS is given by

$$\mathrm{SINR}_i^{\mathrm{C}} = \frac{P_c |g_{i,i}|^2 \|u_i\|^{-\alpha}}{\sum\limits_{x_j \in \Phi_D} e_j P_d |g_{j,i}|^2 l_{j,i}^{-\alpha} + \sum\limits_{u_k \in \Phi_U \setminus \{u_i\}} P_c |g_{k,i}|^2 l_{k,i}^{-\alpha} + \sigma^2}, \tag{2.2}$$

where $g_{j,i}$ is the channel fading from transmitter j to i-th BS, following the same distribution as $h_{j,i}$; $l_{j,i}$ denotes the distance from transmitter j to the typical BS.

In the remainder, we assume that the background noise is negligible compared to the interference, thus the SINR in (2.1) and (2.2) will be replaced by the signal-to-interference ratio (SIR) since $\sigma^2 \to 0$. This is justified in current wireless networks, which are typically interference limited [94]. Background noise can be included in the subsequent analytical framework with little extra work.

2.2 D2D Access Control and Link Activation Scheme

In order to alleviate the interference problem introduced by spectrum sharing in D2D underlaid cellular networks, we propose a D2D access control and link activation mechanism, which involves two main methods: (i) imposing guard zones around cellular BSs; (ii) using SIR-aware thresholding for D2D link activation; these two schemes are described below.

2.2.1 Cellular Exclusion Zones

Any effective and reasonable design of D2D underlaid cellular networks should guarantee that devices engaging in D2D communication lie in close proximity

of each other and that there is sufficient spatial separation between cellular and inband D2D transmissions. One way to achieve the latter is by creating guard zones around cellular users or BSs and controlling the spacing that occurs in dense network deployments.

A first element of our proposed scheme is the use of cellular guard zones around the BSs where no D2D transmitters can lie in. By doing that, cellular uplink transmissions are protected from excessive interference due to D2D communication. This guard zone surrounding the macro BSs imposes that no other device is physically present. In other words, the guard zone creates a minimum separation among macro BSs and D2D devices. Imposing cellular guard zones will create holes around the BSs. The distribution of potential D2D transmitters can then be modeled by a *Poisson hole process* (PHP) $\Phi_H = \{x_i \in \Phi_D : \|x_i - y_j\| > \delta, \forall i \in \mathbb{N}_+, \forall j \in \mathbb{N}_+\}$, where $\|x_i - y_j\|$ denotes the Euclidean distance from the i-th D2D transmitter to the j-th BS, and δ denotes the guard zone radius. This point process model captures the spatial separation and the deactivation of D2D devices in the network in consideration. The density of Φ_H will then be [49]

$$\lambda_H = \lambda_D \cdot \exp\left(-\lambda_M \pi \delta^2\right). \tag{2.3}$$

2.2.2 SIR-Aware Opportunistic Access Control

For the potential D2D transmitters located outside the cellular guard zones, we propose a distributed opportunistic link scheduling protocol to determine the D2D links that are qualified to access the cellular spectrum.

Previous work on distributed opportunistic access control has mainly focused on received signal strength (RSS) (RSS) or channel-aware thresholding [17, 35, 41]. Driven by the fact that local channel state information (CSI) can be obtained by sending training sequences to the receiver, the activation probability is then calculated as the probability that the RSS or SNR (channel strength) is above a certain threshold. For i.i.d. Rayleigh fading, the access probability for D2D links under threshold-based channel-aware scheduling is given by

$$p_{\text{ac}} = \mathbb{P}(|h_d|^2 d^{-\alpha} > G_{\min}) = \exp(-G_{\min} d^{\alpha}), \tag{2.4}$$

where $G_{\min}$ is an optimized threshold. In that case, the set of active D2D links will form a homogeneous PPP as the thresholding operation results in independent thinning of a homogeneous PPP. Existing results on the interference distribution and the outage probability in Poisson networks can be directly applied (e.g. see [17]). The main drawback of this approach is that a potential D2D link with very strong received signal but which receives very strong interference may be activated, hence resulting in an unsuccessful transmission due to decoding errors. In other words, a D2D link with high SNR/RSS but low SINR might be activated, having marginal or even detrimental effect on the sum rate.

For that, we propose a distributed SIR-aware opportunistic access scheme that takes into account both received signal strength and interference level. A potential D2D link is allowed to access the spectrum when its estimated SIR is above a prescribed threshold[2]. This scheme can use measured or estimated SIR metrics; SIR estimation can be performed prior to thresholding by allowing all D2D transmitters to transmit a test signal to their associated receivers and calculate the received SIR. Here, the received SIR only depends on the received signal strength and the aggregate interference level, thus a node does not require CSI knowledge of each interfering link. Alternatively, advanced SIR estimation techniques based on sounding reference signal (SRS) or demodulation reference signal (DRS) can be applied [95]. Note that current and future mobile systems (e.g. 3GPP LTE-A, 5G) consider several methods for aggregate interference and SIR estimation. In that case, the proposed SIR-aware link activation scheme may use a two-stage protocol: in the first stage, each potential D2D link estimates its link quality (e.g. a certain function of SIR) at the beginning of each time transmission interval, and in the second stage, only the qualified D2D links, i.e. those with high estimated SIR, are active for the rest of the transmission interval. Evidently, the first phase should consume a negligibly small fraction of the whole resource block.

The aforementioned decentralized scheduling protocol extends beyond the cellular guard zone, where D2D devices may be present but they are deactivated by the thresholding operation. As we show below, the proposed distributed link activation scheme offers additional protection beyond that offered by the cellular guard zone, and - if properly optimized - it may increase the network throughput. More formally, a potential D2D transmitter requesting access must (i) be outside the cellular guard zones and (ii) have an estimated SIR exceeding a predefined threshold. Let G denote the SIR threshold, the transmission mode (active or not) of each potential D2D transmitter $x_i \in \Phi_D$ is

$$e_i = \mathbb{1}\left\{\mathrm{SIR}_i^{\mathrm{D}} > G, x_i \in \Phi_H\right\}. \tag{2.5}$$

Note that due to dependent thinning, the distribution of active D2D transmitters, denoted by $\Phi_A = \{x_i \in \Phi_H : \mathrm{SIR}_i^{\mathrm{D}} > G, \forall i \in \mathbb{N}_+\}$, is neither homogeneous PPP nor PPP.

2.3 Performance Analysis

The objective of this section is to investigate the effect of the design parameters of the proposed access scheme on the network performance, namely the area spectral efficiency (ASE) of the D2D network and the cellular coverage probability. Based on the results of this section, we provide in Section 3.3 the optimal system operating parameters as a means to maximize the D2D area spectral efficiency, keeping the cellular link quality above a certain quality level.

[2]As mentioned in the previous section, we employ SIR instead of SINR as we consider an interference-limited network.

2.3.1 Step 1: Cellular Exclusion Zones

After the first step of the proposed scheme, the locations of potential D2D transmitters follow a PPP Φ_H with intensity given in (2.3). The SIR distributions of both D2D and cellular links are evidently determined from the cellular guard zone radius. In this section, we analyze the success and the coverage probability in the D2D and cellular tier, respectively.

D2D Link Success Probability

The D2D link success probability is defined as the probability that the SIR of a randomly chosen D2D link is higher than a prescribed SIR target β. Building on previous analytical results for Poisson networks using stochastic geometry [88, 96], we have

$$\begin{aligned} p_{\text{suc}}^{\text{D}}(\beta) &= \mathbb{E}_0^! \left[\mathbb{P}(\text{SIR}_i^{\text{D}} > \beta)\right] \\ &= \mathcal{L}_{I_{cd}}\left(\beta d^\alpha \frac{P_c}{P_d}\right) \mathcal{L}_{I_{dd}}\left(\beta d^\alpha\right), \qquad (2.6) \\ &\overset{(a)}{\approx} \exp\left[-\frac{\pi d^2 \beta^{\frac{2}{\alpha}}}{\text{sinc}\left(\frac{2}{\alpha}\right)} \left(\left(\frac{P_c}{P_d}\right)^{\frac{2}{\alpha}} \lambda_M + \lambda_D\right)\right]. \qquad (2.7) \end{aligned}$$

where $\mathbb{E}_0^!$ is the expectation with respect to the reduced Palm distribution conditioned on having the typical receiver at the origin. The terms $I_{cd} = \sum_{u_k \in \Phi_U} |h_{k,i}|^2 d_{k,i}^{-\alpha}$ and $I_{dd} = \sum_{x_j \in \Phi_H \setminus \{x_i\}} |h_{j,i}|^2 d_{j,i}^{-\alpha}$ denote the interference (with normalized transmit power) caused at a D2D receiver by concurrent cellular and D2D transmissions, respectively. $\mathcal{L}_{I_{cd}}(s) = \mathbb{E}[e^{-sI_{cd}}]$ and $\mathcal{L}_{I_{dd}}(s) = \mathbb{E}[e^{-sI_{dd}}]$ are the Laplace transforms of interference I_{cd} and I_{dd}, respectively.

Here, (a) comes from approximating $\mathcal{L}_{I_{dd}}(s)$ by its lower bound, given by[3]

$$\mathcal{L}_{I_{dd}}(s) \approx \exp\left(-\frac{\pi \lambda_D s^{\frac{2}{\alpha}}}{\text{sinc}\left(\frac{2}{\alpha}\right)}\right), \qquad (2.8)$$

which is derived using the dominant interferer approach, i.e. counting for the interferer (normally closest or strongest) whose interference contribution alone is sufficient to cause outage. From the D2D success probability, we can obtain the area spectral efficiency (ASE) of the D2D network, which is the average number of successful transmissions of a certain rate that can be supported per unit area and has units of bit/s/Hz/m^2. The D2D ASE in our considered network with cellular guard zones can be written as

$$\mathcal{T}_D(\beta) = \lambda_H p_{\text{suc}}^{\text{D}}(\beta) \log_2(1+\beta). \qquad (2.9)$$

[3]It is shown in [48] that this lower bound gives more accurate result on the coverage/outage probability than the prior approximation by a PPP with the matching density λ_H, when the guard zone (hole) radius is not extremely large.

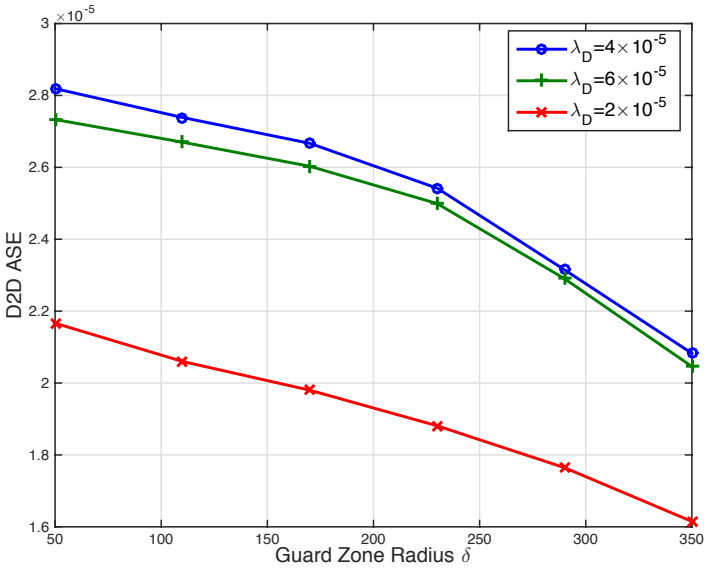


Figure 2.2: ASE of the D2D network vs. cellular guard zone radius δ. The initial density of the potential D2D links is $\lambda_D = \{2, 4, 6\} \times 10^{-5}$.

This metric will be used as the main performance metric to be optimized in the following sections, as a means to quantify the benefit in terms of spatial reuse of spectrum resources with underlaying D2D communications.

In Fig. 2.2 we plot the ASE of the D2D network as a function of the cellular guard zone radius δ. As expected, the D2D ASE is reduced when the guard zone range is increased. Interestingly though, the D2D ASE does not necessarily increase when the D2D link density augments, mainly due to the excessive interference in ultra dense D2D network deployments.

Cellular Coverage Probability

We define the cellular coverage as the probability of a random cellular link having SIR higher than a target γ, i.e.

$$\begin{aligned} p_{\text{cov}}^{\text{C}}(\gamma) &= \mathbb{E}_0^! \left[\mathbb{P}(\text{SIR}_i^{\text{C}} > \gamma)\right] \\ &= \mathbb{E}_l \left[\mathcal{L}_{I_{cc}}\left(\gamma l^{\alpha}\right) \mathcal{L}_{I_{dc}}\left(\gamma l^{\alpha} \frac{P_d}{P_c}\right)\right], \end{aligned} \tag{2.10}$$

where $I_{cc} = \sum_{u_k \in \Phi_U \setminus \{u_i\}} |g_{k,i}|^2 l_{k,i}^{-\alpha}$ and $I_{dc} = \sum_{x_j \in \Phi_H} |g_{j,i}|^2 l_{j,i}^{-\alpha}$ denotes the cellular interference and D2D interference to the typical cellular receiver with normalized transmit power, respectively. Assuming nearest BS association, the pdf of the cellular link distance l is

$$f_l(x) = 2\pi\lambda_M x \cdot e^{\pi\lambda_M x^2}. \tag{2.11}$$

Definition 1. *Consider the aggregate interference to a typical receiver at the origin* $I_\Pi = \sum_{x_i \in \Pi} |h_i|^2 \|x_i\|^{-\alpha}$*, where* Π *represents the spatial distribution of the interfering nodes. If* Π *is generated from a homogeneous PPP with density* λ_Π *and with minimum distance* $r_{\min}$ *to the typical receiver, i.e.,* $\|x_i\| \geq r_{\min}$*, we define a modified Laplace transform of* I_Π *as*

$$\begin{aligned} \mathcal{L}_I^1(s, \lambda_\Pi, r_{\min}) &= \mathbb{E}_{h_i,\Pi}\left[\exp\left(-s\sum_{i\in\Pi}|h_i|^2 r_i^{-\alpha}\right)\right] \\ &= \exp\left(-2\pi\lambda_\Pi \int_{r_{\min}}^{\infty} \frac{sv^{-\alpha}}{1+sv^{-\alpha}} v\,dv\right). \end{aligned} \tag{2.12}$$

Since each uplink cellular user is randomly distributed in the Voronoi cell of its associated BS. Due to the pairwise correlation among active cellular users in a given time slot, the distribution of the interfering uplink users can be modeled by a softcore process, which is intractable [97]. Here, we consider the uplink cellular interference as coming from PPP-distributed interfering nodes outside a circle centered at the typical BS with the same area as its Voronoi cell.

Proposition 1. *The Laplace transform of interference* I_{cc} *can be approximated by*

$$\mathcal{L}_{I_{cc}}(s) \approx \mathcal{L}_I^1(s, \lambda_M, d_{\min}) \tag{2.13}$$

where the pdf of $d_{\min}$ *is given by*

$$f_{d_{\min}}(r) = 2\frac{(3.5\pi\lambda_M)^{3.5}}{\Gamma(3.5)} r^6 \exp(-3.5\pi\lambda_M r^2). \tag{2.14}$$

Proof. See Appendix A.1. □

Fig. **??** compares the simulated and theoretical SIR CCDF of cellular uplink users while considering only cellular interference I_{cc}. We can see that our approximation in Proposition 1 gives relatively accurate result in terms of the SIR distribution in uplink cellular networks .

Remark 1. *According to the nearest BS association, each uplink user is uniformly distributed in the Voronoi cell of its connected BS. It is worth noticing that the user being connected to the nearest BS is not equivalent to that the BS is associated to the nearest user, as assumed in [98]. The nearest interfering uplink user might be closer to the typical BS than its own tagged user. The above proposition captures the discrepancy between these two association conditions and provides a tight approximation for the uplink cellular coverage probability.*

Due to the minimum distance δ between the D2D transmitters in Φ_H and the typical cellular receiver (BS), in terms of the received interference at the

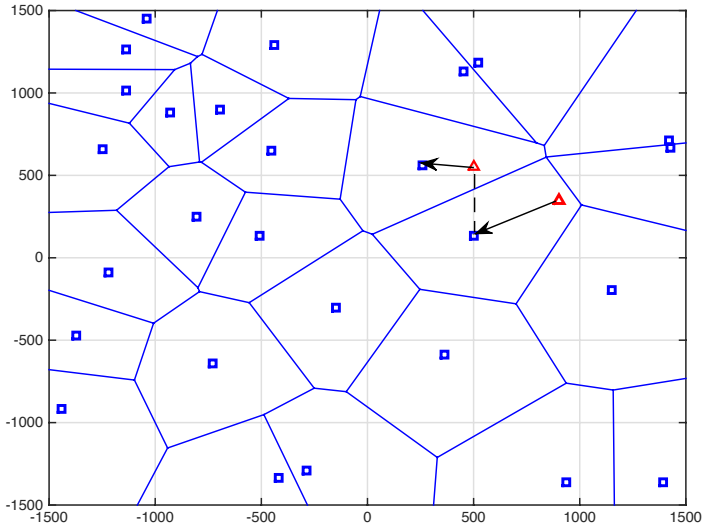


Figure 2.3: Voronoi tessellation of a cellular network with BSs distributed according to a homogeneous PPP. BSs are represented by blue squares. Red triangles represent two random users served by two BSs nearby.

typical BS, we can roughly consider the distribution of the D2D transmitters outside distance δ as a PPP with density λ_H. Thus, we have

$$\mathcal{L}_{I_{dc}}(s) \approx \mathcal{L}_I^1(s, \lambda_H, \delta). \tag{2.15}$$

Substituting (2.15) and the approximation of $\mathcal{L}_{I_{cc}}(s)$ proposed in Proposition 1 into (2.10), we have the cellular coverage probability given as

$$p_{\text{cov}}^C(\gamma) \approx \int_0^\infty f_l(x) \int_0^\infty f_{d_{\min}}(r) \mathcal{L}_I^1\left(\gamma x^\alpha, \lambda_M, r\right) \mathcal{L}_I^1(\gamma x^\alpha \frac{P_d}{P_c}, \lambda_H, \delta) \mathrm{d}r \mathrm{d}x, \tag{2.16}$$

where $f_{d_{\min}}(r) = 2\frac{(3.5\pi\lambda_M)^{3.5}}{\Gamma(3.5)} r^6 \exp(-3.5\pi\lambda_M r^2)$, $f_l(x) = 2\pi\lambda_M x \cdot e^{\pi\lambda_M x^2}$.

In Fig. 2.4 we compare the simulated cellular coverage probability with the theoretical results obtained from (2.16), which shows the accuracy of the cellular coverage analysis. Combined with Fig. 2.2, we conclude that increasing the guard zone radius δ eliminates the potential improvement in terms of D2D area spectral efficiency, but offers better protection to the cellular users. Moreover, we see that setting guard zones alone is not efficient in D2D underlaid cellular networks, which motivates us to introduce the SIR-aware D2D link activation scheme in order to achieve the highest possible D2D ASE for any value of D2D link density.

2.3.2 Step 2: SIR-aware Opportunistic Access

Denoting the set of active D2D transmitters by $\Phi_A = \{x_i \in \Phi_H : \text{SIR}_i^{\text{D}} > G, \forall i \in \mathbb{N}_+\}$ with average density λ_A, the success probability of a typical

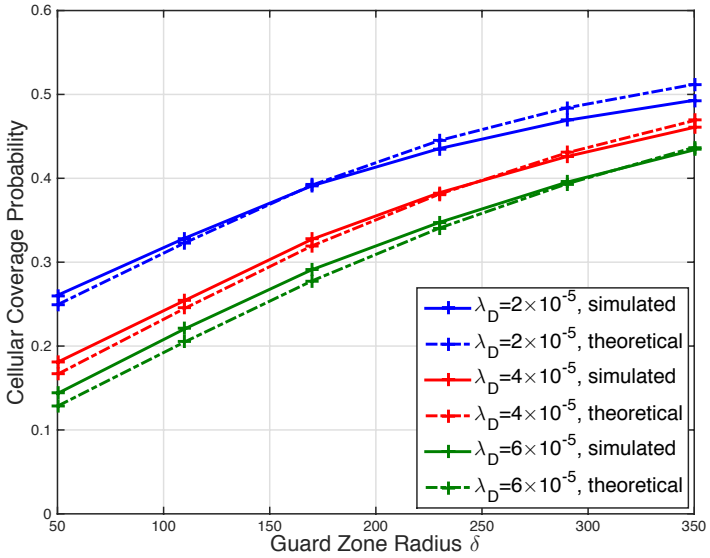


Figure 2.4: Cellular coverage probability vs. guard zone radius δ. The initial density of the potential D2D links is $\lambda_D = \{2, 4, 6\} \times 10^{-5}$ /m^2. Other system parameters are as in Table 2.1.

active D2D link is a conditional probability given that the i-th D2D pair could be active, i.e. the transmitter does not fall within the cellular guard zones and its estimated SIR_i exceeds the threshold G. Given the SIR threshold β for successful D2D transmission, the (conditional) success probability is $\mathbb{P}(\text{SIR}^{\text{D}}_{i\in\Phi_A} > \beta | \text{SIR}^{\text{D}}_{i\in\Phi_H} > G)$. The ASE of the D2D network can be expressed as

$$\mathcal{T}_D(\beta) = \lambda_A \mathbb{P}(\text{SIR}^{\text{D}}_{i\in\Phi_A} > \beta | \text{SIR}^{\text{D}}_{i\in\Phi_H} > G) \log_2(1+\beta). \tag{2.17}$$

From our analysis in Section 2.3.1, for the potential D2D transmitters in Φ_H, the D2D access (activation) probability p_s is the same as the D2D link success probability with G as the SIR target. From (2.7), we have

$$\begin{aligned} p_s &= \mathbb{P}[\text{SIR}^{\text{D}}_{i\in\Phi_H} > G] \\ &\approx \exp\left[-\frac{\pi d^2 G^{\frac{2}{\alpha}}}{\text{sinc}\left(\frac{2}{\alpha}\right)} \left(\lambda_D + \left(\frac{P_c}{P_d}\right)^{\frac{2}{\alpha}} \lambda_M \right) \right]. \end{aligned} \tag{2.18}$$

Note that p_s is a mean value by averaging over the fading statistics and all realizations of PPP Φ_H. For a specific PPP realization or conditioned on Φ_H, each D2D link experiences different SIR and thus should in principle be configured with different access probability depending on its location and surroundings, i.e. the locations of nearby D2D transmitters in this realization. In other words, when there are many interfering nodes in the vicinity of this D2D link, this link has lower access probability than a link in an area isolated from nearby interferers due to the fact that it has potentially lower SIR. So

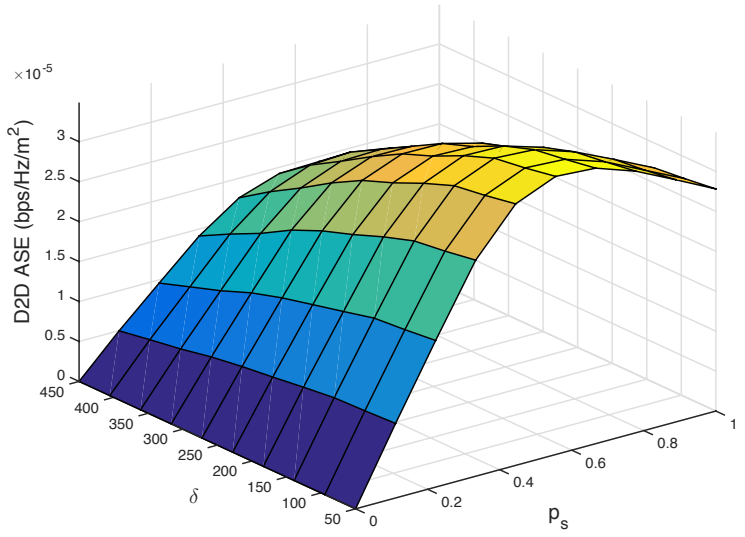


Figure 2.5: ASE of D2D network vs. (δ, p_s)

for each realization of Φ_H, p_s actually represents the proportion of D2D links that are allowed to access the spectrum.

Applying the proposed SIR-aware opportunistic access control results in dependent thinning of the PPP Φ_H, thus the set of active D2D transmitters Φ_A is hard or impossible to define (it is neither PPP nor PPP). For that, we resort to the approximation that Φ_A is a PPP with intensity given by

$$\lambda_A \simeq p_s \lambda_H = p_s \lambda_D \cdot \exp(-\lambda_M \pi \delta^2). \tag{2.19}$$

Rewriting the ASE of the D2D network as a function of the guard zone radius δ and D2D average access probability p_s, we obtain

$$\mathcal{T}_D(\delta, p_s) \simeq p_s \lambda_D \cdot \exp(-\lambda_M \pi \delta^2) \log_2(1+\beta) \cdot \mathbb{P}(\mathrm{SIR}^{\mathrm{D}}_{i \in \Phi_A} > \beta | \mathrm{SIR}^{\mathrm{D}}_{i \in \Phi_H} > G). \tag{2.20}$$

From (2.16), the cellular coverage probability when the locations of active D2D links follow Φ_A with intensity λ_A is given by

$$p^C_{\mathrm{cov}}(\gamma) \approx \int_0^\infty f_l(x) \int_0^\infty f_{d_{\min}}(r) \mathcal{L}^1_I\left(\gamma x^\alpha, \lambda_M, r\right) \mathcal{L}^1_I(\gamma x^\alpha \frac{P_d}{P_c}, \lambda_H p_s, \delta) \mathrm{d}r \mathrm{d}x. \tag{2.21}$$

In order to understand how δ and p_s affect the network performance, we plot the ASE of the D2D tier and the cellular coverage probability in Fig. 2.5 and Fig. 2.6, respectively. The density of the parent process (initial D2D density) is $\lambda_D = 4 \times 10^{-5}$ /m^2 and all other parameters are set as in Table 2.1. The active D2D transmitters are selected by sorting the estimated SIR value of all potential D2D links in Φ_H and choosing the p_s percentage of D2D links with highest estimated SIR. From Fig. 2.5 we observe that larger δ leads to

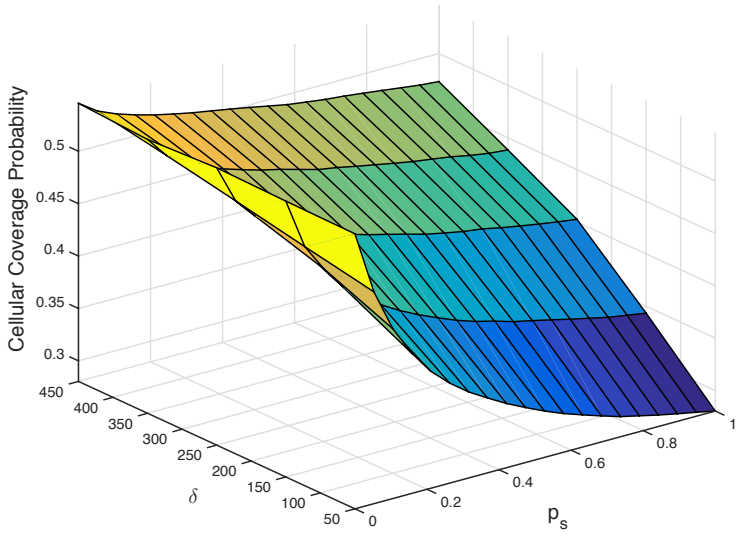


Figure 2.6: Cellular coverage probability vs. (δ, p_s)

lower D2D ASE. For a given value of δ, there always exists an optimal p_s for which the D2D underlaid network maximizes its ASE. As for the cellular coverage, from Fig. 2.6, expectedly, $p_{\text{cov}}^{\text{C}}$ increases with δ and decreases with p_s due to interference reduction. Combining together these two figures, we can easily understand that there exists an optimal point for which the D2D ASE is maximized while guaranteeing that the cellular coverage probability is above a certain threshold, if p_s and δ are properly tuned.

2.4 D2D Throughput Optimization under Cellular Coverage Constraints

In this section, we aim at optimizing the two key operating parameters of the proposed opportunistic access scheme in order to maximize the D2D area spectral efficiency while keeping the cellular link quality above a certain level. The optimization problem is cast as follows

$$(\delta^*, p_s^*) = \underset{(\delta, p_s)}{\arg\max}\ \mathcal{T}_D, \tag{2.22}$$

subject to

$$\begin{aligned} \delta &\in [0, \infty], \\ p_s &\in [0, 1], \\ p_{\text{cov}}^{\text{C}} &\geq (1-\mu) p_{\max}^{\text{C}}, \end{aligned} \tag{2.23}$$

where $\mu \in [0, 1]$ is the maximum coverage degradation coefficient, and $p_{\max}^{\text{C}}$ is the cellular coverage probability without D2D interference (single-tier net-

work). From (2.10), when $\lambda_H = 0$, we have

$$p_{\max}^{\mathrm{C}} \approx \int_0^\infty f_l(x) \int_0^\infty f_{d_{\min}}(r) \mathcal{L}_I^1 \left(\gamma x^\alpha, \lambda_M, r\right) \mathrm{d}r\mathrm{d}x. \tag{2.24}$$

Then the condition in (2.23) can be rewritten as

$$\begin{aligned} \int_0^\infty f_l(x) \int_0^\infty f_{d_{\min}}(r) \mathcal{L}_I^1 \left(\gamma x^\alpha, \lambda_M, r\right) \mathcal{L}_I^1 (\gamma x^\alpha \frac{P_d}{P_c}, p_s\lambda_H, \delta) \mathrm{d}r\mathrm{d}x \\ \geq (1-\mu) \int_0^\infty f_l(x) \int_0^\infty f_{d_{\min}}(r) \mathcal{L}_I^1 \left(\gamma x^\alpha, \lambda_M, r\right) \mathrm{d}r\mathrm{d}x. \end{aligned} \tag{2.25}$$

2.4.1 Decoupled Optimization

A joint design of δ and p_s seems cumbersome to be obtained, mainly due to the involved expressions for the coverage probability and the area spectral efficiency. In order to solve the above optimization problem, we take on a decoupled approach and proceed with the following procedure:

1. For a random value of δ, search for

$$p_s^*(\delta) = \underset{p_s \in [0,1]}{\arg\max}\ \mathcal{T}_D(p_s, \delta), \tag{2.26}$$

   where $\mathcal{T}_D(p_s, \delta) = p_s\lambda_H \log_2(1+\beta)\mathbb{P}(\mathrm{SIR}_{i\in\Phi_A}^{\mathrm{D}} > \beta | \mathrm{SIR}_{i\in\Phi_H}^{\mathrm{D}} > G)$ with $\lambda_H = \lambda_D \cdot \exp(-\lambda_M \pi \delta^2)$.

2. Replace p_s in (2.25) by the $p_s^\star(\delta)$ obtained in the first step, calculate numerically the minimum guard zone radius $\delta^\star$ by solving the following equation

$$p_{\mathrm{cov}}^{\mathrm{C}}(\delta^\star, p_s^\star) = (1-\mu) p_{\max}^{\mathrm{C}}. \tag{2.27}$$

3. Substitute the value of $\delta^\star$ in (2.26) and obtain the optimized access probability $p_s^*(\delta^\star)$.

The values of $(\delta^\star, p_s^\star)$ solving the decoupled optimization problem are clearly not optimal; however, our simulation results provided in the following section show that the solutions of the decoupled approach are very close to the optimal solution of the joint optimization. In the remainder of this section, we focus on deriving the optimal access probability $p_s^\star$ as the solution to (2.26), as well as the optimal SIR threshold $G^\star$ according to the relation between $G^\star$ and $p_s^\star$ given in (2.18).

2.4.2 SIR Threshold Optimization for Given δ

From the definition of the D2D ASE given in (2.20), we see that the conditional D2D success probability $\mathbb{P}(\mathrm{SIR}_{i\in\Phi_A}^{\mathrm{D}} > \beta | \mathrm{SIR}_{i\in\Phi_H}^{\mathrm{D}} > G)$ concerns two

dependent events. A potential D2D link with high SIR during the first stage of our SIR-aware protocol is very likely to have high SIR once allowed to be active. Although it seems hard or impossible to obtain a neat expression for the conditional probability, we approximate the optimal access probability p_s as the crossing point between the following two regimes:

- if $G \gg \beta$, which implies $p_s \to 0$, the set of nodes $\mathcal{A} = \left\{x_i \in \Phi_H : \mathrm{SIR}_i^{\mathrm{D}} > G\right\}$ can be approximately seen as a subset of $\mathcal{B} = \left\{x_i \in \Phi_A : \mathrm{SIR}_i^{\mathrm{D}} > \beta\right\}$, thus

$$\mathbb{P}(\mathrm{SIR}^{\mathrm{D}}_{i\in\Phi_A} > \beta | \mathrm{SIR}^{\mathrm{D}}_{i\in\Phi_H} > G) \simeq 1 \tag{2.28}$$

- if $G \ll \beta$, which implies $p_s \to 1$, the set of nodes $\mathcal{B} = \left\{x_i \in \Phi_A : \mathrm{SIR}_i^{\mathrm{D}} > \beta\right\}$ can be approximately seen as a subset of $\mathcal{A} = \left\{x_i \in \Phi_H : \mathrm{SIR}_i^{\mathrm{D}} > G\right\}$, thus

$$\begin{aligned}\mathbb{P}(\mathrm{SIR}^{\mathrm{D}}_{i\in\Phi_A} > \beta | \mathrm{SIR}^{\mathrm{D}}_{i\in\Phi_H} > G) &\simeq \frac{\mathbb{P}(\mathrm{SIR}^{\mathrm{D}}_{i\in\Phi_A} > \beta)}{\mathbb{P}(\mathrm{SIR}^{\mathrm{D}}_{i\in\Phi_H} > G)} \\ &= \frac{1}{p_s}\exp\left[-\frac{\pi d^2 \beta^{\frac{2}{\alpha}}}{\mathrm{sinc}\left(\frac{2}{\alpha}\right)}\left(p_s\lambda_D + (P_c/P_d)^{\frac{2}{\alpha}}\lambda_M\right)\right].\end{aligned} \tag{2.29}$$

Therefore, the ASE of the D2D tier is written as a function of p_s as

$$\mathcal{T}_D(p_s) = \begin{cases} \lambda_H p_s \log_2(1+\beta) & p_s \to 0 \\ \\ \lambda_H e^{-\xi\beta^{\frac{2}{\alpha}}(p_s\lambda_D + \kappa\lambda_M)}\log_2(1+\beta) & p_s \to 1, \end{cases} \tag{2.30}$$

where $\xi = \frac{\pi d^2}{\mathrm{sinc}\left(\frac{2}{\alpha}\right)}$ and $\kappa = \left(\frac{P_c}{P_d}\right)^{\frac{2}{\alpha}}$.

The approximately optimal access probability $p_s^\star$ and the approximately optimal SIR threshold $G^\star$ are given in the following proposition:

Proposition 2. *The approximately optimal access probability for the proposed SIR-aware opportunistic access scheme (based on the conditional D2D success probability) is given by*

$$p_s^\star \simeq \min\left\{\frac{\mathcal{W}\left(\lambda_D\xi\beta^{\frac{2}{\alpha}}e^{-\kappa\lambda_M\xi\beta^{\frac{2}{\alpha}}}\right)}{\lambda_D\xi\beta^{\frac{2}{\alpha}}}, 1\right\}, \tag{2.31}$$

where $\mathcal{W}$ denotes Lambert W function. The optimal SIR threshold in this case is approximately given as

$$G^\star \simeq \left[\frac{-\ln p_s^\star}{\xi(\lambda_D + \kappa\lambda_M)}\right]^{\frac{\alpha}{2}}, \tag{2.32}$$

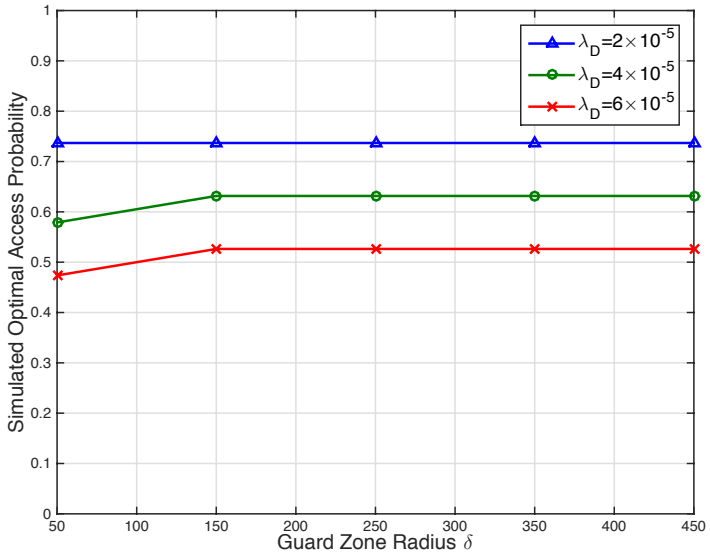


Figure 2.7: Simulated optimal access probability that gives the highest D2D ASE for different cellular guard zone radius δ. In the simulations, only the p_s percentage of D2D links with the highest estimated SIR are allowed to be active. The optimal value of p_s are obtained through exhaustive search.

where $\xi = \frac{\pi d^2}{\mathrm{sinc}\left(\frac{2}{\alpha}\right)}$ *and* $\kappa = \left(\frac{P_c}{P_d}\right)^{\frac{2}{\alpha}}$.

Proof. See Appendix A.2. □

Remark 2. *The derived* $p_s^\star$ *is independent of* δ *because of the approximation used in* (2.7) *when the exclusion zones do not overlap, i.e.* $\delta < \frac{1}{2\sqrt{\pi\lambda_M}}$. *The cellular guard zone range affects only the density of active D2D links and has little impact on the optimal D2D success probability. In order to validate this assumption, we plot in Fig. 2.7 the simulated optimal* p_s *obtained by exhaustive search that satisfies* (2.26) *for different values of* δ. *It evinces that the optimal access probability in terms of D2D ASE maximization is not very sensitive to cellular guard zone radius* δ. *Hence, the decoupled optimization may give approximately optimal values of* p_s *and* δ *if properly performed.*

Based on Remark 2, the decoupled optimization algorithm proposed in Section 2.4.1 can be further simplified into two steps:

1. Obtain $p_s^\star$ from Proposition 2.
2. Obtain $\delta^\star$ by solving $p_{\mathrm{cov}}^{\mathrm{C}}(\delta^\star, p_s^\star) = (1-\mu)p_{\max}^{\mathrm{C}}$ with numerical methods.

The optimized parameters $p_s^\star$ and $\delta^\star$ can be easily calculated at the cellular BSs, which will also be responsible of broadcasting the optimized SIR threshold $G^\star$ to every D2D transmitter outside the cellular guard zones for the access decision.

Table 2.1: Simulation Parameters for D2D Undelaid Cellular Network with Cellular Coverage Constraints

Parameters	**Values**
Macrocell BS density (λ_M)	10^{-6} /m^2
D2D link density (λ_D)	$[2, 10] \times 10^{-5}$ /m^2
D2D link length (d)	50 (m)
Pathloss exponent (α)	4
D2D SIR threshold (β)	5 dB
Cellular SIR threshold (γ)	0 dB
Cellular user transmit power (P_c)	10 (mW)
D2D user transmit power (P_d)	0.1 (mW)
Cellular degradation coefficient (μ)	30%

2.5 Simulation Results

In this section, we assess the performance of the proposed access control scheme for D2D underlaid cellular networks. Simulations are performed on a square region of surface 3000×3000 m^2. Both cellular BSs and potential D2D transmitters are distributed according to a homogeneous PPP with intensity λ_M and λ_D, respectively. The uplink users are uniformly distributed in each Voronoi cell covered by the nearest BS. Each D2D receiver is placed at a random direction around its transmitter with a fixed distance d. Fig. 2.8 shows a snapshot of the network layout with $\lambda_D = 2 \times 10^{-5}$ /m^2 and with cellular guard zone radius $\delta = 250$ m. Rayleigh fading is considered for both cellular and D2D links with $\mathbb{E}[|h|^2] = 1$. All other parameters are set according to Table 2.1.

All results are obtained by averaging over 4000 realizations. The following access strategies are also simulated for comparison and for evincing the performance gains of the proposed scheme:

- Only guard zone (GZ) scheme: all potential D2D links outside the cellular guard zones in Φ_H are active. The guard zone radius δ is chosen to satisfy the cellular coverage constraints. This basically corresponds to the first step of our proposed access scheme.
- Channel-aware access control (AC) with cellular guard zones: the link activation scheme in [17] is applied together with cellular guard zones that satisfy the cellular coverage constraints.

2.5.1 Proposed Access Control with Optimized (p_s, δ)

In Figs. 2.9 and 2.10 we present the optimized D2D access probability p_s and the cellular guard zone radius δ for different D2D link density values. Here,

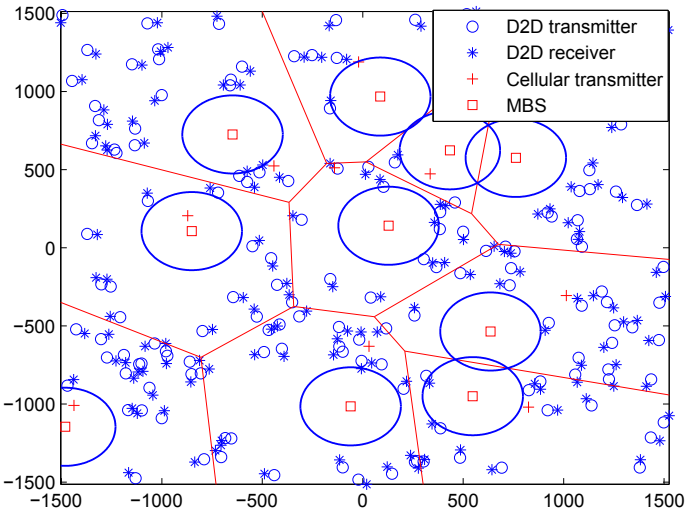


Figure 2.8: A snapshot of a multi-cell D2D underlaid cellular network with cellular guard zones around macrocell BSs . Potential D2D link density $\lambda_D = 2 \times 10^{-5}$. Only one cellular user in each Voronoi cell is active at a time.

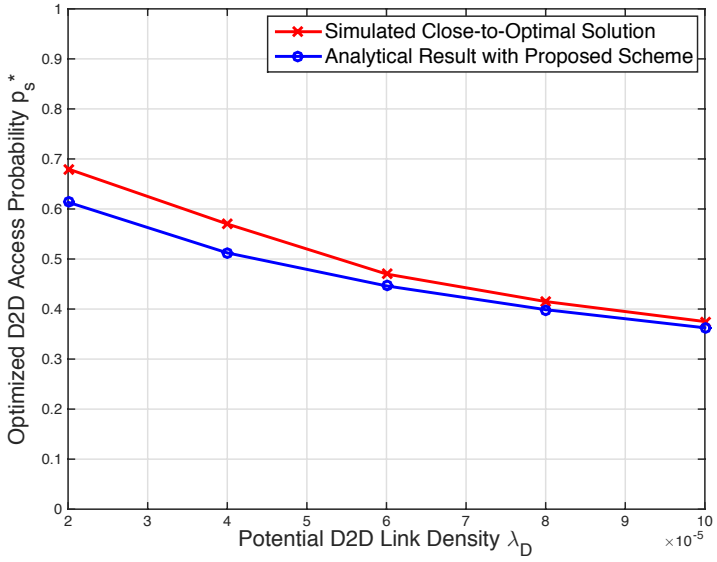


Figure 2.9: Optimized average access probability $p_s^\star$ vs. potential D2D link density λ_D.

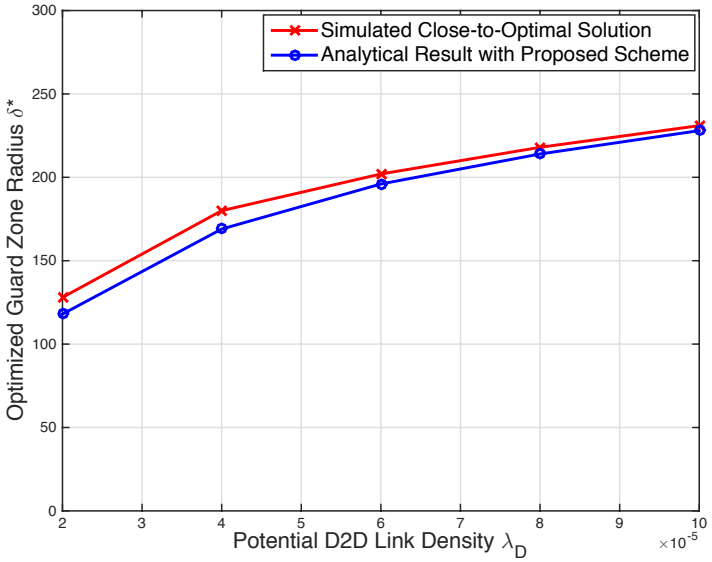


Figure 2.10: Optimized cellular guard zone radius $\delta^\star$ vs. potential D2D link density λ_D.

the simulated close-to-optimal solutions are obtained by exhaustively searching within the two-dimensional space (δ, p_s) with 150 evaluation points within the first dimension $\delta \in [0, 450]$ m and 40 evaluation points within the second dimension $p_s \in [0, 1]$. The analytical results of $p_s^\star$ and $\delta^\star$ are obtained from Proposition 2 and from (2.27), respectively. We see that our analytical results of $p_s^\star$ and $\delta^\star$) are relatively close to the simulated close-to-optimal solutions in dense D2D regime. The gap between the analytical and simulation results in sparse D2D regime is mainly due to the assumptions we use in order to derive the optimized access probability. When the D2D density λ_D is small, the analytical values of $p_s^\star$ and $\delta^\star$ are smaller than the simulated close-to-optimal values.

2.5.2 D2D ASE with Optimized (p_s, δ)

In Fig. 2.11, we evaluate the ASE performance of the D2D tier applying the proposed distributed access control protocol. The results are obtained with $p_s^\star$ and $G^\star$ as in Proposition 2 and guard zone radius $\delta^\star$ that satisfies (2.27). The cellular coverage probability without D2D interference is $p_{\max}^{\mathrm{C}} = 0.5552$, implying that the minimum cellular coverage is $p_{\mathrm{cov}}^{\mathrm{C}} \geq (1-\mu)p_{\max}^{C} = 0.3886$.

For comparison, we plot the simulated close-to-optimal ASE obtained through exhaustive search, demonstrating that the decoupled approach for optimizing the proposed scheme gives very close performance to the optimum. Compared to alternative access control schemes, we observe that our proposed method improves the aggregate throughput and provide evident performance gain. We also see that the SIR-aware access scheme improves the network

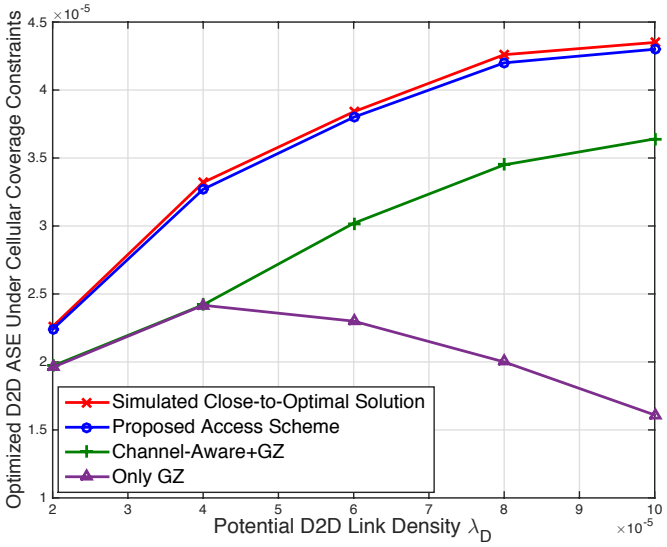


Figure 2.11: Optimized D2D ASE vs. potential D2D link density λ_D with different access control methods.

performance for any range of D2D densities, while the channel-aware method exhibits gains compared to the only GZ scheme starts for densities starting from $\lambda_D = 4 \times 10^{-5}$. This showcase the importance of taking into account the correlation between the estimated SIR and the real SIR of active D2D links in order to maximize the D2D throughput.

2.5.3 Average Sum Rate and Cellular Coverage with Optimized (p_s, δ)

In Table 2.2 we show the average sum rate per area (bps/Hz/m^2) of the D2D tier (denoted by R_{D}) and of the cellular tier (denoted by R_{C}), as well as the cellular coverage probability, achieved with our proposed $p_s^\star$ and $\delta^\star$ for a given potential D2D link density ($\lambda_D = 6 \times 10^{-5}$). The results are compared with the channel-aware scheme, a scheme implementing only guard zones (step 1 of proposed scheme) and a baseline scheme with no access control. The results evince the performance gains by setting $p_s^\star$ and $\delta^\star$ according to the decoupled optimization approach. Note that even though the objective of this paper and hence of our optimization problem was to maximize the D2D ASE under cellular coverage constraints, using $p_s^\star$ and $\delta^\star$ can also improve the average sum rate of D2D network.

Table 2.2: Comparison between different D2D access schemes

	D2D Sum Rate R_{D} ($\times 10^{-5}$)	Cellular Sum Rate R_{C} ($\times 10^{-6}$)	Cellular Coverage $p_{\mathrm{cov}}^{\mathrm{C}}$
Proposed Scheme	7.95	1.637	0.39064
Channel-Aware AC	6.55	1.64	0.3966
Only Guard Zones	5.71	1.656	0.3978
No AC	7.05	0.455	0.094

2.6 Summary and Concluding Remarks

In this chapter, we proposed a decentralized access control scheme for D2D underlaid cellular networks, which combines SIR-aware link activation with cellular guard zones. Using tools from stochastic geometry, we characterized the impact of the SIR threshold and the exclusion region range on the area spectral efficiency of D2D communications and on the cellular coverage probability. A tractable approach was proposed in order to find the optimal SIR threshold and guard zone radius that maximize the ASE of the D2D tier while guaranteeing sufficient cellular coverage probability.

From this chapter we see that very large throughput gains can be achieved in D2D underlaid cellular networks using distributed SIR-aware scheduling in conjunction with cellular exclusion regions. The combination of the two techniques yields significant performance improvement while keeping the merit of distributed medium access control methods, which does not require centralized control over the entire network. Besides the proposed D2D access control scheme, our results in the network coverage analysis in both D2D and cellular tiers can be helpful for future work related to the SIR analysis in guard zone-based cognitive networks. Future work would investigate the effect of multiple antennas at the BSs, which can further improve the cellular coverage. When the D2D users are not initially paired, the joint optimization of device association, mode selection and interference avoidance can be another direction for further extension of this work.

Note that in this chapter our optimization constraints are based on the cellular coverage probability, which is directed linked to the SIR at the cellular receiver. However, with the increasing wireless demand for multimedia content, users' experience of network service is very sensitive to the delay, which cannot be characterized by the simple SIR or SINR measurement. Therefore, in D2D underlaid cellular networks or other spectrum sharing systems with delay-sensitive primary (cellular) user, how to manage the access of the lower-priority users becomes an interesting problem to investigate. In the next chapter, we will present a delay-aware shared access protocol for a general spectrum sharing system with users with different priorities.

Chapter 3

Priority-based Shared Access with Delay Constraints

In contrast to the previous assumption of always backlogged cellular uplink user in device-to-device (D2D) underlaid cellular networks, in this chapter we assume bursty packet arrivals at the cellular user and we impose constraints on the cellular user delay which consists of both queueing delay and transmission delay. The network in study represents a more general setting/scenario and can be modeled as a *shared access network with priorities*. The primary user can be considered as the cellular uplink user, which has higher priority. The secondary nodes are similar to the underlaid D2D users which have lower priority than the cellular user.

We assume bursty packet arrivals at the primary transmitter and saturation at the secondaries, with mutipacket reception (MPR) capabilities at the receivers. The secondaries use a spatial Aloha random access scheme allowing them to access the channel with certain probabilities [34]. In this work we enhance this random access scheme by considering that the access probabilities of the secondaries depend on the queue size of the primary, thus, the activities of the secondaries can be adjusted to alleviate congestion at the primary. We study the throughput of the secondary network and the primary average delay, as well as the impact of the secondary node access probability and transmit power. We formulate an optimization problem to maximize the throughput of the secondary network under delay constraints for the primary node, which in the case that no congestion control is performed has a closed form expression providing the optimal access probability. Our numerical results illustrate the impact of network operating parameters on the performance of the proposed delay-aware shared access protocol with priorities.

This chapter is organized as follows. In Section 3.1 we present the network topology and the priority-based shared access protocol. In Section 3.2 we define several basic performance metrics, which will be used in Section 3.3 for the optimization of the secondary throughput under primary delay constraints. The numerical results are presented in Section 3.4, and we give the concluding

remarks in Section 3.5.

3.1 Network Model and Shared Access Protocol

3.1.1 Network Topology

We consider a shared access network, in which one primary source-destination pair and many secondary communication pairs share the same spectrum, as shown in Fig.3.1. The network region we study is a circular disk $\mathcal{C}$ with radius R. The primary receiver is centered at the origin of $\mathcal{C}$. The primary transmitter is located at fixed location with distance d_p to the primary receiver, which is common in infrastructure-based communication. We assume that the secondary transmitters are distributed in the two-dimensional Euclidean plane $\mathbb{R}^2$ according to a Poisson point process (PPP) $\Phi_s = \{x_i \in \mathbb{R}^2, \forall i \in \mathbb{N}^+\}$ with intensity λ_s, where x_i denotes the location of the i-th secondary transmitter [1]. Their associated receivers are distributed at isotropic directions with fixed distance d_s from their transmitters. For each realization of the PPP, the number of secondary transmitters in our network region $\mathcal{C}$ is a Poisson random variable with mean value $\lambda_s \pi R^2$. The time is slotted and each packet transmission occupies one time slot. We assume that the receivers have mutipacket reception (MPR) and that the secondary nodes can transmit simultaneously with the primary node [99].

The primary source has an infinite capacity queue Q for storing arriving packets of fixed length. The arrival process at the primary transmitter is modeled as a Bernoulli process with average rate λ packets per slot. The secondary node queue is assumed to be saturated, i.e., it always has a packet waiting to be transmitted.

3.1.2 Priority Based Shared Access Protocol

We consider the following priority-based protocol, which is an extension of that proposed in [100]. The primary node transmits a packet whenever backlogged, while the secondary nodes access the channel with a probability that depends on the queue size of the primary node, such that will not deteriorate the performance of the primary user. Denote Q the queue size in the primary node, the activity of the primary and secondary transmitters in a time slot are controlled in the following cases:

[1] Our analysis in this chapter can be easily extended to the case with randomly distributed primary user, with additional spatial averaging over the possible locations of the primary user. However, due to the additional complication on the signal-to-interference-plus-noise ratio (SINR) distribution, some of the analytical results might be no longer available.

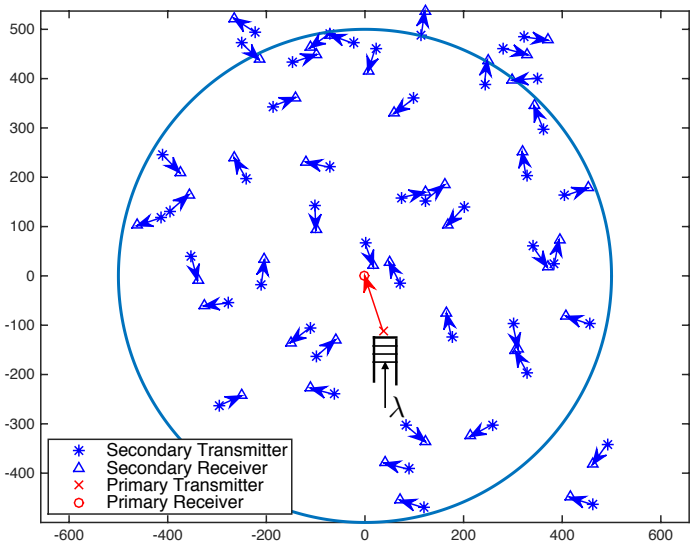


Figure 3.1: One snapshot of the network topology: one primary receiver centered at the origin with PPP distributed secondary transmitters under a given density (here $\lambda_s = 5 \times 10^{-5}$).

- *Case 1*: When $Q = 0$, the primary transmitter does not have packet to transmit, thus remains silent. Secondary transmitters randomly access the channel with probability q_1.
- *Case 2*: When $1 \leq Q \leq M$, the primary transmitter transmits one packet. Secondary transmitters randomly access the channel with probability q_2. [2]
- *Case 3*: When $Q > M$, the primary transmitter transmits one packet. Secondary transmitters remain silent.

For brevity we use PT and PR to denote the primary transmitter and receiver respectively, and ST and SR for denoting the secondary transmitter and receiver.

The threshold M plays the role of a congestion limit for the primary node, meaning that when the queue reaches this size, then the STs do not attempt to transmit any packet. When $M \to \infty$, the protocol model is simplified to the case without congestion control.

Note that we use two random access probabilities for the secondary nodes because the SRs experience different interference levels depending on whether

[2]Note that when the network topology is static, assigning progressively larger q_2 for secondary users further from the origin (the primary receiver) would improve the secondary throughput compared to the case with the same access probability. For the sake of tractability and taking into account of the user mobility, in this work we assume that all the secondary transmitters have the same access probability q_2.

the PT is active or not. Thus, the optimal access probabilities in these two cases need to be investigated separately.

3.1.3 Physical Model Successful Transmission Analysis

The MPR physical model is a generalized form of the packet erasure model. At the receiver side, a packet can be decoded correctly by the receiver if the received signal-to-interference-plus-noise ratio (SINR) exceeds a prescribed threshold θ. Given a set $\mathcal{T}$ of nodes transmitting during the same time slot, the received SINR at the i-th receiving node is given by

$$\mathrm{SINR}_i = \frac{P_i |h_{i,i}|^2 d_{i,i}^{-\alpha}}{\sum_{j \in \mathcal{T} \setminus \{i\}} P_j |h_{j,i}|^2 d_{j,i}^{-\alpha} + \sigma^2},$$

where P_i denotes the power of the transmitting node i; $h_{j,i}$ denotes the small-scale channel fading from the transmitter j to the receiver i, which follows $\mathcal{CN}(0,1)$ (Rayleigh fading); $d_{j,i}$ denotes the distance between the transmitter j to the receiver i. Here we assume a standard distance-dependent power law pathloss attenuation $d^{-\alpha}$, where $\alpha > 2$ denotes the pathloss exponent. σ^2 denotes the background noise power.

Let P_1 and P_2 be the transmit powers of the PT and the STs, respectively. In the following we refer to the primary node by node 0, while the secondary nodes are labeled with index $i \geq 1$. Denote x_0 the location of the PT and recall that the distribution of the STs is given by Φ_s, then we have $\mathcal{T} \subseteq \{x_0 \cup \Phi_s\}$. Note that in this work when we refer to the set of locations of the transmitting nodes, it means the set of transmitting nodes at these locations.

Following the description of our access protocol presented in Section 3.1.2, to derive the success probability of the primary and secondary nodes we need to consider three cases.

3.1.4 Case 1

When $Q = 0$, the PT is silent and the STs attempt packet transmission with probability q_1. Denote Φ_a^1 the locations of active STs, as a result of independent thinning [85], Φ_a^1 follows a homogeneous PPP with intensity $q_1 \lambda_s$. Hence, we have the active transmitter set as $\mathcal{T} = \Phi_a^1$.

Without loss of generality, we consider an arbitrary (typical) active secondary pair i in our network region. Denote $p_{2/2}$ the success probability of the

typical secondary pair when only the STs from Φ_a^1 are active, we have

$$\begin{aligned} p_{2/2} &= \mathbb{P}[\mathrm{SINR_i} > \theta \mid \mathcal{T} = \Phi_{\mathrm{a}}^1] \\ &= \mathbb{P}\left[\frac{P_2|h_{i,i}|^2 d_s^{-\alpha}}{\sigma^2 + \sum_{j\in\Phi_a^1\setminus\{i\}} P_2|h_{j,i}|^2 d_{j,i}^{-\alpha}} > \theta\right] \\ &\overset{(a)}{=} \exp\left(-\frac{\pi q_1 \lambda_s d_s^2 \theta^{\frac{2}{\alpha}}}{\mathrm{sinc}(2/\alpha)}\right) \exp\left(-\frac{\theta\sigma^2 d_s^{\alpha}}{P_2}\right). \end{aligned} \tag{3.1}$$

Here, (a) comes from $|h_{i,i}|^2 \sim \exp(1)$ and the probability generating functional (PGFL) of the PPP [88]. For a specific realization of the PPP, $p_{2/2}$ represents the percentage of active secondary pairs having successful transmission. It can also be seen as the probability of the typical active secondary pair to have successful transmission, averaging over different realizations of the PPP.

3.1.5 Case 2

When $1 \leq Q \leq M$, both the PT and part of the STs are active. Similarly, with independent thinning probability q_2, the locations of active STs follow another homogeneous PPP, denoted by Φ_a^2, with intensity $q_2\lambda_s$. In that case, the active transmitter set contains both the PT and the active STs, i.e., $\mathcal{T} = \{x_0 \cup \Phi_a^2\}$.

Denote $p_{1/1,2}$ and $p_{2/1,2}$ the success probabilities of the primary and secondary pairs when both types of nodes are active. With the help of existing results on the interference and outage distribution in PPP networks [85], we have the success probability of the primary transmission when the secondary network is active, given as

$$\begin{aligned} p_{1/1,2} &= \mathbb{P}\left[\mathrm{SINR_0} > \theta \mid \mathcal{T} = \{\mathrm{x_0} \cup \Phi_{\mathrm{a}}^2\}\right] \\ &= \mathbb{P}\left[\frac{P_1|h_{0,0}|^2 d_p^{-\alpha}}{\sigma^2 + \sum_{j\in\Phi_a^2} P_2|h_{j,0}|^2 d_{j,0}^{-\alpha}} > \theta\right] \\ &= \exp\left[-\frac{\pi q_2 \lambda_s \left(\theta\frac{P_2}{P_1}\right)^{2/\alpha} d_p^2}{\mathrm{sinc}(2/\alpha)}\right] \exp\left(-\frac{\theta\sigma^2 d_p^{\alpha}}{P_1}\right). \end{aligned} \tag{3.2}$$

For the active secondary nodes, considering an arbitrary (typical) active secondary pair i, we obtain the success probability in the following proposition.

Proposition 3. *The success probability of the typical secondary pair, when the active transmitters are $\mathcal{T} = \{x_0 \cup \Phi_a^2\}$, is given by*

$$p_{2/1,2} \simeq \exp\left[-\frac{\pi q_2 \lambda_s d_s^2 \theta^{\frac{2}{\alpha}}}{\mathrm{sinc}(2/\alpha)}\right] \frac{\exp\left(-\frac{\theta\sigma^2 d_s^{\alpha}}{P_2}\right)}{1 + \frac{d_s^2}{\mathbb{E}[d_{0,i}]^2}\left(\theta\frac{P_1}{P_2}\right)^{\frac{2}{\alpha}}}, \tag{3.3}$$

where $\mathbb{E}[d_{0,i}] = \int_0^{2\pi} \frac{1}{2\pi} \int_0^R \frac{2r}{R^2} \sqrt{r^2 + d_p^2 - 2rd_p\cos\varphi}\, dr\, d\varphi$.

Proof. See Appendix B.1. □

3.1.6 Case 3

When $Q > M$, only the PT is active. Denote $p_{1/1}$ the success probability of the primary pair when all the STs are silent, we have

$$\begin{aligned} p_{1/1} &= \mathbb{P}[\mathrm{SINR}_0 > \theta \mid \mathcal{T} = \mathrm{x}_0] = \mathbb{P}\left[\frac{\mathrm{P}_1|\mathrm{h}_{0,0}|^2\mathrm{d}_\mathrm{p}^{-\alpha}}{\sigma^2} > \theta\right] \\ &= \exp\left(-\frac{\theta\sigma^2 d_p^\alpha}{P_1}\right). \end{aligned} \tag{3.4}$$

Note that $p_{1/1} > p_{1/1,2}$ and $p_{2/2} > p_{2/1,2}$ always hold.

3.2 Performance Analysis

In this section, we define and analyze several relevant metrics for the performance evaluation of the proposed priority-based protocol with congestion control.

3.2.1 Throughput of the Secondary Network

For the considered shared access network, we aim at evaluating the throughput of the secondary network, *abbreviated as secondary throughput*, which is the number of packets per slot that can be successfully transmitted by the active secondary nodes to their destinations. In order to be consistent with the PPP model where the secondary nodes are generated with a certain density λ_s, we define the secondary throughput as the throughput of the secondary network per unit area, given as

$$T_s = \lambda_s \mathbb{P}[\mathrm{SINR}_{\mathrm{i}\in\Phi_\mathrm{s}} > \theta]. \tag{3.5}$$

Recall that the active STs is with density $q_1\lambda_s$ when the primary queue is empty, i.e., $Q = 0$. When the primary queue is $1 \le Q \le M$, then the active STs have density $q_2\lambda_s$. Hence, we have

$$\begin{aligned} T_s =& \mathbb{P}[Q=0]\cdot q_1\lambda_s\mathbb{P}[\mathrm{SINR}_{\mathrm{i}\in\Phi_\mathrm{a}^1} > \theta \mid \mathrm{Q}=0] \\ &+ \mathbb{P}[1\le Q\le M]\cdot q_2\lambda_s\mathbb{P}[\mathrm{SINR}_{\mathrm{i}\in\Phi_\mathrm{a}^2} > \theta \mid 1\le \mathrm{Q}\le \mathrm{M}] \\ =& \lambda_s\left\{\mathbb{P}[Q=0]\cdot q_1 p_{2/2} + \mathbb{P}[1\le Q\le M]\cdot q_2 p_{2/1,2}\right\}, \end{aligned} \tag{3.6}$$

where $p_{2/2}$ and $p_{2/1,2}$ are given in (3.1) and (3.3), respectively.

3.2.2 Primary Delay

The delay of the primary user depends on the service rate and the packet arrival rate. The service rate of the primary given a certain SINR target can be defined as the percentage of successfully transmitted packets per time slot. Dividing the cases by the primary queue size greater or less than M, when $1 \leq Q \leq M$, we have the primary service rate given by

$$\mu_1 = p_{1/1,2}. \tag{3.7}$$

When $Q > M$, the service rate is

$$\mu_2 = p_{1/1}. \tag{3.8}$$

Combining the two cases, we have the average service rate of the primary, denoted by $\bar{\mu}$, given by

$$\bar{\mu} = \frac{\mathbb{P}[1 \leq Q \leq M]\mu_1 + \mathbb{P}[Q > M]\mu_2}{\mathbb{P}[Q \geq 1]}. \tag{3.9}$$

The delay per packet at the primary node consists of the queueing delay and the transmission delay from the PT to the PR. From Little's law, we obtain the queueing delay which is related to the average queue size per packet arrival. The transmission delay is inversely proportional to the average service rate [51].

Denote $\bar{D}_p$ the primary average delay per packet, we have

$$\bar{D}_p = \frac{\bar{Q}}{\lambda} + \frac{1}{\bar{\mu}}, \tag{3.10}$$

where $\bar{Q}$ and $\bar{\mu}$ are the average queue size and the average service rate of the primary, which will be analyzed with closed-form expressions in Section 3.2.3.

3.2.3 Analysis of the Primary Queue and Delay

From the definition of the metrics in Section 3.2, we see that the secondary throughput and the primary delay depends on the state of the primary queue size. Therefore, we need to derive first $\mathbb{P}[Q = 0]$ and $\mathbb{P}[1 \leq Q \leq M]$.

We model the primary queue as a discrete time markov chain (DTMC), which describes the queue evolution and is presented in Fig. 3.2. Each state is denoted by an integer and represents the queue size. The packet arrival rate is always λ. The service rate is $\mu_1 = p_{1/1,2}$ when $1 \leq Q \leq M$, and is $\mu_2 = p_{1/1}$ when $Q > M$. From our analysis in Section 3.1.3, we know that $\mu_2 > \mu_1$. All the metrics related to the rate are measured by the average number of packets per time slot.

Denote π the stationary distribution of the DTMC, where $\pi(i) = \mathbb{P}[Q = i]$ is the probability that the queue has i packets in its steady state. We have the following lemma.

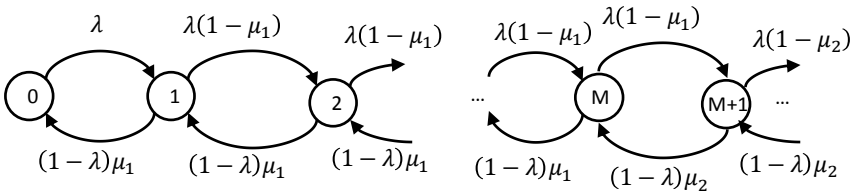


Figure 3.2: The Discrete Time Markov Chain which models the queue evolution at the primary node.

Lemma 1. *The stationary distribution of the DTMC described in Fig. 3.2 is given in the following cases:*

- *For* $1 \leq Q \leq M$*, we have*

$$\pi(i) = \frac{\lambda^i (1-\mu_1)^{i-1}}{(1-\lambda)^i \mu_1^i} \pi(0); \tag{3.11}$$

- *For* $i > M$*, we have*

$$\pi(i) = \frac{\lambda^i (1-\mu_1)^M (1-\mu_2)^{i-M-1}}{(1-\lambda)^i \mu_1^M \mu_2^{i-M}} \pi(0), \tag{3.12}$$

where $\pi(0)$ *is the probability that the queue is empty, given by*

$$\pi(0) = \begin{cases} \frac{(\mu_1-\lambda)(\mu_2-\lambda)}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\left[\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right]^M (\mu_2-\mu_1)} & \text{if } \lambda \neq \mu_1 \\ \\ \frac{\mu_2-\mu_1}{\mu_1 + (\mu_2-\mu_1)\frac{M+1-\mu_1}{1-\mu_1}} & \text{if } \lambda = \mu_1. \end{cases} \tag{3.13}$$

The queue is stable if and only if $\lambda < \mu_2$*.*

Proof. See Appendix B.2. □

Remark 3. *If without congestion control, the service rate is always* μ_1*. Obviously the condition to have stable queue is* $\lambda < \mu_1$*. The congestion control threshold* M *increases the queue stability region to* $\lambda < \mu_2$*, implying that the maximum allowed arrival rate at the PT becomes higher. On the other hand, less opportunity will be given to the secondary nodes to be active, because no secondary transmission is allowed when the primary queue size exceeds* M*.*

In order to simplify the equations, we define $\xi \triangleq \frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}$. In the remainder of this work we will assume that $\lambda \neq \mu_1$, however the general expressions of our results hold also for $\lambda = \mu_1$, but one should replace the $\pi(0)$ with the corresponding expression in this case.

Based on the results in Lemma 1, we have the following probabilities related to the primary queue size.

Lemma 2. *When the primary queue is stable and $\lambda \neq \mu_1$, the probability to have $1 \leq Q \leq M$ is*

$$\mathbb{P}\left[1 \leq Q \leq M\right] = \frac{\lambda\left(1-\xi^M\right)(\mu_2-\lambda)}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2-\mu_1)}. \tag{3.14}$$

The probability to have $Q > M$ is

$$\mathbb{P}\left[Q > M\right] = \frac{\lambda\xi^M(\mu_1-\lambda)}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2-\mu_1)}. \tag{3.15}$$

Proof. See Appendix B.3. □

We give the average queue size and the average delay of the primary in the following theorem.

Theorem 1. *The average queue size of the primary is given by*

$$\bar{Q} = \frac{N_1 + N_2}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2-\mu_1)}, \tag{3.16}$$

where

$$N_1 = \lambda(1-\lambda)\mu_1\frac{\mu_2-\lambda}{\mu_1-\lambda}\left[M\xi^{M+1} - \xi^M(M+1)+1\right], \tag{3.17}$$

and

$$N_2 = \xi^M\lambda(\mu_1-\lambda)\left[M + \frac{(1-\lambda)\mu_2}{\mu_2-\lambda}\right]. \tag{3.18}$$

The primary average delay is given by

$$\bar{D}_p = \frac{\bar{Q}}{\lambda} + \frac{\mu_2 - \lambda - \xi^M(\mu_2-\mu_1)}{\left(1-\xi^M\right)(\mu_2-\lambda)\mu_1 + \xi^M(\mu_1-\lambda)\mu_2}. \tag{3.19}$$

Proof. See Appendix B.4. □

Remark 4. *For a certain packet arrival rate λ at the PT, $\bar{D}_p$ is independent of q_1. The primary queue size augments with q_2 because of the lower service rate μ_1, which leads to higher queueing delay. The transmission delay also increases with q_2. As a result, $\bar{D}_p$ is an increasing function of q_2. Similarly, we know that $\bar{D}_p$ also increases with P_2.*

3.3 Secondary Throughput Optimization with Primary Delay Constraints

In our considered shared access network, spectrum sharing between the primary and secondary users can be exploited in order to bring secondary throughput gains at the expense of increasing interference to the PR. In order to protect the quality-of-service (QoS) of the primary user, the secondary interference

must be kept below a certain level, which corresponds to the thresholds on the ST access probability q_2 and transmit power P_2.

In this section, we analyze the secondary throughput as a function of q_2 and P_2 with respect to the primary delay constraints.

General Case

From the definition of the secondary throughput in (3.6), with the help of the results in Lemma 1 and Lemma 2, we have

$$\begin{aligned} T_s &= \lambda_s \left(\mathbb{P}[Q=0]\cdot q_1 p_{2/2} + \mathbb{P}[1\le Q\le M]\cdot q_2 p_{2/1,2}\right) \\ &= \lambda_s \frac{(\mu_2-\lambda)\left[q_1 p_{2/2}(\mu_1-\lambda) + q_2 p_{2/1,2}\lambda\left(1-\xi^M\right)\right]}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2-\mu_1)} \\ &= \lambda_s \frac{(p_{1/1}-\lambda)\left[q_1 p_{2/2}(p_{1/1,2}-\lambda) + q_2 p_{2/1,2}\lambda\left(1-\xi^M\right)\right]}{p_{1/1,2}p_{1/1} - \lambda p_{1/1,2} - \lambda\xi^M(p_{1/1}-p_{1/1,2})}. \end{aligned} \tag{3.20}$$

Considering the secondary throughput T_s as a function of the access probability q_1, it is obvious that there exists an optimal value $q_1^* = \underset{q_1\in[0,1]}{\arg\max}\ T_s$, which is equivalent to $q_1^* = \underset{q_1\in[0,1]}{\arg\max}\ q_1 p_{2/2}$, where $p_{2/2}$ is given in (3.1). From [17,34] we have that the optimal access probability q_1 of the STs when the PT is silent is given by

$$q_1^* = \min\left\{\frac{\mathrm{sinc}(\frac{2}{\alpha})}{\pi\lambda_s\theta^{\frac{2}{\alpha}}d_s^2}, 1\right\}, \tag{3.21}$$

which depends only on the ST density λ_s, secondary link distance d_s and the pathloss exponent α. Setting q_1^* in (3.20), when the PT transmit power P_1 and the packet arrival rate λ are fixed, the secondary throughput depends only on the access probability q_2 and the transmit power P_2.

As mentioned in Section 3.2.3, the primary average delay is an increasing function of q_2 and P_2. When $\lambda < \mu_2$, i.e., the primary queue is stable, the delay constraints of the primary user can be translated to the feasible region of the two variables (q_2, P_2), defined as

$$\mathcal{R}_{\mathcal{F}} = \{(q_2, P_2) : \bar{D}_p < D_{\max}\}. \tag{3.22}$$

where $D_{\max}$ is the threshold of the primary average delay .

In order to achieve the maximum secondary throughput while keeping the primary average delay below the threshold, we formulate the following optimization problem:

$$(q_2^*, P_2^*) = \underset{(q_2,P_2)}{\arg\max}\ T_s, \tag{3.23}$$

subject to

$$\begin{aligned} q_2 &\in [0,1], \\ P_2 &\in (0, P_{2,\max}], \\ \bar{D}_p(q_2, P_2) &< D_{\max}, \end{aligned}$$

where $P_{2,\max}$ is the maximum available power for a ST.

Due to the complexity of the analytical results related to the primary queue, it is difficult to solve the above optimization problem in closed form. Hence, first we investigate the particular case without congestion control, i.e., $M \to \infty$. The solution to the optimization problem in the general case is evaluated numerically in Section 3.4.

Case with no Congestion Control ($M \to \infty$)

Without congestion control, the activity of the primary and secondary nodes is simplified into two cases:

- When $Q = 0$, the PT remains silent. STs randomly access the channel with probability q_1.
- When $Q \geq 1$, the PT transmits one packet. STs randomly access the channel with probability q_2.

Following the primary queue analysis in Lemma 3.2, we have the probability to have i packets in the primary queue when it is in the steady state, given as

$$\pi(i) = \frac{\lambda^i (1-\mu_1)^{i-1}}{(1-\lambda)^i \mu_1^i} \pi(0), \tag{3.24}$$

where

$$\pi(0) = \mathbb{P}[Q=0] = 1 - \frac{\lambda}{\mu_1} = 1 - \frac{\lambda}{p_{1/1,2}}. \tag{3.25}$$

The primary queue is stable if and only if $\lambda < \mu_1$. Thus the feasible region of (q_2, P_2) is defined by

$$\mathcal{R}_{\mathcal{F}} = \{(q_2, P_2) : \bar{D}_p < D_{\max}, p_{1/1,2} > \lambda\}, \tag{3.26}$$

The secondary throughput becomes

$$\begin{aligned} T_s &= \lambda_s \left(\mathbb{P}[Q=0] \cdot q_1 p_{2/2} + \mathbb{P}[Q \geq 1] \cdot q_2 p_{2/1,2} \right) \\ &= \lambda_s \left[\left(1 - \frac{\lambda}{p_{1/1,2}} \right) q_1 p_{2/2} + \frac{\lambda}{p_{1/1,2}} q_2 p_{2/1,2} \right]. \end{aligned} \tag{3.27}$$

It is straightforward that the optimal value of q_1 is the same as in the case with congestion control, given in (3.21). Inserting q_1^* in (3.27) and denoting

$c^* = q_1^* p_{2/2}(q_1^*)$ the optimal per-node secondary throughput when $Q = 0$, the secondary throughput T_s can be written as a function of q_2 as follows

$$\begin{aligned} T_s &= \lambda_s \left[c^* \left(1 - \frac{\lambda}{p_{1/1,2}(q_2)} \right) + \frac{\lambda}{p_{1/1,2}(q_2)} q_2 p_{2/1,2}(q_2) \right] \\ &= \lambda_s \left[c^* + \frac{\lambda}{p_{1/1,2}(q_2)} \left(q_2 p_{2/1,2}(q_2) - c^* \right) \right]. \end{aligned} \tag{3.28}$$

Our objective is to find the optimal access probability q_2^* that maximizes the secondary throughput for fixed P_2 under the primary delay constraints. For that, the optimization problem is redefined as follows.

$$q_2^* = \arg\max_{q_2} \; T_s, \tag{3.29}$$

subject to

$$\begin{aligned} q_2 &\in [0,1], \\ \bar{D}_p(q_2) &< D_{\max}, \\ p_{1/1,2}(q_2) &> \lambda. \end{aligned}$$

The following lemma provides the global optimal value of q_2 without considering the primary delay constraints.

Lemma 3. *When $\frac{P_2}{P_1} < (d_s/d_p)^\alpha$ is verified, the global optimal value of $q_2 \in [0,1]$ that maximizes the secondary throughput in (3.28) is given by*

$$q_2^o = \min\left\{ \left[-\frac{W\left(\frac{\lambda_s \kappa_1 \kappa_2 c_{12}^*}{\kappa_1 - \kappa_2} e^{\frac{\kappa_1}{\kappa_1 - \kappa_2}} \right)}{\lambda_s \kappa_1} + \frac{1}{\lambda_s (\kappa_1 - \kappa_2)} \right]^+, 1 \right\}, \tag{3.30}$$

where W denotes the Lambert W function, $[z]^+ = \max\{z, 0\}$. $\kappa_1 = \frac{\pi d_s^2 \theta^{2/\alpha}}{\mathrm{sinc}(2/\alpha)}$, $\kappa_2 = \frac{\pi d_p^2 \left(\theta \frac{P_2}{P_1}\right)^{2/\alpha}}{\mathrm{sinc}(2/\alpha)}$, and $c_{12}^* = q_1^* p_{2/2}(q_1^*) \left[1 + \frac{d_s^2}{\mathbb{E}[d_{0,i}]^2} \left(\theta \frac{P_1}{P_2} \right)^{\frac{2}{\alpha}} \right]$ are constant parameters related to the network setting.

Proof. See Appendix B.5. □

Remark 5. *The value of ST power P_2 has a significant impact to the global optimal value of q_2. When P_2 is very high, the primary transmission can be severely harmed by excess interference. We assume here the practically relevant constraint that P_2 satisfies $\frac{P_2}{P_1} < (d_s/d_p)^\alpha$. This choice not only simplifies our analysis on the throughput optimization, but also reflects the evolution of wireless networks in deployments where D2D/machine-to-machine (M2M) communication with very low power nodes could coexist with the traditional high-rate mobile users [101].*

The average queue size of the primary in this case is

$$\bar{Q} = \sum_{i=0}^{+\infty} i\pi(i) = \frac{\lambda(1-\lambda)}{\mu_1 - \lambda}. \tag{3.31}$$

The primary average delay is thus given by

$$\bar{D}_p = \frac{\bar{Q}}{\lambda} + \frac{1}{\mu_1} = \frac{1-\lambda}{\mu_1 - \lambda} + \frac{1}{\mu_1}. \tag{3.32}$$

Then we obtain the feasible region of (q_2, P_2) in the following lemma.

Lemma 4. *With respect to the maximum average delay $D_{\max}$ of the primary and the queue stability condition, the feasible region of (q_2, P_2) is given by*

$$\mathcal{R}_{\mathcal{F}} \triangleq \left\{ (q_2, P_2) : q_2 < \min\left\{ \frac{\ln(p_{1/1}/\lambda)}{\lambda_s \kappa_2}, \frac{\ln(p_{1/1}/\eta_1)}{\lambda_s \kappa_2} \right\} \right\}, \tag{3.33}$$

where $\eta_1 = \frac{(D_{\max}-1)\lambda + 2 + \sqrt{(D_{\max}-1)^2\lambda^2 - 4\lambda + 4}}{2D_{\max}}$, κ_2 is defined in Lemma 3.

Proof. See Appendix B.6. □

Theorem 2 provides the optimal q_2 which maximizes T_s within the feasible region $\mathcal{R}_{\mathcal{F}}$, as the solution to the optimization problem defined in (3.29).

Theorem 2. *The optimal access probability q_2^* that maximizes the secondary throughput under primary delay constraints is given by*

$$q_2^* = \min\left\{ q_2^o, \frac{\ln(p_{1/1}/\lambda)}{\lambda_s \kappa_2}, \frac{\ln(p_{1/1}/\eta_1)}{\lambda_s \kappa_2} \right\}, \tag{3.34}$$

where q_2^o is given in (3.30)*, η_1 is defined in Lemma 4.*

Proof. See Appendix B.7. □

3.4 Numerical Results

In this section we evaluate the secondary throughput as a function of the two variables (q_2, P_2) within their feasible region that satisfies the delay constraints of the primary user. The primary delay and the feasible region boundary are also presented, showing the impact of the priority-based protocol design on the network performance. The values of the parameters are given in Table 3.1.

In Fig. 3.3, we plot the success probabilities $p_{1/1}$, $p_{1/1,2}$, $p_{2/2}$ and $p_{2/1,2}$ as a function of the ST access probability q_1 or q_2 for ST power set to $P_2 = 0.01$ mW. The numerical values are obtained from (3.4), (3.2), (3.1) and (3.3), respectively. Recall that $p_{1/1}$ is a constant value, $p_{1/1,2}$ and $p_{2/1,2}$ depend only on q_2, and $p_{2/2}$ depends only on q_1. As expected, when the secondary network is active, the success probabilities decrease rapidly with q_1 and q_2 increasing, as a result of the increased interference level.

Table 3.1: System Parameters for Priority Based Shared Access Network with Primary Delay Constraints

Parameters	**Values**
ST density (λ_s)	2×10^{-4}
Secondary link distance (d_s)	40 m
Primary link distance (d_p)	300 m
Cell size (R)	500 m
Pathloss exponent (α)	4
PT power (P_1)	100 mW
Maximum ST power ($P_{2,\max}$)	0.02 mW
Noise power (σ^2)	-113.97 dBm
SINR target (θ)	0 dB
Average delay threshold ($D_{\max}$)	3.5 time slots/packet

3.4.1 General Case

In Fig. 3.4 and Fig. 3.5, we plot the secondary throughput under the primary delay constraints. The values of T_s are obtained from (3.20) within the feasible region of (q_2, P_2) defined in (3.22). The results are presented with congestion threshold $M = \{1, 3\}$ and the packet arrival rate $\lambda = \{0.3, 0.7\}$. Knowing that $\mu_2 = p_{1/1} = 0.9997$, we choose $\lambda < 0.9997$ in order to satisfy the queue stability condition.

Our first remark is that the secondary throughput is not a monotonic function of q_2 and P_2. There exists an optimal point that gives the maximum T_s among the feasible choices of (q_2, P_2). We also observe a ceiling effect, i.e. once P_2 reaches a certain level, e.g., $P_2 = 0.006$ in Fig. 3.4, T_s has very small variation with respect to variations of P_2. This result implies that in order to have throughput gains, the necessary power for the secondary transmission should be quite low. Thus, the condition we used in Lemma 3 is validated.

Comparing the sub-figures in Fig. 3.4 we observe that larger M provides higher potential improvement for the secondary throughput, as the secondary links are more likely to be active. In order to validate our conclusion, in Table 3.2 we give the numerical values of the optimal solution (q_2^*, P_2^*) as well as the maximum secondary throughput achieved with different λ and M. We can see that for the same λ, larger M increases the maximum achievable secondary throughput. However, when M is large enough, the improvement of the maximum achievable secondary throughput becomes minor. This results justifies the necessity of our analysis in Section 3.3 with $M \to \infty$, which can be considered as an approximation for the case with large values of M.

Furthermore, in Fig. 3.6 we draw the boundary of the feasible region $\mathcal{R}_{\mathcal{F}}$

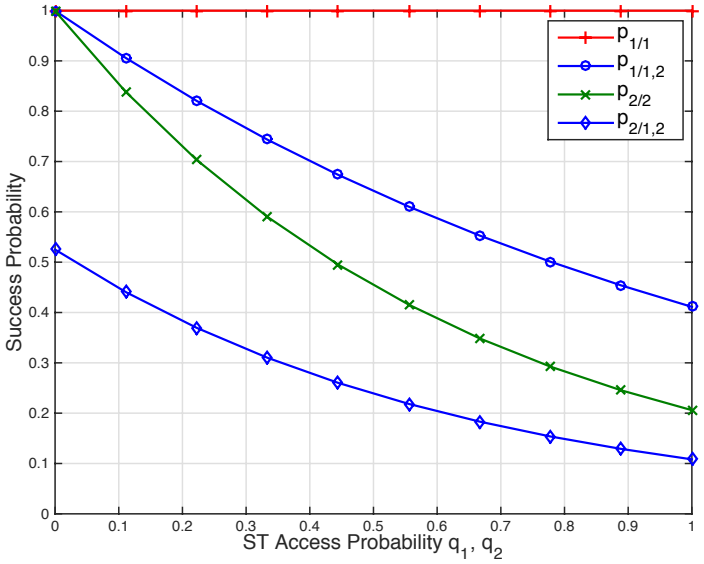


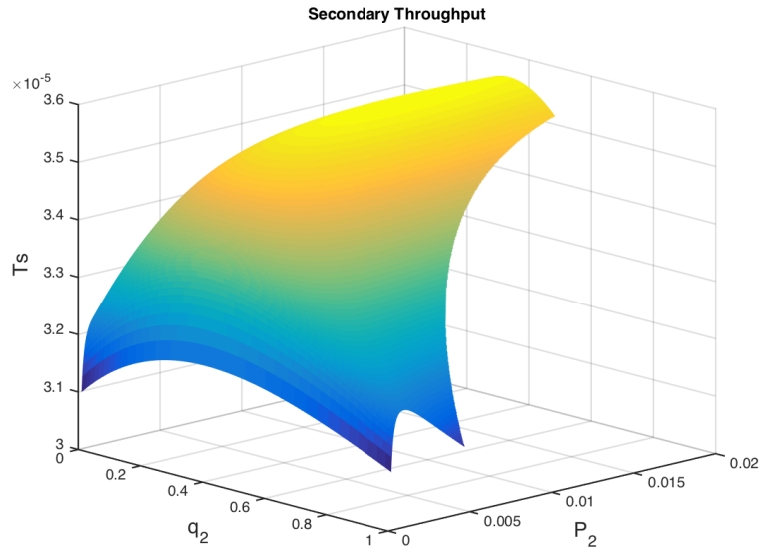
Figure 3.3: Success probabilities of the primary and secondary transmissions vs. ST access probability. $p_{1/1}$ is constant, $p_{2/2}$ is a function of q_1, $p_{1/1,2}$ and $p_{2/1,2}$ are functions of q_2. The ST power is set to $P_2 = 0.01$ mW.

for the four cases with $M = \{1, 3\}$ and $\lambda = \{0.3, 0.7\}$ respectively. The possible values of (q_2, P_2) that satisfy the primary delay constraints are situated below each plot. We observe that, larger M leads to more restricted feasible region, because in this case the congestion control is weaker, thus causes higher primary delay. Interestingly, we remark that the feasible region with $\lambda = 0.7$ and $M = 1$ is larger than that with $\lambda = 0.3$ and $M = 1$. This means that for the same values of (q_2, p_2), the primary average delay obtained with $\lambda = 0.7$ is actually smaller than the case with $\lambda = 0.3$. This is mainly due to the benefits of the congestion control in protecting the primary node transmission when the queue size is large. With high packet arrival rate, i.e., $\lambda = 0.7$, the probability of having $Q > M$ is very high, thus the STs will remain silent with high probability. In that case both the queueing delay and the transmission delay of the primary user will be reduced.

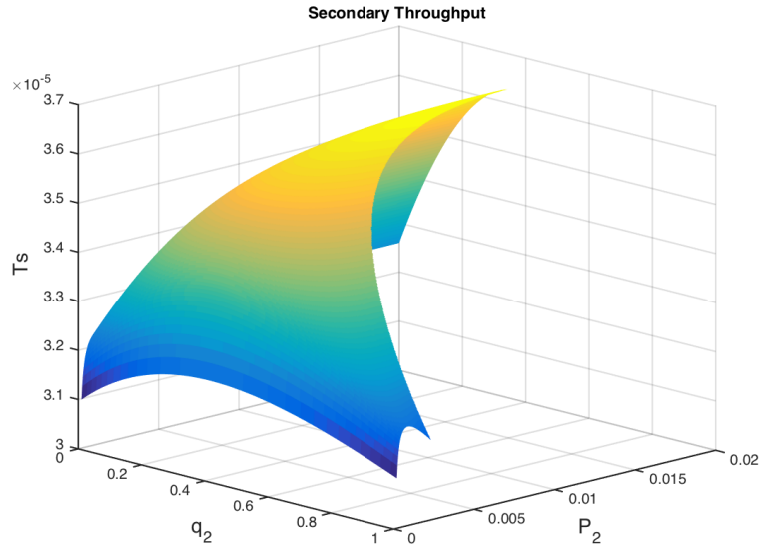
In order to further understand the influence of M and λ on the primary delay, in Fig. 3.7 we plot the primary average delay as a function of the ST access probability q_2 for different values of M and λ. The ST power is set to $P_2 = 0.01$ mW. Note that all the results are obtained with $\lambda < p_{1/1}$ in order to satisfy the queue stability condition. We have the following observations:

1. With q_2 increasing, which corresponds to the case of the PT service rate μ_1 decreasing, the primary delay increases rapidly at first, then saturating. The higher the arrival rate λ is, the lower saturated delay it gives.

2. When q_2 is relatively small, which means relatively high service rate



(a) $\lambda = 0.3$, $M = 1$

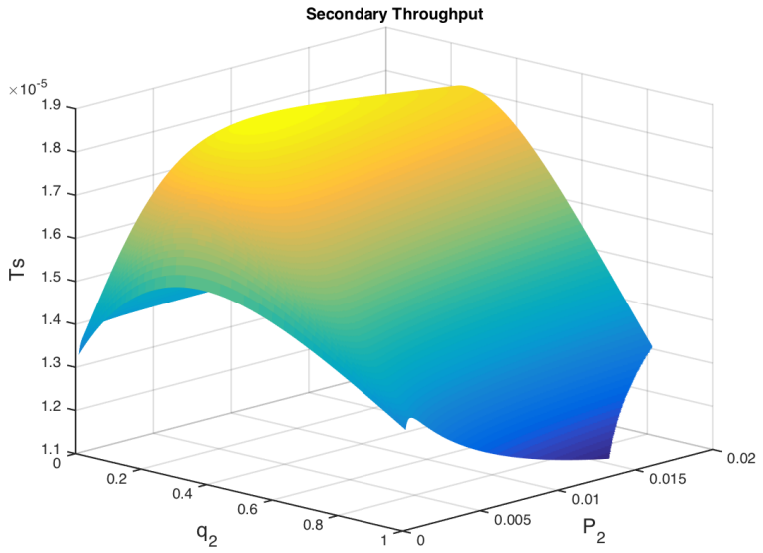


(b) $\lambda = 0.3$, $M = 3$

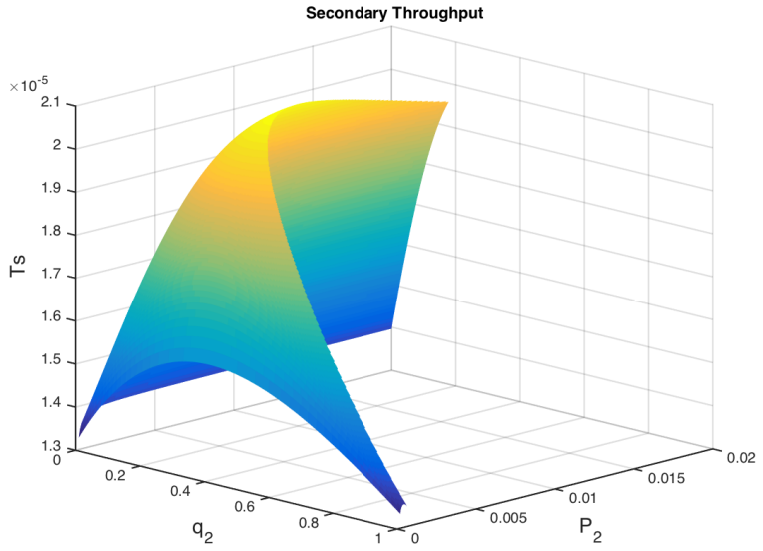
Figure 3.4: Secondary throughput vs. (q_2, P_2) under primary delay constraints. $\lambda = 0.3$.

μ_1, the primary delay is higher in the case with higher arrival rate λ. However, when q_2 is relatively high, depending on the value of M, this trend can be contrasting, e.g., in the case with $M = 1$, when $q_2 > 0.46$, the primary average delay is lower than in the case with higher λ.

3. For given values of λ and q_2, larger M results in higher primary average delay. Combined with the feasible region defined in (3.22), it is obvious



(a) $\lambda = 0.7$, $M = 1$



(b) $\lambda = 0.7$, $M = 3$

Figure 3.5: Secondary throughput vs. (q_2, P_2) under primary delay constraints. $\lambda = 0.7$.

that with larger M, q_2 should be smaller so that the service rate μ_1 will be sufficiently high in order to satisfy the delay constraints of the primary user.

The main takeaway messages we have from these results are:

1. With larger M, the maximum secondary throughput is higher. However,

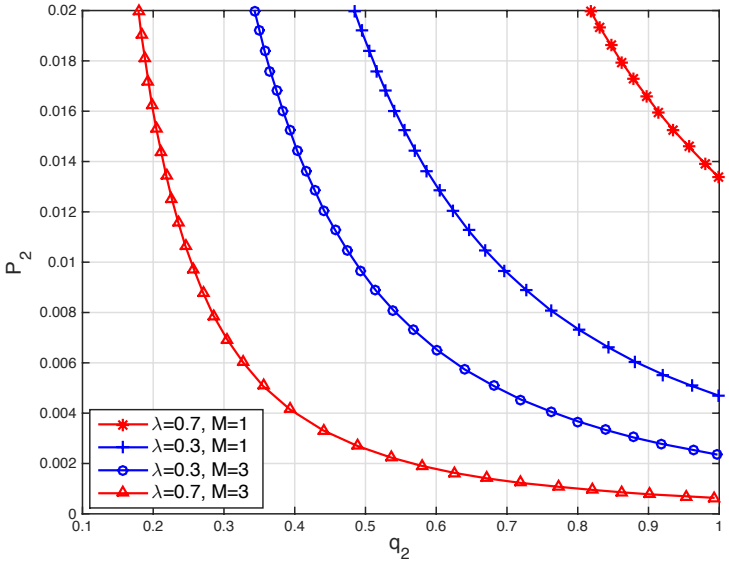


Figure 3.6: The boundary of the feasible region of (q_2, P_2) with $M = \{1, 3\}$ and $\lambda = \{0.3, 0.7\}$. Below each curve is the feasible region $\mathcal{R}_{\mathcal{F}}$ with respective values of λ and M.

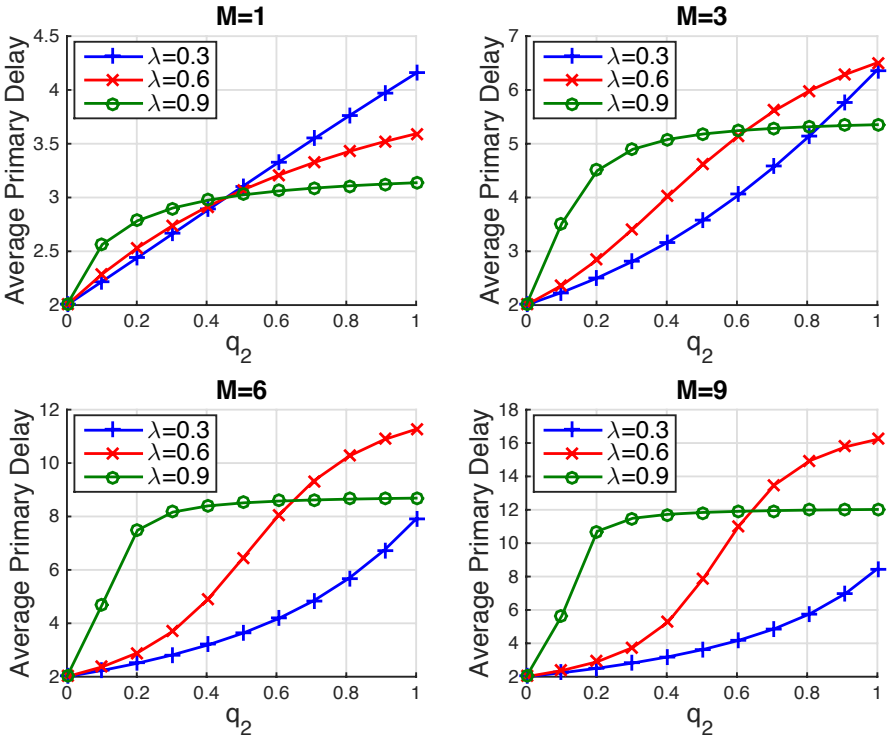


Figure 3.7: Primary average delay $\bar{D}_p$ vs. ST access probability q_2. $P_2 = 0.01$ mW. $M = \{1, 3, 6, 9\}$. $\lambda = \{0.3, 0.6, 0.9\}$

Table 3.2: Optimal System Settings with Secondary Throughput Maximization

λ	M	q_2^*	P_2^*(mW)	$T_s^*(\times 10^{-5})$
0.7	1	0.29	0.0062	1.87
	3	0.301	0.0071	2.08
	6	0.278	0.0058	2.10
0.5	1	0.323	0.0094	2.76
	3	0.345	0.012	2.91
	6	0.336	0.011	2.921
0.3	1	0.349	0.0124	3.57
	3	0.372	0.017	3.63
	6	0.368	0.0167	3.633

larger M put tighter constraints on the feasible values of (q_2, P_2).

2. With higher arrival rate λ at the PT, both the ST access probability and transmit power should be set to be lower. By doing so, the primary user achieves higher service rate, thus the queue size decreases faster, which in turn gives higher chance to the STs to transmit during the next time slot.

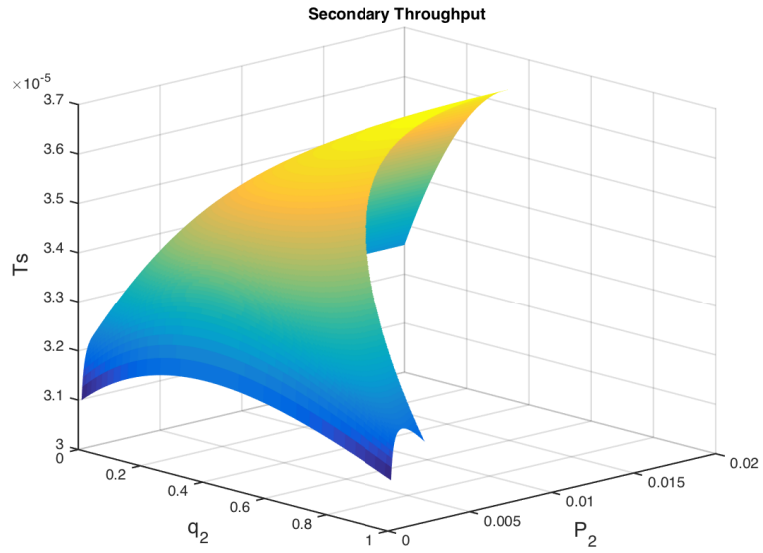
3. When the primary user is very sensitive to the delay, smaller M is more beneficial in order to increase the primary transmission rate.

3.4.2 Case with no Congestion Control ($M \to \infty$)

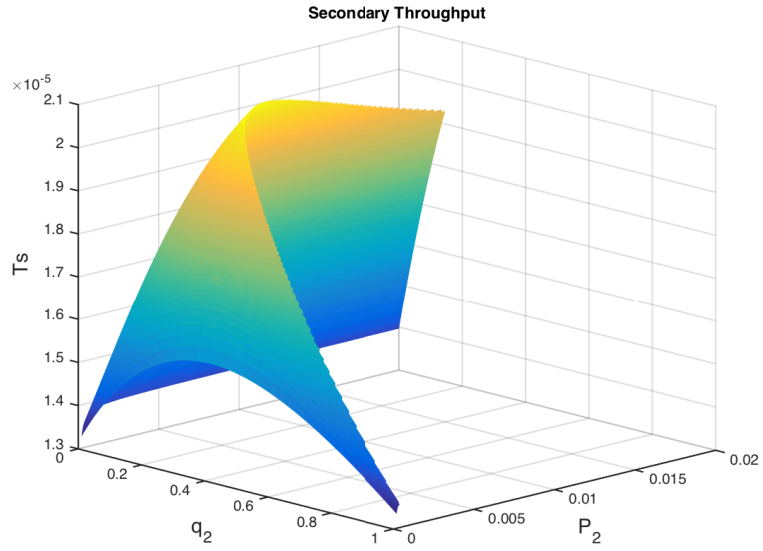
In Fig. 3.8, we plot the secondary throughput under the primary delay constraints in the case without congestion control. The results are presented for $\lambda = \{0.3, 0.7\}$. The evolution of the secondary throughput follows the same trend as observed in the general case presented in Fig. 3.4 and Fig. 3.5.

Fig. 3.9 shows the theoretical boundary of the feasible region $\mathcal{R}_{\mathcal{F}}$ derived in Lemma 4 in comparison to the real boundary obtained by exhaustive search of the feasible values of (q_2, P_2) under primary delay constraints. The results confirm the accuracy of our theoretical analysis on the feasible region of (q_2, P_2).

Fig. 3.10 shows the optimal access probability q_2^* obtained with Theorem 2 in comparison to the the real optimal values obtained by exhaustive search of q_2 that maximizes the secondary throughput with respect to the primary delay constraints. This illustrate the accuracy of our analytical results in Theorem 2. Another observation is that with $\lambda = 0.3$, q_2^* has values close to 0.4. When λ is higher, q_2^* declines rapidly with P_2 after P_2 reaches a certain value. This result is expected because above a certain value of P_2, the optimal q_2 is equal to the maximum feasible value of q_2 which is at the boundary of $\mathcal{R}_{\mathcal{F}}$.



(a) $\lambda = 0.3$



(b) $\lambda = 0.7$

Figure 3.8: Secondary throughput vs. (q_2, P_2) under primary delay constraints, in the simplified case without congestion control.

3.5 Summary and Concluding Remarks

In this chapter we investigated a delay-aware shared access protocol in a spectrum sharing network with one primary node and randomly distributed secondary nodes. The different priorities are given in the sense that the secondary user activities must be adjusted in order to prevent congestion at the primary

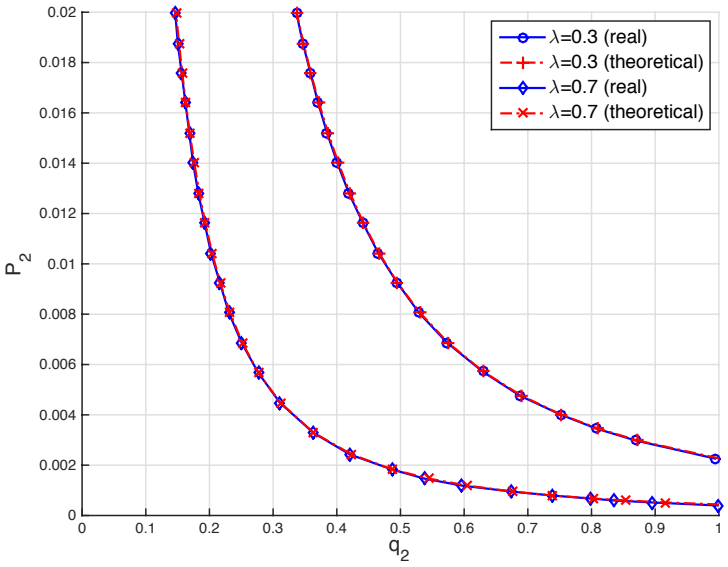


Figure 3.9: The boundary of the feasible region of (q_2, P_2) with $\lambda = \{0.3, 0.7\}$. The real values are obtained with exhaustive search.

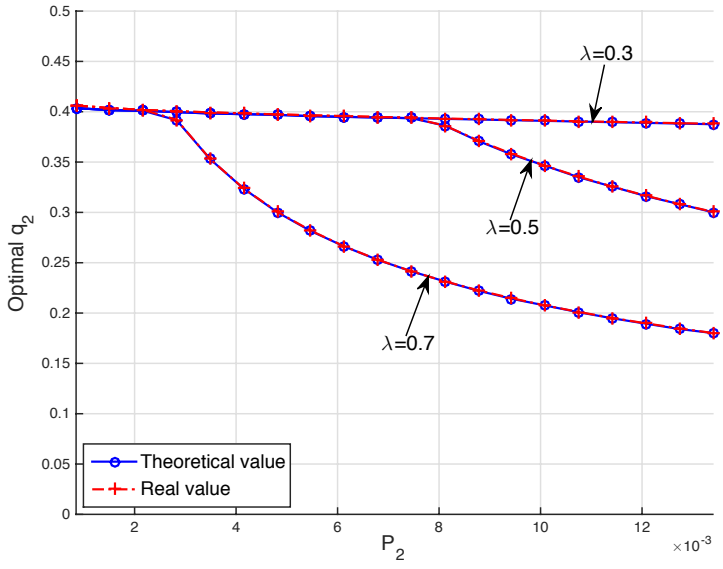


Figure 3.10: Optimal access probability q_2^* vs. P_2. The real values are obtained with exhaustive search.

user. We studied the maximum throughput of the secondary network under the constraints of the primary average delay, and the impact of the access protocol parameters on the throughput and delay performance of such network.

The main contribution of this chapter is to analyze the performance of a priority-based spectrum sharing system using tools from stochastic geometry, as well as to extend prior work on throughput optimization in shared access networks to the case with primary delay constraints. Our analysis in this chapter can be applied in any heterogeneous networks with priority-based users. The theoretical analysis presented in this chapter also provides an efficient approach to optimize the design parameters in the shared access protocol in order to achieve spectrum reuse gain in the presence of delay-sensitive primary users.

As summary, in Part I of this thesis, our research results have made one step forward in the distributed D2D access control with cellular coverage/delay constraints. Depending on the packet arrivals at the cellular (primary) user, we proposed different distributed access protocols for the D2D (secondary) users, namely SIR-aware opportunistic access with cellular guard zones, and delay-aware priority-based shared access protocol, respectively. The simple yet efficient opportunistic/random access scheme with optimized access probability offers balanced resource reuse between the cellular and D2D users, which can be easily implemented in D2D underlaid cellular networks or other priority-based spectrum sharing networks. In the next part of this thesis, we will switch to another research direction that helps to offload cellular data traffic with reduced energy consumption, namely proactive caching at network edge.

Part II

Proactive Caching at Network Edge

Chapter 4

Stochastic Wireless Caching Networks

In Part I of this thesis, we have investigated device-to-device (D2D) underlaid cellular networks and studied the distributed access control schemes to manage the interference issue due to the concurrent cellular and D2D transmissions in the same frequency band. In this part, we investigate an alternative cellular traffic offloading technique, namely proactive caching at the network edge. The research directions presented in Part I and Part II involve different layers of a wireless system, but can be well coupled together to enhance the advantages of device-centric and content-aware communications. For instance, when exploiting storage capacity at user devices for proactive content placement and enabling D2D communication between nearby users, local user requests can be handled with largely reduced energy consumption and improved area spectral efficiency.

Theoretical studies of proactive caching in wireless networks have occurred in many different scenarios with various objectives and analytical approaches. From the perspective of user request prediction, there have been many works investigating the content popularity pattern and the impact of social ties and physical proximity on the prediction of content dissemination. From the perspective of content placement, one might consider random caching strategies with stochastically distributed caching helpers, or study the optimal content placement with static network topology by solving combinatorial problems. However, before studying specific problems in wireless caching systems, there are some fundamental question that need to be answered, i.e.,

1. Where to cache content?
2. Which caching strategy to apply?
3. How to exploit cooperation opportunity instead of considering independent caching decisions at each caching helper?

In this chapter, we target at some basic comparisons between different types of caching network structures and different caching strategies. The first section

of this chapter focus on comparing the performance of D2D caching and small cell caching regarding some basic performance metrics. In the second section we study the optimal content placement in stochastic wireless D2D caching networks, with the performance comparison of two caching strategies: maximizing the cache hit probability and maximizing the density of cache-served requests.

4.1 Small Cell Caching vs. D2D Caching

At first sight, D2D caching and small cell (SC) caching may bring comparable gains in terms of enhanced network throughput, improved area spectral efficiency and energy efficiency. However, one must be aware of the following differences between D2D caching and SC caching systems:

- **Cache capacity**. Caching entities deployed in small base stations (SBSs) can have very large storage capability thanks to the low cost of storage units. In contrast, user devices, such as cellphones and tablets, have relatively small storage capacity and can only serve a rather small amount of requests generated by devices in their proximity.

- **Transmit power and coverage**. User devices normally transmit with much less power than SBSs, which in turn corresponds to smaller covering range.

- **Density of cache-served requests**. D2D communication usually involves small transmission distances. In networks with dense users, in the case of D2D caching, more simultaneous links are allowed to coexist in the same region sharing the spectrum resources as compared to the case of SC caching. Moreover, a special case in D2D caching is when a user finds its requested file stored in its own device. In that case, the user request is satisfied without any delay and cost.

- **Power consumption**. Additional to the power consumption required by cache-assisted D2D or small cell transmission, the power cost in the small cell backhaul is also a significant part. When a user request can not be served locally, either through D2D communication or from the SBS caches, the requested content will be retrieved from the core-network through the small cell backhaul. In general the backhaul power consumption is much higher compared to the transmission power cost.

In this section, we take into account the aforementioned differences and compare the performances of caching in user devices (i.e., D2D caching) and caching in the SBSs (i.e., SC caching) based on some key performance metrics.

4.1.1 Spatial Distribution and User Association

We consider a small cell network (SCN) where the SBSs are distributed in the two-dimensional Euclidean plane $\mathbb{R}^2$ according to a homogeneous Poisson point process (PPP)Φ_{s} with intensity λ_{s}. Mobile users are distributed according to another independent homogeneous PPP Φ_{u} with intensity λ_{u}. Caching capabilities can be enabled either on user devices for potential D2D communication, referred to as D2D caching, or by installing storage units at the SBSs, coined as SC caching.

Each mobile user makes a random request with probability $\rho \in [0, 1]$. As a result, the active users to be served form a homogeneous PPP $\Phi_{\mathrm{u}}^{\mathrm{r}}$ with intensity $\rho\lambda_{\mathrm{u}}$ (independent thinning). The inactive users form another homogeneous PPP $\Phi_{\mathrm{u}}^{\mathrm{t}}$ with intensity $(1-\rho)\lambda_{\mathrm{u}}$. They can serve as potential D2D transmitters in the case with D2D caching mode, or remain silent if D2D communication is not enabled.

Depending on whether cache capability is enabled at the devices or at the edge/SBSs, when an active user requests for a file, the following cases may happen:

- With only D2D caching, if the requested file is not cached in its own device, the user searches for the file in the nearby devices within a certain distance. If there is more than one D2D transmitter which has the requested file, the file is transmitted from the nearest one. Otherwise, the user connects to the nearest SBS to download the file from the core network through the backhaul.

- With only SC caching, the active user always connects to the nearest SBS. If its associated SBS has the file cached inside, the SBS transmits the file directly to the user. Otherwise the file is downloaded from the core network through the backhaul and then transmitted to the user.

We assume spectrum sharing among concurrent transmissions in such network, i.e. both D2D and small cell communication links receive interference from coexisting transmitters.

4.1.2 Request Distribution and Caching Policies

We consider a finite content library $\mathcal{F} = \{f_1, \cdots, f_N\}$ for the user requests, where f_i is the i-th most popular file and N is the library size. All files are assumed to have equal size, which is normalized to one. We use the standard Zipf law for the popularity distribution, meaning that the request probability of the i-th most popular file is

$$p_i = \frac{\Omega}{i^{\gamma}}, \tag{4.1}$$

where $\Omega = \left(\sum_{j=1}^{N} j^{-\gamma} \right)^{-1}$ is the normalization factor and γ is the shape parameter of Zipf law, which defines the correlation level of user requests. High values of γ means that most of the requests are generated from a few most popular files. For a user making a random request, p_i can be seen as the probability that the requested file is f_i.

Given knowledge of the content popularity distribution, depending on whether D2D caching or SC caching is adopted, we apply the following caching policies:

- With only D2D caching, since a random active user will most likely have multiple potential D2D transmitters, each device will independently cache files subject to its capacity-limited storage according to a common probability distribution, in order to increase the content diversity within the search distance [71]. This scheme was originally proposed in [78], referred as *geographic caching* strategy. The optimal caching probabilities are determined by optimizing some predefined objective function;
- With only SC caching, since an active user always connects to the nearest SBS, there are no overlapping coverage areas of different SBSs. Thus, we apply the conventional "most popular content (MPC)"policy, meaning that all SBSs cache the same most popular files within their cache capacity.

The details of the caching policies for both cases are presented as follows.

D2D Caching with Probabilistic Content Placement

With probabilistic content placement strategy, each user device independently caches file f_i with a certain probability q_i. Denote by M_{d} the cache size of a user device, with the assumption that each file has equal unit size, every user device can cache at most M_{d} files. Denote $\mathbf{q} = [q_1, \ldots, q_N]$ the caching probabilities of file $i \in [1, N]$, we have $\sum_{i=1}^{N} q_i \leq M_{\mathrm{d}}$ due to the cache storage limit. As a result of independent thinning, the distribution of potential D2D transmitters who have the file f_i follows a homogeneous PPP $\Phi_{\mathrm{u},i}^{\mathrm{t}}$ with intensity $q_i(1-\rho)\lambda_{\mathrm{u}}$.

Let R_{d} be the maximum search/discovery distance of a user device for establishing D2D communication. The probability that no potential D2D transmitters are found when f_i is being requested is equivalent to the probability of having no points from $\Phi_{\mathrm{u},i}^{\mathrm{t}}$ in the searching area (void probability). The probability of not finding f_i within the discovery distance R_{d} is given by

$$p_{\mathrm{m}}^{i}(R_{\mathrm{d}}) = e^{-\pi(1-\rho)\lambda_{\mathrm{u}} q_i R_{\mathrm{d}}^2}, \tag{4.2}$$

which can be seen as user request that can not be served by D2D helpers within distance R_{d}. Based on the prior studies in [68], we can obtain the

optimal caching probability vector $\mathbf{q} = [q_1, \cdots, q_N]$ by solving the following optimization problem:

$$\min_{\mathbf{q}} \sum_{i=1}^{N} p_i e^{-\pi(1-\rho)\lambda_u q_i R_{\mathrm{d}}^2} \tag{4.3}$$

subject to:

$$\sum_{i=1}^{N} q_i - M_{\mathrm{d}} \leq 0 \tag{4.4}$$

$$q_i \in [0, 1]. \tag{4.5}$$

We use the optimal dual-solution searching (ODSA) algorithm proposed in [68] in order to find the optimal $\mathbf{q}$. In the remainder of the section it is assumed that all the results related to D2D caching are obtained using the optimal caching probabilities.

SC Caching with MPC Policy

Denote by M_{s} the cache capacity of a SBS. According to the MPC caching policy, only files f_i with popularity order $i \in [1, M_{\mathrm{s}}]$ would be cached in each SBS.

4.2 Performance Analysis and Comparison

The potential gain of wireless content caching is mainly captured by the cache hit probability, which gives opportunity to handle user requests and to deliver content without having to retrieve it from the core network. Furthermore, there are potential spatial reuse gains by establishing proximity-based cache-assisted communication links sharing the same spectrum. In this section, we provide analytical results on several key performance metrics for cache-enabled cellular networks with either D2D caching or SC caching.

4.2.1 Cache Hit Probability

The cache hit probability is the probability for a random active user to find its requested file in local caches.

D2D Caching

With D2D caching, a cache hit request may happen in two cases:

- A user requesting for a file may find it stored in its own cache, we call this "self-request";

- When the requested file is not cached in its own device, the user finds it in the cache space of its nearby potential D2D transmitters within a distance R_{d}.

Denoting by $p_{\mathrm{self}}^{\mathrm{d}}$ the self-request probability of a random user, we have

$$p_{\mathrm{self}}^{\mathrm{d}} = \sum_{i=1}^{N} p_i q_i. \tag{4.6}$$

The probability of an active user being served by a nearby potential D2D transmitter is given by

$$p_{\mathrm{r}}^{\mathrm{d}} = \sum_{i=1}^{N} p_i \left(1 - q_i\right) \left(1 - e^{-\pi(1-\rho)\lambda_{\mathrm{u}} q_i R_{\mathrm{d}}^2}\right), \tag{4.7}$$

where $p_{\mathrm{hit},i}^{\mathrm{d}} = 1 - e^{-\pi(1-\rho)\lambda_{\mathrm{u}} q_i R_d^2}$ is the probability to have at least one potential D2D transmitters within distance R_d with file f_i cached.

Therefore, the cache hit probability is the sum of (4.6) and (4.7), given by

$$\begin{aligned} p_{\mathrm{hit}}^{\mathrm{d}} &= p_{\mathrm{self}}^{\mathrm{d}} + p_{\mathrm{r}}^{\mathrm{d}} \\ &= 1 - \sum_{i=1}^{N} p_i \left(1 - q_i\right) e^{-\pi(1-\rho)\lambda_{\mathrm{u}} q_i R_{\mathrm{d}}^2}. \end{aligned} \tag{4.8}$$

SC Caching

The cache hit probability in the case with SC caching is simply the probability that a user finds its requested file stored in the cache of its associated SBS. According to the MPC caching policy, we have that

$$p_{\mathrm{hit}}^{\mathrm{s}} = \sum_{i=1}^{M_{\mathrm{s}}} p_i. \tag{4.9}$$

4.2.2 Success Probability of Cache-assisted Transmission

When an active user finds its requested file cached in either the devices in proximity or the nearest SBS, it is not guaranteed that the cache-assisted transmission of the file will be successful. We calculate here the success probability of a typical cache-assisted transmission conditioning on having a receiver at the origin.

For a given realization of the network, we assume having K established cache-assisted communication links with $\mathbb{E}[K] \leq \rho\lambda_{\mathrm{u}}$. For a random link

$i \in [1, K]$, the received signal-to-interference-plus-noise ratio (SINR) is given by

$$\mathrm{SINR}_i = \frac{P_i |h_{i,i}|^2 d_{i,i}^{-\alpha}}{\sigma^2 + \sum_{j \in T \setminus \{i\}} P_j |h_{j,i}|^2 d_{j,i}^{-\alpha}},$$

where $P_i = \{P_\mathrm{d}, P_\mathrm{s}\}$ denotes the transmit power of either a D2D transmitter or a SBS, depending on whether caching capabilities are enabled on the mobile devices or at the SBSs; $h_{j,i}$ denotes the small-scale channel fading from the transmitter j to the receiver i, which follows $\mathcal{CN}(0,1)$ (Rayleigh fading); $d_{j,i}$ denotes the distance between the transmitter j to the receiver i; σ^2 denotes the background noise power. Note that interference comes not only from cache-assisted transmissions, but also from the SBS-user links when the requested file of the user is not locally cached. Thus, in the D2D caching case, the set of active transmitters belong to $T \subseteq \{\Phi_\mathrm{s} \cup \Phi_\mathrm{u}^\mathrm{t}\}$. In the other case with SC caching, $T \subseteq \Phi_\mathrm{s}$, because active users can only be served by the SBSs. We assume interference-limited network, in which the background thermal noise is negligible. In the remainder of this section the success probability is given as a function of the signal-to-interference ratio (SIR).

D2D Caching

For a random active user requesting for a file, as given in (4.7), with probability p_r^d the requested file is not cached in its own device, but in its nearby devices. Therefore, the density of cache-assisted communication links is $\rho\lambda_\mathrm{u} p_\mathrm{r}^\mathrm{d}$. Note that multiple users might find the same nearest D2D transmitter. In that case only one user can connect to this device, others have to search for another D2D transmitter. Denote by $\Phi_\mathrm{t}^\mathrm{d}$ the set of active D2D transmitters; although the resulting set is not distributed according to a homogeneous PPP, the density of $\Phi_\mathrm{t}^\mathrm{d}$ is given by

$$\lambda_\mathrm{t}^\mathrm{d} = \rho\lambda_\mathrm{u} p_\mathrm{r}^\mathrm{d}. \tag{4.10}$$

The set of users that cannot find their requested files in local caches will be served by the SBSs. The density of users to be served by the SBSs is $\lambda_\mathrm{r}^\mathrm{s} = \rho\lambda_\mathrm{u}(1 - p_\mathrm{hit}^\mathrm{d})$. According to the nearest SBS association, a Poisson-Voronoi tessellation is generated. From [102], the void probability of a typical Voronoi cell can be approximated as

$$p_\mathrm{void} \simeq \left(1 + \frac{\lambda_\mathrm{r}^\mathrm{s}}{3.5\lambda_\mathrm{s}}\right)^{-3.5}. \tag{4.11}$$

The density of active SBSs is thus given by

$$\begin{aligned}\lambda_\mathrm{t}^\mathrm{s} &= \lambda_\mathrm{s}\left(1 - p_\mathrm{void}\right) \\ &\simeq \lambda_\mathrm{s}\left(1 - \left(1 + \frac{\rho\lambda_\mathrm{u}(1 - p_\mathrm{hit}^\mathrm{d})}{3.5\lambda_\mathrm{s}}\right)^{-3.5}\right).\end{aligned} \tag{4.12}$$

Conditioning on having a typical D2D receiver at the origin with its associated transmitter at distance d_i, and assuming a homogeneous PPP for both active D2D transmitters $\Phi_{\mathrm{t}}^{\mathrm{d}}$ and active SBSs $\Phi_{\mathrm{t}}^{\mathrm{s}}$, the success probability is given as [88]

$$\begin{aligned} p_{\mathrm{suc}}^{\mathrm{d}} &= \mathbb{P}\left[\frac{P_{\mathrm{d}}|h_{i,i}|^2 d_i^{-\alpha}}{\sum\limits_{j\in\Phi_{\mathrm{t}}^{\mathrm{d}}\setminus\{i\}} P_{\mathrm{d}}|h_{j,i}|^2 d_{j,i}^{-\alpha} + \sum\limits_{k\in\Phi_{\mathrm{t}}^{\mathrm{s}}} P_{\mathrm{s}}|h_{k,i}|^2 d_{k,i}^{-\alpha}} > \theta\right] \\ &= \mathbb{E}\left[\mathcal{L}_{I_{\mathrm{d}}}\left(\theta d_i^{\alpha}\right)\cdot\mathcal{L}_{I_{\mathrm{s}}}\left(\theta\frac{P_{\mathrm{s}}}{P_{\mathrm{d}}}d_i^{\alpha}\right)\right] \\ &= \mathbb{E}\left[\exp\left(-\frac{\pi d_i^2\theta^{\frac{2}{\alpha}}}{\mathrm{sinc}(2/\alpha)}\left(\lambda_{\mathrm{t}}^{\mathrm{d}}+\lambda_{\mathrm{t}}^{\mathrm{s}}\left(P_{\mathrm{s}}/P_{\mathrm{d}}\right)^{\frac{2}{\alpha}}\right)\right)\right], \end{aligned} \tag{4.13}$$

where $\mathcal{L}_{I_{\mathrm{x}}}(s) = \mathbb{E}\left[\exp\left(-sI_{\mathrm{x}}\right)\right]$ is the Laplace transform of interference I_{x} and θ is the SIR threshold for successful D2D transmission. The expectation is over the distribution of d_i and over the content library $\mathcal{F}$.

When a cache hit occurs, the distribution of the D2D link distance d_i depends on the popularity order of the requested file. If f_i is requested by the typical user, conditioning on having at least one potential D2D transmitters within distance R_{d}, the probability density function (PDF) of the D2D link distance d_i is given by

$$f_{d_i}(r) = \begin{cases} \frac{2\pi(1-\rho)\lambda_{\mathrm{u}}q_i r}{1-e^{-\pi(1-\rho)\lambda_{\mathrm{u}}q_i R_{\mathrm{d}}^2}}e^{-\pi(1-\rho)\lambda_{\mathrm{u}}q_i r^2} & 0\le r\le R_{\mathrm{d}} \\ 0 & r > R_{\mathrm{d}}. \end{cases} \tag{4.14}$$

Then we have the approximate success probability when a cache hit of file f_i happens, given by

$$\begin{aligned} p_{\mathrm{suc,i}}^{\mathrm{d}} = \int_0^{R_{\mathrm{d}}} &\exp\left[-\frac{\pi r^2\theta^{\frac{2}{\alpha}}}{\mathrm{sinc}(2/\alpha)}\left(\rho\lambda_{\mathrm{u}}p_{\mathrm{r}}^{\mathrm{d}}+\lambda_{\mathrm{t}}^{\mathrm{s}}\left(P_{\mathrm{s}}/P_{\mathrm{d}}\right)^{\frac{2}{\alpha}}\right)\right] \\ &\cdot\frac{2\pi(1-\rho)\lambda_{\mathrm{u}}q_i r}{1-e^{-\pi(1-\rho)\lambda_{\mathrm{u}}q_i R_{\mathrm{d}}^2}}e^{-\pi(1-\rho)\lambda_{\mathrm{u}}q_i r^2}\mathrm{d}r, \end{aligned} \tag{4.15}$$

where $p_{\mathrm{r}}^{\mathrm{d}}$ and $\lambda_{\mathrm{t}}^{\mathrm{s}}$ are given in (4.7) and (4.12), respectively.

SC Caching

In the case with SC caching, users are always connected to the nearest SBSs in both cache hit and cache miss events. Denote by $\tilde{\Phi}_{\mathrm{t}}^{\mathrm{s}}$ the set of active SBSs, similarly, using the void probability of Voronoi cell in (4.11), the density of active SBSs is given by

$$\tilde{\lambda}_{\mathrm{t}}^{\mathrm{s}} \simeq \lambda_{\mathrm{s}}\left(1-\left(1+\frac{\rho\lambda_{\mathrm{u}}}{3.5\lambda_{\mathrm{s}}}\right)^{-3.5}\right). \tag{4.16}$$

Conditioning on having the typical receiver at the origin with its associated SBS at distance d_s and using nearest SBS association, we have the pdf of d_s given as

$$f_{d_s}(r) = 2\pi\lambda_{\mathrm{s}} r \cdot e^{-\pi\lambda_{\mathrm{s}} r}. \tag{4.17}$$

For a given SIR threshold θ, the success probability of the cache-assisted small cell transmission is given by

$$\begin{aligned} p_{\mathrm{suc}}^{\mathrm{s}} &= \mathbb{P}\left[\frac{P_{\mathrm{s}}|h_{i,i}|^2 d_s^{-\alpha}}{\sum\limits_{k\in\tilde{\Phi}_{\mathrm{t}}^{\mathrm{s}}\setminus\{i\}} P_{\mathrm{s}}|h_{k,i}|^2 d_{k,i}^{-\alpha}} > \theta\right] \\ &= \int_0^\infty 2\pi\lambda_{\mathrm{s}} r \cdot e^{-\pi\lambda_{\mathrm{s}} r^2} \exp\left(-\frac{\pi\tilde{\lambda}_{\mathrm{t}}^{\mathrm{s}} r^2 \theta^{\frac{2}{\alpha}}}{\mathrm{sinc}(2/\alpha)}\right) \mathrm{d}r, \end{aligned} \tag{4.18}$$

where $\tilde{\lambda}_{\mathrm{t}}^{\mathrm{s}}$ is given in (4.16).

4.2.3 Density of Cache-served Requests

A user request is said to be "served" if the requested file is found in local caches and if the transmission of the file is successful. Based on the above results, we calculate the density of cache-served requests, which is the average number of requests that can be successfully and simultaneously handled by the local cache per unit area.

D2D Caching

In the D2D caching case, a random user request can be served either by self-request or through proximal D2D communication. Denote by $\mu_{\mathrm{suc}}^{\mathrm{d}}$ the density of cache-served requests, we have

$$\begin{aligned} \mu_{\mathrm{suc}}^{\mathrm{d}} &= \rho\lambda_u \left(p_{\mathrm{self}}^{\mathrm{d}} + \sum_{i=1}^{N} p_i(1-q_i) p_{\mathrm{hit},i}^{\mathrm{d}} p_{\mathrm{suc,i}}^{\mathrm{d}}\right) \\ &= \rho\lambda_u \sum_{i=1}^{N} p_i \left[q_i + (1-q_i)\left(1 - e^{-\pi(1-\rho)\lambda_{\mathrm{u}} q_i R_d^2}\right) p_{\mathrm{suc,i}}^{\mathrm{d}}\right], \end{aligned} \tag{4.19}$$

where $p_{\mathrm{suc,i}}^{\mathrm{d}}$ is given in (4.15).

SC Caching

In the SC caching case, the maximum number of cache-assisted transmissions in a given time slot is limited by the density of the SBSs. The probability for

a SBS to have at least one active user in its cell requesting for files that are stored in its cache is given by

$$p_{\mathrm{t}}^{\mathrm{s}} = 1 - \left(1 + \frac{\rho\lambda_{\mathrm{u}} p_{\mathrm{hit}}^{\mathrm{s}}}{3.5\lambda_{\mathrm{s}}}\right)^{-3.5}. \tag{4.20}$$

The density of cache-assisted transmission is $\lambda_{\mathrm{s}} p_{\mathrm{t}}^{\mathrm{s}}$. Then, the density of cache-served requests, which is the density of successful cache-served small cell transmission, is given by

$$\mu_{\mathrm{suc}}^{\mathrm{s}} = \lambda_{\mathrm{s}} p_{\mathrm{t}}^{\mathrm{s}} p_{\mathrm{suc}}^{\mathrm{s}} = \lambda_{\mathrm{s}} \left(1 - \left(1 + \frac{\rho\lambda_{\mathrm{u}} p_{\mathrm{hit}}^{\mathrm{s}}}{3.5\lambda_{\mathrm{s}}}\right)^{-3.5}\right) p_{\mathrm{suc}}^{\mathrm{s}}, \tag{4.21}$$

where $p_{\mathrm{suc}}^{\mathrm{s}}$ is given in (4.18).

4.2.4 Power Consumption

For a random user request, in a cache hit event, the consumed power for content delivery contains only the transmit power of either the D2D transmitter or the associated SBS. In a cache miss event, the requested file is first fetched from the core network via the backhaul and then transmitted from the nearest SBS to the user. Thus, additional energy is consumed at the backhaul. Denote P_{b} the backhaul power consumption required to handle a user request at a single SBS; we study below the power consumption per user request with either D2D caching or SC caching.

For the case with D2D caching, recall that when self-request occurs, no energy is consumed to serve the request. A random user request has probability $p_{\mathrm{r}}^{\mathrm{d}}$ to be served by a nearby D2D transmitter, and probability $1 - p_{\mathrm{hit}}^{\mathrm{d}}$ to be served by the nearest SBS. We have the consumed power per user request given as

$$P_{\mathrm{avg}}^{\mathrm{d}} = p_{\mathrm{r}}^{\mathrm{d}} P_{\mathrm{d}} + (1 - p_{\mathrm{hit}}^{\mathrm{d}})(P_{\mathrm{s}} + P_{\mathrm{b}}). \tag{4.22}$$

For the case with SC caching, when a random user request occurs, the transmission power of its nearest SBS is always consumed for both cache hit and cache miss events. The backhaul power is additionally consumed with probability $1 - p_{\mathrm{hit}}^{\mathrm{s}}$. Thus we have

$$P_{\mathrm{avg}}^{\mathrm{s}} = P_{\mathrm{s}} + (1 - p_{\mathrm{hit}}^{\mathrm{s}}) P_{\mathrm{b}}. \tag{4.23}$$

4.3 Numerical Results

For numerical evaluation, we set the SBS density at $\lambda_{\mathrm{s}} = 10^{-5}$ and the user density at $\lambda_{\mathrm{s}} = K \times 10^{-4}$, where K is a proportion factor. We choose $\rho = 0.2$,

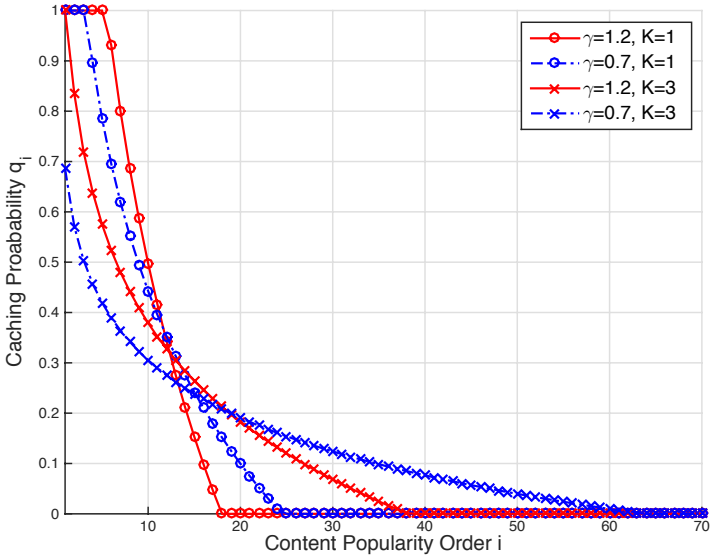


Figure 4.1: Optimal caching probability vector $\mathbf{q} = [q_1, \cdots, q_N]$ of D2D caching.

meaning that 20% of the users will request for a file in the content library $\mathcal{F}$. The cache storage size at a SBS and at a user device are $M_{\mathrm{s}} = 10^4$ and $M_{\mathrm{d}} = 10$, respectively. The content library has size $N = 10^5$, with popularity distribution given by Zipf law with shape parameter $\gamma = 0.7$ for the case with low popularity skewness, and $\gamma = 1.2$ for the other case. The searching distance of a user device to establish D2D link is $R_{\mathrm{d}} = 75$ m. The transmit power of a SBS and a user device are $P_{\mathrm{s}} = 100$ mW and $P_{\mathrm{d}} = 2$ mW, respectively. Backhaul power consumption at the SBS to handle a user request is $P_{\mathrm{b}} = 10P_{\mathrm{s}} = 1$ W. The target SIR of successful transmissions is chosen as $\theta = 0$ dB. We present numerical results for $K \in [1, 9]$, in order to compare the performance between caching at the user devices and caching at the SBSs for different user density regimes.

In Fig. 4.1, we plot the optimal caching probability q_i of file f_i, $i \in [1, N]$, obtained by solving the optimization problem in (4.3). Comparing the results with $\gamma = 1.2$ and with $\gamma = 0.7$, we see that for lower γ, users tend to cache more different files, because the user requests are more diverse due to the low popularity concentration level. We observe the same trend when user density is higher, e.g., $K = 3$. This is reasonable because with higher user density the probability to establish D2D communication is higher. Thus, more different files should be cached in user devices in order to serve more requests by cache-assisted D2D transmission.

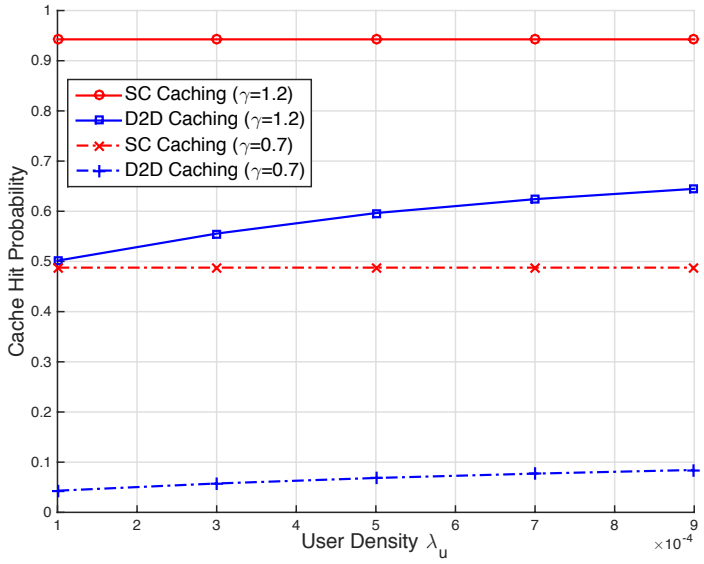


Figure 4.2: Cache hit probability vs. user density λ_s.

4.3.1 Cache Hit Probability

Fig. 4.2 shows the cache hit probability obtained with (4.8) and (4.9) for the case with D2D caching and with SC caching, respectively. As expected, the cache hit probability for both cases are higher with lower γ. Furthermore, we see that caching at the SBSs results in much higher cache hit probability than caching at the mobile devices, as a result of the larger cache capacity.

4.3.2 Density of Cache-served Requests

The density of cache-served requests measures how many requests can be successfully handled simultaneously using the local caches. From Fig. 4.3, we see that D2D caching outperforms SC caching for higher γ, especially in the high user density regime. In the case with lower γ, the performance of SC caching is slightly better in the sparse user regime, whereas the performance of D2D caching outperforms SC caching when the user density increases. The advantage of D2D caching is mainly due to two reasons:

- When the user density is high, the number of potential D2D transmitters are in general larger than the number of SBSs, which allows to have more simultaneous cache-assisted transmission links.

- Self-request of D2D caching gives opportunity to handle a large amount of requests made by the users when the content popularity is highly concentrated.

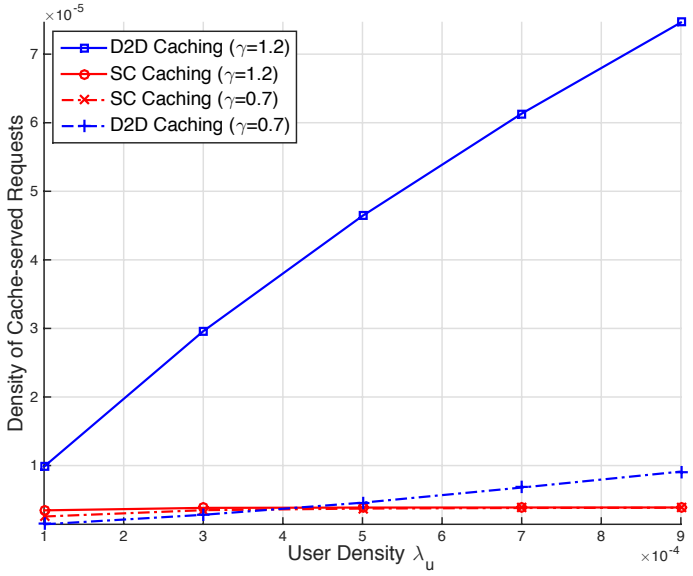


Figure 4.3: Density of cache-served requests vs. user density λ_s.

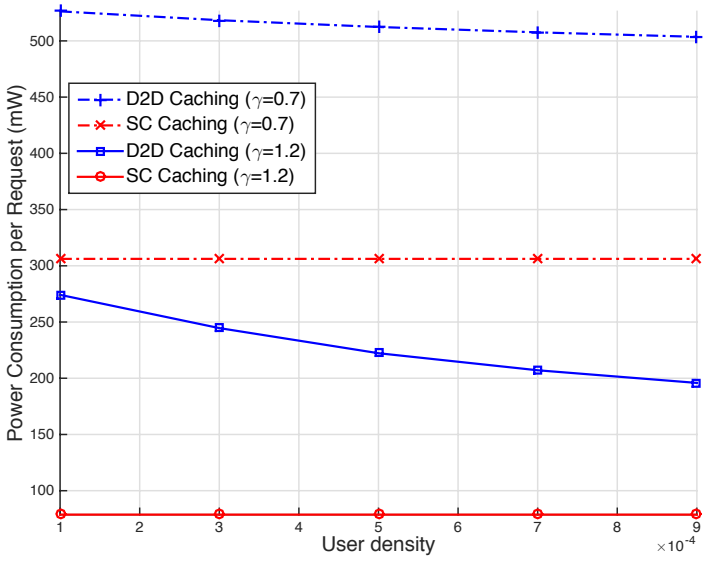


Figure 4.4: Power consumption per user request vs. user density λ_s.

4.3.3 Power Consumption

Fig. 4.4 plots the average power consumption per user request, showing that SC caching is expectedly more energy efficient than the case with D2D caching, which is mainly because of the high cache hit probability. Compared to the consumed power for fetching the content through the backhaul, the transmit power of both user devices and SBSs are relatively low. Hence, the higher the probability to serve a user request locally, the less power is needed. We also observe that the power consumption per user request with D2D caching decreases with the user density. This is because the number of potential D2D transmitters within the discovery distance of a user increases when the user density increases, thus giving higher probability to serve the user request by cache-assisted D2D transmission.

Combining these results, we have the following takeaway messages:

- In networks with high user density, D2D caching gives the opportunity to serve more user requests simultaneously through short-distance cache-assisted D2D communication ;

- Caching at SBSs results in much higher cache hit probability because storage units at the SBSs have much larger capacity than at the mobile devices. As a result, SC caching is also more energy efficient because less power is consumed on average at the backhaul in order to download a file from the core network.

4.4 Cache Hit Optimal vs. Throughput Optimal

In this section, we limit our analysis to wireless D2D caching networks with spatially distributed user devices. We investigate probabilistic caching placement in a stochastic wireless D2D caching network with two different objectives: 1) to maximize the cache hit probability, which is the probability that a random user request can be served locally, either by its own cache or by its neighbor devices within a certain distance, and 2) to maximize the cache-aided throughput, which is the average density of successfully served requests by local caches, including when a user request is served by its own device or by a nearby device through D2D transmission. The optimal solutions are compared, as well as their ability to successfully handle user requests by local caches.

Similar to the previous section when comparing D2D caching with small cell caching, in this section we define the cache hit probability and the cache-aided throughput as the performance metrics that will be used for the optimization of caching probabilities. Note that the case of finding the requested file of a device in its own cache storage is often overlooked in helper-based D2D caching

networks in the literature. This is the major difference between the cache-related performance study in this chapter and the prior work in [67, 68, 80].

From (4.8), we have the cache hit probability given by

$$p_{\text{hit}}^{\text{d}} = 1 - \sum_{i=1}^{N} p_i \left(1 - q_i\right) e^{-\pi(1-\rho)\lambda_{\text{u}} q_i R_{\text{d}}^2}. \tag{4.24}$$

We define another metric that measures the average number of requests that can be successfully and simultaneously handled by the local caches per unit area, namely the cache-aided throughput (per area). Assume that the transmission of each file with equal size takes the same amount of time, one slot for instance. In the self request case, the request is automatically served with probability one, while in the D2D cache hit case, the success probability of content delivery depends on the received SINR. Thus, we have the cache-aided throughput given by

$$\mathcal{T} = \rho\lambda_{\text{u}} \left[\sum_{i=1}^{N} p_i q_i \cdot 1 + \sum_{i=1}^{N} p_i (1 - q_i) p_{\text{hit},i}^{\text{d}} \cdot p_{\text{suc},i}^{\text{d}} \right], \tag{4.25}$$

where $p_{\text{suc},i}^{\text{d}}$ is the success probability of D2D transmission for file f_i, $\rho\lambda_{\text{u}}$ is the density of user requests in a given time slot.

From the analytical results presented in Section 4.2.3, after removing the items related to the SBS interference, we can easily have

$$p_{\text{hit},i}^{\text{d}} = 1 - e^{-\pi(1-\rho)\lambda_{\text{u}} q_i R_d^2}, \tag{4.26}$$

and

$$p_{\text{suc,i}}^{\text{d}} = \int_0^{\infty} f_{d_i}(r) \exp\left(-\frac{\pi\rho\lambda_{\text{u}} p_{\text{hit}}^{\text{d}} r^2 \theta^{\frac{2}{\alpha}}}{\text{sinc}(2/\alpha)} \right) e^{-\frac{\theta\sigma^2 r^{\alpha}}{P_{\text{d}}}} \text{d}r, \tag{4.27}$$

where

$$f_{d_i}(r) = \begin{cases} \frac{2\pi(1-\rho)\lambda_{\text{u}} q_i r}{1 - e^{-\pi(1-\rho)\lambda_{\text{u}} q_i R_{\text{d}}^2}} e^{-\pi(1-\rho)\lambda_{\text{u}} q_i r^2} & 0 \le r \le R_{\text{d}} \\ 0 & r > R_{\text{d}}, \end{cases} \tag{4.28}$$

and $p_{\text{r}}^{\text{d}} = \sum_{i=1}^{N} p_i \left(1 - q_i\right) \left(1 - e^{-\pi(1-\rho)\lambda_{\text{u}} q_i R_{\text{d}}^2}\right)$. Substituting (4.26), (4.27) and (4.28) in (4.25), we obtain the cache-aided throughput averaged over all the files in the content library.

4.5 Optimization of Probabilistic Caching Placement

In this section we study the optimal caching probabilities $\mathbf{q} = [q_1, \ldots, q_N]$ by cache hit maximization and by cache-aided throughput optimization, respectively.

4.5.1 Cache Hit Maximization

Based on (4.24), the optimization problem in order to maximize the cache hit probability is defined as

$$\begin{aligned} \max_{\mathbf{q}} \quad & p_{\text{hit}}^{\text{d}} = 1 - \sum_{i=1}^{N} p_i \left(1 - q_i\right) e^{-\pi(1-\rho)\lambda_{\text{u}} q_i R_{\text{d}}^2} \\ \text{s.t.} \quad & 0 \leq q_i \leq 1 \text{ for } i \in [1, N], \\ & \sum_{i=1}^{N} q_i \leq M_{\text{d}}. \end{aligned} \tag{4.29}$$

The second order derivative of the objective function is strictly negative, thus $p_{\text{hit}}^{\text{d}}$ is a concave function of q_i for $i \in [1, N]$. Consider the following Lagrangian function

$$\begin{aligned} \mathcal{L}(\mathbf{q}, \mu) = -1 + \sum_{i=1}^{N} p_i \left(1 - q_i\right) e^{-\pi(1-\rho)\lambda_{\text{u}} q_i R_{\text{d}}^2}. \\ + \mu \left(\sum_{i=1}^{N} q_i - M_{\text{d}} \right), \end{aligned} \tag{4.30}$$

where μ is the non-negative Lagrangian multiplier. We solve this optimization problem by applying the Karush-Kuhn-Tucker (KKT) conditions. From $\frac{\partial \mathcal{L}}{\partial q_i} = 0$, we have

$$q_i(\mu_i) = -\frac{\mathcal{W}\left\{ \frac{\mu}{p_i} \exp\left[1 + \pi(1-\rho)\lambda_{\text{u}} R_{\text{d}}^2\right] \right\}}{\pi(1-\rho)\lambda_{\text{u}} R_{\text{d}}^2} + 1 + \frac{1}{\pi(1-\rho)\lambda_{\text{u}} R_{\text{d}}^2}, \tag{4.31}$$

where $\mathcal{W}$ denotes the Lambert W function [103]. Combined with the condition $0 \leq q_i \leq 1$, let $[x]^+ = \max\{x, 0\}$, we have

$$q_i^\star = \min\left\{ [q_i(\mu^\star)]^+, 1 \right\}, \tag{4.32}$$

where $\mu^\star$ can be obtained by the bisection search method under the other KKT condition $\sum_{i=1}^{N} q_i^\star = M_{\text{d}}$.

4.5.2 Cache-aided Throughput Maximization

Due to the complicated expression of $\mathcal{T}$, the optimal caching probabilities that maximize the cache-aided throughput are difficult to obtain, even with numerical methods. Consider the following approximation

$$\mathbb{E}_{d_i}[\exp\left(-\eta d_i^\delta\right)] \approx \exp\left(-\eta \mathbb{E}[d_i^2]^{\delta/2}\right), \tag{4.33}$$

the success probability $p_{\text{suc},i}^{\text{d}}$ in (4.13) can be approximated by

$$\hat{p}^{\mathrm{d}}_{\mathrm{suc},i} \approx \exp\left[-\frac{\pi\rho\lambda_{\mathrm{u}}p^{\mathrm{d}}_{\mathrm{hit}}\mathbb{E}[d_i^2]\theta^{2/\alpha}}{\mathrm{sinc}(2/\alpha)}\right]\exp\left[-\frac{\theta\sigma^2\mathbb{E}[d_i^2]^{\alpha/2}}{P_{\mathrm{d}}}\right]. \tag{4.34}$$

From the PDF of d_i in (4.14), we can obtain $\mathbb{E}[d_i^2]$ as follows.

$$\begin{aligned}\mathbb{E}[d_i^2] &= \int_0^{R_{\mathrm{d}}} r^2 \frac{2\pi(1-\rho)\lambda_{\mathrm{u}}q_i r}{1-e^{-\pi(1-\rho)\lambda_{\mathrm{u}}q_i R_{\mathrm{d}}^2}} e^{-\pi(1-\rho)\lambda_{\mathrm{u}}q_i r^2}\mathrm{d}r \\ &= \frac{1}{\pi(1-\rho)\lambda_{\mathrm{u}}q_i} - \frac{R_{\mathrm{d}}^2}{e^{\pi(1-\rho)\lambda_{\mathrm{u}}q_i R_{\mathrm{d}}^2}-1}.\end{aligned} \tag{4.35}$$

When $q_i \to 0$, we obtain $\lim_{q_i\to 0}\mathbb{E}[d_i^2] = R_{\mathrm{d}}^2/2$ by applying L'Hôpital's rule.

Then we have the approximated cache-aided throughput as

$$\hat{\mathcal{T}} = \rho\lambda_{\mathrm{u}}\left[\sum_{i=1}^{N} p_i q_i + \sum_{i=1}^{N} p_i(1-q_i)p^{\mathrm{d}}_{\mathrm{hit},i}\cdot\hat{p}^{\mathrm{d}}_{\mathrm{suc},i}\right], \tag{4.36}$$

where $\hat{p}^{\mathrm{d}}_{\mathrm{suc},i}$ is given in (4.34). Our objective is to find $\mathbf{q}^\star = \max_{\mathbf{q}} \hat{\mathcal{T}}$, subject to $0 \le q_i \le 1$ and $\sum_{i=1}^{N} q_i \le M_{\mathrm{d}}$.

This problem is non-convex as it can be seen numerically. Providing an analytical solution to this problem is difficult. Therefore for the the cache-aided throughput maximization we solve it numerically with Simulated Annealing.

4.6 Numerical and Simulation Results

For numerical evaluation, we consider the user density between $\lambda_{\mathrm{u}} = [10^{-4}, 10^{-3}]$ /m^2. $\rho = 50\%$ of the users will request for a random file in $\mathcal{F}$ according to the request probabilities $\mathbf{p} = [p_1, \ldots, p_N]$, which follows the Zipf distribution with parameter $\gamma = \{0.5, 1.2\}$. The rest 50% of users act as potential D2D transmitters helping to serve the user requests locally. The device cache capacity is $M_{\mathrm{d}} = 2$ files. The content library has size $N = 20$ files.[1] The D2D searching distance is $R_{\mathrm{d}} = 75$ m. The device transmission power and the background noise power are $P_{\mathrm{d}} = 0.1$ mW and $\sigma^2 = -110$ dBm, respectively. The target SINR of successful D2D transmissions is $\theta = 0$ dB.

In Fig. 4.5 and Fig. 4.6 we compare the cache-hit-optimal (Section 4.5) and throughput-optimal (Section 4.5.2) caching probabilities $\mathbf{q}^\star$ in sparse and dense user environments, respectively. The optimal caching probabilities of file f_i for $i = 1, \ldots, N$ are plotted as a function of the popularity order i. Interestingly, we observe that with sparse users, throughput-optimal and cache-hit-optimal caching probabilities are very close, while with dense users, each device tends

[1]Note that in reality the content library size is very large. Here we take $N = 20$ files to avoid high complexity of the optimization problem. Similar choices can also be found in [80] and [74].

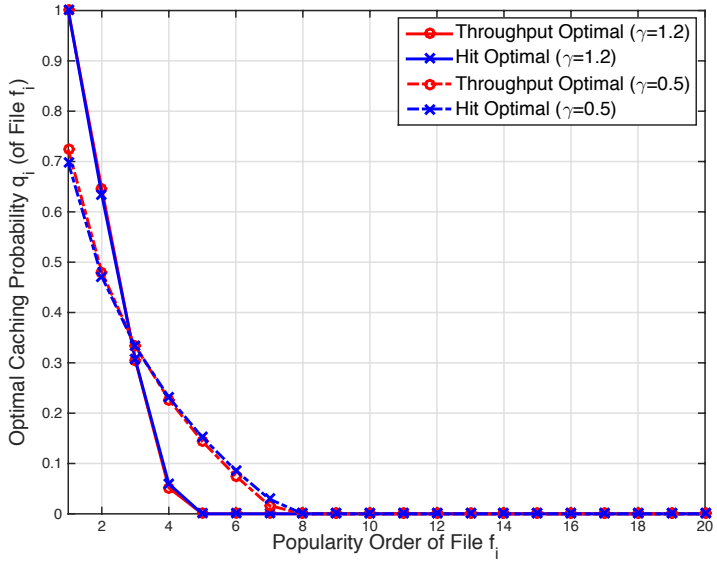


Figure 4.5: Optimal caching probabilities with sparse devices, i.e., $\lambda_u = 10^{-4}$.

to cache the most popular files with higher probability in order to increase the cache-aided throughput. For instance, in Fig. 4.6, $q_1^\star$, $q_2^\star$ and $q_3^\star$ in the throughput-optimal case are much higher than in the cache-hit-optimal case. The intuition behind this is that *in dense user regime, due to the excessive D2D interference that leads to very low D2D success probability, users' caching strategy tends to be more "selfish" in the sense that self-request matters more than cache-assisted D2D transmission.*

In Fig. 4.7 we plot the simulated cache-aided throughput obtained with the throughput-optimal caching probabilities. In order to validate the accuracy of the approximation we used in (4.34), we plot the theoretical values of the approximated cache-aided throughput $\hat{\mathcal{T}}$, which turn out to have negligible error. For the comparison of different caching strategies, we also plot the simulated cache-aided throughput when applying $\mathbf{q}^\star$ obtained with cache hit probability optimization and with the conventional "cache the most popular content" (MPC) strategy. It is obvious that with the throughput-optimal strategy that is aware of the D2D success probability, the achieved cache-aided throughput can be significantly improved compared to the cache hit optimization and the MPC strategies. The gain is more profound in the dense user regime. Another interesting remark is that with dense users and highly concentrated content popularity ($\gamma = 1.2$), the cache-aided throughput with MPC gives better performance than the cache-hit-optimal case, meaning that it is more beneficial to increase the chance of "self-request" than increasing the total cache hit ratio. As summary, our results validate the necessity of taking into account the transmission reliability of cache-aided D2D communication while searching for the optimal content placement.

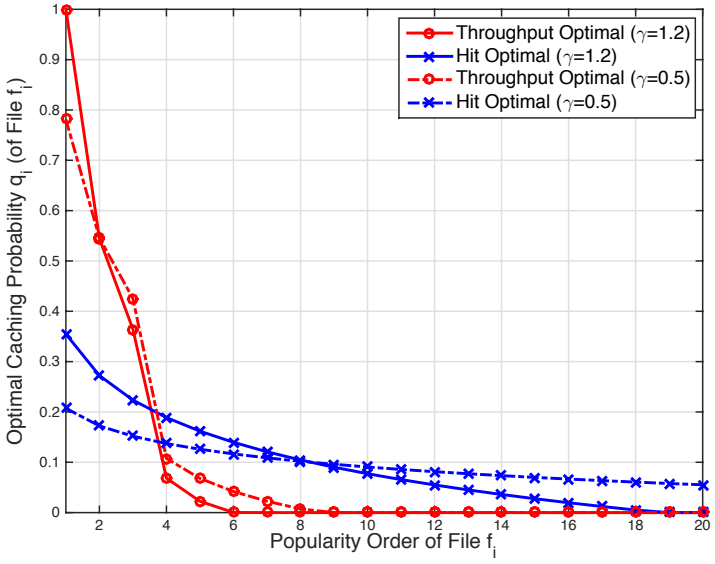


Figure 4.6: Optimal caching probabilities with dense devices, i.e., $\lambda_u = 10^{-3}$.

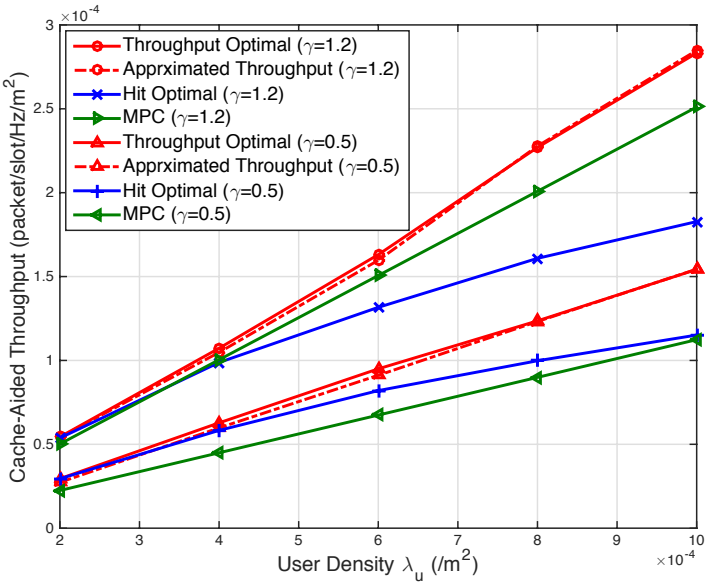


Figure 4.7: Simulated cache-aided throughput vs. user device density λ_u. $\gamma = \{0.5, 1.2\}$

4.7 Summary and Concluding Remarks

In this chapter, we studied the probabilistic caching placement strategy in two types of wireless caching architectures and with two types of optimization objectives. Based on the numerical and simulation results, we have the following remarks and takeaway messages:

- In a network with dense D2D users, caching at user devices results in higher throughput-related performance because proximity-based D2D communication gives better spectrum utilization.
- Caching at SBSs in general gives higher cache hit probability as a result of the high storage capacity at the SBSs. Consequently, backhaul load and power consumption can also be reduced.
- By taking into account the success probability of D2D transmission in the optimal probabilistic caching placement, the density of successfully served user request by local caches can be significantly improved compared to the conventional cache hit optimization method, especially in the dense user regime.

Though in this chapter we have considered relatively simple caching network structures with basic performance evaluation based on the received SIR distribution, our presented results bring insights on the design of wireless caching networks. The main caching strategy we used in this chapter is the probabilistic content placement, which means that each SBS or user device decides to cache content independently according to a certain probability distribution. When nearby caching helpers such as SBSs can cooperate to serve users as an opportunity enabled by existing coordinated multi-point (CoMP) techniques, how to adjust the caching strategy among the cooperative SBSs becomes an interesting subject to study. To address this question, in the next chapter we will present a cooperative caching and transmission strategy in cluster-centric SCNs, and study the impact of network parameters on the cache-related performance in such networks.

Chapter 5

Small Cell Cooperative Caching

When multiple base stations (BSs) have overlapping coverage area, considering independent caching strategies at each BS might lead to inefficient usage of the overall cache space among the BSs that might cooperate to serve users in the overlapped coverage area. In the literature, there are several studies on the joint design of cooperative caching and physical layer (PHY) transmission [83, 84]. The existing coordinated multi-point (CoMP) techniques, such as coordinated scheduling and coordinated beamforming (CS/CB) and joint transmission (JT) among the the BSs in nearby cells, bring PHY cooperation gain that can be further incorporated with cache-level cooperative content placement.

In this chapter, we consider a cluster-centric small cell network (SCN) with combined design of cooperative caching and transmission policy. Small base stations (SBSs) are grouped into disjoint clusters, in which clustered cache space is utilized as an entity. We propose a combined caching scheme where part of the available cache space is reserved for caching the most popular content in every SBS, while the remaining is used for cooperatively caching different partitions of the less popular content in different SBSs, as a means to increase local content diversity. Depending on the availability and placement of the requested content, CoMP technique with either joint transmission (JT) or parallel transmission (PT) is used to deliver content to the served user. We show an inherent tradeoff between transmission diversity and content diversity in our cooperation design. We also study optimal cache space assignment for two objective functions: maximization of the cache service performance and the energy efficiency. Simulation results show that the proposed scheme achieves performance gain by leveraging cache-level and signal-level cooperation and adapting to the network environment and user requirements.

This chapter is organized as follows. We present the network model and cooperation schemes in Section 5.1. In Section 5.2 we define the successful content delivery probability (SCDP) as the main performance metric and give analytical results for JT and PT transmission schemes. Based on numerical results of the successful content delivery probability (SCDP) and cache hit

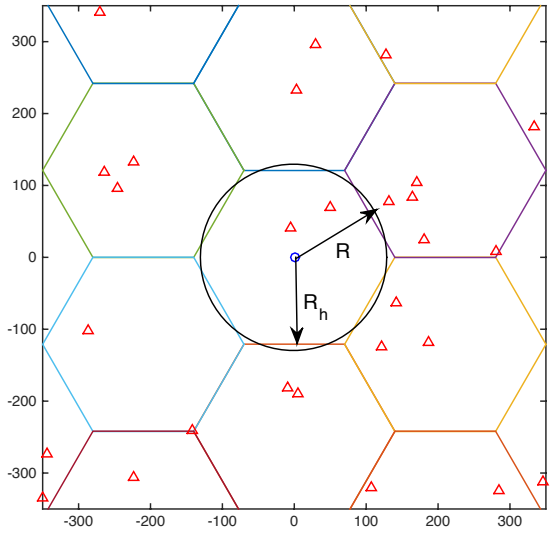


Figure 5.1: A snapshot of the cluster-centric SCN topology. The clusters are defined by a hexagonal grid, wherein SBSs (red triangles) are distributed according to a homogeneous PPP. A cluster of interest is considered for performance analysis with cluster center at the origin.

probability, in Section 5.3 we show the tradeoff between transmission diversity and content diversity. We then pose two optimization problems for the optimal cache space assignment. Simulation results are presented in Section 5.4 and Section 5.5 concludes this chapter.

5.1 Network Model

5.1.1 Small Cell Clustering

We consider a cache-enabled SCN where SBSs are distributed according to a homogeneous PPP $\Phi_b = \{b_i \in \mathbb{R}^2, \forall i \in \mathbb{N}^+\}$ with intensity λ_b. Nearby SBSs are grouped into disjoint clusters modeled using a hexagonal grid with inter-cluster center distance equal to $2R_\mathrm{h}$ [104], as shown in Fig. 5.1. SBSs belonging to the same cluster can cooperate to serve users inside the cluster. The total cache (storage) capacity in a cluster is considered as an entity and cache placement decisions are performed at the central controllers, which are located at the center of each cluster, denoted by $\mathcal{H} = \{y_j \in \mathbb{R}^2, \forall j \in \mathbb{N}^+\}$. The area of each cluster is given by $\mathcal{A} = 2\sqrt{3}R_\mathrm{h}^2$. For a random cluster, the probability mass function (PMF) of the number n of SBSs inside, which follows a Poisson distribution with mean $\lambda_b\mathcal{A}$, is given by

$$\mathbb{P}(n = K) = e^{-2\sqrt{3}\lambda_b R_\mathrm{h}^2} \frac{\left(2\sqrt{3}\lambda_b R_\mathrm{h}^2\right)^K}{K!}. \tag{5.1}$$

In clusters with no SBS inside, i.e., $n = 0$, users connect to the nearest SBS to download the requested content. For simplicity, we do not consider the case of

empty clusters.[1]

Conditioning on having K SBSs in the cluster of interest with cluster center y_0 at the origin, the in-cluster SBS distribution follows a binomial point process (BPP), which consists of K uniformly and independently distributed SBSs in the hexagonal cluster. The distance distribution between randomly distributed nodes and the cell center for hexagonal cell can be found in [105]. For analytical convenience, we approximate the cluster area to a circle with the same area, i.e., with radius $R = R_{\mathrm{h}}\sqrt{\frac{2\sqrt{3}}{\pi}}$, as shown in Fig. 5.1. The set of cooperative SBSs inside the cluster of interest is thus defined as $\mathcal{C} = \{b_i \in \Phi_b \cap \mathcal{B}(y_0, R)\}$, where $\mathcal{B}(y_0, R)$ denotes the ball centered at y_0 with radius R. This approximation turns out to have negligible impact on the performance of the network under study [105]. Consider a user located at the origin (cell center), the distances from the cooperative SBSs to the user are denoted by $\mathbf{r} = [r_1, r_2, \ldots, r_K]$. The K cooperative SBSs can be approximately seen as the K closest SBSs to the cluster-center user.

5.1.2 Cache Placement Strategies

We consider a finite content library $\mathcal{F} = \{f_1, \ldots, f_N\}$, where N is the library size and f_m is the m-th most popular file with normalized size equal to 1. Each user makes independent request for a file with probability according to a given popularity pattern, e.g., Zipf distribution, which is a commonly used distribution for video popularity [52–54, 78, 106]. Suppose we have the request probability of each file in $\mathcal{F}$ denoted by $\mathbf{p} = \{p_1, \ldots, p_N\}$. With Zipf distribution, the request probability of the m-th most popular file is given as [2]

$$p_m = \left(m^{\gamma} \sum_{n=1}^{N} n^{-\gamma} \right)^{-1}, \tag{5.2}$$

where γ is the shape parameter, denoting the popularity skewness [60].

Due to finite caching capacity, each SBS can store up to M files. In a cluster with K cooperating SBSs, the total available storage capacity is KM. Each file is divided into K equal-size partitions [107]. In our cluster-centric SCN model, we consider a combined "most popular content (MPC)" and "largest content diversity (LCD)" caching strategy with partition-based caching to distribute partitions of content to the SBSs in the same cluster. Specifically, a proportion ρ of cache space in each SBS is used for caching the most popular files, and the rest $1 - \rho$ proportion is reserved for disjointly placing different partitions

[1] In this work, our main interest is the cache content placement in a cluster-centric SCN. When the cluster is empty, there is no cache placement to perform. Hence, this case can be ignored in our analysis.

[2] As the group of users belonging to each cluster changes dynamically in mobile small cell networks, it is difficult to design the cache placement according to the specific content preferences of specific users in the cluster. Hence, our analysis is performed based on a typical user with averaged content preference, whose distribution is given by Zipf-law distribution.

of the less popular files in different SBSs to increase the content diversity.[3] Hence, files f_m with popularity order $1 \leq m \leq \lfloor \rho M \rfloor$ are cached in every SBS inside the cluster (i.e., MPC-based caching). For files f_m with $\lfloor \rho M \rfloor < m \leq \lfloor \rho M \rfloor + K(M - \lfloor \rho M \rfloor)$, every SBS has one different partition of each file (i.e., LCD-based caching). For $m > \lfloor \rho M \rfloor + K(M - \lfloor \rho M \rfloor)$, the files are not cached. In total $\lfloor \rho M \rfloor + K(M - \lfloor \rho M \rfloor)$ different files can be cached inside a cluster.

For a random request within the content library $\mathcal{F}$, the *cache hit* probability, i.e., the probability to find the requested file stored in the local cache, is given by

$$P_{\text{hit}}(\rho) = \sum_{m=1}^{\lfloor \rho M \rfloor + K(M - \lfloor \rho M \rfloor)} p_m = \sum_{m=1}^{\lfloor \rho M \rfloor + K(M - \lfloor \rho M \rfloor)} \left(m^{\gamma} \sum_{n=1}^{N} n^{-\gamma} \right)^{-1}. \tag{5.3}$$

Obviously, the cache hit probability is a monotonically decreasing function of ρ. To increase the chance of cache hit, more cache (storage) space should be reserved for the LCD-based caching.

5.1.3 Transmission Schemes

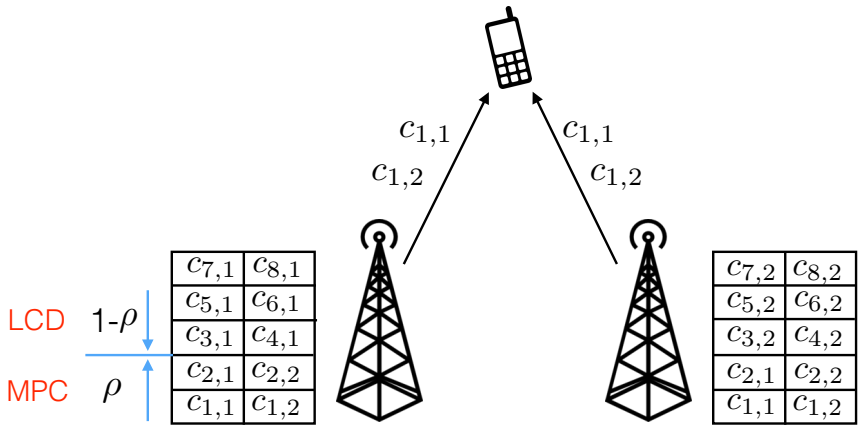
In this work, we assume single antenna at both SBSs and user devices. Hence, in each frequency/time block, only one user in each cluster can be served. If simultaneous content requests arrive at the SBSs inside the same cluster, these requests can be handled using orthogonal multiple access methods, such as time division multiple access (TDMA) and frequency division multiple access (FDMA).

In the cluster of interest with K SBSs, when a random user requests for a file in $\mathcal{F}$, the availability and placement of the requested file enables different transmission schemes, depending on whether the requested file is in the MPC or the LCD range. We study here two transmission schemes, which are designed according to the cache scheme related to the requested file, namely joint transmission (JT) and parallel transmission (PT), as described below.

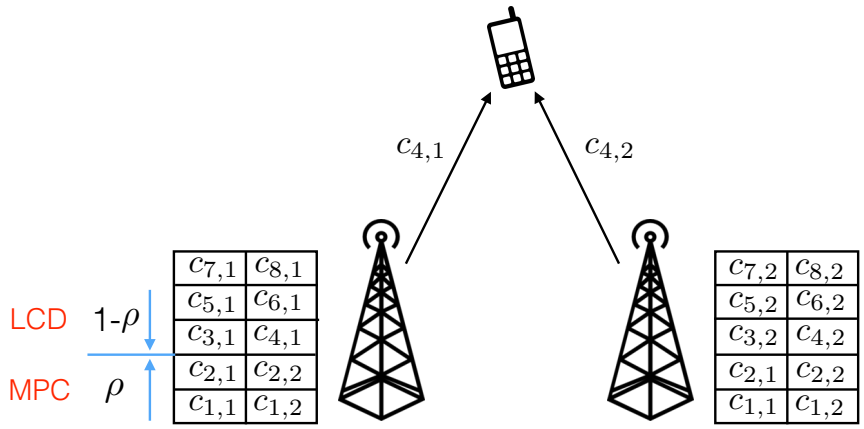
Joint Transmission

If the requested file f_m is in the MPC range, i.e., the popularity order is between $1 \leq m \leq \lfloor \rho M \rfloor$, K SBSs in the cluster have the same entire file. Hence, the requested file is jointly transmitted to the user as a means to enhance the content delivery reliability, i.e., increase the received SINR, as shown in Fig. 5.2(a). We denote this case as JT cooperation scheme.

[3] Content diversity represents the disparity of cached content in each SBS inside the same cluster. It can be seen as the ratio between the number of different files cached within a cluster and the maximum number of files that can be cached inside the cluster. The higher content diversity is, the more different files one can find inside the same cluster.



(a) Joint transmission



(b) Parallel transmission

Figure 5.2: Illustration of combined MPC and LCD caching strategy when $K = 2$. Cooperative SBSs use joint transmission/parallel transmission schemes when a user requests for a file falling into the MPC or the LCD range. Here, $c_{i,j}$ denotes the j-the partition of the i-th file.

Parallel Transmission

If the requested file f_m is in the LCD range, i.e., the popularity order is between $\lfloor \rho M \rfloor < m \leq \lfloor \rho M \rfloor + K(M - \lfloor \rho M \rfloor)$, cooperating SBSs inside the same cluster have disjoint partitions of the requested file, thus joint transmission is not possible in this case. The different partitions need to be transmitted to the user at the same time by parallel (multiple) streams, one from each cooperating SBS, as shown in Fig. 5.2(b). We denote this case as PT cooperation scheme. There are two ways of frequency allocation: i) parallel transmission with orthogonal spectrum assignment (PT-OS) case, where each SBS uses $\frac{1}{K}$ of the overall available spectrum to transmit the stored partition of the requested file to the user; and ii) parallel transmission with successive decoding based spectrum sharing (PT-SS) case, where K SBSs concurrently transmit K streams containing different partitions of the requested file to the user using the same available spectrum. In the PT-SS case, at the receiver, successive decoding with successive interference cancellation (SIC) is used to decode the signal according to the received signal power order [108, 109]. More explicitly, the strongest signal is decoded first and extracted from the received signal, then proceed to the next decoding layer for the next strongest signal, and so on.

Transmission for Cache Miss Case

If the requested file is not cached in local cluster, a *cache miss* event occurs. In this case, all SBSs fetch the requested file from the core network through backhaul links and jointly transmit the file to the user to reduce the delivery latency. The power consumption consists of the required power for fetching content from the core network to the cooperative SBSs and the transmission power for delivering content from the SBSs to the user. The backhauling process increases not only end-to-end delivery delay but also energy consumption, compared to the case of serving user requests by local caches [110, 111]. By considering those impacts, the energy efficiency is investigated in Section 5.3.3.

5.2 PHY Successful Content Delivery Analysis

In this section, we study a key metric for the performance of our considered cluster-centric SCN, namely the successful content delivery probability (SCDP). We give analytical results on the SCDPs of JT, PT-SS and PT-OS cases for a user located at the cluster center. Note that taking the cluster-center user as a reference is mainly done for analytical tractability, but it can be seen as an upper bound on theSCDP for randomly located users inside the cluster.

5.2.1 SCDP Definition

Assuming that each file contains S bits, the successful delivery of a file is defined by the event that S bits are successfully delivered using bandwidth W and time T. Note that in the JT and PT cases, the number of bits delivered from each SBS is different. In the JT case, each SBS sends S bits to the user using the same bandwidth. Hence, at the receiver, the received signals from K SBSs are superimposed and considered as a single stream. TheSCDP is defined as a function of the received SINR (SINR), given as[4]

$$p_{\mathrm{d},K}^{\mathrm{JT}} = \mathbb{P}\left[WT\log_2(1+\mathrm{SINR}) > S \mid K\right]. \tag{5.4}$$

In the PT-SS case, each SBS sends $\frac{S}{K}$ bits to the user employing SIC by sharing the same W bandwidth. The decodability of the received streams depends on the SINR of each stream and rate requirement. Decoding K streams using SIC is theoretically feasible if all K streams achieves higher rate than the target rate for successful transmission[108]. Hence, we have

$$p_{\mathrm{d},K}^{\text{PT-S}} = \mathbb{P}\left[\bigcap_{i\in\{1,\ldots,K\}} WT\log_2(1+\mathrm{SINR}_i) > \frac{S}{K} \;\middle|\; K\right], \tag{5.5}$$

where SINR_i is the received SINR of the stream containing the i-th partition of the requested file. In the PT-OS case, each SBS sends $\frac{S}{K}$ bits by using $\frac{W}{K}$ bandwidth each. The delivery rate is bounded by the stream with the lowest achievable rate, so theSCDP is defined as

$$p_{\mathrm{d},K}^{\text{PT-O}} = \mathbb{P}\left[\frac{W}{K}T\log_2(1+\min_{i\in[1,\ldots,K]}\{\mathrm{SINR}_i\}) > \frac{S}{K} \;\middle|\; K\right]. \tag{5.6}$$

We denote $R_d = \frac{S}{T}$ (bit/s) as the target rate for successful content delivery. In terms of SINR requirement, theSCDP can be rewritten as

$$p_{\mathrm{d},K}^{\mathrm{JT}} = \mathbb{P}\left[\mathrm{SINR} > 2^{\frac{R_d}{W}} - 1 \;\middle|\; K\right], \tag{5.7}$$

$$p_{\mathrm{d},K}^{\text{PT-S}} = \mathbb{P}\left[\bigcap_{i\in\{1,\ldots,K\}} \mathrm{SINR}_i > 2^{\frac{R_d}{KW}} - 1 \;\middle|\; K\right], \tag{5.8}$$

$$p_{\mathrm{d},K}^{\text{PT-O}} = \mathbb{P}\left[\min_{i\in[1,\ldots,K]}\{\mathrm{SINR}_i\} > 2^{\frac{R_d}{W}} - 1 \;\middle|\; K\right]. \tag{5.9}$$

[4]TheSCDP represents the probability of guaranteeing the required content delivery rate $\frac{S}{WT}$ bit/s/Hz, which in turn is defined from the QoS requirements of users. This is similar to the successful transmission probability [112], i.e., complementary outage probability, which is also used in [84].

5.2.2 SCDP of MPC-JT strategy

For the cluster-center user located at $y_0 = (0, 0)$, when it requests for file f_m with $1 \leq m \leq \lfloor \rho M \rfloor$, which is in the MPC range, coordinated joint transmission is used to combine coherently the received signals from cooperating SBSs. Hence, over each symbol duration time, the cooperating SBSs transmit the same symbol s. Assuming equal transmit power P_t for every SBS and a standard distance-dependent power law pathloss attenuation, i.e., $r^{-\alpha}$, where $\alpha > 2$ is the pathloss exponent, the channel output at the user is

$$y = \sum_{b_i \in \mathcal{C}} \sqrt{P_t} {r_i}^{-\frac{\alpha}{2}} h_i s + \sum_{b_j \in \Phi_b \setminus \{\mathcal{C}\}} \sqrt{P_t} {r_j}^{-\frac{\alpha}{2}} h_j s_j + n, \tag{5.10}$$

where h_l denotes the small-scale Rayleigh fading from the l-th SBS to the user, which follows $h_l \sim \mathcal{CN}(0, 1)$; r_l denotes the distance from the l-th SBS to the user; s_l denotes the transmitted symbol of the l-th SBS; and n denotes the background thermal noise.

Considering an interference-limited network and neglecting the background thermal noise, the signal-to-interference ratio (SIR) of received signal is given by

$$\mathrm{SIR}_{\mathrm{JT}} = \frac{\left| \sum\limits_{b_i \in \mathcal{C}} h_i {r_i}^{-\frac{\alpha}{2}} \right|^2}{\sum\limits_{b_j \in \Phi_b \setminus \{\mathcal{C}\}} |h_j|^2 {r_j}^{-\alpha}}. \tag{5.11}$$

Using (5.7) and (5.11), we can obtain theSCDP of JT case as follows.

Lemma 5. *For the cluster-center user with target SIR $\theta_1 = 2^{\frac{R_b}{W}} - 1$, theSCDP of JT case with K cooperating SBSs is given by*

$$p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1) \simeq \int_0^R \cdots \int_0^R \mathcal{L}_{I|R}\left(\frac{\theta_1}{\sum_{i=1}^K x_i^{-\alpha}} \right) \prod_{i=1}^K \frac{2x_i}{R^2} \mathrm{d}x_1 \cdots \mathrm{d}x_K, \tag{5.12}$$

where $\mathcal{L}_{I|x}(s)$ is the Laplace transform of the interference coming from SBSs located outside of $\mathcal{B}(0, x)$, given by

$$\mathcal{L}_{I|x}(s) = \exp\left(-\pi \lambda_b s^{\frac{2}{\alpha}} \int_{\frac{x^2}{s^{2/\alpha}}}^{\infty} \frac{1}{1 + w^{\frac{2}{\alpha}}} \mathrm{d}w \right). \tag{5.13}$$

Proof. See Appendix C.1. □

5.2.3 SCDP of LCD-PT strategy

When the cluster-center user requests for file f_m with $\lfloor \rho M \rfloor < m \leq \lfloor \rho M \rfloor + K(M - \lfloor \rho M \rfloor)$, which is in the LCD range, parallel streams containing different partitions of the requested file are simultaneously sent to the user. Considering different spectrum usages, we study SCDPs for PT-SS and PT-OS cases separately in this section.

PT-SS

In the PT-SS case, over each symbol duration time, K SBSs transmit K different symbols (one symbol in each partition) $[s_1, s_2, \ldots, s_K]$ to the user at the origin. If all SBSs use the same transmit power P_t as in the JT case, the channel output at the receiver is

$$y = \sum_{b_j \in \mathcal{C}} \sqrt{P_t} {r_i}^{-\frac{\alpha}{2}} h_i s_i + \sum_{b_j \in \Phi_b \setminus \{\mathcal{C}\}} \sqrt{P_t} {r_j}^{-\frac{\alpha}{2}} h_j s_j + n. \tag{5.14}$$

In order to decode multiple streams simultaneously, we use SIC with respect to a certain order of received signal. The detailed analysis of SIC based on power ordering statistics is out of the scope of this work and has been studied in [109]. For the ease of analysis, we consider here the case where the user decodes different information streams based on the distance order [113]. After approximating the cluster area by the circle $\mathcal{B}(0, R)$, the decoding order will be from the nearest SBS to the K-th nearest SBS to the cluster-center user. We define $\widetilde{\mathbf{r}} = [\widetilde{r}_1, \ldots, \widetilde{r}_K]$ the distance vector with increasing distance order, where $\widetilde{r}_k, k \in [1, K]$ is the distance from the k-th nearest SBS to the cluster-center user.

When decoding the information from the k-th nearest SBS, all signals coming from closer SBSs $\{b_1, \ldots, b_{k-1}\}$ need to be successfully decoded and canceled. In this case, the interference comes from $K - k$ remaining SBSs inside the cluster and PPP distributed SBSs outside the cluster. Due to the conditioned number K, the interference distribution is different from the case with PPP-distributed SBSs. For the tractability analysis, we assume that at the k-th decoding step with $k \in [1, K-1]$, the distribution of interfering SBSs outside $\mathcal{B}(0, \widetilde{r}_k)$ still follows a homogeneous PPP. The SIR of the k-th stream with SIC is thus given as

$$\mathrm{SIR}_k \simeq \frac{|h_k|^2 \widetilde{r}_k^{-\alpha}}{\sum\limits_{b_j \in \Phi_b \setminus \mathcal{B}(0, \widetilde{r}_k)} |h_j|^2 {r_j}^{-\alpha}}. \tag{5.15}$$

At the last decoding step, all in-cluster interfering signals are canceled. Out-of-cluster SBSs have minimum distance R to the user. Hence, for the last decoded stream, we have

$$\mathrm{SIR}_K \simeq \frac{|h_K|^2 \widetilde{r}_K^{-\alpha}}{\sum\limits_{b_j \in \Phi_b \setminus \mathcal{B}(0, R)} |h_j|^2 {r_j}^{-\alpha}}. \tag{5.16}$$

Using (5.8), (5.15), and (5.16), we now obtain theSCDP of PT-SS case as follows.

Lemma 6. *For the cluster-center user with target SIR* $\theta_2 = 2^{\frac{R_d}{KW}} - 1$*, theSCDP*

of PT-SS case with K cooperating SBSs is given by

$$p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2) \simeq \int\limits_{0<x_1<\cdots<x_K<R} \frac{2K\cdot x_K}{R^2}\mathcal{L}_{I|R}\left(\theta_2 x_K^{\alpha}\right)\prod_{k=1}^{K-1}\frac{2k\cdot x_k}{R^2}\mathcal{L}_{I|x_k}\left(\theta_2 x_k^{\alpha}\right)\mathrm{d}x_1\cdots\mathrm{d}x_K, \tag{5.17}$$

where $\mathcal{L}_{I|x}(s)$ is defined in (5.13).

Proof. See Appendix C.2. □

PT-OS

In the PT-OS case, different SBSs transmit different partitions of the requested content through orthogonal frequency bandwidth. For the information stream transmitted from the i-th SBS, we have the channel output as

$$y_i = \sqrt{P_t}{r_i}^{-\frac{\alpha}{2}}h_i s_i + \sum_{b_j\in\Phi_b\setminus\{\mathcal{C}\}}\sqrt{P_t}{r_j}^{-\frac{\alpha}{2}}h_j s_j + n. \tag{5.18}$$

Then the received SIR of the i-th stream is

$$\mathrm{SIR}_i = \frac{|h_i|^2\, {r_i}^{-\alpha}}{\sum\limits_{b_j\in\Phi_b\setminus\mathcal{C}}|h_j|^2{r_j}^{-\alpha}}. \tag{5.19}$$

Using (5.9) and (5.19), we obtain theSCDP of PT-OS case as follows.

Lemma 7. *For the cluster-center user with target SIR $\theta_1 = 2^{\frac{R_d}{W}}-1$, theSCDP of PT-OS case with K cooperating SBSs is given by*

$$p_{\mathrm{d},K}^{\text{PT-O}}(\theta_1) \simeq \int_0^R\cdots\int_0^R\prod_{i=1}^{K}\frac{2x_i}{R^2}.\mathcal{L}_{I|R}\left(\theta_1 x_i^{\alpha}\right)\mathrm{d}x_1\cdots\mathrm{d}x_K, \tag{5.20}$$

where $\mathcal{L}_{I|x}(s)$ is defined in (5.13).

Proof. See Appendix C.3. □

5.3 Optimization of Cache Utilization Strategy

In this section, we first show the inherent tradeoff between transmission diversity and content diversity based on our analysis in Section 5.2. We then define two optimization problems in order to provide the optimal cache space assignment for the proposed combined caching strategy.

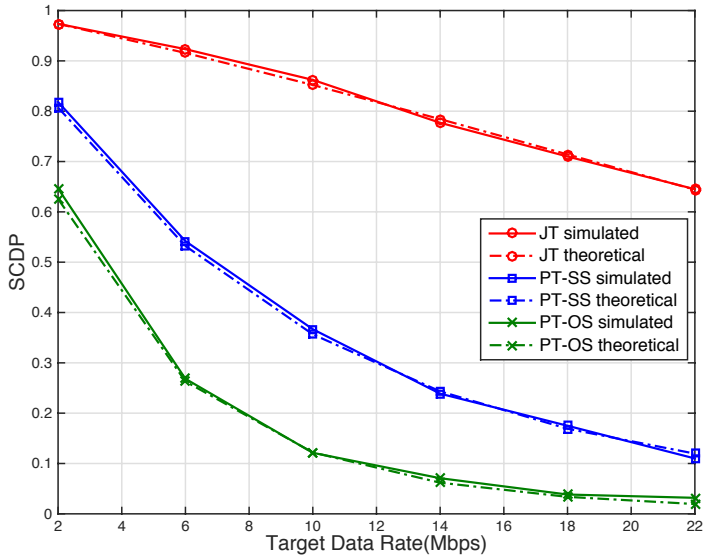


Figure 5.3: Theoretical and simulation results of SCDP of JT, PT-SS and PT-OS transmission schemes.

5.3.1 Transmission Diversity vs. Content Diversity

In Fig. 5.3, we plot both numerical and simulation results ofSCDP of the three transmission schemes discussed in Section 5.2 as a function of the target rate. The theoretical values are obtained from (5.12), (5.17) and (5.20) for the JT, PT-SS, and PT-OS cases, respectively. For the simulation results, the values of used parameters are given in Table 5.1 of Section 5.4. The number of cooperative SBSs is chosen as $K = 3$, which is close to the average number of SBSs per cluster according to our network settings. Fig. 5.4 shows the cache hit probability given in (5.3) as a function of the percentage of cache space assigned for MPC caching strategy in each SBS, ρ.

From Fig. 5.3, we first see that simulation results ofSCDP match well with numerical results. We also observe that JT always achieves higherSCDP than PT cases, evincing the benefit of MPC caching and JT transmission scheme in terms of higher transmission reliability. In addition, PT-SS always has higherSCDP than PT-OS, because spectrum sharing with SIC gives better reuse of communication resources for the parallel transmission. Therefore, in the following, we only consider PT-SS as the transmission scheme when the requested content falls in LCD range. Hence, when we refer to PT transmission scheme, it means PT-SS scheme.

From Fig. 5.4, we may observe that for lower ρ we have better cache hit ratio due to higher content diversity, achieved by assigning more space for LCD caching. From those two figures, we see that higher ρ increases the chance for joint transmission, which helps to improve the transmission reliability. With lower ρ, more different files will be cached in the cluster, thus offering higher

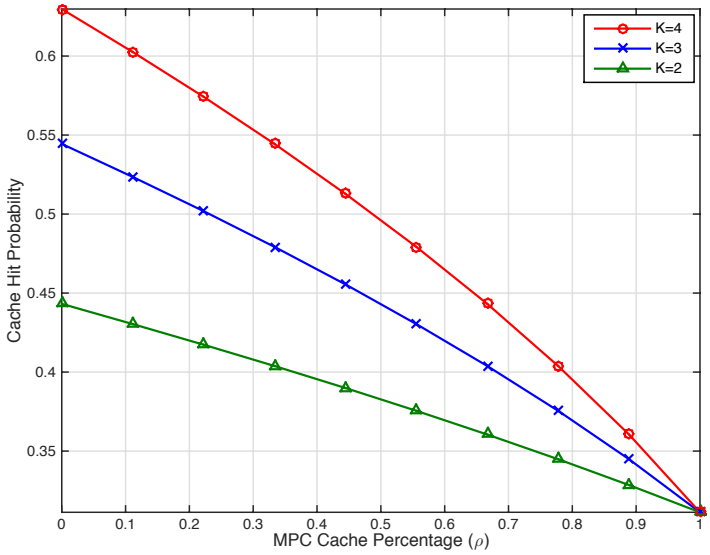


Figure 5.4: Cache hit probability vs. MPC cache percentage (ρ) for $K = 2, 3, 4$. $\gamma = 0.5$.

cache hit probability. In other words, there is a tradeoff between transmission diversity and content diversity. The cluster-centric cache utilization design should be able to leverage both diversity gains and adapt to the network environment and requirements. For instance, when the transmission rate requirement is high, caching the same most popular files in every SBS is preferable. Alternatively, increasing content diversity brings more opportunities to handle local requests by the cache.

5.3.2 Optimal Design for Cache Service Performance

Since the MPC cache percentage ρ in each SBS affects both local content diversity and transmission reliability, we seek here the optimal ρ that maximizes the percentage of requests successfully served by local caches, namely the cache service probability. A user request can be successfully served by local caches only when: 1) the requested file is cached inside the cluster, 2) the content delivery from the cooperative SBSs to the user is successful. We then define the cache service probability as follows.

Definition 2. *In the cluster-centric SCN with proposed combined caching strategy, the average cache service probability is given as*

$$p_{\text{sv}} = \sum_{K=1}^{\infty} \mathbb{P}(n = K) f(\rho|K), \tag{5.21}$$

where $f(\rho|K)$ is the cache service probability conditioning on having K SBSs

inside the cluster, given by

$$f(\rho|K) = p_{\text{CH, M}}(\rho)\, p^{\text{JT}}_{\text{d},K}(\theta_1) + p_{\text{CH, L}}(\rho)\, p^{\text{PT-S}}_{\text{d},K}(\theta_2). \tag{5.22}$$

$p^{\text{JT}}_{\text{d},K}(\theta)$ *and* $p^{\text{PT-S}}_{\text{d},K}(\theta)$ *are given in* (5.12) *and* (5.17)*, respectively. Here,* $p_{\text{CH, M}}(\rho)$ *and* $p_{\text{CH, L}}(\rho)$ *are probabilities to have the requested file cached in MPC and LCD ranges, respectively, given by*

$$p_{\text{CH, M}}(\rho) = \sum_{m=1}^{\lfloor \rho M \rfloor} p_m, \tag{5.23}$$

$$p_{\text{CH, L}}(\rho) = \sum_{\lfloor \rho M \rfloor + 1}^{\lfloor \rho M \rfloor + K(M - \lfloor \rho M \rfloor)} p_m, \tag{5.24}$$

where p_m *is defined in* (5.2)*.*

Since each cluster performs cooperative caching independent of other clusters, for a random cluster with K SBSs, the objective is to maximize its in-cluster cache service probability, that is, to find ρ which maximizes $f(\rho|K)$.

When $\gamma < 1$ and $M \ll N$, since $p_m \propto 1/m^{\gamma}$, we have [5] [114]

$$g(L) = \sum_{m=1}^{L} p_m \approx (L/N)^{1-\gamma}. \tag{5.25}$$

Using this approximation, $p_{\text{CH, M}}(\rho)$ and $p_{\text{CH, L}}(\rho)$ can be approximated by two continuous functions of ρ, given by

$$p_{\text{CH, M}}(\rho) \simeq \left(\frac{M}{N}\right)^{1-\gamma} \rho^{1-\gamma} = \widetilde{p}_{\text{CH, M}}(\rho), \tag{5.26}$$

$$p_{\text{CH, L}}(\rho) \simeq \left(\frac{M}{N}\right)^{1-\gamma} \left\{ [\rho(1-K)+K]^{1-\gamma} - \rho^{1-\gamma} \right\} = \widetilde{p}_{\text{CH, M}}(\rho). \tag{5.27}$$

Then, the cache service probability in (5.22) is simplified as

$$f(\rho|K) \simeq \left(\frac{M}{N}\right)^{1-\gamma} \rho^{1-\gamma} p^{\text{JT}}_{\text{d},K}(\theta_1) + \left(\frac{M}{N}\right)^{1-\gamma} \left[(\rho(1-K)+K)^{1-\gamma} - \rho^{1-\gamma} \right] p^{\text{PT-S}}_{\text{d},K}(\theta_2). \tag{5.28}$$

Using (5.28), we can obtain the optimal ρ as follows.

Lemma 8. *In a cluster-centric SCN with proposed combined caching strategy, knowing that there are* K *cooperative SBSs in the cluster, the optimal percentage of cache space assigned for MPC caching is given as*

$$\rho^* = \arg\max_{\rho \in [0,1]} f(\rho|K) \simeq \min \left\{ K \left[\left(\frac{K-1}{\frac{p^{\text{JT}}_{\text{d},K}(\theta_1)}{p^{\text{PT-S}}_{\text{d},K}(\theta_2)} - 1} \right)^{1/\gamma} + K - 1 \right]^{-1}, 1 \right\}, \tag{5.29}$$

where $p^{\text{JT}}_{\text{d},K}(\theta_1)$ *and* $p^{\text{PT-S}}_{\text{d},K}(\theta_2)$ *are given in* (5.12) *and* (5.17)*, respectively.*

[5] In a realistic scenario, $\gamma < 1$ captures better the heavy-tailed distribution of user requests than larger values of γ. For this reason, in this work, we do not study the case with $\gamma \leq 1$.

Proof. See Appendix C.4. □

Remark 6. *From* (5.29)*, we can see that the ratio* $\frac{p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1)}{p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2)}$ *is critical for the optimal cache assignment. When the transmission reliability of JT scheme is much higher than that of PT scheme, i.e.,* $p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1) \gg p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2)$*, we have* $\rho^* \simeq 1$*, meaning that most of the cache space would be used to store the most popular contents. When* $\frac{p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1)}{p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2)} \simeq 1$*,* $\rho^* \simeq 0$*, then increasing the content diversity becomes more beneficial.*

Inside each cluster, based on its knowledge about the number of in-cluster SBSs and out-of-cluster interfering SBS density, the central controllers will be able to compute the optimal percentage of cache space for MPC caching and assist the cache placement in each cooperative SBS.

5.3.3 Optimal Design for Energy Efficiency

When a user requests for a file, depending on the availability of this file in local caches and the placement strategy, both the delivery rate and power consumption will be different. If the requested file is not in local caches, the SBSs serving the user needs to download the file from the core network throughout backhaul. In that case, energy is consumed at the backhaul and there is additional delay of downloading from the core network to the SBSs. As a result, the energy consumption and the content delivery rate are determined according to our cache utilization design, more explicitly, they depend on ρ in the combined caching scheme. In our network model, the energy efficiency can be defined as the effective delivery rate per unit energy consumption, where the effective delivery rate is the number of successfully delivered bits per second, similar to [115].

When the requested file is stored in local caches (i.e., cache hit case), the effective delivery rate is defined as $R_d p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1)$ and $R_d p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2)$ for the MPC-JT and LCD-PT cases, respectively, where $\theta_1 = 2^{\frac{R_d}{W}} - 1$ and $\theta_2 = 2^{\frac{R_d}{KW}} - 1$ are the corresponding target SIRs. When the requested file is not in local caches (i.e., cache miss case), we need to consider the backhaul delay $T_{\mathrm{bh}} (< T)$ to define the effective delivery rate. For delivering the requested file within the time slot T, the maximum transmission time should be $T' = T - T_{\mathrm{bh}} = \beta T$, where $\beta = 1 - \frac{T_{\mathrm{bh}}}{T}$ is the fraction of reduced transmission time due to backhaul delay. As mentioned in Section 5.1.3, in the cache miss case, the requested file is downloaded from the core network to every in-cluster SBS and joint transmission will be used to serve the user. Hence, the effective delivery rate in this case becomes $R_d p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_3)$ with $\theta_3 = 2^{\frac{R_d}{\beta W}} - 1$.

By taking the aforementioned three cases into account, the average effective date rate can be given as

$$\widetilde{R}_{\mathrm{avg}} = p_{\mathrm{CH,\,M}}(\rho) R_d \, p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1) + p_{\mathrm{CH,\,L}}(\rho) R_d \, p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2) + p_{\mathrm{CM}}(\rho) R_d \, p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_3), \tag{5.30}$$

where $p_{\text{CH, M}}(\rho)$ and $p_{\text{CH, L}}(\rho)$ are defined in (5.23) and (5.24), respectively, and $p_{\text{CM}}(\rho)$ is the probability of not having the request file cached inside the cluster (i.e., cache miss probability), given by

$$p_{\text{CM}}(\rho) = 1 - \sum_{m=1}^{\lfloor \rho M \rfloor + K(M - \lfloor \rho M \rfloor)} p_m. \tag{5.31}$$

For the cache hit case, the consumed power for content delivery contains only the transmit power of the K SBSs if we ignore other static power consumption for the baseband processing, etc. For the cache miss case, the requested file is fetched from the core network through backhaul, and then transmitted from the K SBSs to the user. Denote P_b as the wireline backhaul power consumption required to handle a user request at a single SBS [116]. Then, we have the average power consumption to serve a user request inside a cluster of K SBSs as [6]

$$P_{\text{avg}} = K\left\{\left[p_{\text{CH, M}}(\rho) + p_{\text{CH, L}}(\rho)\right] P_t + p_{\text{CM}}(\rho)(P_t + P_b)\right\} = KP_t + KP_b p_{\text{CM}}(\rho), \tag{5.32}$$

which is averaged over the three cases. From (5.30) and (5.32), we can define the energy efficiency as follows.

Definition 3. *In the cluster-centric SCN with proposed combined caching strategy, the average energy efficiency is given as*

$$\eta_{\text{EE}} = \sum_{K=1}^{\infty} \mathbb{P}(n = K)\eta(\rho|K), \tag{5.33}$$

where $\eta(\rho|K)$ is the energy efficiency conditioning on having K SBSs inside the cluster, given by

$$\eta(\rho|K) = \frac{\widetilde{R}_{\text{avg}}}{P_{\text{avg}}} = \frac{R_d\left[p_{\text{CH,M}}(\rho)p_{\text{d},K}^{\text{JT}}(\theta_1) + p_{\text{CH,L}}(\rho)p_{\text{d},K}^{\text{PT-S}}(\theta_2) + p_{\text{CM}}(\rho)p_{\text{d},K}^{\text{JT}}(\theta_3)\right]}{KP_t + KP_b p_{\text{CM}}(\rho)}. \tag{5.34}$$

Here, $p_{\text{d},K}^{\text{JT}}(\theta)$ and $p_{\text{d},K}^{\text{PT-S}}(\theta)$ are given in (5.12) *and* (5.17)*, respectively, and $p_{\text{CH,M}}(\rho)$, $p_{\text{CH,L}}(\rho)$ and $p_{\text{CM}}(\rho)$ are given in* (5.23)*,* (5.24)*, and* (5.31)*, respectively.*

Inside a cluster with K cooperative SBSs, the optimal cache utilization strategy that maximizes the energy efficiency is given by finding $\rho^* = \arg\max_{\rho \in [0,1]} \eta(\rho|K)$.

[6] Here, we do not consider the static power consumption for the baseband processing, site cooling, etc., since this part of consumed power is the same for MPC, LCD and cache miss cases. Adding the static power in the average power consumption is equivalent to having higher transmit power P_t for each SBS in (5.32).

Table 5.1: Simulation Parameters of Cluster-centric Cooperative SCNs

Parameters	**Values**
SBS density (λ_b)	$10^{-4}/\mathrm{m}^2$
Half cluster center distance (R_h)	100 m
Pathloss exponent (α)	4
SBS transmit power (P_t)	1 W
Backhaul power per request per SBS (P_b)	10 W
Available bandwidth (W)	10 MHz
SBS cache capacity (M)	5000
Content library size (N)	10^5
Zipf shape parameter (γ)	$\{0.5, 0.9\}$
Transmission time fraction (β)	$\{0.3, 0.95\}$

Similarly, with the help of the approximation in (5.25) for the case when $\gamma < 1$ and $M \ll N$, we get

$$p_{\mathrm{CM}}(\rho) \simeq 1 - \left(\frac{M}{N}\right)^{1-\gamma} [\rho(1-K)+K]^{1-\gamma} = \widetilde{p}_{\mathrm{CM}}(\rho). \tag{5.35}$$

Putting (5.26), (5.27) and (5.35) into (5.34), we obtain the approximated energy efficiency $\widetilde{\eta}(\rho|K)$ as a continuous function of ρ, given as

$$\widetilde{\eta}(\rho|K) \simeq \frac{R_d \left[\widetilde{p}_{\mathrm{CH,M}}(\rho) p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1) + \widetilde{p}_{\mathrm{CH,L}}(\rho) p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2) + \widetilde{p}_{\mathrm{CM}}(\rho) p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_3)\right]}{KP_t + KP_b \widetilde{p}_{\mathrm{CM}}(\rho)}. \tag{5.36}$$

Due to the above involved expression, we cannot have a closed-form solution for $\rho^* = \arg\max_{\rho\in[0,1]} \widetilde{\eta}(\rho|K)$ directly. However, with the help of existing standard optimization methods, we can still have numerical values for the optimal ρ that maximizes $\widetilde{\eta}(\rho|K)$. Note that the accuracy of the optimal ρ obtained using the approximated energy efficiency is verified in Section 5.4.

5.4 Simulation Results

In this section, we validate the performance analysis of our cooperative caching and transmission design in cluster-centric SCNs using simulations. The performance is compared with that of cases using only MPC and LCD type caching schemes.

Simulations are performed in a square area of $10^3 \times 10^3$ m^2. The hexagonal cluster of interest has its cluster center at the origin with distance between two cluster centers equal to $2R_\mathrm{h} = 200$ m. The approximated circle for the

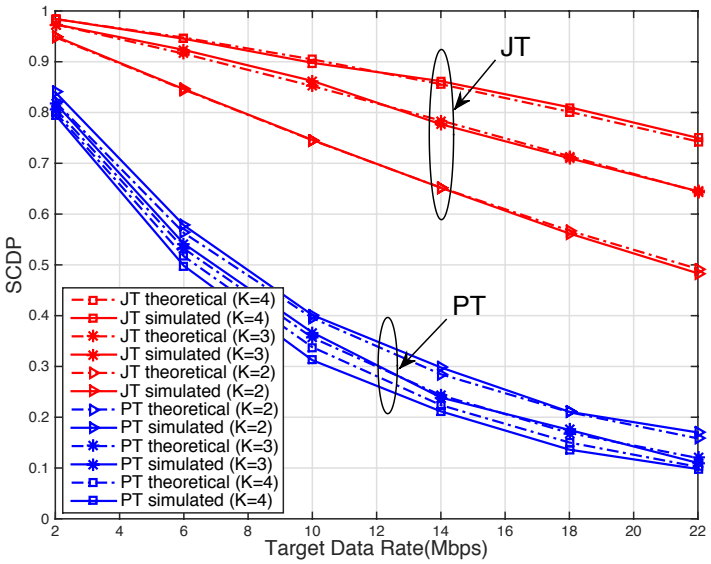


Figure 5.5: Theoretical and simulation results of the SCDP of JT and PT transmission schemes with $K = \{2, 3, 4\}$.

cluster area has radius $R = R_{\mathrm{h}}\sqrt{\frac{2\sqrt{3}}{\pi}} \simeq 105$ m. SBSs are distributed according to a homogeneous PPP. All the channel fading follows Rayleigh fading with $|h_i|^2 \sim \exp(1)$. The values of parameters used for simulation are given in Table 5.1. Simulation results are obtained by averaging over 40000 realizations.

Remind that we do not consider the case when there is no SBS in a reference cluster. With our network settings, from (5.1) we have $\mathbb{P}(n = 0) = e^{-2\sqrt{3}\lambda_b R_{\mathrm{h}}^2} = 0.03$, meaning that only for 3% of realizations we have empty reference cluster. Therefore, excluding empty clusters does not have much impact on the overall network performance.

5.4.1 Successful Content Delivery Probability

Fig. 5.5 shows the theoretical and simulation results ofSCDP of JT and PT (PT-SS) transmission schemes when assuming to have $K = \{2, 3, 4\}$ SBSs inside the cluster of interest. It first validates the accuracy of our analysis in (5.12) and (5.17), especially when K is the close to the average number of SBSs per cluster, i.e., $K = 3$. It also proves that the circular approximation of the cluster area has negligible impact on theSCDP analysis. We notice that the error gap in the PT case becomes slightly larger when the conditioned number K is further from the average value $\mathbb{E}[K] = 2\sqrt{3}\lambda_b R_{\mathrm{h}}^2$. This is mainly due to the PPP approximation that we use for the interference distribution in (5.15). When the density of SBSs inside the cluster conditioning on having K SBSs is comparable to the density of PPP distributed out-of-cluster SBSs, the approximation in (5.15) is reasonable. Otherwise the mismatch between

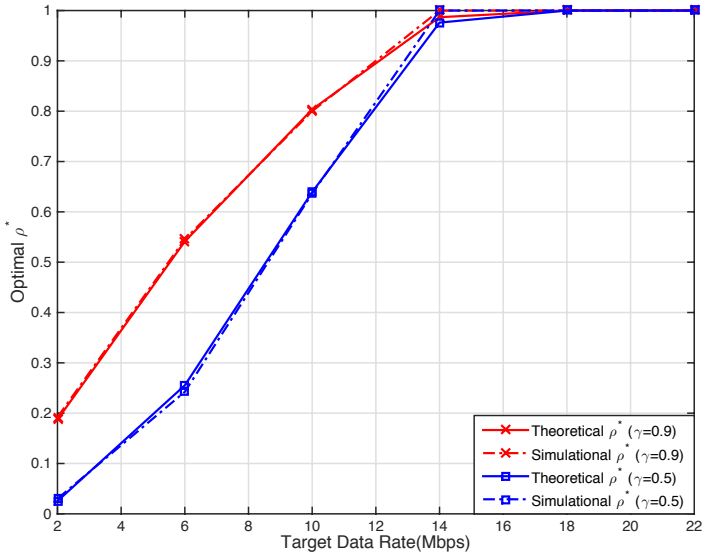


Figure 5.6: Optimal percentage of MPC type caching, ρ^*, obtained by cache service probability maximization, when using the proposed combined caching scheme. Both theoretical and simulation results are obtained with $K = 3$ and $\gamma = \{0.5, 0.9\}$.

the conditioned SBS density inside the cluster and the density of out-of-cluster SBSs causes approximation error in the SIR analysis.

We also observe that, in the JT case, higher K yields higher SCDP, but for PT cases,SCDP is lower when K is larger. This is because in the JT case, more cooperative SBSs gives stronger received signal, thus higher SIR value. In the PT case, theSCDP is defined as the product of success probability of multiple streams. When the number of parallel transmitting streams increases, theSCDP will be relatively lower.

5.4.2 Optimization Study of the Combined Caching Strategy

In the cluster-centric network, each cluster makes caching decisions independently based on its knowledge about network status inside and outside the cluster, so the optimal percentage for MPC caching, ρ^*, is computed in each cluster according to the number of cooperative SBSs K. In this section all the theoretical and simulation results are obtained conditioning on having a certain number K of SBSs inside the cluster of interest.

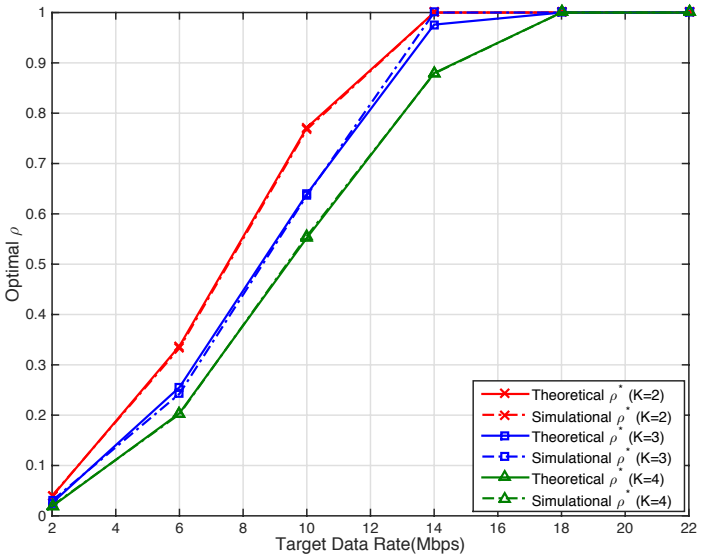


Figure 5.7: Optimal percentage of MPC type caching, ρ^*, obtained by cache service probability maximization, when using the proposed combined caching scheme. The results are obtained and presented with $K = \{2, 3, 4\}$ and $\gamma = 0.5$.

Cache Service Probability Maximization

In Fig. 5.6, we plot the optimal ρ obtained in (5.29), which maximizes the cache service probability, as a function of the target data rate. The number of in-cluster SBSs is chosen as $K = 3$. The theoretical optimal values are compared with the real optimum values obtained from the exhaustive search of ρ that maximizes the cache service probability defined in (5.22). We see that ρ^* in (5.29) gives accurate estimation of the real optimum result. We can also see that as expected, ρ^* increases with the target rate, because for higher SIR requirement, the transmission reliability is more important for the cache service performance, thus MPC type caching is more favorable. The content popularity skewness also affects the optimal MPC cache percentage. When the content popularity is more concentrated, i.e., $\gamma = 0.9$, the potential benefit from caching more different files is limited, and in this case, the optimal ρ is expected to be higher, as also shown in Fig. 5.6.

In Fig. 5.7, we plot the theoretical and simulated values of the optimal ρ conditioning on having $K = \{2, 3, 4\}$ SBSs inside the cluster of interest. The results are obtained with $\gamma = 0.5$. Beside the accuracy of the theoretical results, we also notice that for larger K, the optimal MPC cache percentage, ρ^*, is smaller. It shows the potential of improved cooperation gain by reserving more cache space for partition-based LCD caching when the number of cooperative SBSs is larger.

Fig. 5.8 shows the average cache service probability of our proposed cooperative caching and transmission design. The results are obtained by averaging

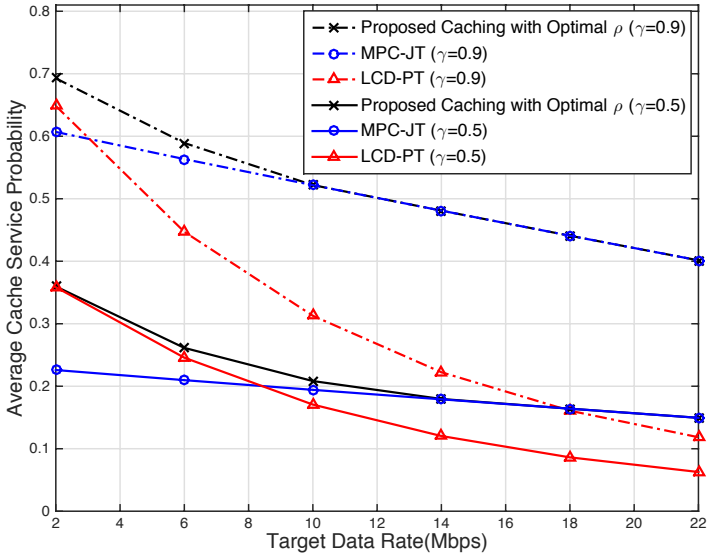


Figure 5.8: Cache service performance of the proposed combined caching scheme with ρ^* given in (5.29), with comparison to the case where only MPC or LCD caching is applied. The results are obtained with $\gamma = \{0.5, 0.9\}$

over different values of K, as given in (5.21). For practical reasons, we choose $K \in [1, 10]$ to get numerical results in the finite range. We see that our proposed caching scheme with optimal ρ derived in (5.29) always gives better performance than the cases when only either MPC or LCD scheme is applied. As expected, the performance of the proposed caching scheme converges to the performance of LCD and MPC schemes in the extreme cases.

Energy Efficiency Maximization

In Fig. 5.9, we plot the optimal ρ obtained by the energy efficiency maximization for different values of backhaul delay. The number of in-cluster SBSs is chosen as $K = 3$. The theoretical results are obtained by numerical evaluation of $\rho^* = \arg\max_{\rho \in [0,1]} \widetilde{\eta}(\rho|K)$ with $\widetilde{\eta}(\rho|K)$ given in (5.36). The real optimal values which maximize the energy efficiency defined in (5.34) are obtained in simulations by exhaustive searching. We can see that the theoretical ρ^* matches well the result obtained in simulation, validating the accuracy of energy efficiency maximization with the approximated expression. We also observe the same trend of ρ^* as in Fig. 5.6. When the SIR target increases, the optimal value of ρ is higher, meaning that more space will be assigned for MPC caching. In terms of the impact of backhaul delay on the value of ρ^*, we see that for higher backhaul delay, i.e., $\beta = 0.3$, ρ^* is lower, meaning that more space will be assigned for LCD caching in order to avoid fetching the requested content through the backhaul. Compared to the case with cache service probability maximization, ρ^* in Fig. 5.9 is always smaller than the ones in Fig. 5.6, es-

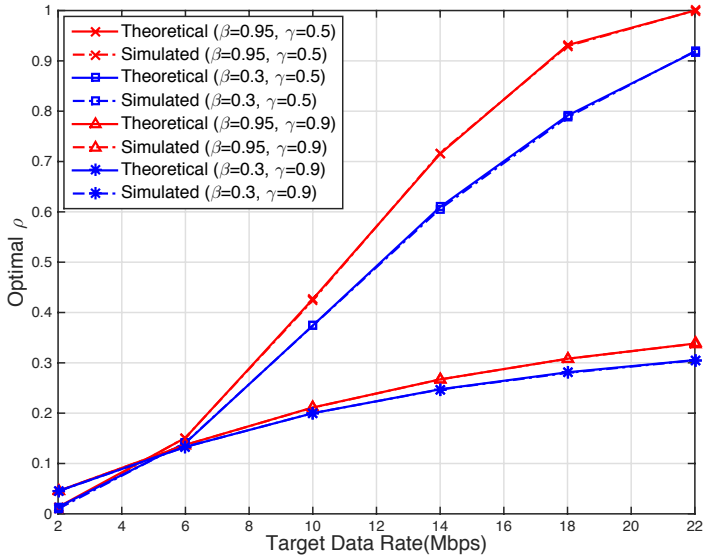


Figure 5.9: Optimal percentage of MPC type caching, ρ^*, obtained by the energy efficiency maximization when using the proposed combined caching scheme. The results are obtained with $\beta = \{0.95, 0.3\}$, representing the cases with very low and high backhaul delay, respectively, $K = 3$, and $\gamma = \{0.5, 0.9\}$.

pecially in the case with $\gamma = 0.9$. We also observe that, when the target rate is relatively high, the optimal ρ obtained with higher γ is much smaller than the one obtained with lower γ. This is because when the popularity is highly concentrated, i.e., $\gamma = 0.9$, the benefit of having more space for MPC caching in terms of average rate improvement becomes limited by taking into account the growth trend of the power consumption. It shows the necessity of reserving more space for LCD caching when taking into account the backhaul energy consumption and delay, which coincides with the rationale behind caching in SCNs for improved energy efficiency.

Fig. 5.10 shows the average energy efficiency defined in (5.33) when using the optimal ρ obtained by the energy efficiency maximization for our proposed caching scheme. The results are compared to the case with only MPC or LCD caching and the baseline result without cache capacity at SBSs. Similar to the case with cache service probability maximization, we observe that our proposed scheme combines the advantage of MPC and LCD caching, thus outperforms the cases where either MPC or LCD caching is applied. Compared to the case without caching, the improvement of energy efficiency is validated, showing the benefits of cooperative caching design in SCNs.

5.4.3 Potential Improvement of Power Control for SIC

An important remark from the presented results is that the potential benefit of cooperative caching using PT-SS scheme is strongly limited by the lack

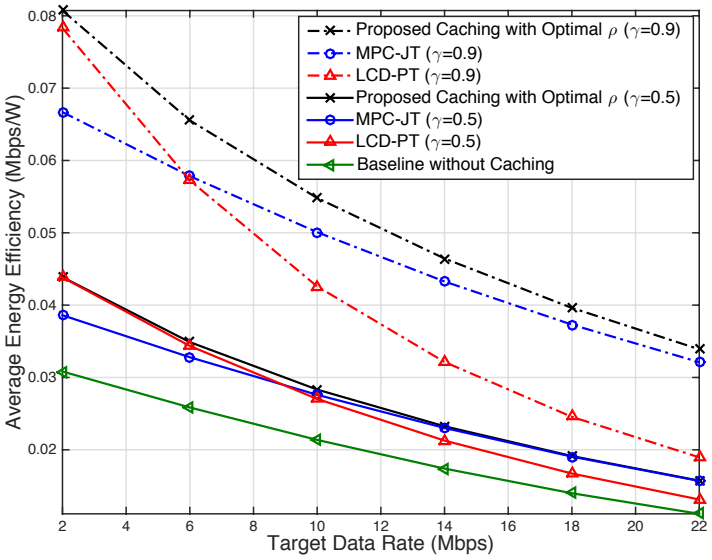


Figure 5.10: Average energy efficiency of the network when using the proposed combined caching scheme, with comparison to the case with either only MPC or LCD caching. The results are obtained with $\gamma = \{0.5, 0.9\}$ and $\beta = 0.5$.

of transmission reliability when high transmission rate is required. It is well known that the performance of SIC can be improved by properly assigning different transmit powers such that users experience the same SIR at each decoding time [117] [118].

We consider a simple power control for PT-SS scheme assuming Channel State Information (CSI) is available within a cluster for every cooperative SBS inside, i.e., the transmit power of the k-th SBS is chosen as $P_{t,k} = \frac{KP_d\theta(1+\theta)^{K-k}/g_k}{\sum_{k=1}^{K}\theta(1+\theta)^{K-k}/g_k}$, where $g_k = |h_k|^2 d_k^{-\alpha}$ is the channel gain of the k-th SBS, and θ is the SIR threshold for successful interference cancellation. $P_{t,k}$ is a normalized value such that $\sum_{k=1}^{K} P_{t,k} = KP_t$ to ensure the same power consumption as the case without power control. We plot the SCDP of PT-SS in the case with and without power control in Fig. 5.11, showing the improvement of SIC performance using power control method. Due to the difficulty of analyzing the interference distribution in this case, the SIR analysis is not discussed in this work. Intuitively, when using power control for SIC, the improved SCDP of PT-SS scheme will result in smaller $\rho^\star$ than the case without power control, implying that more different files can be cached within the cluster.

5.4.4 Cluster-Center User vs. Randomly Located User

Remind that we evaluate the performance of our proposed cooperative caching and transmission design based on cluster-center user assumption. As presented in Section 5.3.2, the optimal MPC cache percentage, ρ^*, depends on the ratio

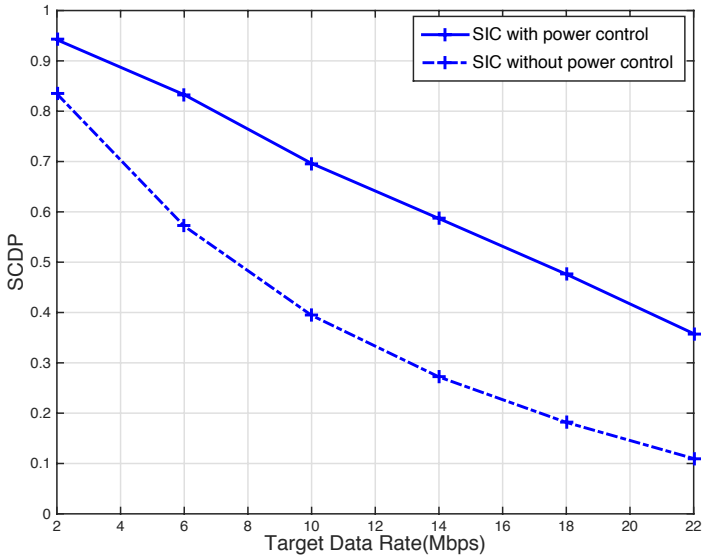


Figure 5.11: SCDP of PT-SS with and without power control for SIC. The number of cooperative SBSs inside the cluster of interest is chosen as $K = 3$.

$\frac{p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1)}{p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2)}$. The optimal solution for randomly located users in the general case requires analytical results on the SCDP of JT and PT schemes for arbitrary users, which are difficult to obtain in a simple or neat form. Fig. 5.12 shows the simulated SCDPs of JT, PT-SS and PT-OS schemes with randomly distributed user inside each cluster. Compared to the SCDP with cluster-center user, despite the difference of simulated values, we observe the same trend of SCDP for the three transmission schemes, i.e., JT $>$ PT-SS $>$ PT-OS in terms of SCDP.

In Fig. 5.13, we present the simulated average cache service probability when users are randomly distributed in each cluster. The simulation results are provided for 4 cases: 1) MPC-JT only; 2) LCD-PT only; 3) proposed caching scheme with ρ^*, obtained with theoretical SCDPs based on the cluster-center user assumption; and 4) proposed caching scheme with ρ^*, obtained with simulated SCDPs based on randomly distributed users. From this figure, we see that the proposed caching scheme achieves higher performance than the cases of MPC-JT only and LCD-PT only, even when the users are randomly distributed. Furthermore, although the theoretical SCDPs obtained with cluster-center user assumption do not provide accurate estimation of the SCDPs of randomly distributed users (as shown in Fig. 5.12), the cache service performance of our proposed caching scheme with the theoretical ρ^* is very close to the one obtained with the simulated SCDPs for randomly distributed users.

In order to validate the performance gain of cooperative transmission, we compare the average cache service probability of the proposed scheme to that of the non-cooperative transmission with MPC, where users are randomly distributed on the 2-dimensional Euclidean plane. In the non-cooperative transmission case, each user receives a file from a single nearest SBS. From Fig. 5.13,

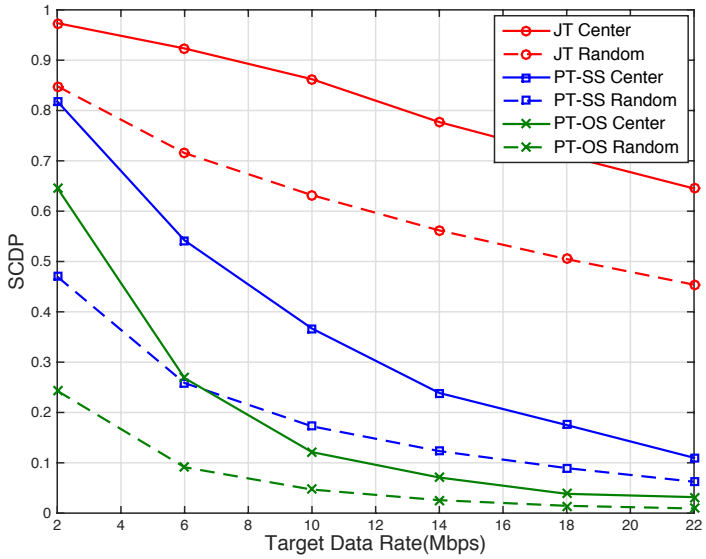


Figure 5.12: Simulated SCDP of JT, PT-SS and PT-OS schemes with randomly distributed users inside each cluster.

it is shown that the cooperative caching and cooperative transmission improve the cache service performance compared to the case with no cooperation.

5.5 Summary and Concluding Remarks

In this chapter, we studied the potential benefits of using cooperative caching and transmission schemes in cluster-centric cache-enabled SCNs. We proposed a combined caching and transmission strategy by cooperatively utilize the cache space of multiple SBSs in the same cluster. Our analysis revealed an inherent tradeoff between transmission diversity and content diversity. Motivated by this tradeoff, we solved two optimization problems, namely maximizing the cache service probability and the energy efficiency, respectively. The optimal solutions were given as a function of network parameters and content popularity characteristics.

The major contribution of this work is to illustrate that when PHY cooperation is enabled among the SBSs, the performance of content caching can be significantly improved if the caching strategy is duly designed. The results presented in this chapter can be extended to general scenarios with more advanced RRM techniques and caching placement design, therefore followed by different modeling and analytical approaches, which is out of the scope of this work. The main takeaway of this work is that base station cooperation can improve the bandwidth usage, which can be further translated into improved cache memory usage.

As summary of Part II of this thesis, we have investigated proactive caching at SBSs and at D2D enabled user devices, in the scenarios with independent probabilistic content placement and with cooperative caching. Our remarks

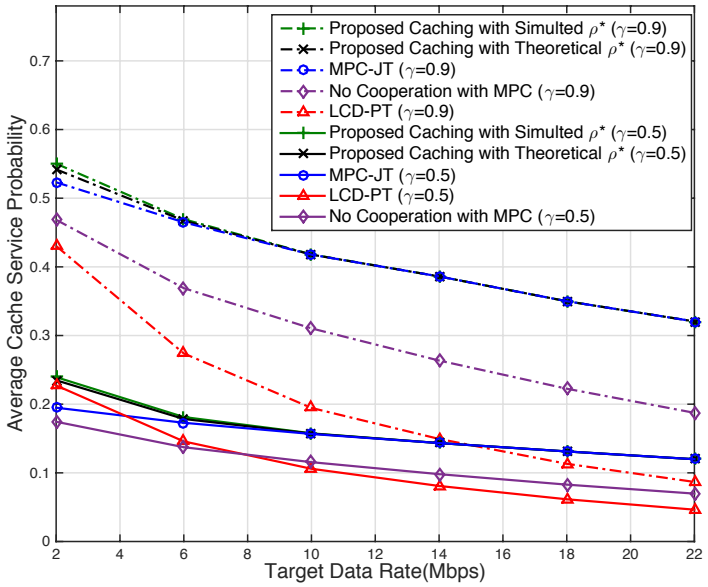


Figure 5.13: Simulated average cache service probability with randomly distributed users in each cluster. The simulated ρ^* is obtained with theoretical SCDPs based on the cluster-center user assumption and the theoretical ρ^* is obtained with simulated SCDPs based on randomly distributed users.

from the results presented in this part provides meaningful insights in the proactive caching methods at the network edge. There is no clear conclusions yet about how and when caching in wireless networks will become reality, which leaves us more questions to answer and more blank space to fill in the design and operation of wireless caching networks.

Part III

Conclusions, Outlook and Appendices

Chapter 6

Conclusions and Outlook

In this chapter, we summarize the contribution of this thesis, and the perspectives regarding the future work related to the topics presented in this thesis.

6.1 General Comments and Conclusions

In this thesis, we have investigated two important techniques for cellular traffic offloading based on user-centric and content-aware communications, namely D2D communication and proactive caching at the network edge.

In the first part of this thesis, **Device-to-Device (D2D) Underlaid Cellular Networks**, we studied distributed access control schemes for D2D underlaid cellular networks with respect to certain constraints on the cellular links: 1) minimum cellular coverage probability, and 2) maximum average cellular delay while assuming bursty packet arrivals at cellular users. Particularly, in **Chapter 2 – Distributed SIR-aware Opportunistic Access Control**, we proposed a SIR-aware opportunistic D2D access scheme combined with cellular guard zones, which aims at achieving the maximum D2D area spectral efficiency (ASE) while offering minimum coverage for the cellular users. The optimal SIR threshold and the minimum cellular guard zones radius were obtained through a decoupled optimization approach. The optimal results can be applied directly in D2D underlaid cellular networks as a distributed optimization scheme which not only depends on the channel condition of each D2D link, but also on the received aggregate interference. In **Chapter 3 – Priority-based Shared Access with Delay Constraints**, we extended our considered network model to a more generalized scenario, namely shared access network with priorities. When assuming bursty packet arrivals at the primary node, we proposed a congestion control protocol and introduced queueing analysis in such network. We analyzed the impact of the design parameters of the shared access protocol on the secondary throughput. For a specific case, we derived in closed-form the optimal access probability of the secondary nodes with respect to the primary delay constraints. The results of this chapter can

be easily applied in other network scenarios with spectrum sharing and with QoS-related constraints on the nodes with higher priority.

In the second part of this thesis, **Proactive Caching at the Network Edge**, we investigated wireless caching in different network structures and with different caching strategies. More explicitly, in **Chapter 4 – Stochastic Wireless Caching Networks**, we studied the probabilistic content placement in stochastic wireless caching networks and compared the performance of D2D caching and small cell caching methods. Our analysis on the performance evaluation of these two caching network structures revealed some basic information regarding where to cache content in a wireless network, which strongly depends on the network conditions and the user request distribution. In this chapter we also investigated the optimization of probabilistic content placement in wireless D2D networks with two different objective functions: maximizing the cache hit probability and the cache-aided throughput, respectively. Our results illustrated the necessity of taking into account the PHY transmission reliability in the optimal content placement with probabilistic caching strategy. In **Chapter 5 – Small Cell Cooperative Caching**, we proposed a cooperative caching scheme in cluster-centric SCNs and studied the cooperation gain with coupled designed of cache utilization and PHY transmission. For our proposed combined "most popular content (MPC)" with "largest content diversity (LCD)" caching scheme, we optimized the cache utilization design in each cluster, as a function of the number of cooperative SBSs, the network parameters and user quality-of-service (QoS) requirements. The superiority of our scheme compared to non-cooperative caching strategy was validated by simulation results in terms of cache service probability and energy efficiency.

6.2 Perspectives and Future Work

There are many unexploited research topics related to my PhD thesis and further extensions of the studies that we have accomplished. In this section, we give a brief overview of the perspectives and possible extensions of this thesis.

In **Chapter 2 – Distributed SIR-aware Opportunistic Access Control**, the extension of this chapter might be carried out by considering the following aspects:

- **Centralized D2D power control**. It is with no doubt that the centralized control over the entire network will results in the optimal performance, which in turn adds the complexity in the implementation of such scheme.

- **Multiple antenna techniques**. Considering multiple antennas at the base stations (BSs) and at user devices will lead to enhanced the cellular and D2D coverage. Other benefits such as intra-cell and inter-cell

interference cancellation also requires different approaches of theoretical analysis on the coexistence between cellular and D2D communication in the same frequency band. The impact of multiple-input multiple-output (MIMO) techniques on the performance of D2D underlaid cellular networks remains an interesting topic to study.

- **Other objective functions for the optimization problem**. Besides the D2D area spectral efficiency, we can also consider other objective function such as the weighted sum rate of D2D underlaid cellular networks.

In **Chapter 3 – Priority-based Shared Access with Delay Constraints**, we might consider the following aspects for the future investigation:

- **Location-dependent access probabilities for the secondary users**. For the primary receiver at fixed location, the average interference caused by the co-channel secondary transmitters strongly depends on the distance. Hence, one possible extension of this work is to assign different access probabilities for the secondary users depending on their distances to the primary receiver. However, due to the user mobility, the signaling overhead cost in this case might become an issue.

- **Quality-based opportunistic access**. In this chapter we considered random access with a certain probability for the convenience of distributed secondary access control. Similar to Chapter 2, we can apply opportunistic access scheme on the secondary users based on certain quality metrics that we choose, e.g., channel gain or estimated SINR.

In **Chapter 4 – Stochastic Wireless Caching Networks**, we can consider the following cases as the extensions of this work:

- **Heterogeneous caching networks**. One possible extension of this work is to consider heterogeneous caching networks with caching capabilities at user devices, at distributed caching helpers and at SBSs. The joint optimization of the caching probabilities in multiple tiers so as to maximize the density of successful served requests remains an open problem to study.

- **Delay performance**. Except the cache hit probability and the density of successful served requests by local caches, referred as *cache-aided throughput* in this work, one can also consider other objective functions for the optimization problem, for instance, the average delay experienced by the user. When considering the retransmission of content once the previous transmission fails, the delay analysis is not trivial.

In **Chapter 5 – Small Cell Cooperative Caching**, one might consider the following aspects to study:

- **Base station clustering and user association model**. In this work we assumed hexagonal grid cluster model with randomly distributed SBS. It is possible to consider other cluster models in stochastic geometry, such as Poisson cluster process (PCP) or a hierarchical clustering model with Poisson superposition. The mathematical tractability of the coverage analysis depends on the cluster model we choose, which will results in different forms of success probabilities.

- **Multiple antennas at the SBSs and at user devices**. By allowing MIMO communication in small cell networks, the extra degree of freedoms (DoFs) bring opportunity to increase the transmission reliability, which will further affects the optimal cache utilization.

- **Multiple thresholds of cache utilization**. The cache utilization design in this work can be further generalized to the case with multiple thresholds for the cache utilization, i.e., a file or a partition/segment of file can be selectively cached by some, but not all the SBSs in the cluster, as proposed in [84].

Appendix A

Distributed SIR-aware Opportunistic Access Control

A.1 Proof of Proposition 1

Due to the asymmetric shape of Voronoi cells, the distribution of the distance from the nearest interfering uplink user to the typical BS is not straightforward. From existing results on Poisson-Voronoi tessellations [119, 120], the area distribution of a Voronoi cell, denoted by $\mathcal{A}$, can be approximated by

$$f_{\mathcal{A}}(a) = \frac{(3.5\lambda_M)^{3.5}}{\Gamma(3.5)} a^{2.5} \exp(-3.5\lambda_M a). \tag{A.1}$$

If then the typical Voronoi cell is approximated by a circle centered at the typical BS with the same area, the distance from the nearest uplink interferer to the typical cellular receiver (BS) is the radius of the circle. Knowing that $\mathcal{A} = \pi d_{\min}^2$, the distribution of the radius $d_{\min}$ is given by

$$f_{d_{\min}}(r) = 2\frac{(3.5\pi\lambda_M)^{3.5}}{\Gamma(3.5)} r^6 \exp(-3.5\pi\lambda_M r^2). \tag{A.2}$$

From Definition 1, assuming that the distribution of uplink users can be approximated by a homogeneous PPP with density λ_M, $\mathcal{L}_{I_{cc}}(s)$ can be derived by the Laplace transform of interference coming from PPP-distributed nodes with minimum distance $d_{\min}$ to the typical receiver. Thus we have

$$\mathcal{L}_{I_{cc}}(s) \approx \mathcal{L}_I^1(s, \lambda_M, d_{\min}), \tag{A.3}$$

where the pdf of $d_{\min}$ is given in (A.2).

A.2 Proof of Proposition 2

Since $\mathcal{T}(p_s)$ increases monotonically with p_s when $p_s \to 0$, and decreases monotonically with p_s when $p_s \to 1$, and is a continuous function, it is reasonable

to consider that the crossing point of these two functions could be approximately the p_s that maximizes $\mathcal{T}(p_s)$. Under this assumption, the optimal access probability $p_s^\star$ should satisfy

$$\begin{aligned} p_s^\star &\simeq \exp\left[-\xi\beta^{\frac{2}{\alpha}}\left(p_s^\star\lambda_D + \kappa\lambda_M\right)\right] \\ \Rightarrow e^{-\xi\beta^{\frac{2}{\alpha}}p_s^\star\lambda_D} &\simeq e^{\xi\beta^{\frac{2}{\alpha}}\kappa\lambda_M}p_s^\star. \end{aligned} \tag{A.4}$$

For a general type of equation $p^{ax+b} = cx + d$, where x is the variable and a, b, c, d, p are constant, when $p > 0$ and $a, c \neq 0$, the solution by using Lambert W function is

$$x = -\frac{\mathcal{W}\left(-\frac{a\ln p}{c}p^{b-\frac{ad}{c}}\right)}{a\ln p} - \frac{d}{c}. \tag{A.5}$$

By solving (A.4) with the help of Lambert W funtion we have

$$p_s^\star \simeq \frac{\mathcal{W}\left(\lambda_D\xi\beta^{\frac{2}{\alpha}}e^{-\kappa\lambda_M\xi\beta^{\frac{2}{\alpha}}}\right)}{\lambda_D\xi\beta^{\frac{2}{\alpha}}}. \tag{A.6}$$

Knowing that $p_s^\star$ should not exceed one, we have

$$p_s^\star \simeq \min\left\{\frac{\mathcal{W}\left(\lambda_D\xi\beta^{\frac{2}{\alpha}}e^{-\kappa\lambda_M\xi\beta^{\frac{2}{\alpha}}}\right)}{\lambda_D\xi\beta^{\frac{2}{\alpha}}}, 1\right\}. \tag{A.7}$$

Substituting it into (2.18), we have that the approximately optimal SIR threshold is given as

$$G^\star \simeq \left[\frac{-\ln p_s^\star}{\xi(\lambda_D + \kappa\lambda_M)}\right]^{\frac{\alpha}{2}}. \tag{A.8}$$

Appendix B

Priority-based Shared Access with Delay Constraints

B.1 Proof of Proposition 3

According to the definition of the success probability, for the typical active secondary pair i, we have

$$\begin{aligned} p_{2/1,2} &= \mathbb{P}\left[\mathrm{SINR_i} > \theta \mid \mathcal{T} = \{\mathrm{x}_0 \cup \Phi_{\mathrm{a}}^2\}\right] \\ &= \mathbb{P}\left[\frac{P_2|h_{i,i}|^2 d_s^{-\alpha}}{\sigma^2 + \sum\limits_{j\in\Phi_a^2\setminus\{i\}} P_2|h_{j,i}|^2 d_{j,i}^{-\alpha} + P_1|h_{0,i}|^2 d_{0,i}^{-\alpha}} > \theta\right] \\ &\overset{(a)}{=} \exp\left[-\theta d_s^{\alpha}\left(\frac{\sigma^2}{P_2} + \frac{P_1}{P_2}|h_{0,i}|^2 d_{0,i}^{-\alpha} + \sum_{j\in\Phi_a^2\setminus\{i\}} |h_{j,i}|^2 d_{j,i}^{-\alpha}\right)\right] \\ &\overset{(b)}{=} \exp\left(-\frac{\theta\sigma^2 d_s^{\alpha}}{P_2}\right) \mathbb{E}\left[\frac{1}{1+\frac{P_1}{P_2}\theta d_s^{\alpha} d_{0,i}^{-\alpha}}\right] \mathcal{L}_{I_s}(\theta d_s^{\alpha}). \end{aligned} \tag{B.1}$$

Here, (a) follows from $|h_{i,i}|^2 \sim \exp(1)$. (b) follows from $|h_{0,i}|^2 \sim \exp(1)$, and the expectation is over $d_{0,i}$. $\mathcal{L}_{I_s}(s) = \mathbb{E}\left[\exp\left(-s \sum\limits_{j\in\Phi_a^2\setminus\{i\}} |h_{j,i}|^2 d_{j,i}^{-\alpha}\right)\right]$ is the Laplace transform of interference coming from active STs with normalized transmit power.

With the help of the approximation $\mathbb{E}\left[\frac{1}{1+\frac{\kappa}{d_{0,i}^{\alpha}}}\right] \simeq \frac{1}{1+\frac{\kappa^{2/\alpha}}{\mathbb{E}[d_{0,i}]^2}}$ in [17], the second term in (B.1) becomes

$$\mathbb{E}\left[\frac{1}{1+\frac{P_1}{P_2}\theta d_s^{\alpha} d_{0,i}^{-\alpha}}\right] \simeq \frac{1}{1+\frac{d_s^2}{\mathbb{E}[d_{0,i}]^2}\left(\theta\frac{P_1}{P_2}\right)^{\frac{2}{\alpha}}}. \tag{B.2}$$

Depending on the distance from the PT to the active SRs, different SRs experience different interference levels caused by the primary transmission. The expectation of $d_{0,i}$ is over all the possible locations of the typical SR inside the network region $\mathcal{C}$.

The distribution of the active SRs depends on the locations of their associated STs, which follows a homogeneous PPP with intensity $q_2\lambda_s$. For an arbitrary active SR, it can be approximately seen as uniformly distributed on the disk $\mathcal{C}$ with radius R. Hence, the pdf of the distance from the typical SR to the origin of $\mathcal{C}$, denoted by d_1, is given by

$$f_{d_1}(r) = \begin{cases} \frac{2r}{R^2} & \text{if } 0 \le r \le R \\ 0 & \text{else.} \end{cases} \tag{B.3}$$

The distance from the PT to the origin is d_p. As shown in Fig. B.1, using the

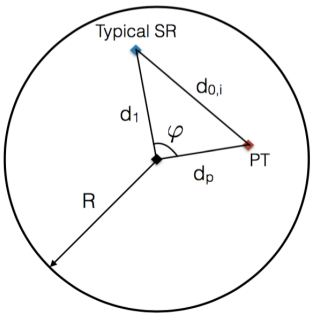


Figure B.1: Geographical locations of the PT and the typical SR on the network region $\mathcal{C}$ with radius R.

law of cosine, we have the distance between the PT and the typical SR given by

$$d_{0,i} = \sqrt{d_1^2 + d_p^2 - 2d_1 d_p \cos\varphi}, \tag{B.4}$$

where φ is a random variable uniformly distributed in $[0, 2\pi]$. Averaging over d_1 and φ, we have the expectation of the distance $d_{0,i}$ given by

$$\mathbb{E}[d_{0,i}] = \int_0^{2\pi} \frac{1}{2\pi} \int_0^R \frac{2r}{R^2} \sqrt{r^2 + d_p^2 - 2rd_p \cos\varphi} \mathrm{d}r\mathrm{d}\varphi. \tag{B.5}$$

The third term in (B.1) is the Laplace transform of interference coming from nodes in $\Phi_a^2 \backslash \{i\}$ with intensity $q_2\lambda_s$. From existing results on the interference distribution in Poisson networks [**?**], we have

$$\mathcal{L}_{I_s}(\theta d_s^\alpha) = \exp\left[-\frac{\pi q_2 \lambda_s d_s^2 \theta^{\frac{2}{\alpha}}}{\mathrm{sinc}(2/\alpha)} \right]. \tag{B.6}$$

Substituting (B.2) and (B.6) in (B.1), together with (B.5), we have

$$p_{2/1,2} \simeq \exp\left[-\frac{\pi q_2 \lambda_s d_s^2 \theta^{\frac{2}{\alpha}}}{\mathrm{sinc}(2/\alpha)}\right] \frac{\exp\left(-\frac{\theta\sigma^2 d_s^\alpha}{P_2}\right)}{1+\frac{d_s^2}{\mathbb{E}[d_{0,i}]^2}\left(\theta\frac{P_1}{P_2}\right)^{\frac{2}{\alpha}}},$$

where $\mathbb{E}[d_{0,i}]$ is given in (B.5). Proposition 3 is obtained.

B.2 Proof of Lemma 1

From the DTMC described in Fig. 3.2, we obtain the following balance equations.

$$\lambda\pi(0) = (1-\lambda)\mu_1\pi(1) \Leftrightarrow \pi(1) = \frac{\lambda}{(1-\lambda)\mu_1}\pi(0)$$

$$\begin{aligned}[\lambda(1-\mu_1)+(1-\lambda)\mu_1]\,\pi(1) &= \lambda\pi(0)+(1-\lambda)\mu_1\pi(2)\\ \Leftrightarrow \pi(2) &= \frac{\lambda^2(1-\mu_1)}{(1-\lambda)^2\mu_1^2}\pi(0).\end{aligned}$$

Summarizing, for $1 \le i \le M$ we have that

$$\pi(i) = \frac{\lambda^i(1-\mu_1)^{i-1}}{(1-\lambda)^i\mu_1^i}\pi(0),$$

and for $i > M$ we obtain

$$\pi(i) = \frac{\lambda^i(1-\mu_1)^M(1-\mu_2)^{i-M-1}}{(1-\lambda)^i\mu_1^M\mu_2^{i-M}}\pi(0).$$

Knowing that

$$\sum_{i=0}^{\infty}\pi(i) = 1, \tag{B.7}$$

combined with the previous expressions, when $\lambda \neq \mu_1$, the probability that the queue is empty is given by

$$\pi(0) = \frac{(\mu_1-\lambda)(\mu_2-\lambda)}{\mu_1\mu_2-\lambda\mu_1-\lambda\left[\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right]^M(\mu_2-\mu_1)}. \tag{B.8}$$

A special case is when $\lambda = \mu_1$. Denote $g(\lambda)$ and $h(\lambda)$ the nominator and the denominator of $\pi(0)$. Since $g(\mu_1) = h(\mu_1) = 0$, (B.8) is no longer valid. By using l'Hôpital's rule, we have

$$\pi(0) = \lim_{\lambda\to\mu_1}\frac{g'(\lambda)}{h'(\lambda)} = \frac{\mu_2-\mu_1}{\mu_1+(\mu_2-\mu_1)\frac{M+1-\mu_1}{1-\mu_1}}. \tag{B.9}$$

Combining the two cases with $\lambda \neq \mu_1$ and $\lambda = \mu_1$, we have (3.13) in Lemma 1.

The condition that the DTMC is aperiodic irreducible Markov chain, which implies that the queue is stable, is $\lambda < \mu_2$. Since $\pi(0)$ is a positive probability, we have an additional condition $0 < \pi(0) < 1$ that λ must satisfy. We consider the following cases:

- If $\lambda < \mu_1 \Rightarrow \frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1} < 1$, the denominator $h(\lambda) > \mu_2(\mu_1 - \lambda)$. Then we have $\pi(0) < \frac{\mu_2 - \lambda}{\mu_2} = 1 - \frac{\lambda}{\mu_2} < 1$. It is also obvious that $\pi(0) > 0$.
- If $\lambda = \mu_1$, from (B.9) we have $0 < \pi(0) < 1$.
- If $\mu_1 < \lambda < \mu_2 \Rightarrow \frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1} > 1$, we have $g(\lambda) < 0$. As for the denominator $h(\lambda)$, it can be proven that

$$\left(\frac{1-\lambda}{1-\mu_1}\right)^M \frac{\mu_2 - \lambda}{\mu_2 - \mu_1} < 1 < \left(\frac{\lambda}{\mu_1}\right)^{M+1}$$
$$\Longrightarrow \mu_1(\mu_2 - \lambda) < \lambda \left[\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right]^M (\mu_2 - \mu_1)$$
$$\Longrightarrow h(\lambda) < 0.$$

  Thus we have $\pi(0) = \frac{g(\lambda)}{h(\lambda)} > 0$. From $\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1} > 1$ we also know that $h(\lambda) < \mu_2(\mu_1 - \lambda) < 0$, then we have $\pi(0) < 1 - \frac{\lambda}{\mu_2} < 1$.

Since in the three cases $0 < \pi(0) < 1$ is always verified, we obtain the necessary and sufficient condition that the queue is stable when $\lambda < \mu_2$.

B.3 Proof of Lemma 2

From the results in Lemma 1, when $\lambda < \mu_2$ and $\xi \triangleq \frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1} \neq 1$, we have

$$\begin{aligned}
\mathbb{P}[1 \leq Q \leq M] &= \sum_{i=1}^{M} \pi(i) = \frac{\pi(0)}{1-\mu_1} \sum_{i=1}^{M} \left[\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right]^i \\
&= \frac{\pi(0)}{1-\mu_1} \frac{\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1} - \left[\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right]^{M+1}}{1 - \frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}} \\
&= \frac{\pi(0)\lambda\left(1-\xi^M\right)}{\mu_1 - \lambda} \\
&= \frac{\lambda\left(1-\xi^M\right)(\mu_2 - \lambda)}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2 - \mu_1)}.
\end{aligned} \tag{B.10}$$

We also have

$$\begin{aligned}
\mathbb{P}[Q > M] &= \sum_{i=M}^{\infty} \pi(i) = 1 - \sum_{i=0}^{M} \pi(i) \\
&= \frac{\lambda\xi^M(\mu_1 - \lambda)}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2 - \mu_1)}.
\end{aligned} \tag{B.11}$$

B.4 Proof of Theorem 1

From the results in Lemma 1, we have the average size of the queue at the PT given by

$$\begin{aligned}\bar{Q} &= \sum_{i=1}^{\infty} i\pi(i) = \sum_{i=1}^{M} i\pi(i) + \sum_{i=1}^{\infty}(M+i)\pi(M+i) \\ &= \sum_{i=1}^{M} i\pi(i) + M\sum_{i=1}^{\infty}\pi(M+i) + \sum_{i=1}^{\infty} i\pi(M+i). \end{aligned} \tag{B.12}$$

When $\lambda < \mu_2$ and $F \triangleq \frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1} \neq 1$, the first term can be derived as follows.

$$\begin{aligned}\sum_{i=1}^{M} i\pi(i) &= \sum_{i=1}^{M} i\pi(0)\frac{\lambda^i(1-\mu_1)^{i-1}}{(1-\lambda)^i\mu_1^i} \\ &= \frac{\pi(0)\lambda}{(1-\lambda)\mu_1}\sum_{i=1}^{M} i\left[\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right]^{i-1} \\ &= \frac{\pi(0)\lambda}{(1-\lambda)\mu_1}\sum_{i=1}^{M}\left(\left[\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right]^{i}\right)' \\ &= \frac{\pi(0)\lambda}{(1-\lambda)\mu_1}\frac{M\xi^{M+1} - \xi^M(M+1)+1}{\left(1-\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right)^2} \\ &= \frac{\lambda(1-\lambda)\mu_1\frac{\mu_2-\lambda}{\mu_1-\lambda}\left[M\xi^{M+1} - \xi^M(M+1)+1\right]}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2-\mu_1)}. \end{aligned} \tag{B.13}$$

For the second term, with the help of (3.15), we have

$$\begin{aligned}M\sum_{i=1}^{\infty}\pi(M+i) &= M\left(1-\mathbb{P}[Q>M]\right) \\ &= \frac{\lambda(\mu_1-\lambda)M\xi^M}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2-\mu_1)}. \end{aligned} \tag{B.14}$$

For the third term, with the help of (3.12), we have

$$\begin{aligned}\sum_{i=1}^{\infty} i\pi(M+i) &= \frac{\xi^M\pi(0)\lambda}{(1-\lambda)\mu_2}\sum_{i=1}^{\infty} i\left[\frac{\lambda(1-\mu_1)}{(1-\lambda)\mu_1}\right]^{i-1} \\ &= \frac{\xi^M\pi(0)\lambda}{(1-\lambda)\mu_2}\frac{1}{\left[1-\frac{\lambda(1-\mu_2)}{(1-\lambda)\mu_2}\right]^2} \\ &= \frac{\lambda(1-\lambda)\mu_2\frac{\mu_1-\lambda}{\mu_2-\lambda}\xi^M}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2-\mu_1)}. \end{aligned} \tag{B.15}$$

Substituting (B.13), (B.14) and (B.15) in (B.12), we have

$$\bar{Q} = \frac{N_1 + N_2}{\mu_1\mu_2 - \lambda\mu_1 - \lambda\xi^M(\mu_2 - \mu_1)}, \tag{B.16}$$

where

$$N_1 = \lambda(1-\lambda)\mu_1\frac{\mu_2 - \lambda}{\mu_1 - \lambda}\left[M\xi^{M+1} - \xi^M(M+1) + 1\right], \tag{B.17}$$

and

$$N_2 = \xi^M\lambda(\mu_1 - \lambda)\left[M + \frac{(1-\lambda)\mu_2}{\mu_2 - \lambda}\right]. \tag{B.18}$$

From the definition of the primary average delay in Section **??** and the expression of the average queue size $\bar{Q}$ given in (B.16), we have

$$\begin{aligned}\bar{D}_p &= \frac{\bar{Q}}{\lambda} + \frac{\mathbb{P}[Q \neq 0]}{\mathbb{P}[1 \leq Q \leq M]\mu_1 + \mathbb{P}[Q > M]\mu_2} \\ &= \frac{\bar{Q}}{\lambda} + \frac{\mu_2 - \lambda - \xi^M(\mu_2 - \mu_1)}{(1-\xi^M)(\mu_2 - \lambda)\mu_1 + \xi^M(\mu_1 - \lambda)\mu_2}.\end{aligned} \tag{B.19}$$

Theorem 1 is obtained.

B.5 Proof of Lemma 3

Define $\kappa_1 = \frac{\pi d_s^2\theta^{2/\alpha}}{\mathrm{sinc}(2/\alpha)}$, $\kappa_2 = \frac{\pi d_p^2\left(\theta\frac{P_2}{P_1}\right)^{2/\alpha}}{\mathrm{sinc}(2/\alpha)}$, and $\kappa_{12} = \frac{1}{1+\frac{d_s^2}{\mathbb{E}[d_{0,i}]^2}\left(\theta\frac{P_1}{P_2}\right)^{\frac{2}{\alpha}}}$ constant parameters related to the network setting, (3.28) becomes

$$\begin{aligned}T_s &= c^* + \frac{\lambda}{p_{1/1}}\frac{q_2\exp(-q_2\lambda_s\kappa_1)\cdot\kappa_{12} - c^*}{\exp(-q_2\lambda_s\kappa_2)} \\ &= c^* + \frac{\lambda\kappa_{12}}{p_{1/1}}\left\{q_2\exp\left[q_2\lambda_s(\kappa_2 - \kappa_1)\right] - \frac{c^*}{\kappa_{12}}\exp(q_2\lambda_s\kappa_2)\right\}.\end{aligned}$$

Define $c_{12}^* = \frac{c^*}{\kappa_{12}} = q_1^* p_{2/2}(q_1^*)\left[1 + \frac{d_s^2}{\mathbb{E}[d_{0,i}]^2}\left(\theta\frac{P_1}{P_2}\right)^{\frac{2}{\alpha}}\right]$, we need to find the optimal value of q_2 that maximizes T_s with respect to $q_2 \in [0,1]$, i.e.,

$$q_2^{\mathrm{o}} = \arg\max_{q_2\in[0,1]}\ q_2\exp\left[q_2\lambda_s(\kappa_2 - \kappa_1)\right] - c_{12}^*\exp(q_2\lambda_s\kappa_2). \tag{B.20}$$

We define the following objective function $f(x) = x\exp\left[x\lambda_s(\kappa_2 - \kappa_1)\right] - c_{12}^*\exp(x\lambda_s\kappa_2)$ with $x \in [0,1]$. First, $f(x)$ is not for sure a concave function. Secondly, maximizing $f(x)$ depends on whether $(\kappa_1 - \kappa_2)$ is positive or negative. Taking the first order derivative of $f(x)$, we have

$$f'(x) = e^{x\lambda_s\kappa_2}\left\{e^{-x\lambda_s\kappa_1}\left[1 - x\lambda_s(\kappa_1 - \kappa_2)\right] - c_{12}^*\lambda_s\kappa_2\right\}. \tag{B.21}$$

When $\frac{P_2}{P_1} < (d_s/d_p)^\alpha$, $\kappa_1 - \kappa_2 > 0$ holds. Obviously $f'(x)$ decreases with x, and we have $\lim_{x\to-\infty} f'(x) = +\infty$ and $\lim_{x\to+\infty} f'(x) = -\infty$. For $x \in (-\infty, +\infty)$, the only critical point of $f(x)$ is the global optimal (maximum) point, which can be easily obtained by using the first order optimality condition. Considering that x is bounded by $x \in [0, 1]$, we can find the optimal point in the following cases.

- If $f'(1) < f'(0) < 0$, $f(x)$ monotonically decreases in $x \in [0, 1]$. The optimal point is at $x^o = 0$.
- If $f'(0) > f'(1) > 0$, $f(x)$ monotonically increases in $x \in [0, 1]$. The optimal point is at $x^o = 1$.
- If $f'(0) > 0 > f'(1)$, the optimal point is at x^o such that $f'(x) = 0$.

From (B.21), the first order optimality condition gives

$$\begin{aligned} & e^{-x\lambda_s\kappa_1}\left[1 - x\lambda_s(\kappa_1 - \kappa_2)\right] - c_{12}^*\lambda_s\kappa_2 = 0. \\ \Longrightarrow\ & e^{x\lambda_s\kappa_1} = -\frac{\kappa_1 - \kappa_2}{c_{12}^*\kappa_2}x + \frac{1}{c_{12}^*\lambda_s\kappa_2}. \end{aligned} \tag{B.22}$$

For a general type of equation $p^{ax+b} = cx + d$, where x is the variable to be solved and a, b, c, d, p are constant, when $p > 0$ and $a, c \neq 0$, the solution by using the Lambert W function is

$$x = -\frac{W\left(-\frac{a\ln p}{c}p^{b-\frac{ad}{c}}\right)}{a\ln p} - \frac{d}{c}. \tag{B.23}$$

Solving (B.22) with the help of the Lambert W function, combined with the condition $q_2 \in [0, 1]$, we have the solution to (B.20) when $\frac{P_2}{P_1} < (d_s/d_p)^\alpha$, given by

$$q_2^{\mathrm{o}} = \min\left\{\left[-\frac{W\left(\frac{\lambda_s\kappa_1\kappa_2c_{12}^*}{\kappa_1-\kappa_2}e^{\frac{\kappa_1}{\kappa_1-\kappa_2}}\right)}{\lambda_s\kappa_1} + \frac{1}{\lambda_s(\kappa_1-\kappa_2)}\right]^+, 1\right\},$$

where $[z]^+ = \max\{z, 0\}$.

When $\frac{P_2}{P_1} \geq (d_s/d_p)^\alpha$, $f'(x)$ is not a monotonic function of x. Therefore, $f(x)$ may have more than one critical points, depending on the shape of $f(x)$ with different network parameters. Here, we disregard the case where $\frac{P_2}{P_1} \geq (d_s/d_p)^\alpha$ in order to have tractable analysis on the optimization problem.

Combining these results, we have the Lemma 3.

B.6 Proof of Lemma 4

The feasibility region $\mathcal{R}_\mathcal{F}$ is defined by the intersection of the queue stability condition and the queue size constraint. (q_2, P_2) should satisfy

1. $\lambda < \mu_1$;
2. $\frac{1-\lambda}{\mu_1-\lambda} + \frac{1}{\mu_1} < D_{\max}$.

From the first condition, we have

$$\lambda < p_{1/1,2}(q_2) \Rightarrow q_2 < \frac{\ln(P_{1/1}/\lambda)}{\lambda_s \kappa_2}. \tag{B.24}$$

From the second condition, we have

$$D_{\max}\mu_1^2 - [(D_{\max}-1)\lambda + 2]\mu_1 + \lambda < 0 \tag{B.25}$$

The solution to the inequality is

$$q_2 < \frac{\ln(P_{1/1}/\eta_1)}{\lambda_s \kappa_2} \quad \text{or} \quad q_2 > \frac{\ln(P_{1/1}/\eta_2)}{\lambda_s \kappa_2} \tag{B.26}$$

where $\eta_1 = \frac{(D_{\max}-1)\lambda+2+\sqrt{(D_{\max}-1)^2\lambda^2-4\lambda+4}}{2D_{\max}}$ and $\eta_2 = \frac{(D_{\max}-1)\lambda+2-\sqrt{(D_{\max}-1)^2\lambda^2-4\lambda+4}}{2D_{\max}}$.

Knowing that $\eta_2 < \lambda$ always holds, the intersection of (B.24) and (B.26) gives

$$q_2 < \min\left\{\frac{\ln(P_{1/1}/\lambda)}{\lambda_s \kappa_2}, \frac{\ln(P_{1/1}/\eta_1)}{\lambda_s \kappa_2}\right\}. \tag{B.27}$$

The feasible region of (q_2, P_2) is thus defined by

$$\mathcal{R}_{\mathcal{F}} = \left\{(q_2, P_2) : q_2 < \min\left\{\frac{\ln(P_{1/1}/\lambda)}{\lambda_s \kappa_2}, \frac{\ln(P_{1/1}/\eta_1)}{\lambda_s \kappa_2}\right\}\right\}. \tag{B.28}$$

B.7 Proof of Theorem 2

When $\frac{P_2}{P_1} < (d_s/d_p)^\alpha$, from (B.21) we have $f'(0) > 0$. Knowing that $f'(x)$ decreases with x, $f(x)$ is either an monotonically increasing function or firstly increases then decreases in $x \in [0, 1]$.

If q_2^o obtained in (3.30) falls within the feasible region $\mathcal{R}_{\mathcal{F}}$ given in (3.33), i.e, $q_2^o < \min\left\{\frac{\ln(P_{1/1}/\lambda)}{\lambda_s \kappa_2}, \frac{\ln(P_{1/1}/\eta_1)}{\lambda_s \kappa_2}\right\}$, the optimal value of q_2 with respect to the delay constraints is q_2^o. Otherwise $f(x)$ is an increasing function in $\mathcal{R}_{\mathcal{F}}$, and the optimal value is the one at the feasible region boundary $\min\left\{\frac{\ln(P_{1/1}/\lambda)}{\lambda_s \kappa_2}, \frac{\ln(P_{1/1}/\eta_1)}{\lambda_s \kappa_2}\right\}$.

Combining the two cases, we obtain Theorem 2.

Appendix C

Small Cell Cooperative Caching

C.1 Proof of Lemma 5

In our network model, we approximately consider the cluster of interest as a circular area $\mathcal{B}(y_0, R)$, where y_0 is the cluster center at the origin. For simplicity, in the following we use $\mathcal{B}(0, R)$ to represent the cluster area. SBSs inside the cluster of interest form the cooperation set, denoted by $\mathcal{C} = \{b_i \in \Phi_b \cap \mathcal{B}(0, R)\}$. Conditioning on having K cooperative SBSs jointly transmitting to the same user, from the definition of SCDP of JT scheme in (5.7) and the SIR expression in (5.11), with target SIR $\theta_1 = 2^{\frac{R_d}{W}} - 1$, we have

$$\begin{aligned} p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1) &= \mathbb{P}\left[\left|\sum_{b_i \in \mathcal{C}} h_i {r_i}^{-\frac{\alpha}{2}}\right|^2 > \theta_1 \sum_{b_j \in \Phi_b \setminus \{\mathcal{C}\}} |h_j|^2 {r_j}^{-\alpha}\right] \\ &= \mathbb{P}\left[\left|\sum_{i=1}^{K} h_i {r_i}^{-\frac{\alpha}{2}}\right|^2 > \theta_1 \sum_{b_j \in \Phi_b \setminus \mathcal{B}(0,R)} |h_j|^2 {r_j}^{-\alpha}\right]. \end{aligned}$$

Knowing that $\left|\sum_{i=1}^{K} h_i {r_i}^{-\frac{\alpha}{2}}\right|^2 \sim \exp\left(1/\sum_{i=1}^{K} r_i^{-\alpha}\right)$ because of the property of the sum of normally distributed random variables, then we have

$$\begin{aligned} p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1) &= \mathbb{E}_{\mathbf{r}}\left[\mathcal{L}_{I|R}\left(\frac{\theta_1}{\sum_{i=1}^{K} r_i^{-\alpha}}\right) \,\middle|\, \mathbf{r}\right] \\ &\simeq \int_{\mathbb{R}^K} \mathcal{L}_{I|R}\left(\frac{\theta_1}{\sum_{i=1}^{K} x_i^{-\alpha}}\right) f_{\mathbf{r}}(x_1, \ldots, x_K) \mathrm{d}x_1 \cdots \mathrm{d}x_K, \end{aligned} \tag{C.1}$$

where $\mathcal{L}_{I|R}(s) = \mathbb{E}\left[\exp\left(-s \sum_{b_j \in \Phi_b \setminus \mathcal{B}(0,R)} |h_j|^2 {r_j}^{-\alpha}\right)\right]$ is the Laplace transform of interference coming from out-of-cluster SBSs; $f_{\mathbf{r}}(x_1, \ldots, x_K)$ denotes the joint probability density function (pdf) of the distances $\mathbf{r} = [r_1, \ldots, r_K]$.

Since K SBSs are independently and uniformly distributed in the cluster approximated by $\mathcal{B}(0,R)$, we have the pdf of the distance r_i from the i-th SBS to the user at the origin as

$$f_{r_i}(x_i) \simeq \begin{cases} \frac{2x_i}{R^2} & 0 \le x_i \le R \\ 0 & x_i > R \end{cases} \tag{C.2}$$

for any $i \in [1,K]$. From the i.i.d. property of BPP, the joint pdf of the link distances $\mathbf{r} = [r_1, \ldots, r_K]$ is

$$f_{\mathbf{r}}(x_1, \ldots, x_K) \simeq \prod_{i=1}^{K} \frac{2x_i}{R^2}, \tag{C.3}$$

with $0 \le x_i \le R$, $\forall i \in [1,K]$. Then (C.1) becomes

$$p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1) \simeq \int_0^R \cdots \int_0^R \mathcal{L}_{I|R}\left(\frac{\theta_1}{\sum_{i=1}^{K} x_i^{-\alpha}}\right) \prod_{i=1}^{K} \frac{2x_i}{R^2} \mathrm{d}x_1 \cdots \mathrm{d}x_K. \tag{C.4}$$

Out-of-cluster interference comes from PPP distributed interfering SBSs with minimum distance R to the cluster-center user. We have the Laplace transform of interference from SBSs located out of $\mathcal{B}(0,x)$, given by

$$\begin{aligned} \mathcal{L}_{I|x}(s) &= \mathbb{E}\left[\exp\left(-s \sum_{b_j \in \Phi_b \setminus \mathcal{B}(0,x)} |h_j|^2 {r_j}^{-\alpha}\right)\right] \overset{(a)}{=} \exp\left(-2\pi\lambda_b \int_x^{\infty} \frac{s v^{-\alpha}}{1+sv^{-\alpha}} v \mathrm{d}v\right) \\ &\overset{(b)}{=} \exp\left(-\pi\lambda_b s^{\frac{2}{\alpha}} \int_{\frac{x^2}{s^{2/\alpha}}}^{\infty} \frac{1}{1+w^{\frac{2}{\alpha}}} \mathrm{d}w\right). \end{aligned} \tag{C.5}$$

Here, (a) follows from the probability generating functional (PGFL) ofPPP, and (b) is obtained by the change of variable $w = \frac{v^2}{s^{2/\alpha}}$. Combining (C.4) and (C.5), we obtain Lemma 5.

C.2 Proof of Lemma 6

From the definition of SCDP of PT-SS scheme in (5.8), success content delivery happens when the K streams after SIC are decodable, i.e., $\mathrm{SIR}_k > \theta_2$ for $k = 1, \ldots, K$, where $\theta_2 = 2^{\frac{R_d}{KW}} - 1$ is the target SIR. Then we have the SCDP of PT-SS scheme, given by

$$p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2) = \mathbb{P}[\mathrm{SIR}_i > \theta_2, \ldots, \mathrm{SIR}_K > \theta_2] = \mathbb{E}_{\widetilde{\mathbf{r}}}\left[\prod_{k=1}^{K} \mathbb{P}\left[\mathrm{SIR}_k > \theta_2\right] \middle| \widetilde{\mathbf{r}}\right]. \tag{C.6}$$

Here the link distance vector $\widetilde{\mathbf{r}} = [\widetilde{r}_1, \ldots, \widetilde{r}_K]$ is with increasing distance order, where $\widetilde{r}_k$ denotes the distance from the k-th nearest SBS to the cluster-center

user. With the approximation of cluster area as a circle $\mathcal{B}(0,R)$, using the results on the distance distribution of BPP distributed points in a circular area [121], we have the pdf of the distance from the furthest in-cluster SBS to the cluster center given as

$$f_{\widetilde{r}_K}(x_K) \simeq \frac{2K}{x_K}\left(\frac{x_K}{R}\right)^{2K}. \tag{C.7}$$

The conditional distribution of the distance $\widetilde{r}_{k-1}$ from the $(k-1)$-th nearest SBS to the cluster center knowing the distance $\widetilde{r}_k = x_k$ from the k-th nearest SBS is given by

$$f_{\widetilde{r}_{k-1}}(x_{k-1}|x_k) = \frac{2}{x_k}\cdot\frac{1}{B(1,k-1)}\left(\frac{x_{k-1}}{x_k}\right)^{2(k-1)-1}, \tag{C.8}$$

where $B(a,b)$ is the Beta function. Knowing that $f_{\widetilde{r}_k}(x_k|x_{k+1},\ldots,x_K) = f_{\widetilde{r}_k}(x_k|x_{k+1})$ because of the i.i.d. property of a BPP, we obtain the joint pdf of the distances from the k-th nearest SBS to the cluster center for $k=1,\ldots,K$ given as

$$f_{\widetilde{\mathbf{r}}}(x_1,\ldots,x_K) = f_{\widetilde{r}_K}(x_K)f_{\widetilde{r}_{K-1}}(x_{K-1}|x_K)\cdots f_{\widetilde{r}_1}(r_1|r_2) \simeq \prod_{k=1}^{K}\frac{2k\cdot x_k}{R^2}. \tag{C.9}$$

At the k-th SIC step with $k\in[1,K-1]$, since we approximately consider the distribution of interfering SBS as a homogeneous PPP, then we have

$$\mathbb{P}\left[\mathrm{SIR}_k > \theta_2 \mid \widetilde{r}_k\right] \simeq \mathbb{P}\left[\frac{|h_k|^2\,\widetilde{r}_k^{-\alpha}}{\sum\limits_{b_j\in\Phi_b\setminus\mathcal{B}(0,\widetilde{r}_k)}|h_j|^2 {r_j}^{-\alpha}} > \theta_2\right] = \mathcal{L}_{I|\widetilde{r}_k}(\theta_2\cdot\widetilde{r}_k^{\alpha}). \tag{C.10}$$

For the last decoded stream, we have

$$\mathbb{P}\left[\mathrm{SIR}_K > \theta_2 \mid \widetilde{r}_K\right] \simeq \mathbb{P}\left[\frac{|h_K|^2\,\widetilde{r}_K^{-\alpha}}{\sum\limits_{b_j\in\Phi_b\setminus\mathcal{B}(0,R)}|h_j|^2 {r_j}^{-\alpha}} > \theta_2\right] = \mathcal{L}_{I|R}(\theta_2\cdot\widetilde{r}_K^{\alpha}). \tag{C.11}$$

Combining (C.10) and (C.11) with the joint pdf in (C.9), (C.6) becomes

$$p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2) \simeq \int\limits_{0<x_1<\cdots<x_K<R} \frac{2K\cdot x_K}{R^2}\mathcal{L}_{I|R}\left(\theta_2 x_K^{\alpha}\right)\prod_{k=1}^{K-1}\frac{2k\cdot x_k}{R^2}\mathcal{L}_{I|x_k}\left(\theta_2 x_k^{\alpha}\right)\mathrm{d}x_1\cdots\mathrm{d}x_K, \tag{C.12}$$

where $\mathcal{L}_{I|x}(s)$ is given in (C.5).

C.3 Proof of Lemma 7

In the PT-OS case, due to the orthogonal spectrum usage among in-cluster SBSs, interference only comes from out-of-cluster SBSs. Under the circular

approximation $\mathcal{B}(0,R)$ of the cluster area, the interfering SBSs have minimum distance R to the cluster-center user. For each received stream i, with target SIR $\theta_1 = 2^{\frac{R_d}{W}} - 1$, we have the CCDF of SIR, given by

$$\begin{aligned}\mathbb{P}\left[\mathrm{SIR}_i > \theta_1 \mid r_i\right] &\simeq \mathbb{P}\left[\frac{|h_i|^2 {r_i}^{-\alpha}}{\sum\limits_{b_j \in \Phi_b \setminus \mathcal{B}(0,R)} |h_j|^2 {r_j}^{-\alpha}} > \theta_1\right] \\ &= \mathcal{L}_{I|R}\left(\theta_1 r_i^{\alpha}\right). \end{aligned} \tag{C.13}$$

Since the instantaneous SIR_i of each stream is independent of each other, $\min\{\mathrm{SIR}_i\} > \theta_1$ is equivalent to the event that all K streams satisfy $\mathrm{SIR}_i > \theta_1$. With the help of the approximated joint pdf of $\mathbf{r} = [r_1, \ldots, r_K]$ in (C.3), we have the SCDP of the PT-OS case, given as

$$\begin{aligned} p_{\mathrm{d},K}^{\text{PT-O}}(\theta_1) &= \mathbb{P}\left[\min_{i\in[1,\ldots,K]}\{\mathrm{SIR}_i\} > \theta_1\right] \simeq \mathbb{E}_{\mathbf{r}}\left[\prod_{k=1}^{K}\mathbb{P}\left[\mathrm{SIR}_k > \theta_1\right] \,\middle|\, \mathbf{r}\right] \\ &= \int_0^R \cdots \int_0^R \prod_{i=1}^{K} \frac{2x_i}{R^2} . \mathcal{L}_{I|R}\left(\theta_1 x_i^{\alpha}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_K, \end{aligned} \tag{C.14}$$

where $\mathcal{L}_{I|x}(s)$ is given in (C.5).

C.4 Proof of Lemma 8

For simplicity, we use $f(\rho) = f(\rho|K)$ when $K \in [2,\infty]$ is a fixed value. We exclude the case with $K = 1$ because it does not require any cache space assignment. The simplified cache service probability $f(\rho)$ in (5.28) is twice differentiable in $\rho \in [0,1]$. The second order derivative is

$$f''(\rho) = \gamma\rho^{-\gamma-1}\left[p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2) - p_{\mathrm{d},K}^{\text{JT}}(\theta_1)\right] - p_{\mathrm{d},K}^{\text{PT-S}}\gamma\left[\rho(1-K)+K\right]^{-\gamma-1}(1-K)^2. \tag{C.15}$$

Knowing that $p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2) < p_{\mathrm{d},K}^{\text{JT}}(\theta_1)$ from the results presented in Section 5.3.1, $f''(\rho)$ is always negative, thus $f(\rho)$ is strictly concave. The first order derivative is

$$f'(\rho) = (p_{\mathrm{d},K}^{\text{JT}}(\theta_2) - p_{\mathrm{d},K}^{\text{PT-S}}(\theta_1))\rho^{-\gamma} + p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2)\left(\rho(1-K)+K\right)^{-\gamma}(1-K). \tag{C.16}$$

Here, $f'(0)$ is positive, and we observe followings.

- If $f'(1) \geq 0$, that is, $\frac{p_{\mathrm{d},K}^{\text{JT}}(\theta_1)}{p_{\mathrm{d},K}^{\text{PT-S}}(\theta_2)} \geq K$, $f(\rho)$ monotonically increases in $\rho \in [0,1]$, and the optimal solution is $\rho^* = 1$.

- If $f'(1) < 0$, that is, $1 < \frac{p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1)}{p_{\mathrm{d},K}^{\mathrm{PT\text{-}S}}(\theta_2)} < K$, the optimal solution is the one that satisfies $f'(\rho) = 0$. Then, we have $\rho^* \simeq K \left[\left(\frac{K-1}{\frac{p_{\mathrm{d},K}^{\mathrm{JT}}(\theta_1)}{p_{\mathrm{d},K}^{\mathrm{PT\text{-}S}}(\theta_2)} - 1} \right)^{1/\gamma} + K - 1 \right]^{-1}$.

By combining both cases together, we get (5.29) in Lemma 8.

Bibliography

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020 white paper," *White paper*, Feb. 2015.

[2] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 20–27, Jun. 2013.

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

[4] T. Q. Quek, G. de la Roche, İ. Güvenç, and M. Kountouris, *Small cell networks: Deployment, PHY techniques, and resource management.* Cambridge University Press, 2013.

[5] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

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

[7] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014.

[8] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tutorials*, vol. 16, no. 4, pp. 1801–1819, Fourthquarter 2014.

[9] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

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

[11] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, G. Feng, and S. Li, "Device-to-device communication underlaying cellular networks," *IEEE Trans. on Commun.*, vol. 61, no. 8, pp. 3541–3551, Aug. 2013.

[12] M. G. Khoshkholgh, Y. Zhang, K. C. Chen, K. G. Shin, and S. Gjessing, "Connectivity of cognitive device-to-device communication underlying cellular networks," *IEEE Journal on Sel. Areas in Commun.*, vol. 33, no. 1, pp. 81–99, Jan. 2015.

[13] C. Xu, L. Song, and Z. Han, *Resource management for device-to-device underlay communication.* Springer, 2014.

[14] N. Naderializadeh and A. S. Avestimehr, "Itlinq: A new approach for spectrum sharing in device-to-device communication systems," in *Proc., IEEE Intl. Symp. on Inform. Theory (ISIT)*, Honolulu, HI, Jun. 2014, pp. 1573–1577.

[15] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014.

[16] Y. Pei and Y. C. Liang, "Resource allocation for device-to-device communication overlaying two-way cellular networks," *IEEE Trans. on Wireless Commun.*, vol. 12, no. 7, pp. 3611–3621, Jul. 2013.

[17] N. Lee, X. Lin, J. G. Andrews, and R. W. Heath, "Power control for D2D underlaid cellular networks: Modeling, algorithms, and analysis," *IEEE Journal on Sel. Areas in Commun.*, vol. 33, no. 1, pp. 1–13, Jan. 2015.

[18] V. Chandrasekhar, J. G. Andrews, T. Muharemovic, Z. Shen, and A. Gatherer, "Power control in two-tier femtocell networks," *IEEE Trans. on Wireless Commun.*, vol. 8, no. 8, pp. 4316–4328, Aug. 2009.

[19] N. Jindal, S. Weber, and J. G. Andrews, "Fractional power control for decentralized wireless networks," *IEEE Trans. on Wireless Commun.*, vol. 7, no. 12, pp. 5482–5492, Dec. 2008.

[20] X. Zhang and M. Haenggi, "Random power control in Poisson networks," *IEEE Trans. on Commun.*, vol. 60, no. 9, pp. 2602–2611, Sep. 2012.

[21] H. Song, J. Y. Ryu, W. Choi, and R. Schober, "Joint power and rate control for device-to-device communications in cellular systems," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 10, pp. 5750–5762, Oct. 2015.

[22] C. H. Yu, O. Tirkkonen, K. Doppler, and C. Ribeiro, "On the performance of device-to-device underlay communication with simple power control," in *IEEE 69th Vehic. Tech. Conf. (VTC)*, Apr. 2009, pp. 1–5.

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

[24] J. Gu, S. J. Bae, B.-G. Choi, and M. Y. Chung, "Dynamic power control mechanism for interference coordination of device-to-device communication in cellular networks," in *3rd Intl. Conference on Ubiquitous and Future Networks (ICUFN)*, Jun. 2011, pp. 71–75.

[25] P. Jänis, C.-H. Yu, K. Doppler, C. Ribeiro, C. Wijting, K. Hugl, O. Tirkkonen, and V. Koivunen, "Device-to-device communication underlaying cellular communications systems," *Intl. Journal of Commun., Network and System Sciences*, vol. 2, no. 3, pp. 169–178, Jun. 2009.

[26] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in LTE-advanced networks," *IEEE Wireless Commun.*, vol. 19, no. 3, pp. 96–104, Jun. 2012.

[27] C. H. Yu and O. Tirkkonen, "Device-to-device underlay cellular network based on rate splitting," in *Proc., IEEE Wireless Networking and Comm. Conf. (WCNC)*, Apr. 2012, pp. 262–266.

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

[29] Y. Kim, F. Baccelli, and G. de Veciana, "Spatial reuse and fairness of mobile ad-hoc networks with channel-aware CSMA protocols," in *Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, Princeton, NJ, May 2011, pp. 360–365.

[30] A. Hasan and J. G. Andrews, "The guard zone in wireless ad hoc networks," *IEEE Trans. on Wireless Commun.*, vol. 6, no. 3, pp. 897–906, Mar. 2007.

[31] D. Torrieri and M. C. Valenti, "Exclusion and guard zones in DS-CDMA ad hoc networks," *IEEE Trans. on Commun.*, vol. 61, no. 6, pp. 2468–2476, Jun. 2013.

[32] H. Min, J. Lee, S. Park, and D. Hong, "Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks," *IEEE Trans. on Wireless Commun.*, vol. 10, no. 12, pp. 3995–4000, Dec. 2011.

[33] X. Chen, L. Chen, M. Zeng, X. Zhang, and D. Yang, "Downlink resource allocation for device-to-device communication underlaying cellular networks," in *IEEE Intl. Symposium on Personal, Indoor and Mobile Radio Commun. (PIMRC)*, Sydney Australia, Sep. 2012, pp. 232–237.

[34] F. Baccelli, B. Blaszczyszyn, and P. Muhlethaler, "Stochastic analysis of spatial and opportunistic ALOHA," *IEEE Journal on Sel. Areas in Commun.*, vol. 27, no. 7, pp. 1105–1119, Sep. 2009.

[35] S. Weber, J. Andrews, and N. Jindal, "The effect of fading, channel inversion, and threshold scheduling on ad hoc networks," *IEEE Trans. on Inform. Theory*, vol. 53, no. 11, pp. 4127–4149, Nov. 2007.

[36] D. Zheng, W. Ge, and J. Zhang, "Distributed opportunistic scheduling for ad hoc networks with random access: An optimal stopping approach," *IEEE Trans. on Inform. Theory*, vol. 55, no. 1, pp. 205–222, Jan. 2009.

[37] L. Lei, X. . Shen, M. Dohler, C. Lin, and Z. Zhong, "Queuing models with applications to mode selection in device-to-device communication underlaying cellular networks," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 12, pp. 6697–6715, Dec. 2014.

[38] K. Doppler, C. H. Yu, C. B. Ribeiro, and P. Janis, "Mode selection for device-to-device communication underlaying an LTE-advanced network," in *Proc., IEEE Wireless Networking and Comm. Conf. (WCNC)*, Sydney, Australia, Apr. 2010, pp. 1–6.

[39] H. ElSawy, E. Hossain, and M. S. Alouini, "Analytical modeling of mode selection and power control for underlay D2D communication in cellular networks," *IEEE Trans. on Commun.*, vol. 62, no. 11, pp. 4147–4161, Nov. 2014.

[40] G. Yu, L. Xu, D. Feng, R. Yin, G. Y. Li, and Y. Jiang, "Joint mode selection and resource allocation for device-to-device communication," *IEEE Trans. on Commun.*, vol. 62, no. 11, pp. 3814–3824, Nov. 2014.

[41] B. Niu, C.-L. Wu, M. Kountouris, and Y. Li, "Distributed opportunistic medium access control in two-tier femtocell networks," in *Proc., IEEE Wireless Networking and Comm. Conf. (WCNC)*, Apr. 2012, pp. 93–97.

[42] C.-H. Liu, "Distributed interferer-channel aware scheduling in large-scale wireless ad hoc networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2013, pp. 213–218.

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

[44] G. George, R. K. Mungara, and A. Lozano, "An analytical framework for device-to-device communication in cellular networks," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 11, pp. 6297–6310, Nov. 2015.

[45] M. Ni, J. Pan, and L. Cai, "Geometrical-based throughput analysis of device-to-device communication in a sector-partitioned cell," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 4, pp. 2232–2244, Apr. 2015.

[46] H. Sun, M. Wildemeersch, M. Sheng, and T. Q. S. Quek, "D2D enhanced heterogeneous cellular networks with dynamic TDD," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 8, pp. 4204–4218, Aug. 2015.

[47] X. Xu, H. Wang, H. Feng, and C. Xing, "Analysis of device-to-device communications with exclusion regions underlaying 5G networks," *Trans. on Emerging Telecommunications Technologies*, vol. 26, no. 1, pp. 93–101, Jan. 2015.

[48] Z. Yazdanshenasan, H. S. Dhillon, M. Afshang, and P. H. J. Chong, "Poisson hole process: Theory and applications to wireless networks," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 11, pp. 7531–7546, Nov. 2016.

[49] C. H. Lee and M. Haenggi, "Interference and outage in Poisson cognitive networks," *IEEE Trans. on Wireless Commun.*, vol. 11, no. 4, pp. 1392–1401, Apr. 2012.

[50] N. Deng, W. Zhou, and M. Haenggi, "Heterogeneous cellular network models with dependence," *IEEE Journal on Sel. Areas in Commun.*, vol. 33, no. 10, pp. 2167–2181, Oct. 2015.

[51] D. Bertsekas and R. Gallager, *Data Networks (2nd ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1992.

[52] E. Bastug, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: modeling and tradeoffs," *EURASIP Journal on Wireless Commun. and Networking*, vol. 2015, no. 1, pp. 1–11, Feb. 2015.

[53] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. on Inform. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[54] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

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

[56] X. Peng, J. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, Dec. 2015.

[57] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[58] E. G. Coffman, Jr. and P. J. Denning, *Operating Systems Theory*. Prentice Hall Professional Technical Reference, 1973.

[59] C. Fricker, P. Robert, and J. Roberts, "A versatile and accurate approximation for LRU cache performance," in *International Teletraffic Congress*, ser. ITC '12. International Teletraffic Congress, 2012, pp. 8:1–8:8. [Online]. Available: http://dl.acm.org/citation.cfm?id=2414276.2414286

[60] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *Proc. IEEE Conf. on Computer Commun. (INFOCOM)*, New York, NY, Mar. 1999, pp. 126–134.

[61] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, "Content-aware user clustering and caching in wireless small cell networks," in *IEEE Intl. Symp. on Wireless Commun. Systems (ISWCS)*, Barcelona, Spain, Aug. 2014, pp. 945–949.

[62] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching and file sharing under heterogeneous file preferences," in *Proc., IEEE Intl. Conf. on Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[63] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Temporal locality in today's content caching: Why it matters and how to model it," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 5, pp. 5–12, Nov. 2013.

[64] ——, "Unravelling the impact of temporal and geographical locality in content caching systems," *IEEE Trans. on Multimedia*, vol. 17, no. 10, pp. 1839–1854, Oct. 2015.

[65] Y. Zhang, E. Pan, L. Song, W. Saad, Z. Dawy, and Z. Han, "Social network aware device-to-device communication in wireless networks," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 1, pp. 177–190, Jan. 2015.

[66] S. H. Chae, J. Y. Ryu, T. Q. S. Quek, and W. Choi, "Cooperative transmission via caching helpers," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, Dec. 2015, pp. 1–6.

[67] J. Song, H. Song, and W. Choi, "Optimal caching placement of caching system with helpers," in *Proc., IEEE Intl. Conf. on Commun. (ICC)*, London, UK, Jun. 2015, pp. 1825–1830.

[68] H. Kang, K. Park, K. Cho, and C. Kang, "Mobile caching policies for device-to-device (D2D) content delivery networking," in *Proc., IEEE Conf. on Computer Commun. Workshops (INFOCOM WKSHPS)*, Toronto, Canada, Apr. 2014, pp. 299–304.

[69] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Fundamentals of cluster-centric content placement in cache-enabled device-to-device networks," *IEEE Trans. on Commun.*, vol. 64, no. 6, pp. 2511–2526, Jun. 2016.

[70] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless content caching for small cell and D2D networks," *IEEE Journal on Sel. Areas in Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.

[71] M. Ji, G. Caire, and A. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc., IEEE Intl. Symp. on Inform. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013, pp. 1461–1465.

[72] ——, "Wireless device-to-device caching networks: basic principles and system performance," *IEEE Journal on Sel. Areas in Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[73] C. Jarray and A. Giovanidis, "The effects of mobility on the hit performance of cached D2D networks," in *Intl. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, Tempe, Arizona, May 2016, pp. 1–8.

[74] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," in *Proc., IEEE Intl. Conf. on Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[75] K. Poularakis and L. Tassiulas, "Code, cache and deliver on the move: A novel caching paradigm in hyper-dense small-cell networks," *IEEE Trans. on Mobile Computing*, vol. PP, no. 99, pp. 1–1, 2016.

[76] ——, "On the complexity of optimal content placement in hierarchical caching networks," *IEEE Trans. on Commun.*, vol. 64, no. 5, pp. 2092–2103, May 2016.

[77] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5g wireless networks," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[78] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc., IEEE Intl. Conf. on Commun. (ICC)*, London, UK, Jun. 2015, pp. 3358–3363.

[79] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. on Inform. Theory*, vol. 61, no. 12, pp. 6833–6859, Dec. 2015.

[80] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching," *to appear in IEEE Trans. on Wireless Commun.*

[81] D. Malak, M. Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Trans. on Commun.*, vol. PP, no. 99, pp. 1–1, 2016.

[82] D. Liu and C. Yang, "Caching policy toward maximal success probability and area spectral efficiency of cache-enabled hetnets," *submitted to IEEE Trans. on Commun.* [Online]. Available: http://arxiv.org/abs/1608.03749

[83] A. Liu and V. K. N. Lau, "Asymptotic scaling laws of wireless ad hoc network with physical layer caching," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 3, pp. 1657–1664, Mar. 2016.

[84] W. C. Ao and K. Psounis, "Distributed caching and small cell cooperation for fast content delivery," in *Proc., ACM Intl. Symp. on Mobile Ad Hoc Networking and Computing (MobiHoc)*, Hangzhou, China, Jun. 2015, pp. 127–136.

[85] M. Haenggi, *Stochastic geometry for wireless networks.* Cambridge University Press, 2012.

[86] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE Journal on Sel. Areas in Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.

[87] J. Andrews, F. Baccelli, and R. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. on Commun.*, vol. 59, no. 11, pp. 3122–3134, November 2011.

[88] M. Haenggi and R. K. Ganti, "Interference in large wireless networks," *Found. Trends Netw.*, vol. 3, no. 2, pp. 127–248, Feb. 2009.

[89] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," *IEEE Journal on Sel. Areas in Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

[90] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Device-to-device modeling and analysis with a modified Matérn hardcore BS location model," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Atlanta, GA, Dec. 2013, pp. 1825–1830.

[91] M. Haenggi, "Mean interference in Hard-Core wireless networks," *IEEE Commun. Letters*, vol. 15, no. 8, pp. 792–794, August 2011.

[92] V. Suryaprakash, J. Møller, and G. Fettweis, "On the modeling and analysis of heterogeneous radio access networks using a Poisson cluster process," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 2, pp. 1035–1047, Feb. 2015.

[93] K. Gulati, B. L. Evans, J. G. Andrews, and K. R. Tinsley, "Statistics of co-channel interference in a field of Poisson and Poisson-Poisson clustered interferers," *IEEE Trans. on Signal Processing*, vol. 58, no. 12, pp. 6207–6222, Dec. 2010.

[94] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks," *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 74 – 81, Apr. 2009.

[95] F. Afroz, R. Subramanian, R. Heidary, K. Sandrasegaran, and S. Ahmed, "SINR, RSRP, RSSI and RSRQ measurements in long term evolution networks," *International Journal of Wireless & Mobile Networks (IJWMN)*, vol. 7, Aug. 2015.

[96] F. Baccelli and B. Blaszczyszyn, "Stochastic geometry and wireless networks: Volume I theory," *Found. Trends Netw.*, vol. 3, no. 3-4, pp. 249–449, Mar. 2009.

[97] H. ElSawy and E. Hossain, "On stochastic geometry modeling of cellular uplink transmission with truncated channel inversion power control," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 8, pp. 4454–4469, Aug. 2014.

[98] T. D. Novlan, H. S. Dhillon, and J. G. Andrews, "Analytical modeling of uplink cellular networks," *IEEE Trans. on Wireless Commun.*, vol. 12, no. 6, pp. 2669–2679, Jun. 2013.

[99] L. Tong, Q. Zhao, and G. Mergen, "Multipacket reception in random access wireless networks: from signal processing to optimal medium access control," *IEEE Commun. Mag.*, vol. 39, no. 11, pp. 108–112, Nov. 2001.

[100] N. Pappas and M. Kountouris, "Throughput of a cognitive radio network under congestion constraints: A network-level study," in *Proc. IEEE Intl. Conf. on Cognitive Radio Oriented Wireless Networks and Commun.*, Oulu, Finland, Jun. 2014, pp. 162–166.

[101] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE Journal on Sel. Areas in Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[102] S. Lee and K. Huang, "Coverage and economy of cellular networks with many base stations," *IEEE Commun. Letters*, vol. 16, no. 7, pp. 1038–1040, Jul. 2012.

[103] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances in Computational Mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[104] K. Huang and J. Andrews, "An analytical framework for multicell cooperation via stochastic geometry and large deviations," *IEEE Trans. on Inform. Theory*, vol. 59, no. 4, pp. 2501–2516, Apr. 2013.

[105] Y. Zhuang, Y. Luo, L. Cai, and J. Pan, "A geometric probability model for capacity analysis and interference estimation in wireless mobile cellular systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Houston, TX, Dec. 2011, pp. 1–6.

[106] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE Journal on Sel. Areas in Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[107] V. Sourlas, P. Georgatsos, P. Flegkas, and L. Tassiulas, "Partition-based caching in information-centric networks," in *Proc., IEEE Intl. Workshop on Network Science for Commun. Networks (NetSciCom)*, Hong Kong, Apr. 2015, pp. 396–401.

[108] S. Sen, N. Santhapuri, R. Choudhury, and S. Nelakuditi, "Successive interference cancellation: Carving out MAC layer opportunities," *IEEE Trans. on Mobile Computing*, vol. 12, no. 2, pp. 346–357, Feb. 2013.

[109] X. Zhang and M. Haenggi, "The performance of successive interference cancellation in random wireless networks," *IEEE Trans. on Inform. Theory*, vol. 60, no. 10, pp. 6368–6388, Oct. 2014.

[110] D. C. Chen, T. Q. S. Quek, and M. Kountouris, "Backhauling in heterogeneous cellular networks: Modeling and tradeoffs," *IEEE Trans. on Wireless Commun.*, vol. 14, no. 6, pp. 3194–3206, Jun. 2015.

[111] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE Journal on Sel. Areas in Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.

[112] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[113] M. Wildemeersch, T. Q. S. Quek, M. Kountouris, A. Rabbachin, and C. H. Slump, "Successive interference cancellation in heterogeneous cellular networks," *IEEE Trans. on Commun.*, vol. 62, no. 12, pp. 4440–4453, Dec. 2014.

[114] S. E. Elayoubi and J. Roberts, "Performance and cost effectiveness of caching in mobile access networks," in *Proc., ACM Conf. on Information-Centric Networking (ICN)*, San Francisco, CA, Sep. 2015, pp. 79–88.

[115] G. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient wireless communications: tutorial, survey, and open issues," *IEEE Wireless Commun.*, vol. 18, no. 6, pp. 28–35, Dec. 2011.

[116] S. Tombaz, P. Monti, K. Wang, A. Vastberg, M. Forzati, and J. Zander, "Impact of backhauling power consumption on the deployment of heterogeneous mobile networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Houston, TX, Dec. 2011, pp. 1–5.

[117] D. Warrier and U. Madhow, "On the capacity of cellular CDMA with successive decoding and controlled power disparities," in *IEEE Vehicular Technology Conference*, vol. 3. IEEE, 1998, pp. 1873–1877.

[118] J. G. Andrews and T. H. Meng, "Optimum power control for successive interference cancellation with imperfect channel estimation," *IEEE Trans. on Wireless Commun.*, vol. 2, no. 2, pp. 375–383, Mar. 2003.

[119] M. Tanemura, "Statistical distributions of Poisson Voronoi cells in two and three dimensions," *Forma*, vol. 18, no. 4, pp. 221–247, Nov. 2003.

[120] F. Jarai-Szabo and Z. Neda, "On the size-distribution of Poisson Voronoi cells," *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, Jul. 2004.

[121] S. Srinivasa and M. Haenggi, "Distance distributions in finite uniformly random networks: Theory and applications," *IEEE Trans. on Veh. Technology*, vol. 59, no. 2, pp. 940–949, Feb. 2010.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Communication Centrée sur les Utilisateurs et les Contenus dans les Réseaux sans Fil

Mots clefs : device-to-device (D2D), caching proactif, géométrie aléatoire, coopération à petites cellules

Résumé : Cette thèse porte sur plusieurs technologies de déchargement cellulaire pour les futurs réseaux sans fil avec l'amélioration envisagée sur la efficacité spatiale du spectre et l'efficacité énergétique. Notre recherche concerne deux directions principales, y compris la communication D2D underlaid dans les réseaux cellulaires et le caching proactif au bord de réseau.
La première partie de cette thèse contient deux chapitres qui présentent nos résultats de recherche sur les réseaux cellulaire avec D2D underlaid. Notre recherche se focalise sur l'accès opportuniste distribué, dont la performance en termes du débit D2D est optimisé dans deux scénarios: 1) en supposant que l'utilisateur cellulaire avec un trafic saturé peut avoir une probabilité de couverture minimale; 2) en supposant que le trafic discontinu à l'utilisateur cellulaire, dont le délai moyen doit être maintenue au-dessous d'un certain seuil. La deuxième partie de cette thèse se focalise sur les méthodes de caching proactif au bord de réseau, y compris le caching aux petites cellules et aux appareils des utilisateurs. Tout d'abord, nous étudions le placement de contenu probabiliste dans différents types de réseaux et avec différents objectifs d'optimisation. Deuxièmement, pour le caching aux petites cellules, nous proposons un schéma coopérative parmi les petites stations de base, qui exploite le gain combiné du caching coopérative et les techniques de multipoint coordonnée. Les modèles de processus ponctuel nous permet de créer la connexion entre la diversité de transmission en couche PHY et la diversité de contenus stockés.

Title : User-Centric Content-Aware Communication in Wireless Networks

Keywords : device-to-device (D2D), proactive caching, stochastic geometry, small cell cooperation

Abstract : This thesis focuses on several emerging technologies towards future wireless networks with envisaged improvement on the area spectral efficiency and energy efficiency. The related research involves two major directions, including device-to-device (D2D) communication underlaid cellular networks and proactive caching at network edge. The first part of this thesis starts with introducing D2D underlaid cellular network model and distributed access control methods for D2D users that reuse licensed cellular uplink spectrum. We aim at optimize the throughput of D2D network in the following two scenarios: 1) assuming always backlogged cellular users with coverage probability constraint, 2) assuming bursty packet arrivals at the cellular user, whose average delay must be kept below a certain threshold. The second part of this thesis focuses on proactive caching methods at network edge, including at small base stations (SBSs) and user devices. First, we study and compare the performance of probabilistic content placement in different types of wireless caching networks and with different optimization objectives. Second, we propose a cooperative caching and transmission strategy in a cluster-centric small cell networks (SCNs), which exploits the combined gain of cache-level cooperation and CoMP technique. Using spatial models from stochastic geometry, we build the connection between PHY transmission diversity and the content diversity in local caches.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France