Distributed caching methods in small cell networks

Ejder Bastug

THESE DE DOCTORAT
DE
L'UNIVERSITÉ PARIS-SACLAY
PREPAREE A
CENTRALESUPÉLEC

ECOLE DOCTORALE N°580
Sciences et technologies de l'information et de la communication (STIC)

Spécialité : Réseaux, Information et Communications

Par

M. Ejder Baştuğ

Les Méthodes de Caching Distribué dans les Réseaux Small Cells

(Distributed Caching Methods in Small Cell Networks)

Thèse présentée et soutenue à Gif-sur-Yvette, le 14 decembre 2015 :

Composition du jury :

M. Hikmet Sari,	CNRS/CentraleSupélec	Examinateur, Président du Jury
M. Giuseppe Caire,	Technische Universität Berlin	Examinateur, Rapporteur
M. Deniz Gunduz,	Imperial College	Examinateur, Rapporteur
M. Laurent Massoulié,	Microsoft Research - INRIA Joint Centre	Examinateur
M. Mérouane Debbah,	Chaire LANEAS - CentraleSupélec	Directeur de Thèse
M. Jean-Claude Belfiore,	Telecom ParisTech	Co-Directeur de Thèse
M. Mehdi Bennis,	Université d'Oulu	Invité
M. Jean-Louis Guénégo,	JLG Consulting	Invité

Titre : Les Méthodes de Caching Distribué dans les Réseaux Small Cells

Mots clés : caching proactif, géométrie stochastique, apprentissage automatique, réseaux cellulaires, 5G

Résumé : Cette thèse explore le caching proactif, l'un des principaux paradigmes des réseaux cellulaires 5G, utilisé en particulier dans les réseaux à petites cellules (RPCs). La capacité de prévision des petites stations de bases couplée avec les récents développements dans le stockage, la sensibilité au contexte et les réseaux sociaux, le caching distribué permet de réduire considérablement les pics de trafic dans la demande des utilisateurs en servant de manière proactive ces derniers en fonction de leurs demandes potentielles, et en stockant les contenus à la fois dans les stations de base et dans les terminaux des utilisateurs. Pour montrer la faisabilité des techniques de caching proactif, nous abordons le problème sous deux angles différents, à savoir théorique et pratique.

Dans la première partie de cette thèse, nous utiliserons des outils de la géométrie stochastique pour modéliser et analyser les gains théoriques résultant du stockage dans les stations de base. Nous nous focalisons en particulier sur 1-) les réseaux "niveau-simple" dans lesquels de petites stations de base ont une capacité de stockage limitée, 2-) Réseaux "niveau-multiples" avec un backhaul à capacité limitée et 3-) Les réseaux "niveau-multiples groupés" à deux topologies différentes: déploiements en fonction de la couverture et en fonction de la capacité. Nous y caractérisons les gains de stockage en termes de débit moyen fourni et de délai moyen, puis nous montrons différents compromis en fonction du nombre de stations de base, de la taille de stockage, du facteur de popularité des contenus et du débit des contenus ciblés.

Dans la seconde partie de la thèse, nous nous focalisons à une approche pratique du caching proactif et nous nous focalisons sur l'estimation du facteur de popularité des contenus et les aspects algorithmiques.

En particulier, 1-) nous établissons dans un premier lieu, les gains du caching proactif à la fois au niveau des stations de base et des terminaux des utilisateurs, en utilisant des outils récents d'apprentissage automatique en exploitant le transfert des communications appareil-à-appareil (AàA); 2-) nous proposons une approche d'apprentissage sur la base de la richesse des informations échangées entre terminaux (que nous désignons par *domaine source*) dans le but d'avoir une meilleure estimation de la popularité des différents contenus et des contenus à stocker de manière stratégique dans les stations de base (que nous désignons par *domaine cible*); 3-) Enfin, pour l'estimation de la popularité des contenus en pratique, nous collectons des données de trafic d'usagers mobiles d'un opérateur de télécommunications sur plusieurs de ses stations de base pendant un certain nombre d'observations. Cette grande quantité de données entre dans le cadre du traitement "Big Data" et nécessite l'utilisation de nouveaux mécanismes d'apprentissage automatique adaptés à ces grandes masses de données. A ce titre, nous proposons une architecture parallélisée dans laquelle l'estimation de la popularité des contenus et celle du stockage stratégique au niveau des stations de base sont faites simultanément.

Nos résultats et analyses fournissent des visions clés pour le déploiement du stockage de contenus dans les petites stations de base, l'une des solutions les plus prometteuses des réseaux cellulaires mobiles hétérogènes 5G.

Title: Distributed Caching Methods in Small Cell Networks

Keywords: proactive caching, stochastic geometry, machine learning, cellular networks, 5G

Abstract: This thesis explores one of the key enablers of 5G wireless networks leveraging small cell network deployments, namely proactive caching. Endowed with predictive capabilities and harnessing recent developments in storage, context-awareness and social networks, peak traffic demands can be substantially reduced by proactively serving predictable user demands, via caching at base stations and users' devices. In order to show the effectiveness of proactive caching techniques, we tackle the problem from two different perspectives, namely theoretical and practical ones.

In the first part of this thesis, we use tools from stochastic geometry to model and analyse the theoretical gains of caching at base stations. In particular, we focus on 1) single-tier networks where small base stations with limited storage are deployed, 2) multi-tier networks with limited backhaul, and) multi-tier clustered networks with two different topologies, namely coverage-aided and capacity-aided deployments. Therein, we characterize the gains of caching in terms of average delivery rate and mean delay, and show several trade-offs as a function of the number of base stations, storage size, content popularity behaviour and target content bitrate.

In the second part of the thesis, we take a more practical approach of proactive caching and focus on content popularity estimation and algorithmic aspects.

In particular: 1) We first investigate the gains of proactive caching both at base stations and user terminals, by exploiting recent tools from machine learning and enabling social-network aware device-to-device (D2D) communications; 2) we propose a transfer learning approach by exploiting the rich contextual information extracted from D2D interactions (referred to as *source domain*) in order to better estimate the content popularity and cache strategic contents at the base stations (referred to as *target domain*); 3) finally, to estimate the content popularity in practice, we collect users' real mobile traffic data from a telecom operator from several base stations in hours of time interval. This amount of large data falls into the framework of big data and requires novel machine learning mechanisms to handle. Therein, we propose a parallelized architecture in which content popularity estimation from this data and caching at the base stations are done simultaneously.

Our results and analysis provide key insights into the deployment of cache-enabled small base stations, which are seen as a promising solution for 5G heterogeneous cellular networks.

*... this thesis is dedicated to my grandfather Osman,
for his encouragements in the early stage of my childhood ...*

Acknowledgments

This PhD was an opportunity to challenge myself. I tried and I was not alone. My deepest thanks to:

Contents

Acronyms

LAC   location area code. 110
LFU   least-frequently used. 49
LRU   least-recently used. 49
LTE   long term evolution. 16, 17

MAB   multi-armed band. 19
MIMO   multiple-input multiple-output. 17, 18, 20, 21

OTT   over-the-top. 103

PDF   probability distribution function. 32, 47, 50,
        66, 149, 150, 162, 167
PDN   packet data network. 109
PGFL   probability generating functional. 150, 163
PHY   physical layer. 31
PPP   Poisson point process. 21, 22, 31, 37, 45, 46,
        59–62, 66, 67, 149–151, 159, 160, 163–166, 169,
        170

QoE   quality-of-experience. 23, 32, 34, 43, 106, 107
QoS   quality-of-service. 16, 21, 49, 70, 81

RHS   right hand side. 53, 54
RMSE   root-mean-square error. 114

SAC   service area code. 110
SBS   small base station. 15–19, 21, 22, 31–34, 36,
        37, 40, 41, 43, 59, 79–82, 84, 85, 89, 91–96,
        98–101, 104–108, 113, 114
SCN   small cell network. 15, 16, 18, 19, 23, 81
SDN   software defined networking. 18
SGCN   Serving GPRS Support Node. 109
SINR   signal-to-interference-plus-noise ratio. 17, 21,
        22, 31, 33, 34, 37, 49, 119, 121, 123, 124, 149
SIR   signal-to-interference ratio. 23, 46, 47, 49, 56,
        57, 62, 65, 66, 123, 157, 162, 167
SNR   signal-to-noise ratio. 21, 72, 74
SON   self-organizing networks. 16
SSD   solid-state disk. 50
SVD   singular value decomposition. 82, 108, 113

TDMA   time-division multiple access. 20, 54, 159
TEID   tunnel endpoint identifier. 110, 111
TL   transfer learning. 91, 92, 96–102

TTT   time-to-trigger. 17

UL   uplink. 17

URI   request-uniform resource identifier. 110–112

UT   user terminal. 16–18, 79–81, 84, 85, 89, 91, 92, 100, 104

List of Figures

List of Tables

List of Tables

Chapter 1

Resumé (French)

1.1 Contexte et Motivation

La récente explosion des smartphones a substantiellement enrichi l'expérience des utilisateurs mobiles qui a conduit au développement de nouveaux services mobiles à savoir la diffusion multimédia, les applications web et les réseaux sociaux inter-connectés. Ce phénomène a été davantage alimenté par la diffusion vidéo mobile qui représente aujourd'hui près de 50 % du trafic de données avec une projection de 500 fois plus dans les dix prochaines années [1]. D'un autre coté, les réseaux sociaux représentent le deuxième plus important volume de trafic de données avec une part avoisinant les 15 % [2]. Ces nouveaux phénomènes ont rapidement alerté les opérateurs de services mobiles au redéploiement de leurs actuels réseaux en développant des techniques plus avancées et sophistiquées pour élargir la couverture du réseau, booster la capacité du réseau, et fournir à faible coût les contenus désirés à proximité des utilisateurs.

Une approche prometteuse pour faire face à ces nouvelles demandes de trafic consiste au développement des réseaux à petites cellules (RPCs) [3]. Les RPCs représentent un nouveau paradigme des réseaux basé sur le déploiement de petites stations de base (PSBs) à faible couverture, faible consommation d'énergie et faible coût en adéquation avec le réseau macro cellulaire sous-jacent. De nos jours, la grande majorité des travaux de recherche se focalisaient sur les problèmes liés à l'auto-organisation, à la coordination de l'interférence entre cellules (CIEC), au déchargement du trafic, et à l'efficacité énergétique, etc (voir [4] et les références relatives). Ces études étaient établies sous l'existence du paradigme réseau *réactif*, dans lequel les demandes de trafic utilisateurs et les flux doivent être servis de manière urgente ou rejetés, ce qui induit des pertes. Pour cela, le paradigme des réseaux à petites cellules est loin de résoudre les problèmes de pic de demande de trafic dont le déploiement à grande échelle nécessite des coûts élevés pour l'acquisition, l'installation des sites et de backhaul. Ces défauts deviendront de plus en plus importants au vu du nombre de plus en plus important de terminaux connectés et de l'avènement des réseaux ultra denses, et continueront à limiter les infrastructures de réseaux cellulaires actuels. Ces observations clés nécessitent un nouveau paradigme réseau, qui va au delà du

déploiement réseaux cellulaires à petites cellules et hétérogènes actuels, en tenant compte des récents développements en stockage, la sensibilité au contexte et les réseaux sociaux [5].

Le nouveau paradigme réseau est *pro-actif* en ce sens que les nœuds aux bords du réseau (à savoir PSBs et terminal utilisateurs (TUs)) prédisent les besoins en information des utilisateurs et pré-stockent intelligemment les contenus stratégiques dans le but de décharger le backhaul en même temps que de satisfaire la qualité de service (QdS) des utilisateurs. Cela va au-delà de l'objet des réseaux cellulaires traditionnels qui avaient été développés en supposant des TUs muets avec capacité de stockage et de traitement limités. De nos jours, les TUs sont plus sophistiqués qu'avant, donnant l'opportunité d'exploiter leurs capacités et celles des RPCs, en stockant les contenus prédits aux bords du réseau. Cela implique des gains considérables en termes de ressources réseau et minimise les dépenses opérationnelles [4].

Comme énoncé précédemment, des résultats récents ont montré que les comportements humains sont corrélés et prédictibles sur un large horizon [6]. A ce titre, les PSBs sont supposées équipées d'unités de stockage et le backhaul à faible débit est utilisé pour leurs larges connexions. Ensuite, comme nous montrerons dans les prochaines sections, stocker de manière pro-active les contenus des utilisateurs dans les PSBs évite d'avoir un backhaul chargé et des utilisateurs insatisfaits. Le caching pro-actif est basé sur l'idée du stockage des contenus populaires au niveaux des PSBs. Pour y parvenir, la popularité des contenus doit être estimée. En utilisant des outils d'apprentissage automatique et en analysant les journaux de l'infrastructure (comme dans [7]), un trésor d'informations cachées sur les utilisateurs peut être obtenu. Ces analyses entrent dans le cadre du phénomène de *big data* où les méthodes de filtrage collaboratif (FC) peuvent être appliquées pour l'inférence.

Dans la suite de cette section, nous donnerons dans un premier lieu un aperçu des avancées passées, futures et récentes dans les RPCs. Ensuite, nous discuterons brièvement l'historique du stockage et résumerons les efforts récents dans le contexte des réseaux mobiles cellulaires. Le plan de la thèse sera présenté en conséquence.

1.1.1   Réseaux à Petites Cellules: Passé, Présent et futures tendances

Les Smartphones ont exponentiellement augmenté le trafic de charge dans les réseaux cellulaires actuels ne présentant aucun signe de lenteur [1,2]. Il est maintenant bien connu qu'une manière effective d'augmenter la capacité du réseau est d'avoir de très petites cellules en réduisant la distance aux utilisateurs [4]. En fait, la densification des cellules est allée de l'ordre de centaines de kilomètres carrés (dans les années 80) à une fraction de mètres carrés ou moins avec l'avènement des hot-spots. Il y a eu récemment un grand intérêt pour le déploiement des relais, des antennes distribuées et de cellules à petits points d'accès (telles que micro/pico/femto cellules) dans les maisons résidentielles, les passages souterrains, les entreprises et les surfaces hot-spot. Ces architectures réseau qui sont à la fois déployés par les opérateurs ou les utilisateurs sont connues sous le nom de réseaux

hétérogènes (ResHéts) ou RPCs [3, 4]. En déployant des nœuds réseaux additionnels à couverture locale et en rapprochant les utilisateurs au réseau, les réseaux à petites cellules peuvent significativement améliorer la réutilisation spatiale et la couverture, booster la capacité et décharger le trafic plus efficacement [3].

Il existe une large littérature sur le sujet des ResHéts et des RPCs abordant plusieurs aspects allant de la gestion des interférences, de l'association des cellules, la modélisation stochastique du réseau, la CIEC, l'efficacité énergétique, les réseaux self organisés (RSOs), la gestion de la mobilité, les réseaux long term evolution (LTE)/Wi-Fi, pour ne citer que ceux là (voir [3] pour une large revue). L'une des clés tirée de ces études est que la la coordination stricte d'interférence entre les macro cellules d'une part et les micro/pico/femto cellules adjacentes d'autre part est nécessaire pour atteindre des gains importants suite à la répartition en cellules. Cela repose sur la disponibilité d'un backhaul à faible latence et à grande capacité [8]. Les approches de modélisation de réseau basées sur les outils de géométrie stochastique ont permis d'établir des bornes inférieures aux gains de performance en termes de capacités globale et par utilisateur. La caractéristique intéressante de ces approches est attribuée au fait que malgré la consommation en temps des niveaux de simulations, de fondamentales visions peuvent être tirées de ces outils, dont un certain nombre ont été corroborés par des essais dans le domaine de l'industrie et des observations provenant de simulations [9]. En parallèle à cela, la gestion de la mobilité a reçu une attention significative aussi bien dans le secteur du sans fil industriel que dans les groupes de recherches et les organismes de standardisation [10]. Contrairement aux réseaux homogènes conventionnels où les TUs utilisent typiquement le même ensemble de paramètres handover (à savoir marge d'hystérésis, temps de vie (TDV), etc.), utiliser le même ensemble de paramètres de handover dans les ResHéts pour toutes les cellules et/ou pour tous les TUs peut dégrader les performances de mobilité. Ceci parce que les macros TUs à grande mobilité pourraient rapidement entrer dans la surface de couverture des petites cellules avant que le TDV optimisé pour les macro-cellule expire, entraînant donc un échec du handover (suite à une dégradation du rapport signal sur interférence plus bruit (RSIB)) [11]. Les stratégies décentralisées de gestion/réduction des interférences dans les scénarios impliquant des interférences co-canal ont aussi été étudiées en détails, où les petites cellules sont capables de s'organiser par elles mêmes sur la base des informations locales et d'optimiser leurs stratégies de transmission (à savoir puissance/fréquence) sur la base de l'échange d'un minimum d'information [12]. Cela conduit à un ensemble de compromis en termes de rapidité/lenteur de convergence au prix d'information partielle/-complète. L'agrégation de porteuse (AP) et ses améliorations uni/multiflux ont également été examinés comme moyens pour booster de plus en plus, la capacité du réseau et de celle par utilisateur, opération dans laquelle les utilisateurs peuvent être simultanément servis sur plusieurs bandes [13]. En plus, avec l'augmentation asymétrique du trafic en uplink (UL) par rapport au downlink (DL), de nouveaux mécanismes d'associations de cellules et d'architectures sont nécessaires pour faire face aux nouveaux types d'interférences entre nœuds (DL-to-UL), ouvrant par conséquent de nouvelles issues de recherche telles que les communications DL/UL flexibles, massif-entrées multiples sorties multiples (massif-EMSM), appareil-à-appareil (AàA), full-duplex, etc. [4] [14]. Enfin, le sujet de la coexis-

tence LTE et Wi-Fi a reçu une attention énorme due à l'aptitude multi-mode des PSBs[1] et la possibilité d'utiliser à la fois des bandes licenciées et non licenciées. A cet égard, la balance de charge dynamique et les mécanismes de pilotage du trafic ont été proposés pour faciliter la disponibilité du Wi-FI à fournir des services de best-effort, charge de trafic, tolérance au délai, etc [15].

Quoique la densification du réseau par de petites cellules constitue la bonne voie à suivre, un certain nombre de défis techniques restent non résolus. En fait, bien qu'il a été montré que la densification avec de petites cellules permet de booster la capacité, ajouter simplement des petites cellules peut être inefficace énergétiquement [16]. De plus, l'utilisation du backhaul ainsi que le positionnement des stations de base sont des facteurs qui impactent considérablement la performance des réseaux cellulaires, et doivent donc être optimisés avant de pouvoir déployer les petites stations de base. L'importance de backhaul est davantage mis en évidence avec la prolifération sans relâche des smartphones avec la grande plage de nouveaux services sans fil (à savoir streaming multimédia, applications web, etc.). De ce fait, de nouvelles approches aux réseaux à petites cellules avec backhaul sous-jacent ont récemment été proposées dans la littérature [17] à savoir comment découpler de façon optimale les plans de contrôle et de données pour faire des cellules plus adaptées aux demandes de trafic dynamiques et l'état du réseau tout en ayant une vue globale du réseau, le déchargement de backhaul via un stockage intelligent aux bords du réseau [18–20], réseau d'accès radio du cloud (RAR-C) [21], réseau défini par logiciel (RDS) [22], virtualisation du réseau et des ressources, réseaux ultra-denses, massif-EMSM, etc. Parmi ces approches, dans cette thèse, nous nous focalisons sur le stockage pro-actif aux bords du réseau comme un moyen de faire face au surchargement de backhaul dans les PSBs, un fait spécialement crucial dans les déploiements de réseaux denses.

1.1.2 Caching: Bref historique et travaux liés

L'idée du caching remonte aux années soixante utilisé pour le design d'algorithmes pour systèmes d'exploitation [23]. D'après [23], la stratégie optimale de retrait de contenu lors de l'arrivée d'une nouvelle requête consiste à retirer de la mémoire, le contenu qui ne sera pas utilisé dans un futur proche. Au delà de ces travaux, il y a eu aussi des études poussées sur le web caching au cours des décennies précédentes dans le but d'améliorer l'évolutivité du world wide web et de décharger le réseau, en stockant des contenus dans des serveurs proxy et/ou des nœuds intermédiaires du réseau (voir [24] pour une littérature brève). De nombreux algorithmes de stockage pour les réseaux fournisseur de contenu (RFC) ont émergé récemment [25], permettant aux fournisseurs de services de réduire les délais d'accès aux contenus demandés par les utilisateurs. Conceptuellement, il existe aussi les réseau centré sur l'informations (RCIs) qui ont but de changer la manière d'accéder aux données sur internet, en nommant de façon unique les contenus et en les distribuant de façon intelligente à travers le réseau que d'avoir traditionnellement une seule source pour

---

[1]Le terme "PSB" sera utilisé de manière interchangeable avec "petite cellule" dans cette thèse.

l'accès aux contenus [26] (voir aussi [27] pour une récente revue). A côté de ces travaux, le problème du caching, comme un moyen de décharger les infrastructures de réseaux sans fil, est récent. Similairement à ce que nous présentons dans la thèse, la littérature croissante est principalement basée sur le caching aux bords du réseau. Une liste exhaustive de littérature récente est donnée dans [18–20, 28–176]. Dans la suite, nous résumons certains de ces travaux sur la base de leurs similarités et directions.

Caching pro-actif et estimation de la popularité des contenus

Le stockage pro-actif dans les RPCs avec parfaite connaissance de la popularité des contenus est donnée dans [39]. Dans [18], en exploitant la sensibilité au contexte, les réseaux sociaux, les communications AàA, les approches de stockage pro-actif pour les RPCs sont étudiées à la fois au niveau des PSBs et au niveau des TUs, montrant que plusieurs gains sont possibles sous des conditions numériques de départ. A cet égard, au lieu d'avoir une parfaite connaissance de la popularité des contenus, une estimation est faite via des outils d'apprentissage automatique (le FC en particulier), en exploitant les corrélations entre les comportements humains et leurs préférences. Ainsi, en ayant cette estimation, la décision de caching est appliquée plus efficacement, impliquant une meilleure performance en termes de satisfaction utilisateur et de décharge du réseau. D'un autre côté, un problème très connu dans la littérature du FC est le problème du "cold-start" qui peut arriver dans le cas d'une estimation avec très peu d'information. Par conséquent, pour booster l'estimation de la popularité de contenu, une approche abordée dans la littérature de l'apprentissage automatique est l'*apprentissage via transfert*, basé sur l'idée de transférer intelligemment de l'information d'un domaine cible vers le domaine source (voir [177] pour une revue). Inspiré de cela, une étude préliminaire sur l'apprentissage via transfert pour le caching dans les RPCs a été faite dans [35]. Malgré que cette approche a ses propres propres limites (transfert négatif par exemple), il a été montré dans [34] que l'estimation de la popularité des contenus via FC peut être améliorée par cette approche. D'autres analyses sont nécessaires pour combiner cette approche avec le caching pro-actif dans les RPCs. En plus, dans le contexte du caching pro-actif , des mesures de centralité pour le placement des contenus sont exploitées dans [43]. A ce titre, un simple traitement de dissémination de contenu est introduit et les résultats préliminaires de performance de la méthode de placement basé sur la centralité sont donnés via des simulations numériques. Comme alternative à ces approches, formulation basée sur la théorie des jeux du problème de caching pro-actif comme jeu de similarité plusieurs-à-plusieurs est introduite dans [41]. Un algorithme de similarité qui atteint un résultat stable est établi pour le problème du caching, montrant que le nombre de demandes satisfaites peut atteindre trois fois la satisfaction due à un caching aléatoire.

Algorithmes d'approximation

L'idée du *FemtoCaching* est donnée dans [19], dans laquelle les PSBs, dans une architecture avec faible débit de backhaul mais avec des unités à haute capacité de stockage,

sont chargées de délivrer les contenus aux utilisateurs à travers des transmissions à courte distance. L'analyse est faite à la fois pour les cas codés et non codés, montrant que l'assignation optimale des contenus est NP-hard, tandis que le cas codé est formulé comme problème convexe qui peut être réduit plus loin à un programme linéaire. Un greedy algorithme pour le cas codé et des résultats numériques sont donnés, montrant que le débit de la vidéo peut être amélioré par un facteur de $3-5$ dans des scénarios réels. Des extensions à ce travail, incluant le cas AàA sont données dans [85,88]. Alternativement, un problème de caching multicast est formulé dans [101] et un algorithme heuristique est proposé à ce titre, montrant que le coût de service peut être réduit jusqu'à 52% en comparaison au cas multicast agnostique.

Le placement optimal de contenus dans une PSB dans une architecture à capacité backhaul limitée est aussi étudié dans [107], montrant que le problème peut être réduit à un problème de knapsack lorsque la distribution de la popularité du contenus est connue. Supposant que la distribution de la popularité de contenus n'est pas connue à l'avance, le problème est formulé comme un problème bande multi armes (BMA) tel que la distribution de la popularité de contenus peut être apprise online et le placement de contenus peut être fait en conséquence. Trois différents algorithmes de mise en cache sont fournis pour montrer le compromis exploration vs exploitation de ce problème. Comme extension, une dérivation des bornes de regret ainsi que des analyses plus poussées des algorithmes à travers des simulations numériques sont présentées dans [108]. Additionnellement, un modèle de caching distribué avec plusieurs PSB est donné dans [109] dans le cadre du problème BMA, montrant que le caching codé peut surmonter le cas non codé. A côté des approches du BMA, une approche d'approximation basée sur le problème de facilité de location est donnée dans [102]. Aussi, pour une certaine demande de trafic, un algorithme de caching distribué basé sur l'algorithme des directions alternées (ADDA) est présenté dans [45].

Gains du caching codé

La formulation du problème de caching vue d'un point de vu théorie de l'information est étudiée par [65]. Là, les gains global et local de caching, dépendant de la mémoire disponible de chaque utilisateur et de la mémoire cumulative de tous les utilisateurs respectivement, sont déterminés sur la base d'un schème de caching codé. La structure proposée consiste en des phases de placement et de livraison: (i) donné pour un setup centralisé où le placement des contenus est géré par un serveur central, (ii) est essentiellement hors ligne puisqu'il n'y a pas de placement de contenu durant la phase de livraison, (iii) a été montré plus performant que les schèmes non codés conventionnels sous popularité de contenus uniforme , et (iv) fonctionne sur un unique lien partagé au lieu de plusieurs réseaux généraux. Ces résultats son ensuite étendus à des popularités de contenus non uniformes dans [66,69], des accès de stockage non uniformes dans [70], des tailles de stockage hétérogènes dans [79], des systèmes de caching on-line dans [68], des réseaux de caching hiérarchiques dans [67] et le cas multi-serveurs dans [76]. En plus, les bornes sont données dans [72, 75], le cas contenu sensible au délai est étudié [71] et les aspects

sécurité du point de vue théorie de l'information sont présentés dans [73]. Dans la même lignée de travaux, une approche décentralisée pour réseaux AàA avec caching aléatoire est étudiée dans [86, 89] en termes de lois de normalisation où un modèle de protocole canal similaire à celui dans [178] est pris en compte. Dans le même ordre d'idées, la performance du placement dans les techniques de caching décentralisées avec un schème de livraison codé est donné dans [87, 90], où le débit moyen est caractérisé par des demandes aléatoires avec des popularités suivant une distribution de Zipf.

Dans le contexte des systèmes de stockage distribués et de codage, les performances de simple caching, les codes de réplication et régénération sont étudiés dans un scénario AàA dans [111], lequel énonce de simples règles de choix de simples techniques de caching et de réplication , dérivés pour minimiser le coût total moyen en termes de consommation d'énergie. D'un autre coté, l'étude des fonctionnalités de la couche physique des systèmes de stockage distribués sans fil est faite dans [113] du point de vue codes de stockage spatial. Basé sur ce travail, un système de stockage sans fil qui communique à travers un canal à évanouissement est étudié dans [112] et un nouveau protocole de transmission est proposé basé sur les codes algébriques spatiaux, dans le but d'améliorer la fiabilité tout en gardant le décodage à un niveau acceptable. Il est montré que le protocole proposé est mieux performant que le simple protocole accès multiple par répartition en temps (AMRT) et entre dans la famille des optimaux compromis des gains diversité-multiplexage (CDM). Alternativement, une approche de codage sur réseau triangularisé pour placement du contenu à mettre en caching est présenté dans [128], dans lequel le placement de contenu non codé et les stratégies de codage en réseau triangularisé sont comparées via des résultats numériques. Additionnellement, un schème de mise en cache codé à travers le canal radio à évanouissement est présenté dans [80], tandis que [77] réduit le problème de caching en en un problème de codage source multi-terminal avec information sur l'état.

Design conjoint

En termes de conception, une approche à double échelle à la fois sur l'optimisation de puissance et le contrôle du caching est donnée dans [117] pour systèmes EMSM opportunistes et coopératifs utilisant le caching. Dans un premier temps, pour les échelles de temps courtes, les expressions complètes pour le contrôle de puissance sont calculées sur la base d'équations de Bellman approximées. Ensuite, pour les échelles de temps longues, le problème de caching est translaté en un problème d'optimisation stochastique convexe et un algorithme de sous-gradient stochastique est fourni comme solution. La solution proposée a été montrée être asymptotiquement optimale pour de large rapport signal sur bruit (RSB) tandis que sa comparaison avec des approches baseline sont faites via des simulations. Une autre solution pour des échelle de temps mixes pour les systèmes EMSM coopératifs est donnée dans [116]. Là, dans le but de minimiser la puissance émise sous contrainte de QdS, le pré-codage EMSM est optimisé pour une échelle de temps court et le contrôle de la mise en cache est fait dans une échelle de temps long terme. En addition à ces approches, l'optimisation conjointe du contrôle de cache et de la gestion du buffer playback pour la diffusion vidéo est donnée dans [118]. Le caching conjointement avec

le beamforming pour des réseaux à backhaul de capacité limitée est étudiée dans [134], et enfin le caching conjointement avec l'alignement d'interférence (AI) dans les systèmes EMSM à canal interférant sous backhaul à capacité limitée est présenté dans [32].

Mobilité

Les aspects de mobilité dans la délivrance des contenus codés sont analysés dans [100] sur la base de modèles de chaînes de Markov à temps discret. Dans le but de minimiser l'utilisation de la station de base principale dans ce modèle, un algorithme d'approximation distribué basé sur des inégalités de large déviation est introduit et des résultats expérimentaux sur des données réelles ont été données pour l'algorithme proposé. Un autre schème de caching qui exploite la mobilité de l'utilisateur est donné dans [129], lequel met en exergue l'influence des paramètres du système sur les gains de délai et est confirmé via différents niveaux de simulations. Les travaux dans [137] et [135] considèrent aussi l'impact de la mobilité dans les réseaux où le caching est effectué.

Consommation d'énergie

Les aspects de consommation d'énergie dans le caching à la fois en termes de consommation de puissance et d'efficacité énergétique sont traités dans [33]. Là, les stations de base avec caching actif sont distribuées selon un processus ponctuel de Poisson (PPP) homogène et l'optimisation est faite en utilisant un modèle de puissance détaillé. D'un autre coté, les aspects mettant l'accent sur l'énergie du caching pro-actif est mise en exergue dans [125], et un mécanisme de poussée effectif pour l'aspect énergétique des petites stations de base alimentées en puissance est proposée dans [124]. Aussi, le caching conjointement à l'activation des stations de base dans les "green" réseaux cellulaires est proposé dans [106].

Aspects de déploiement

Concernant les aspects de déploiement des PSBs avec caching activé et backhaul à capacité limitée, une étude est donnée dans [36]. Dans cette étude, les PSBs avec caching activé sont stochastiquement distribuées pour l'analyse en lieu et place des modèles "grids" traditionnels. Les expressions de la probabilité d'outage et du débit de livraison de contenu moyen sont dérivés en fonction du RSIB, de l'intensité des PSBs, du débit binaire du contenu cible, de la taille de la mise en cache et de la forme de la distribution de la popularité de contenu. Suivant les travaux dans [36], les résultats dans [132] montrent que stocker les contenus les plus populaires est bénéfique seulement dans des scénarios particuliers de déploiement. D'un autre côté, pour les systèmes de communication AàA avec caching activé, une autre approche stochastique est présentée dans [53], en se basant sur deux métriques de performance qui quantifient les fractions locale et globale des contenus demandés qui ont été servis. Aussi, une autre étude sur les noeuds avec caching activé, stochastiquement distribués est donnée dans [121]. Sachant que le coût est défini

comme une fonction de la distance, le coût moyen pour obtenir le contenu complet sous des stratégies d'allocation de contenus codés ou non codés est investi. Comme extension de [121], un compromis coût moyen de déploiement de mise en cache vs. le taux moyen de reconstruction de contenus stockés est analysé dans [122].

1.2 Plan de la thèse et contributions

Cette thèse contient trois parties. Dans la partie I, nous nous focalisons sur la modélisation et l'analyse de performance des réseaux cellulaires avec mise en cache activée en utilisant des outils de géométrie stochastique. En particulier :

Dans le chapitre 3 (Réseaux Cellulaires Niveaux Simples), nous considérons un modèle de réseau où les PSBs ont des capacités de caching comme moyens d'éviter le chargement de backhaul et de satisfaire en même temps les demandes des utilisateurs. Les PSBs sont stochastiquement distribuées dans le plan suivant un PPP, et servent les utilisateurs soit (i) en apportant les contenus depuis internet à travers un backhaul à débit fini ou (ii) en les servant des caches locaux. Nous dérivons les expressions closed-form de la probabilité d'outage et du débit moyen délivré en fonction du RSIB, de la densité de PSB, du débit binaire des contenus cibles, de la taille de stockage, la longueur du contenu et la popularité des contenus. Nous analysons ensuite l'impact des paramètres opérationnels clés sur la performance du système. Il est montré qu'une certaine probabilité d'outage peut être atteinte soit en augmentant le nombre de stations de base soit la taille totale de stockage.

Dans le chapitre 4 (Réseaux Cellulaires à Niveaux Multiples), nous considérons un réseau hétérogène à multiple niveaux où les noeuds à chaque niveau sont modélisés comme PPP homogène. En particulier, nous supposons un réseau hétérogène à quatre niveaux constitué de terminaux mobiles (utilisateurs), de petites cellules avec caching activé, des macro cellules et des routeurs centraux. Le réseau est sujet à des délais en downlink, backhaul et caches. Supposant que les petites stations de base sont en mesure de stocker du contenu à l'avance, nous caractérisons ensuite les délais moyens des utilisateurs connectés aux macro et petites stations de base. En particulier, en vue de modéliser les modèles d'accès spatio-temporels des utilisateurs, nous considérons des popularités de contenus *fixes*, des popularités *dépendant de la distance* et *dépendant de la charge*. En ayant une parfaite connaissance de ces popularités de contenus, nous employons ensuite trois différentes stratégies de caching qui reposent essentiellement sur la popularité de contenus et la randomisation. A la fin de ce chapitre, nous validons nos résultats à travers des simulations numériques et tirons plusieurs conclusions par rapport à ce type de réseaux hétérogènes.

Dans le Chapitre 5 (Réseaux Cellulaires Clusterisés), nous considérons un réseau à niveaux multiples qui consiste en des terminaux d'utilisateurs mobiles, de stations de base en clusters avec mise en cache activée, des macro cellules et des routeurs centraux. Le déploiement des petites stations de base suit deux différents processus de

clustering à savoir déploiements 1) *basés sur la couverture* et 2) *basés sur la capacité*. Dans la première topologie, les petites stations de base sont modélisées par un processus de "trou" Poisson qui permet d'être dans le trou couverture des macro cellules. Dans la seconde topologie, les petites stations de base avec caching activé sont modélisées par un processus de cluster Matérn, sont ainsi installées en hot-spots dans la région de couverture, permettant d'améliorer la capacité dans des scénarios de réseaux denses. Dans les deux topologies, nous caractérisons le débit moyen délivré des utilisateurs connectés aux macro et aux petites cellules en clusters. Cette métrique de débit moyen délivré capture des paramètres de couche physique tels que la taille de stockage, le débit binaire ciblé, le backhaul limité. Bien que les expressions de débit moyen délivré reposent sur des approximations (puisque ces processus de point sont raisonnablement difficiles à manier et des dépendances apparaissent à l'intérieur de chaque processus), nous montrons à travers des simulations numériques que plusieurs visions clés peuvent être tirées. Un modèle hiérarchique est aussi présenté dans le but de montrer les manipulations potentielles sur ces réseaux en clusters.

Dans la seconde partie de cette thèse à savoir Partie II, nous prenons une approche plus pratique pour examiner l'apport du caching. En particulier:

Dans le chapitre 6 (Caching Proactif), nous explorons le nouveau paradigme du caching *proactif* dans les RPCs qui vient suite aux récents développements dans le stockage, la sensibilité au contexte et les réseaux sociaux. En particulier, nous examinons deux études de cas qui exploitent la structure sociale et spatiale du réseau, où le caching proactif joue un rôle crucial. En premier lieu, dans le but d'éviter la congestion de backhaul, nous proposons un mécanisme par lequel les contenus/fichiers sont pro-activement mis en cache durant des périodes à moindre pic de demandes sur la base de la popularité des contenus et des corrélations entre les utilisateurs et des modèles d'accès aux contenus. En second lieu, en se basant sur les réseaux sociaux et les communications AàA, nous proposons une procédure exploitant la structure sociale du réseau en prédisant l'ensemble des utilisateurs influant, à mettre en cache de manière pro-active des contenus stratégiques et à les disséminer de leurs liens sociaux à travers les communications AàA. Avec cette approche nous montrons que des gains importants peuvent être obtenus, avec des décharges de backhaul et des hauts ratios d'utilisateurs satisfaits atteignant les 22% et 26%, respectivement.

Dans le chapitre 7 (Apprentissage par Transfert), nous proposons une nouvelle procédure de caching basée sur l'*apprentissage par transfert* effectué au niveau de chaque petite station de base. Ceci est fait en exploitant la riche information contextuelle (à savoir l'historique des vues des utilisateurs, des liens sociaux, etc.) extrait des interactions AàA, référé comme *domaine source*. Cette primo information est incorporée au dit *domaine cible* où le but est de mettre en cache de manière optimale des contenus stratégiques au niveau des petites stations de base en fonction du stockage, de la popularité de contenus estimée, de la charge de trafic, et de la capacité de backhaul. Il est montré que l'approche proposée surmonte les célèbres problèmes de sparsité de données et "cold start", en apportant des gains significatifs en termes de Qualité d'Expérience (QdE) des utilisateurs et de décharge du réseau avec des gains atteignant les 22% dans un setting constitué de quatre petites

stations de base.

Dans le chapitre 8 (Big Data pour Caching), Comme énoncé précédemment, les réseaux cellulaires mobiles deviennent de plus en plus complexes à gérer puisque les déploiements classiques/techniques d'optimisation et les solutions actuelles (à savoir densification des cellules, acquisition de plus de spectre,etc.) sont inefficaces en termes de coût et sont donc considérés comme des bouche-trou. Cela a conduit au développement de nouvelles approches qui prennent levier sur les récents développements en stockage/mémoire, la sensibilité au contexte, le edge/cloud computing, et entrent dans le cadre du *big data*. Par contre, le big data en lui-même est un autre phénomène complexe à aborder et est souvent lié aux célèbres 4V: vélocité, voracité, volume et variété. Dans ce chapitre, nous adressons les problèmes d'optimisation dans les réseaux cellulaires sans fil $5G$ via la notion de caching pro-actif dans les stations de base. En particulier, nous examinons les gains de caching pro-actif en termes de décharge de backhaul et de satisfactions des requêtes tout en abordant la question de l'estimation de la popularité avec la large masse de données existante. Dans le but d'estimer la popularité des contenus, nous collectons en premier lieu le trafic de données d'utilisateurs mobiles d'un opérateur de téléphonie Turque sur plusieurs de ses stations de base pendant un certain nombre d'observations. Ensuite, une analyse est effectuée localement sur une plateforme big data et les gains de caching proactif dans les stations de base sont étudiés via des simulations numériques. Il arrive que plusieurs gains sont possibles en fonction du niveau des informations disponibles et de la taille de stockage. Par exemple, avec 10% de contenus estimés et 15.4 Gigaoctets de taille de stockage (87% de la taille totale du catalogue), le caching pro-actif fournit 100% de satisfaction des requêtes et décharge de 98% le backhaul lorsque l'on considère 16 stations de base.

Enfin, la Partie III inclut nos conclusions et futurs travaux liés aux travaux présentés dans la thèse. Notons que chaque chapitre ci-dessus contient ses propres notations mathématiques.

1.3 Publications

Liste des publications au cours de cette thèse sont énumérés ci-dessous. Les résultats / détails qui sont entièrement ou partiellement fournies dans ce manuscrit sont marqués avec ∗.

Chapitres de Livres

[31]∗ E. Baştuğ, M. Bennis, and M. Debbah, "Proactive Caching in 5G Small Cell Networks", Towards 5G: Applications, Requirements and Candidate Technologies, Wiley, In Press (2015). **(Chapitres 1 et 2)**

Papiers de Journaux

[179]* E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "Modelling and Delay Analysis of Geographical Caching Methods in Cellular Networks", *(être soumis à)* IEEE Journal on Selected Areas in Communications, 2016. **(Chapitre 4)**

[180]* E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Modelling and Analysis of Geographical Caching Methods in Clustered Cellular Networks", *(être soumis à)* IEEE Transactions on Wireless Communications, 2016. **(Chapitre 5)**

[181] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah and A. Conte, "Energy Consumption Aspects of Cache-Enabled 5G Wireless Networks", *(être soumis à)* IEEE Transactions on Wireless Communications, 2016.

[182] F. Dilmi, E. Baştuğ, and M. Debbah, "FlexibleEarth3D : Un kit de visualisation pour les simulations des réseaux 5G", *(être soumis à une revue nationale)*, 2016.

[29]* E. Zeydan, E. Baştuğ, M. Bennis, M. Abdel Kader, A. Karatepe, A. Salih Er, and M. Debbah, "Big Data Caching for Networking: Moving from Cloud to Edge", IEEE Communications Magazine, Soumis (2015). **(Chapitre 8)**

[30]* E. Baştuğ, M. Bennis, E. Zeydan, M. Abdel Kader, A. Karatepe, A. Salih Er, and M. Debbah, "Big Data Meets Telcos: A Proactive Caching Perspective", Journal of Communications and Networks, Special Issue on Big Data Networking-Challenges and Applications, vol. 17, no. 6, pp. 549–558, December 2015. **(Chapitre 8)**

[36]* E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled Small Cell Networks: Modeling and Tradeoffs", EURASIP Journal on Wireless Communications and Networking, Special Issue on Technical Advances in the Design and Deployment of Future Heterogeneous Networks, vol. 2015, no. 1, pp. 41, 2015. **(Chapitre 3)**

[42]* K. Hamidouche, E. Baştuğ, M. Bennis, and M. Debbah, "Le caching proactif dans les réseaux cellulaires 5G", La Revue de l'Electricité et de l'Electronique (REE), vol. 2014-4, 2014. **(Chapitre 1)**

[18]* E. Baştuğ, M. Bennis, and M. Debbah, "Living on the Edge: The role of Proactive Caching in 5G Wireless Networks", IEEE Communications Magazine, vol 52, no 8, p. 82-89, 2014. **(Chapitres 2 et 6)**

[183] M. Maso, E. Baştuğ, L. S. Cardoso, M. Debbah, and Ö. Özdemir, "Reconfigurable Cognitive Transceiver for Opportunistic Networks", EURASIP Journal on Advances in Signal Processing, vol. 2014, no. 1, 2014.

Papiers de Conférences

[184]* E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "Deployment Cost and Delay of Caching in Two-Tiered Networks", *(être soumis à une conférence)*, 2016. **(Chapitre 4)**

[185]* E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Benefits of Edge Caching in Coverage and Capacity-aided Heteregeneous Networks", *(être soumis à une conférence)*, 2016. **(Chapitre 5)**

[186] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, ""Energy Consumption Aspects of Cache-Empowered Heterogeneous Networks: Optimization and Analysis", *(être soumis à une conférence)*, 2016.

[187] F. Dilmi, E. Baştuğ, and M. Debbah, "FlexibleEarth3D: A Visualization Toolkit for 5G Networks Simulations", *(être soumis à une conférence)*, 2016.

[28]* M. Abdel Kader, E. Baştuğ, M. Bennis, E. Zeydan, A. Karatepe, A. Salih Er, and M. Debbah, "Leveraging Big Data Analytics for Cache-Enabled Wireless Networks", IEEE Global Communications Conference (GLOBECOM) Workshop, San Diego, CA, USA, December 2015. **(Chapitre 8)**

[32] M. Deghel, E. Baştuğ, M. Assaad, and M. Debbah, "On the benefits of Edge Caching for MIMO Interference Alignment", IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'15), Stockholm, Sweden, June-July 2015.

[45] A. Abboud, E. Baştuğ, K. Hamidouche, and M. Debbah, "Distributed Caching in 5G Networks: An Alternating Direction Method of Multipliers Approach", IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'15), Stockholm, Sweden, June-July 2015.

[33] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, "Caching on the Edge: a Green Perspective for 5G Networks", IEEE International Conference on Communications (ICC'15), London, UK, June 2015.

[34]* E. Baştuğ, M. Bennis, and M. Debbah, "A Transfer Learning Approach for Cache-Enabled Wireless Networks", International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt'15), Mumbai, India, May 2015. **(Chapitre 7)**

[43] E. Baştuğ, K. Hamidouche, W. Saad, and M. Debbah, "Centrality-Based Caching for Mobile Wireless Networks", 1st KuVS Workshop on Anticipatory Networks, Stuttgart, Germany, September, 2014.

[35]* E. Baştuğ, M. Bennis, and M. Debbah, "Anticipatory Caching in Small Cell Networks: A Transfer Learning Chapitre", 1st KuVS Workshop on Anticipatory Networks, Stuttgart, Germany, September 2014. **(Chapitre 7)**

[37]* E. Baştuğ, M. Bennis, and M. Debbah, "Cache-enabled Small Cell Networks: Modeling and Tradeoffs", 11th International Symposium on Wireless Communication Systems (ISWCS), Barcelona, Spain, August 2014. **(Chapitre 3)**

[38]* E. Baştuğ, M. Bennis, and M. Debbah, "Social and Spatial Proactive Caching for Mobile Data Offloading", Small Cell and 5G Networks (SmallNets) workshop in conjuction with IEEE International Conference on Communications (ICC), Sydney, Australia, June 2014. **(Chapitre 6)**

[188] E. Baştuğ, A. Menafoglio, and T. Okhulkova, "Polynomial Chaos Expansion for an Efficient Uncertainty and Sensitivity Analysis of Complex Numerical Models", ESREL 2013, Amsterdam, Netherlands, September-October 2013.

[39] E. Baştuğ, JL. Guénégo, and M. Debbah, "Proactive Small Cell Networks", 20th International Conference on Telecommunications (ICT), Casablanca, Morocco, May 2013.

[40] E. Baştuğ, JL. Guénégo, and M. Debbah, "Cloud Storage for Small Cell Networks", IEEE International Conference on Cloud Networking (CloudNet), Paris, France, November 2012.

[189] B. Mawlawi, E. Baştuğ, C. Nerguizian, S. Azarian, and M. Debbah, "Non-Invasive Green Small Cell Network", 46th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California, USA, November 2012.

Papiers Blanc

[190] Claudio Cicconetti et. al., "5G radio network architecture", Radio Access and Spectrum - FP7 Future Internet Cluster, 2014.

Chapter 2

Introduction

2.1 Background and Motivation

The recent proliferation of smartphones has substantially enriched the mobile user experience, leading to a vast array of new wireless services, including multimedia streaming, web-browsing applications and socially-interconnected networks. This phenomenon has been further fueled by mobile video streaming, which currently accounts for almost 50% of mobile data traffic, with a projection of 500-fold increase over the next 10 years [1]. At the same time, social networking is already the second largest traffic volume contributor with a 15% average share [2]. This new phenomenon has urged mobile operators to redesign their current networks and seek more advanced and sophisticated techniques to increase coverage, boost network capacity, and cost-effectively bring contents closer to users.

A promising approach to meet these unprecedented traffic demands is via the deployment of small cell networks (SCNs) [3]. SCNs represent a novel networking paradigm based on the idea of deploying short-range, low-power, and low-cost small base stations (SBSs) underlaying the macrocellular network. To date, the vast majority of research works has been dealing with issues related to self-organization, inter-cell interference coordination (ICIC), traffic offloading, energy-efficiency, etc (see [4] and references therein). These studies were carried out under the existing *reactive* networking paradigm, in which users' traffic requests and flows must be served urgently upon their arrival or dropped causing outages. Because of this, the existing small cell networking paradigm falls short of solving peak traffic demands whose large-scale deployment hinges on expensive site acquisition, installation and backhaul costs. These shortcomings are set to become increasingly acute, due to the surging number of connected devices and the advent of ultra-dense networks, which will continue to strain current cellular network infrastructures. These key observations mandate a *novel* networking paradigm which goes beyond current heterogeneous small cell deployments leveraging the latest developments in storage, context-awareness, and social networking [5].

This novel network paradigm is *proactive* in the sense that the nodes at the edge of

the network (i.e., SBSs and user terminals (UTs)) predicts users' context information and pre-store intelligently strategic contents, in order offload the backaul and satisfy users' quality-of-service (QoS). This goes beyond the scope of traditional cellular networks which they have been designed assuming *dumb* UTs with limited storage and processing features. Nowadays, UTs are much more sophisticated than before, giving the opportunity to exploit their capabilities in conjunction with SCNs by storing the predicted contents at the network edge. This in turn yields significant gains in terms of network resources, minimizing operational and capital expenditures [4].

As stated before, recent results have shown that the human behaviour is correlated and predictable to a large extent [6]. Therefore, SBSs are assumed to be equipped with storage units and the low-speed backhaul is used for their broadband connections. Then, as we will show in the latter sections, proactively caching users' contents at SBSs alleviates the backhaul load and incurs higher users' satisfaction. The proactive caching procedure is based on the idea of storing the popular contents at the SBSs. To achieve this, the popularities of the contents have to be estimated. Using tools from machine learning and analysing the infrastructure logs (such as in [7]), a trove of hidden information about users' behaviour can be revealed. Analysing these traces falls into the *big data* phenomenon where collaborative filtering (CF) methods can be successfully applied for inference.

In the following of this section, we first give an overview of past, recent and future advancements in SCNs. Afterwards, we briefly discuss the history of caching and summarize recent efforts in the context of mobile cellular networks. The outline of the thesis will be given accordingly.

2.1.1   Small Cell Networks: Past, Present and Future Trends

Smartphones have exponentially increased the traffic load in current cellular networks showing no signs of slowing down [1, 2]. It is now well understood that a very effective way to increase network capacity is making cells smaller by reducing the distance to the users [4]. Indeed, cell densification has gone from the order of hundreds of square kilometers (back in the eighties) to a fraction of a square meter or less with the advent of hotspots. There has been recently a great interest to deploy relays, distributed antennas and small cellular access points (such as micro/pico/femto cells) in residential homes, subways, enterprises, and hot-spot areas. These network architectures, which are either operator-deployed or user-deployed are referred to heterogeneous networks (HetNets) or SCNs [3, 4]. By deploying additional network nodes within local-area range and making the network closer to end-users, small cells can significantly improve spatial reuse and coverage, boost capacity, and offload traffic more efficiently [3].

There exists a comprehensive literature on the topic of HetNets and SCNs tackling various aspects from interference management, cell association, stochastic network modeling, ICIC, energy-efficiency, self-organizing networkss (SONs), mobility management, long term evolution (LTE)/Wi-Fi interworking, among others (see [3] for a comprehensive survey). One of the key take-away drawn from these studies is that tight interference

coordination among macro and femto/picocell tiers is necessary for achieving cell split-ting gains. This hinges on the availability of low-latency and high-capacity backhauls [8]. Network modeling approaches based on stochastic geometric tools have shown reasonably-close performance gains (i.e., lower bound) in terms of system-wide and per-user capacities. Their attractive feature is attributed to the fact that unlike time-consuming system-level simulations, fundamental insights can be gleaned from these tools, some of which have been corroborated by industry field trials and observations from detailed simulations [9]. In parallel to that, mobility management has received significant attention from the wire-less industry, research community, and standardization bodies [10]. Unlike conventional homogeneous networks where UTs typically use the same set of handover parameters (i.e., hysteresis margin, time-to-trigger (TTT), etc.), using the same set of handover pa-rameters in HetNets for all cells and/or for all UTs may degrade mobility performance. This is because high-mobility macro UTs may run deep inside coverage areas of small cells before the TTT optimized for macro cells expires, thus incurring handover failure (due to degraded signal-to-interference-plus-noise ratio (SINR)) [11]. Decentralized inter-ference management/mitigation strategies in co-channel interference scenarios have also been studied in details, whereby small cells are able to self-organize based on local informa-tion and optimize their transmission strategies (i.e., power/frequency) based on minimum information exchange [12]. This leads to a number of tradeoffs in terms of faster/slower convergence at the cost of partial/full information. Carrier aggregation (CA) and its single/multiflow enhancements have also been investigated as a means of further boost-ing network capacity and per-user throughput, in which users may be served on several bands simultaneously [13]. Furthermore, with the increasing traffic asymmetry in the uplink (UL) as compared to the downlink (DL), novel cell association mechanisms and architectures are needed to cope with new types of inter-node interferences (DL-to-UL), thereby opening new avenues for research such as flexible DL/UL communication, massive multiple-input multiple-output (MIMO), device-to-device (D2D), full-duplexing, etc. [4] [14]. Finally, the topic of LTE and Wi-Fi coexistence has received tremendous attention due to the multi-mode capability of SBSs[1] and the possibility of using both licensed and unlicensed bands. Therein, dynamic load balancing and traffic steering mechanisms have been proposed leveraging the availability of Wi-Fi for best-effort services, traffic load, delay tolerance, etc [15].

While small cell densification is clearly the way to go, a number of technical challenges remain unsolved. Indeed, while small cell densification was shown to boost capacity, simply adding small cells may turn out to be energy-inefficient [16]. In addition, backhaul optimization and the optimal location of small cells represent one of the main limiting factors before a full rollout of small cells takes place. The importance of the backhaul is further underscored with the unabated proliferation of smartphones with the vast array of new wireless services (i.e., multimedia streaming, web-browsing applications, etc.). As a result, novel approaches to backhaul-aware small cell networking have been recently proposed in the literature [17] such as how to optimally decouple control and data planes to make cells more adaptive to traffic dynamics and network state while having a global

---

[1]The term "SBS" will be used interchangeably with "small cell" in this thesis.

view of the network, backhaul offloading via smart edge caching [18–20], cloud radio access network (C-RAN) [21], software defined networking (SDN) [22], resource/network virtualization, ultra-dense networks, massive MIMO, etc. Among these approaches, in this thesis, we focus on proactive edge caching as a way of dealing with backhaul offloading in SCNs, which is especially crucial in dense deployments.

2.1.2 Caching: A Brief History and Related Works

Indeed, the idea of caching goes back to the sixties in the context of algorithm design in operating systems [23]. According to [23], the optimal content removing strategy in the case of new content arrival is to evict the content from the memory which is not going to be requested in the near future. Beside this line of work, there has been also extensive studies on web caching schemes in the past decades, aiming to improve the scalability of world wide web and offloading the network, by caching contents in the proxy servers and/or intermediate nodes of the network (see [24] for a brief literature). Numerous caching algorithms for content delivery network (CDN) have emerged in the recent years [25], allowing content providers to reduce access delays to the requested contents. Conceptually, there exist also information-centric networks (ICNs) which aim to change the way of accessing the contents on the internet, by uniquely naming the contents and smartly distribute these across the network, rather than traditionally having one source for the content access [26] (see also [27] for a recent survey). Beside these line of works, the caching problem as a way of offloading the wireless communications infrastructure is recent. Similar to what we present in this thesis, the growing literature is mostly based on caching at the edge of network. An exhaustive list of recent literature is given in [18–20, 28–176]. In the following, we summarize some of these works based on their similarities and directions.

Proactive Caching and Content Popularity Estimation

Proactive caching in SCNs with perfect knowledge of the content popularity is given in [39]. In [18], exploiting context-awareness, social networks, D2D communications, the proactive caching approaches for SCNs are studied both at the SBSs and UTs, showing that several gains are possible under the given numerical setup. Therein, instead of perfect knowledge of the content popularity, an estimation is done via machine learning tools (the CF in particular), by exploiting correlations of human behaviour on their preferences. Thus, having such an estimation, the caching decision is applied more efficiently, yielding better performance in terms of the users' satisfaction and offloading of the network. On the other hand, a well-known problem in the CF literature is the cold-start problem which can occur in the case of estimation with very few amount of information. Therefore, to boost the content popularity estimation, one approach harnessing the machine learning literature is *transfer learning*, based on the idea of *smartly* transferring information from a target domain to a source domain (see [177] for a survey). Inspired from this, a preliminary study on transfer learning for caching in SCNs is conducted in [35]. Even though it has

naturally its own challenges (i.e., negative transfer), it is shown in [34] that the content popularity estimation via CF can be improved by this approach. Further investigations are needed to combine this approach with the proactive caching in SCNs. Additionally, in the context of proactive caching, the centrality measures for the content placement are exploited in [43]. Therein, a simple content dissemination process is introduced and the preliminary performance results of this centrality-based content placement methods are given via numerical simulations. Alternative to these proactive approaches, a game theoretical formulation of the proactive caching problem as a many-to-many matching game is introduced in [41]. A matching algorithm that reaches a pairwise stable outcome is provided for the caching problem, showing that the number of satisfied requests can be reach up to three times the satisfaction of a random caching policy.

Approximation Algorithms

The idea of *FemtoCaching* is given in [19], in which the SBSs (helpers) with low-rate backhaul but high storage units are in charge of delivering the contents to the users via short-range transmissions. The analysis is carried out both for coded and uncoded cases, showing that the optimum content assignment is NP-hard, whereas the coded case is formulated as a convex problem that further can be reduced to a linear program. A greedy algorithm for coded case and numerical results are provided, showing that video throughput can be improved by a factor $3-5$ in realistic settings. Extensions to this work, including D2D case, is given in [85, 88]. Alternatively, a multicast aware caching problem is formulated in [101] and a heuristic algorithm is provided for that purpose, showing that servicing cost can be reduced down to 52% compared to the multicast-agnostic case.

Optimal content placement in a SBS with limited backhaul capacity is also studied in [107], showing that the problem can be reduced to a knapsack problem when the content popularity distribution is known. Assuming that the content popularity distribution is not known in advance, the problem is formulated as a multi-armed band (MAB) problem so that the content popularity distribution can be learned online and content placement can be done. Three different caching algorithm is provided to show the exploration vs. exploitation trade-offs of this problem. As an extension, a derivation of regret bounds and more extensive analysis of the algorithms through numerical simulations are presented in [108]. Additionally, a distributed caching model with multiple SBS is given in [109] in the framework of MAB problem, showing that coded caching can outperform the uncoded case. Beside MAB approaches, an approximation framework based on the facility location problem is given in [102]. Also, for a given traffic demand, a distributed caching algorithm based on alternating direction method of multipliers (ADMM) is presented in [45].

Coded Caching Gains

Information-theoretic formulation of the caching problem is studied by [65]. Therein, local and global caching gains, which depend on the available memory of each user and cumulative memory of all users respectively, are derived based on a coded caching scheme. The

proposed scheme consists of placement and delivery phases (i) is given for a *centralized* setup where the content placement is handled by a central server, (ii) is essentially *offline* as there is no content placement during the delivery phase, (iii) is shown to outperform conventional uncoded schemes under *uniform content popularities*, and (iv) works in a single shared link instead of *more general networks*. These results are then extended to non-uniform content popularities in [66, 69], non-uniform cache access in [70], heterogeneous cache sizes in [79], online caching systems in [68], hierarchical caching networks in [67] and multi-server case in [76]. Moreover, the improved bounds are given in [72, 75], delay-sensitive content case is studied in [71] and the information-theoretic security aspects are shown in [73]. With similar line to these works, a decentralized approach for D2D networks with random coded caching is studied in [86, 89] in terms of scaling laws where a protocol channel model similar to [178] is taken into account. In the same vein, the performance of decentralized random caching placement with a coded delivery scheme is given in [87, 90], where the expected rate is characterized for random demands with Zipf popularity distribution.

In the context of distributed storage systems and coding, the performance of simple caching, replication and regenerating codes is studied in a D2D scenario in [111], in which a simple decision rule for choosing simple caching and replication is derived for minimizing the expected total cost in terms of energy consumption. On the other hand, the study of the physical layer functionality of wireless distributed storage systems is given in [113] from point of space-time storage codes. Based on that work, a wireless storage system that communicates over a fading channel is studied in [112] and a novel protocol for the transmission is proposed based on algebraic space-time codes, in order to improve the system reliability while keeping the decoding at a feasible level. It is shown that the proposed protocol performs better than the simple time-division multiple access (TDMA) protocol and falls behind the optimal diversity-multiplexing gain tradeoff (DMT). Alternatively, a triangular network coding approach for cache content placement is presented in [128], in which the uncoded content placement and the triangular network coding strategies are compared in a numerical setup. Additionally, a coded caching scheme over wireless fading channel is presented in [80], whereas [77] casts the caching problem into a multi-terminal source coding problem with side information.

Joint Designs

In terms of joint designs, a two time-scale joint optimization of power and cache control is given in [117] for cache-enabled opportunistic cooperative MIMO. First, for the short time scales, the closed-form expressions for the power control are derived from an approximated Bellman equation. Then, for the long time scales, the caching problem is translated into a convex stochastic optimization problem and a stochastic subgradient algorithm is provided for its solution. The proposed solution is shown to be asymptotically optimal for high signal-to-noise ratio (SNR) whereas its comparison with baseline approaches are done via simulations. Another mixed time-scale solution for cooperative MIMO is given in [116]. Therein, in order to minimize the transmit power under the QoS constraint, the

MIMO precoding is optimized in the short time scale and cache control is done in the long time scale. Additional to these approaches, the joint optimization of cache control and playback buffer management for video streaming is given in [118]. The joint caching and beamforming for backhaul limited caching networks is studied in [134], and finally the joint caching and interference alignment (IA) in MIMO interference channel under limited backhaul capacity is presented in [32].

Mobility

Mobility aspects of coded content delivery is analyzed in [100] based on a discrete-time Markov chain model. In order to minimize the probability of using the main base station in this model, a distributed approximation algorithm based on large deviation inequalities is introduced and numerical experiments on a real world dataset are conducted for the proposed algorithm. Another caching scheme that exploits users' mobility is given in [129], in which the influence of the system parameters on the delay gains are investigated via the system level simulations. The works in [137] and [135] also consider the impact of mobility in cache-enabled networks.

Energy Consumption

Energy consumption aspects of caching both in terms of area power consumption and energy efficiency are investigated in [33]. Therein, the cache-enabled base stations are distributed according to a homogeneous Poisson point process (PPP) and the optimization is done using a detailed power model. On the other hand, energy harvesting aspects of proactive caching is highlighted in [125], and an effective push mechanism for energy harvesting powered small-cell base stations is proposed in [124]. Also, a joint caching and base station activation for green cellular networks is proposed in [106].

Deployment Aspects

Concerning the deployment aspects of cache-enabled SBSs with limited backhaul, a study is given in [36]. In that study, the cache-enabled SBSs are stochastically distributed for the analysis rather than the traditional grid models. The expressions for the outage probability and average content delivery rate are derived as a function of the SINR, SBSs intensity, target content bitrate, cache size and shape of content popularity distribution. Following the work in [36], the results in [132] shows that storing the most popular contents is beneficial only in some particular deployment scenarios. On the other hand, for cache-enabled D2D communications, another stochastic framework is shown in [53], by relying on two performance metrics that quantify the local and global fraction of served content requests. Yet another study for the stochastically distributed cache-enabled nodes is given in [121]. Given the fact that the cost is defined as a function of distance, the expected cost of obtaining the complete content under coded as well as uncoded content allocation

strategies is investigated. As an extension to [121], the expected deployment cost of caches vs. the expected content retrieval from the caches is analyzed in [122].

2.2   Thesis Outline and Contributions

This thesis contains three parts. In Part I, we focus on the modeling and performance analysis of cache-enabled cellular networks by using tools from stochastic geometry. In particular:

In Chapter 3 (Single-Tier Cellular Networks), we consider a network model where SBSs have caching capabilities as a means to alleviate the backhaul load and satisfy users' demand. The SBSs are stochastically distributed over the plane according to a PPP, and serve their users either (i) by bringing the contents from the Internet through a finite rate backhaul or (ii) by serving them from the local caches. We derive closed-form expressions for the outage probability and the average delivery rate as a function of the SINR, SBS density, target content bitrate, storage size, content length and content popularity. We then analyze the impact of key operating parameters on the system performance. It is shown that a certain outage probability can be achieved either by increasing the number of base stations or the total storage size.

In Chapter 4 (Multi-Tier Cellular Networks), we consider a multi-tier heterogeneous network where nodes in each tier are modeled as a homogeneous PPP. In particular, we suppose a four-tier heterogeneous network consists of mobile terminals (users), cache-enabled small cells, macro cells and central routers. The network is subject to delays in downlink, backhaul and caches. Assuming that small base stations are able to cache contents in advance, we then characterize average delay of users connecting to macro and small base stations. In particular, in order to model the spatio-temporal access patterns of users, we consider *fixed* content popularity, *distance-dependent* and *load-dependent* content popularities. Having perfect knowledge of these type of content popularities, we then employ three different caching strategies which essentially rely on content-popularity and randomization. In the final part of this chapter, we shall validate our results via numerical simulations and draw several conclusions for such a heterogeneous network.

In Chapter 5 (Clustered Cellular Networks), we consider a multi-tier network which consists of mobile user terminals, clustered cache-enabled base stations, macro cells and central routers. The deployment of small cells follows two different clustering processes, namely 1) *coverage-aided* and 2) *capacity-aided* deployments. In the first topology, small base stations are modeled by a Poisson hole process which in turn allows them to be in the coverage hole of macro cells. In the second topology, cache-enabled small base stations are modeled by a Matérn cluster process, thus are installed in hot-spots of the area with the aim of improving capacity in dense user scenarios. In both topologies, we characterize average delivery rate of users connecting to macro and clustered small cells. This average delivery rate metric captures physical layer parameters such as signal-to-interference ratio (SIR) as well as storage size, target bitrate and limited backhaul. Even

though the expressions of average delivery rate rely on approximations (since these point processes are reasonably hard to handle and dependence occurs within each process), we shall show by numerical simulations that several key insights can be still gathered. A hierarchical model is also presented in order to show potential manipulations on this clustered networks.

In the second part of the thesis, namely in Part II, we take a more practical approach to investigate the gains of caching. In particular:

In Chapter 6 (Proactive Caching), we explore the novel paradigm of *proactive* caching in SCNs that leverages the latest developments in storage, context-awareness, and social networking. In particular, we examine two case studies which exploit the spatial and social structure of the network, where proactive caching plays a crucial role. Firstly, in order to alleviate backhaul congestion, we propose a mechanism whereby contents/files are proactively cached during off-peak demands based on content popularity and correlations among users and content access patterns. Secondly, leveraging social networks and D2D communications, we propose a procedure that exploits the social structure of the network by predicting the set of influential users to (proactively) cache strategic contents and disseminate them to their social ties via D2D communications. With this approach, we show that important gains can be obtained, with backhaul offloadings and higher ratios of satisfied users reaching up to 22% and 26%, respectively.

In Chapter 7 (Transfer Learning), we propose a novel *transfer learning*-based caching procedure carried out at each small cell base station. This is done by exploiting the rich contextual information (i.e., users' content viewing history, social ties, etc.) extracted from D2D interactions, referred to as *source domain*. This prior information is incorporated in the so-called *target domain* where the goal is to optimally cache strategic contents at the small cells as a function of storage, estimated content popularity, traffic load and backhaul capacity. It is shown that the proposed approach overcomes the notorious data sparsity and cold-start problems, yielding significant gains in terms of users' quality-of-experience (QoE) and backhaul offloading, with gains reaching up to 22% in a setting consisting of four small cell base stations.

In Chapter 8 (Big Data for Caching), As stated earlier, mobile cellular networks are becoming increasingly complex to manage while classical deployment/optimization techniques and current solutions (i.e., cell densification, acquiring more spectrum, etc.) are cost-ineffective and thus seen as stopgaps. This calls for development of novel approaches that leverage recent advances in storage/memory, context-awareness, edge/cloud computing, and falls into framework of *big data*. However, the big data by itself is yet another complex phenomenon to handle and comes with its notorious 4V: velocity, voracity, volume and variety. In this chapter, we address these issues in optimization of $5G$ wireless networks via the notion of proactive caching at the base stations. In particular, we investigate the gains of proactive caching in terms of backhaul offloadings and request satisfactions, while tackling the large-amount of available data for content popularity estimation. In order to estimate the content popularity, we first collect users' mobile traffic data from a Turkish telecom operator from several base stations in hours of time interval.

Then, an analysis is carried out locally on a big data platform and the gains of proactive caching at the base stations are investigated via numerical simulations. It turns out that several gains are possible depending on the level of available information and storage size. For instance, with 10% of content ratings and 15.4 Gbyte of storage size (87% of total catalog size), proactive caching achieves 100% of request satisfaction and offloads 98% of the backhaul when considering 16 base stations.

Finally, Part III includes our conclusions and future works related to work presented in this thesis. We note that each chapter above contains its own mathematical notation.

2.3 Publications

List of publications during the course of this PhD are listed below. The results/details which are either fully or partially provided in this manuscript are marked with ∗.

Book Chapters

[31]∗ E. Baştuğ, M. Bennis, and M. Debbah, "Proactive Caching in 5G Small Cell Networks", Towards 5G: Applications, Requirements and Candidate Technologies, Wiley, In Press (2015). **(Chapters 1 and 2)**

Journal Articles

[179]∗ E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "Modeling and Delay Analysis of Geographical Caching Methods in Cellular Networks", *(to be submitted to a special issue in)* IEEE Journal on Selected Areas in Communications, 2016. **(Chapter 4)**

[180]∗ E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Modeling and Analysis of Geographical Caching Methods in Clustered Cellular Networks", *(to be submitted to)* IEEE Transactions on Wireless Communications, 2016. **(Chapter 5)**

[181] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah and A. Conte, "Energy Consumption Aspects of Cache-Enabled 5G Wireless Networks", *(to be submitted to)* IEEE Transactions on Wireless Communications, 2016.

[182] F. Dilmi, E. Baştuğ, and M. Debbah, "FlexibleEarth3D : Un kit de visualisation pour les simulations des réseaux 5G", *(to be submitted to a national journal)*, 2016.

[29]∗ E. Zeydan, E. Baştuğ, M. Bennis, M. Abdel Kader, A. Karatepe, A. Salih Er, and M. Debbah, "Big Data Caching for Networking: Moving from Cloud to Edge", IEEE Communications Magazine, Submitted (2015). **(Chapter 8)**

[30]* E. Baştuğ, M. Bennis, E. Zeydan, M. Abdel Kader, A. Karatepe, A. Salih Er, and M. Debbah, "Big Data Meets Telcos: A Proactive Caching Perspective", Journal of Communications and Networks, Special Issue on Big Data Networking-Challenges and Applications, vol. 17, no. 6, pp. 549–558, December 2015. **(Chapter 8)**

[36]* E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled Small Cell Networks: Modeling and Tradeoffs", EURASIP Journal on Wireless Communications and Networking, Special Issue on Technical Advances in the Design and Deployment of Future Heterogeneous Networks, vol. 2015, no. 1, pp. 41, 2015. **(Chapter 3)**

[42] K. Hamidouche, E. Baştuğ, M. Bennis, and M. Debbah, "Le caching proactif dans les réseaux cellulaires 5G", La Revue de l'Electricité et de l'Electronique (REE), vol. 2014-4, 2014. **(Chapter 1)**

[18]* E. Baştuğ, M. Bennis, and M. Debbah, "Living on the Edge: The role of Proactive Caching in 5G Wireless Networks", IEEE Communications Magazine, vol 52, no 8, p. 82-89, 2014. **(Chapters 2 and 6)**

[183] M. Maso, E. Baştuğ, L. S. Cardoso, M. Debbah, and Ö. Özdemir, "Reconfigurable Cognitive Transceiver for Opportunistic Networks", EURASIP Journal on Advances in Signal Processing, vol. 2014, no. 1, 2014.

Conference Papers

[184]* E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "Deployment Cost and Delay of Caching in Two-Tiered Networks", *(to be submitted to a conference)*, 2016. **(Chapter 4)**

[185]* E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Benefits of Edge Caching in Coverage and Capacity-aided Heteregeneous Networks", *(to be submitted to a conference)*, 2016. **(Chapter 5)**

[186] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, ""Energy Consumption Aspects of Cache-Empowered Heterogeneous Networks: Optimization and Analysis", *(to be submitted to a conference)*, 2016.

[187] F. Dilmi, E. Baştuğ, and M. Debbah, "FlexibleEarth3D: A Visualization Toolkit for 5G Networks Simulations", *(to be submitted to a conference)*, 2016.

[28]* M. Abdel Kader, E. Baştuğ, M. Bennis, E. Zeydan, A. Karatepe, A. Salih Er, and M. Debbah, "Leveraging Big Data Analytics for Cache-Enabled Wireless Networks", IEEE Global Communications Conference (GLOBECOM) Workshop, San Diego, CA, USA, December 2015. **(Chapter 8)**

[32] M. Deghel, E. Baştuğ, M. Assaad, and M. Debbah, "On the benefits of Edge Caching for MIMO Interference Alignment", IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'15), Stockholm, Sweden, June-July 2015.

[45] A. Abboud, E. Baştuğ, K. Hamidouche, and M. Debbah, "Distributed Caching in 5G Networks: An Alternating Direction Method of Multipliers Approach", IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'15), Stockholm, Sweden, June-July 2015.

[33] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, "Caching on the Edge: a Green Perspective for 5G Networks", IEEE International Conference on Communications (ICC'15), London, UK, June 2015.

[34]* E. Baştuğ, M. Bennis, and M. Debbah, "A Transfer Learning Approach for Cache-Enabled Wireless Networks", International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt'15), Mumbai, India, May 2015. **(Chapter 7)**

[43] E. Baştuğ, K. Hamidouche, W. Saad, and M. Debbah, "Centrality-Based Caching for Mobile Wireless Networks", 1st KuVS Workshop on Anticipatory Networks, Stuttgart, Germany, September, 2014.

[35]* E. Baştuğ, M. Bennis, and M. Debbah, "Anticipatory Caching in Small Cell Networks: A Transfer Learning Approach", 1st KuVS Workshop on Anticipatory Networks, Stuttgart, Germany, September 2014. **(Chapter 7)**

[37]* E. Baştuğ, M. Bennis, and M. Debbah, "Cache-enabled Small Cell Networks: Modeling and Tradeoffs", 11th International Symposium on Wireless Communication Systems (ISWCS), Barcelona, Spain, August 2014. **(Chapter 3)**

[38]* E. Baştuğ, M. Bennis, and M. Debbah, "Social and Spatial Proactive Caching for Mobile Data Offloading", Small Cell and 5G Networks (SmallNets) workshop in conjuction with IEEE International Conference on Communications (ICC), Sydney, Australia, June 2014. **(Chapter 6)**

[188] E. Baştuğ, A. Menafoglio, and T. Okhulkova, "Polynomial Chaos Expansion for an Efficient Uncertainty and Sensitivity Analysis of Complex Numerical Models", ESREL 2013, Amsterdam, Netherlands, September-October 2013.

[39] E. Baştuğ, JL. Guénégo, and M. Debbah, "Proactive Small Cell Networks", 20th International Conference on Telecommunications (ICT), Casablanca, Morocco, May 2013.

[40] E. Baştuğ, JL. Guénégo, and M. Debbah, "Cloud Storage for Small Cell Networks", IEEE International Conference on Cloud Networking (CloudNet), Paris, France, November 2012.

[189] B. Mawlawi, E. Baştuğ, C. Nerguizian, S. Azarian, and M. Debbah, "Non-Invasive Green Small Cell Network", 46th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, California, USA, November 2012.

White Papers

[190] Claudio Cicconetti et. al., "5G radio network architecture", Radio Access and Spectrum - FP7 Future Internet Cluster, 2014.

Part I

Modeling and Performance Analysis

Chapter 3

Single-Tier Cellular Networks

3.1 Overview

The main contribution of this chapter is to formulate the caching problem in a scenario where stochastically distributed SBSs are equipped with storage units but have the limited backhaul capacity. In particular, we build on a tractable system model and define its performance metrics (outage probability and average delivery rate) as functions of SINR, number of SBSs, target content bitrate, storage size, content length and content popularity distribution. By coupling the caching problem with physical layer (PHY) in this way and relying on recent results from [191], we show that a certain outage probability can be achieved either by 1) increasing number of SBSs while the total storage size budged is fixed, or 2) increasing the total storage size while the number of SBSs is fixed. To the best of our knowledge, our work differs from the previous works in terms of studying deployment aspects of cache-enabled SBSs. Similar line of work in terms of analysis with stochastic geometry tools can be found in [53, 121]. However, the system model and performance metrics are different than what is studied here.

The rest of this chapter is structured as follows. We describe our system model in Section 3.2. The performance metrics and main results are given in Section 3.3. In the same section, much simpler expressions are obtained by making specific assumptions on the system model. We validate these results via numerical simulations in Section 3.4 and discuss the impact of parameters on the performance metrics. Then, a tradeoff between the number of deployed SBSs and total storage size is given in Section 3.5. Finally, our conclusions and future perspectives are given in Section 3.6.

3.2 System Model

The cellular network under consideration consists of SBSs, whose locations are modeled according to a PPP Φ with density λ. The broadband connection to these SBSs is provided by a central scheduler (CS) via wired backhaul links. We assume that the broadband

connection is finite and fixed, thus the backhaul link capacity of each SBS is a decreasing function of λ. This in practice means that deploying more SBSs in a certain area yields sharing the total broadband capacity among backhaul links. We will define this function more precisely in the next sections.

We suppose that every SBS has a storage unit with capacity S nats (1 bit $= \ln(2) = 0.693$ nats), thus they cache users' most popular contents given in a catalog. The size of each content in the catalog has a length of L nats and bitrate requirement of T nats/sec/Hz. We note that the assumption on content length is for ease of analysis. Alternatively, the contents in the catalog can be divided into chunks with the same length. The content popularity distribution of this catalog is a right continuous and monotonically decreasing probability distribution function (PDF), denoted as $f_{\mathrm{pop}}(f, \gamma)$. The parameter f here corresponds to a point in the support of a content and γ is the shape parameter of the distribution. We assume that this distribution is identical among all users.

Every user equipped with a mobile user terminal is associated with the nearest SBS, where its location falls into a point in a Poisson-Voronoi tessellation on the plane. In this model, we only consider the downlink transmission and overhead due to the content requests of users via uplink is neglected. In the downlink transmission, a tagged SBS transmits with the constant transmit power $1/\mu$ Watts, and the standard unbounded power-law pathloss propagation model with exponent $\alpha > 2$ is used for the environment. The tagged SBS and tagged user experience Rayleigh fading with mean 1. Hence, the received power at the tagged user, located r-meters far away from its tagged SBS, is given by $hr^{-\alpha}$. The random variable h here follows an Exponential distribution with mean $1/\mu$, represented as $h \sim \mathrm{Exponential}(\mu)$.

Once users are associated with their closest SBSs, we assume that they request some contents (or chunks) randomly according to the content popularity distribution $f_{\mathrm{pop}}(f, \gamma)$. When requests reach to the SBSs via uplink, the users are served immediately, either getting the content from the Internet via backhaul or being served from the local cache, depending on the availability of the content therein. If a requested content is available in the local cache of the SBS, a *cache hit* event occurs, otherwise a *cache miss* event is said to be occurred. According to what we have explained so far, a sketch of the network model is given in Figure 3.1.

In general, the performance of our system depends on several factors. To meet the QoE requirements, the downlink rate provided to the requested user has to be equal or higher than the content bitrate T, so that the user does not observe any interruption during its experience. Although this requirement can be achieved in the downlink, yet another bottleneck can be the rate of the backhaul in case of cache misses. In the following, we define our performance metrics which take into account the aforementioned situations. We then present our main results in the same section.
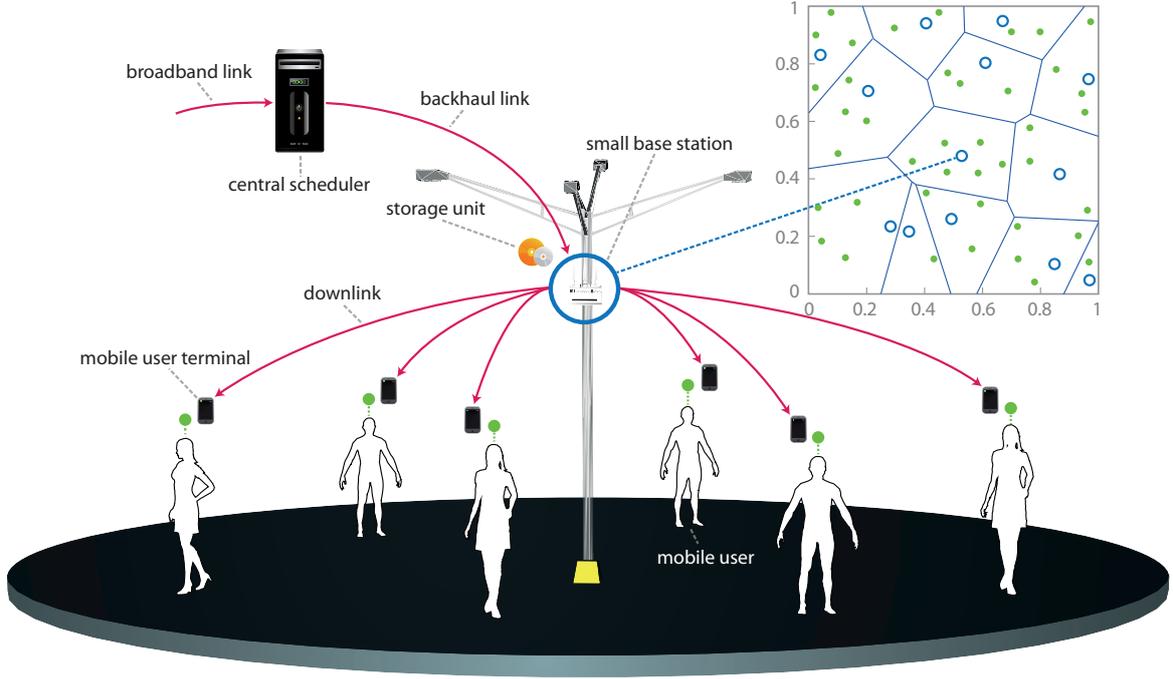
Figure 3.1: An illustration of the considered network model. The top right side of the figure shows a snapshot of PPP per unit area where the SBSs are randomly located. A closer look to communication structure of a cache-enabled SBS is shown in the main figure.

3.3 Performance Metrics and Main Results

Performance metrics of interest in our system model are the *outage probability* and *average delivery rate*. We start by defining these metrics for the downlink. From now on, without loss of generality, we refer to the user o as *typical* user, which is located at the origin on the plane.

We know that the downlink rate depends on the SINR. The SINR of user o which is located at a random distance r far away from its SBS b_o is given by:

$$\text{SINR} \triangleq \frac{hr^{-\alpha}}{\sigma^2 + I_r}, \tag{3.1}$$

where

$$I_r \triangleq \sum_{i \in \Phi/b_o} g_i R_i^{-\alpha}, \tag{3.2}$$

is the total interference experienced from all other SBSs except the connected SBS b_o. Assume that the *success probability* is the probability of the downlink rate exceeding the content bitrate T and the probability of requested content being in the local cache. Then, the outage probability can be given as the complementary of the success probability as

follows:

$$p_{\text{out}}(\lambda, T, \alpha, S, L, \gamma) \triangleq 1 - \underbrace{\mathbb{P}\Big[\ln(1 + \text{SINR}) > T, f_o \in \Delta_{b_o}\Big]}_{\text{success probability}}, \qquad (3.3)$$

where f_o is the requested content by the typical user, and Δ_{b_o} is the local cache of serving SBS b_o. Indeed, such a definition of the outage probability comes from a simple observation. Ideally, if a requested content is in the cache of the serving SBS (thus the limited backhaul is not used) and if the downlink rate is higher than the content bitrate T (thus the user does not observe any interruption during the playback of the content), we then expect the outage probability to be close to zero. Given this explanation and the assumptions made in the previous section, we state the following theorem for outage probability.

Theorem 1 (Outage probability). *The typical user has an outage probability from its tagged base station which can be expressed as:*

$$p_{\text{out}}(\lambda, T, \alpha, S, L, \gamma) = 1 - \pi\lambda \int_0^\infty \int_0^{S/L} e^{-\pi\lambda v\beta(T,\alpha) - \mu(e^T - 1)\sigma^2 v^{\alpha/2}} f_{\text{pop}}(f, \gamma)\mathrm{d}f\mathrm{d}v, \quad (3.4)$$

where $\beta(T, \alpha)$ is given by:

$$\beta(T, \alpha) = \frac{2\left(\mu(e^T - 1)\right)}{\alpha} \mathbb{E}_g\left[g^{\frac{2}{\alpha}}\left(\Gamma\left(-\frac{2}{\alpha}, \mu\left(e^T - 1\right)g\right) - \Gamma\left(-\frac{2}{\alpha}\right)\right)\right], \qquad (3.5)$$

where $\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}\mathrm{d}t$ is the upper incomplete Gamma function and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ is the Gamma function.

Proof. The proof is provided in Appendix A.1. $\qquad\qquad\square$

Yet another useful metric in our system model is the delivery rate, which we define as follows:

$$\tau \triangleq \begin{cases} T, & \text{if } \ln(1 + \text{SINR}) > T \text{ and } f_o \in \Delta_{b_o}, \\ C(\lambda), & \text{if } \ln(1 + \text{SINR}) > T \text{ and } f_o \notin \Delta_{b_o}, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{nats/sec/Hz} \qquad (3.6)$$

where $C(\lambda)$ is the backhaul capacity provided to the SBS for single frequency in the downlink.[1] The definition above can be explained as follows. If the downlink rate is higher than the threshold T (namely the bitrate of the requested content) and the requested content is available in the local cache, the rate T is dedicated to the user by the tagged SBS, which in turn is sufficient for QoE. On the other hand, if the downlink rate is higher than T but the requested content does not exist in the local cache of the tagged SBS, the delivery rate will be limited by the backhaul link capacity $C(\lambda)$, for which we assume that $C(\lambda) < T$. Given this definition for the delivery rate, we state the following theorem.

---

[1]Without loss of generality, more realistic values of delivery rate can be obtained by making a proper SINR gap approximation and considering the total wireless bandwidth instead of 1 Hz.

Theorem 2 (Average delivery rate). *The typical user has an average delivery rate from its tagged base station which can be expressed as:*

$$\bar{\tau}(\lambda, T, \alpha, S, L, \gamma) = \pi\lambda \int_0^\infty e^{-\pi\lambda v \beta(T,\alpha) - \mu(e^T - 1)\sigma^2 v^{\alpha/2}} \mathrm{d}v \times$$

$$\left(C(\lambda) + (T - C(\lambda)) \int_0^{S/L} f_{\mathrm{pop}}(f, \gamma)\mathrm{d}f \right), \quad (3.7)$$

where $\beta(T, \alpha)$ has the same definition as in Theorem 1.

Proof. The proof is deferred to Appendix A.2. □

What we provided above are the general results. The exact values of outage probability and average delivery rate can be obtained by specifying the distribution of the interference, the backhaul link capacity $C(\lambda)$ and the content popularity distribution $f_{\mathrm{pop}}(f, \gamma)$. If this treatment does not yield closed form expressions, numerical integration can be done as a last resort for evaluating the functions. In the next section, as an example, we derive special cases of these results after some specific assumptions, which in turn yield much simpler expressions.

3.3.1 Special Cases

Assumption 1. *The following assumptions are given for the the system model:*

1. *The noise power σ^2 is higher than 0, and the pathloss component α is 4.*

2. *Interference is Rayleigh fading, which in turn $g_i \sim \mathrm{Exponential}(\mu)$.*

3. *The capacity of backhaul links is given by:*

$$C(\lambda) \triangleq \frac{C_1}{\lambda} + C_2, \quad (3.8)$$

where $C_1 > 0$ and $C_2 \geq 0$ are some arbitrary coefficients such that $C(\lambda) < T$ holds.

4. *The content popularity distribution of users is characterized by a power law [192] such as:*

$$f_{\mathrm{pop}}(f, \gamma) \triangleq \begin{cases} (\gamma - 1) f^{-\gamma}, & f \geq 1, \\ 0, & f < 1, \end{cases} \quad (3.9)$$

where $\gamma > 1$ is the shape parameter of the distribution.

The assumption $C(\lambda) < T$ comes from the observation that the high-speed fiber-optic backhaul links might be very costly in densely deployed SBSs scenarios. Therefore, we assume that $C(\lambda)$ is lower than the bitrate of content. On the other hand, we characterize the content popularity distribution with a power law. Indeed, this comes from the observation that many real world phenomena can be characterized by power laws (i.e. distribution of contents in web proxies, distribution of word counts in natural languages) [192]. According to our system model and the specific assumptions made in Assumption 1, we state the following results.

Proposition 1 (Outage probability). *The typical user has an outage probability from its tagged base station which can be expressed as:*

$$p_{\text{out}}(\lambda, T, 4, S, L, \gamma) = 1 - \frac{\pi^{\frac{3}{2}}\lambda}{\sqrt{\frac{e^T-1}{\text{SNR}}}}\exp\left(\frac{(\lambda\pi(1+\rho(T,4)))^2}{4(e^T-1)/\text{SNR}}\right) \times$$

$$Q\left(\frac{\lambda\pi(1+\rho(T,4))}{\sqrt{2(e^T-1)/\text{SNR}}}\right)\left(1-\left(\frac{L}{L+S}\right)^{\gamma-1}\right), \quad (3.10)$$

where $\rho(T,4) = \sqrt{e^T-1}\left(\frac{\pi}{2}-\arctan\left(\frac{1}{\sqrt{e^T-1}}\right)\right)$ *and the standard Gaussian tail probability is given as* $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-y^2/2}\mathrm{d}y.$

Proof. The proof is given in Appendix A.3. □

Proposition 2 (Average delivery rate). *The typical user has an average delivery rate from its tagged base station which can be expressed as:*

$$\bar{\tau}(\lambda, T, 4, S, L, \gamma) = \frac{\pi^{\frac{3}{2}}\lambda}{\sqrt{\frac{e^T-1}{\text{SNR}}}}\exp\left(\frac{(\lambda\pi(1+\rho(T,4)))^2}{4(e^T-1)/\text{SNR}}\right) \times$$

$$Q\left(\frac{\lambda\pi(1+\rho(T,4))}{\sqrt{2(e^T-1)/\text{SNR}}}\right)\left(T+\left(\frac{C_1}{\lambda}+C_2-T\right)\left(\frac{L}{L+S}\right)^{\gamma-1}\right), \quad (3.11)$$

where $\rho(T,4)$ *and* $Q(x)$ *has the same definition as in Proposition 1.*

Proof. The proof is given in Appendix A.4. □

The expressions obtained for special cases are cumbersome but fairly easy to compute and does not require any integration. Note that $Q(x)$ function given in the expressions is a well-known function and can be computed by using lookup tables or standard numerical packages.

3.4   Validation of the Proposed Model

So far we have provided the results for outage probability and average delivery rate. In this section, we validate these results via Monte Carlo simulations. The numerical results shown here are obtained by averaging out over 1000 realizations. In each realization, the SBSs are distributed according to a PPP. The content requests, signal and interfering powers of the typical user are drawn randomly according to the corresponding probability distributions. The outage probability and average delivery rate are then calculated by considering SINR and cache hit statistics. We note that all simulation curves match the theoretical ones. However, a slight mismatch is observed due to the fact that more precise discretization of continuous variables is avoided for affordable simulation times. As alluded to previously, the target content bit rate as well as average delivery rate are in units of nats/sec/Hz. On the other hand, the storage size and content lengths are in units of nats.

3.4.1   Impact of storage size

The storage size of SBSs is one critical parameter in our system model. The effect of the storage size on the outage probability and the average delivery rate is plotted in Figures 3.2 and 3.3, respectively. Each curve represents a different value of target content bit rate. We observe that the outage probability reduces whereas the average delivery rate increases, as we increase the storage size. Such behaviour, observed both in theoretical and simulation curves, confirms our initial intuition.

3.4.2   Impact of the number of base stations

The evolution of outage probability with respect to the number of base stations is depicted in Figure 3.4. As the base station density increases, the outage probability decreases. This decrement in outage probability can be improved further by increasing the storage size of SBSs.

3.4.3   Impact of target content bitrate

Yet another important parameter in our setup is the target content bitrate T. Figure 3.5 shows its impact on the outage probability for different values of storage size. Clearly, increasing the target content bitrate results in higher outage probability. However, this performance reduction can be compensated by increasing the storage size of SBSs. The impact of storage size reduces, as T increases.
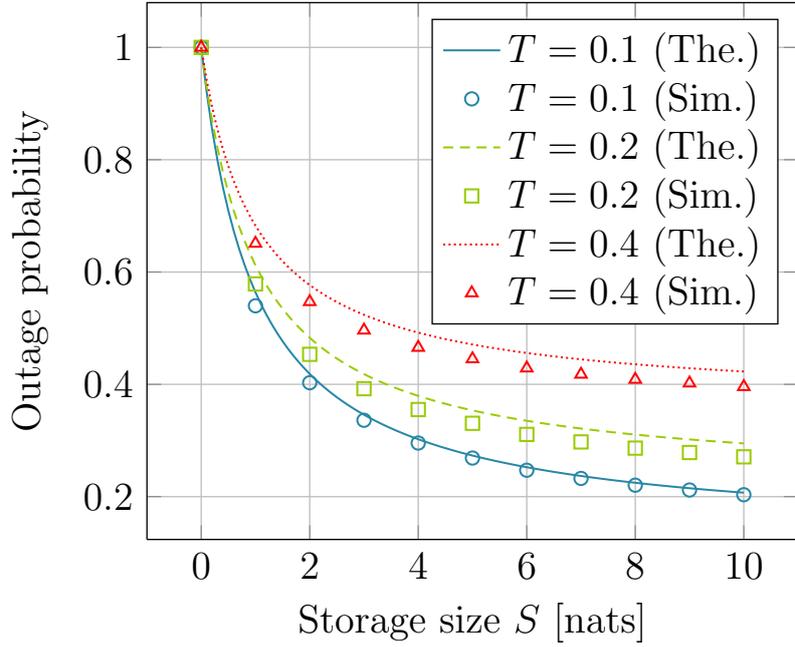
Figure 3.2: The evolution of outage probability with respect to the storage size. SNR = 10 dB, $\lambda = 0.2$, $\gamma = 2$, $L = 1$ nats, $\alpha = 4$, $C_1 = 0.0005$, $C_2 = 0$.
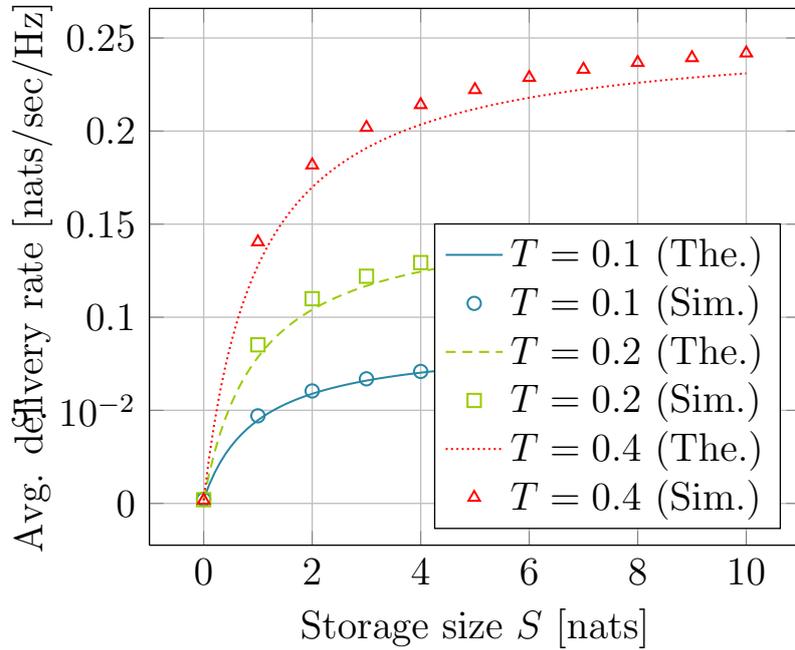
Figure 3.3: The evolution of average delivery rate with respect to the storage size. SNR = 10 dB, $\lambda = 0.2$, $\gamma = 2$, $L = 1$ nats, $\alpha = 4$, $C_1 = 0.0005$, $C_2 = 0$.
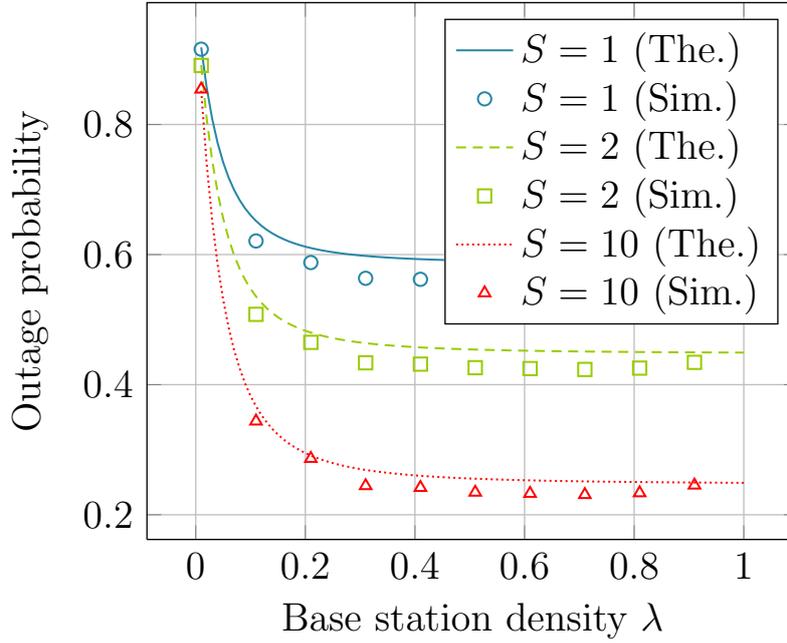
Figure 3.4: The evolution of outage probability with respect to the base station density. SNR = 10 dB, $T = 0.2$, $\gamma = 2$, $L = 1$ nats, $\alpha = 4$, $C_1 = 0.0005$, $C_2 = 0$.
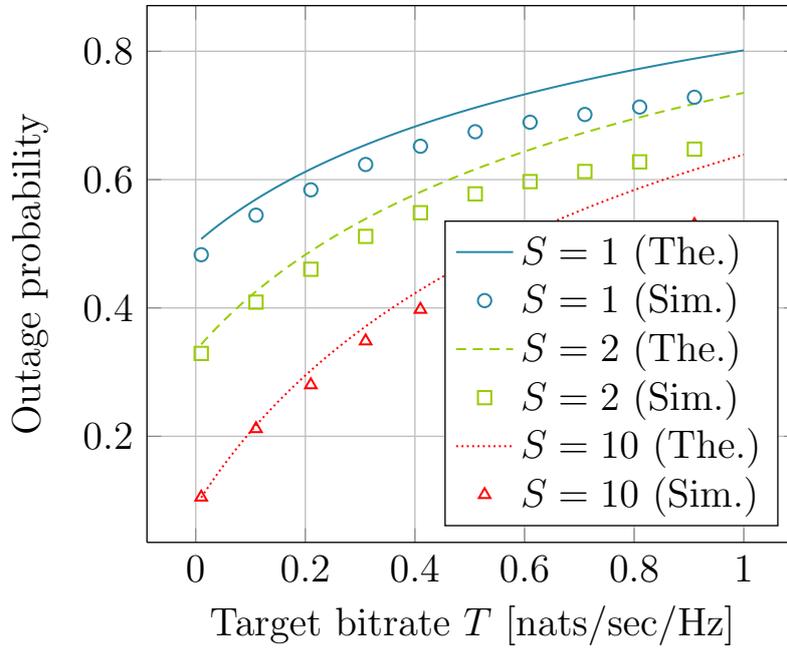
Figure 3.5: The evolution of outage probability with respect to the target file bitrate. SNR = 10 dB, $\lambda = 0.2$, $\gamma = 2$, $L = 1$ nats, $\alpha = 4$, $C_1 = 0.0005$, $C_2 = 0$.

3.4.4 Impact of content popularity shape

Another crucial parameter in our setup is the shape of the content popularity distribution, parameterized by γ. The impact of the parameter γ on the outage probability, for different storage sizes, is given in Figure 3.6. Generally, a higher value of γ means that only a small portion of contents is highly popular compared to the rest of the contents. On the contrary, lower values of γ correspond to a more uniform behavior on the popularity distribution. Therefore, as γ increases, the outage probability reduces due to reduced requirement in terms of storage size. However, in very low and high values of γ, the impact on the outage probability is not high compared to the intermediate values.
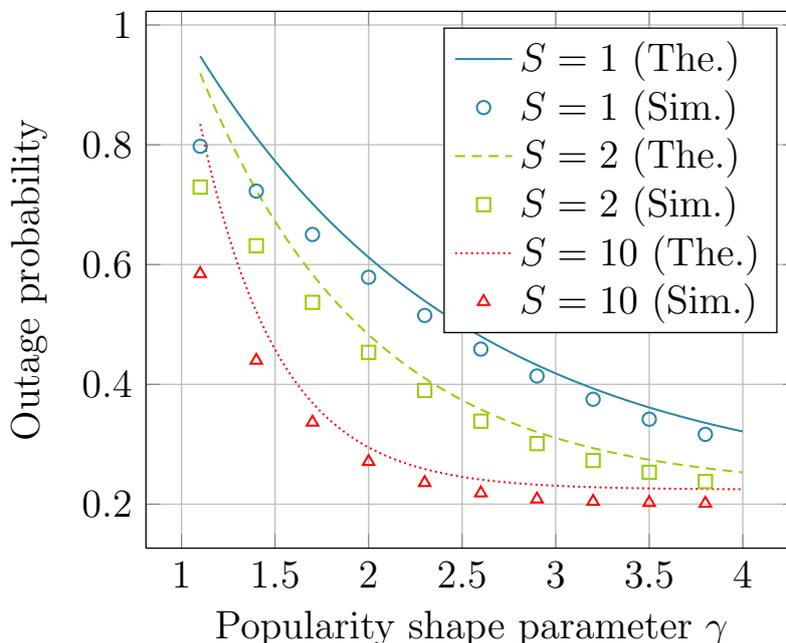
Figure 3.6: The evolution of outage probability with respect to the popularity shape parameter γ. SNR = 10 dB, $\lambda = 0.2$, $\gamma = 2$, $L = 1$ nats, $\alpha = 4$, $C_1 = 0.0005$, $C_2 = 0$.

3.5 David vs. Goliath: More SBSs with less storage or less SBSs with more storage?

In the previous section, we have validated our results via numerical simulations and discussed the impact of several parameters on the outage probability and average delivery rate. On top of those, we are now interested in finding a tradeoff between the SBS density and the total storage size for a fixed set of parameters. We start by making an analogy with well-known David and Goliath story to examine the tradeoff between the SBS density

and total storage size.[2] More precisely, we aim to answer the following question: Should we increase storage size of current SBSs (**David**) or deploy more SBSs with less storage (**Goliath**) in order to achieve a certain success probability? The answer is indeed useful for the realization of such a scenario. Putting more SBSs in a given area may be not desirable due to increased deployment and operation costs (**Evil**). Therefore, increasing the storage size of already deployed SBSs may incur less cost (**Good**). To characterize this tradeoff, we first define the optimal region as follows:

Definition 1 (Optimal region). *An outage probability $p^\dagger$ is said to be achievable if there exist some parameters $\lambda, T, \alpha, S, L, \gamma$ satisfying the following condition:*

$$p_{\mathrm{out}}(\lambda, T, \alpha, S, L, \gamma) \leq p^\dagger.$$

The set of all achievable $p^\dagger$ forms the optimal region.

The optimal region can be tightened by restricting parameters $\lambda, T, \alpha, S, L, \gamma$ to some intervals. A detailed analysis on this is left for future work. Hereafter, we restrict ourselves to find the optimal SBS density for a fixed set of parameters. In such a case, optimal SBS density can be readily obtained by plugging these fixed parameters into p_{out} and solving the equation either analytically or numerically (i.e. bisection method [194]). In the following, we obtain a tradeoff curve between the SBSs density and total storage size, by solving these equations systematically in the form of optimization problem.

Definition 2 (SBS density vs. total storage size tradeoff). *Define the average total storage as $S_{\mathrm{total}} = \lambda S$, and fix T, α, L and γ to some values in the optimal region given in Definition 1. Denote also $\lambda^\star$ as the optimal SBS density for a given S_{total}. Then, $\lambda^\star$ is obtained by solving the following optimization problem:*

$$\begin{aligned}
&\underset{\lambda}{\text{minimize}} && \lambda && (3.12)\\
&\text{subject to} && p_{\mathrm{out}}(\lambda, T, \alpha, S_{\mathrm{total}}/\lambda, L, \gamma) \leq p^\dagger. && (3.12a)
\end{aligned}$$

The set of all achievable pairs $(\lambda^\star, S_{\mathrm{total}})$ characterize a tradeoff between the SBS density and total storage size.

Figures 3.7 and 3.8 show two different configurations of the tradeoff. In these plots, to achieve a certain outage probability (i.e. $p^\dagger = 0.3$), we see that it is sufficient to decrease the number of SBSs by increasing the total storage size. Alternatively, the total storage size can be decreased by increasing the number of SBSs. Moreover, for different values of parameter of interest (i.e. $T \in \{0.1, 0.2\}$ or $L \in \{1, 2\}$), there is also a scaling and shifting in this tradeoff. Regardless of this scaling and shifting, we see that David wins victory against Goliath.

---

[2]David vs. Goliath refers to the underlying resource sharing problem which arises in a variety of scenarios including massive MIMO vs. Small Cells [193].
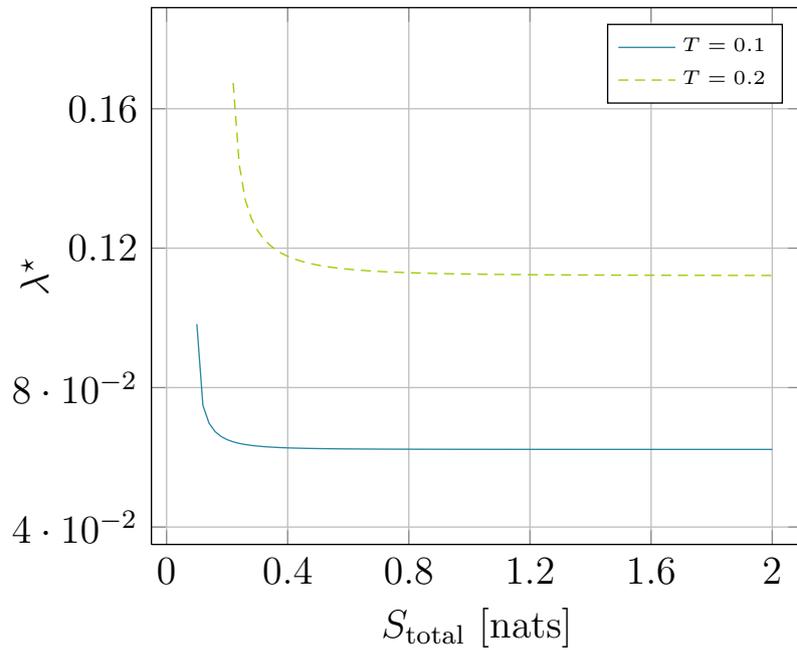
Figure 3.7: The trade-off between SBSs density and total storage size for different file target bitrates. SNR = 10 dB, $\alpha = 4$, $L = 1$ nats, $\gamma = 3$ and $p^\dagger = 0.3$.
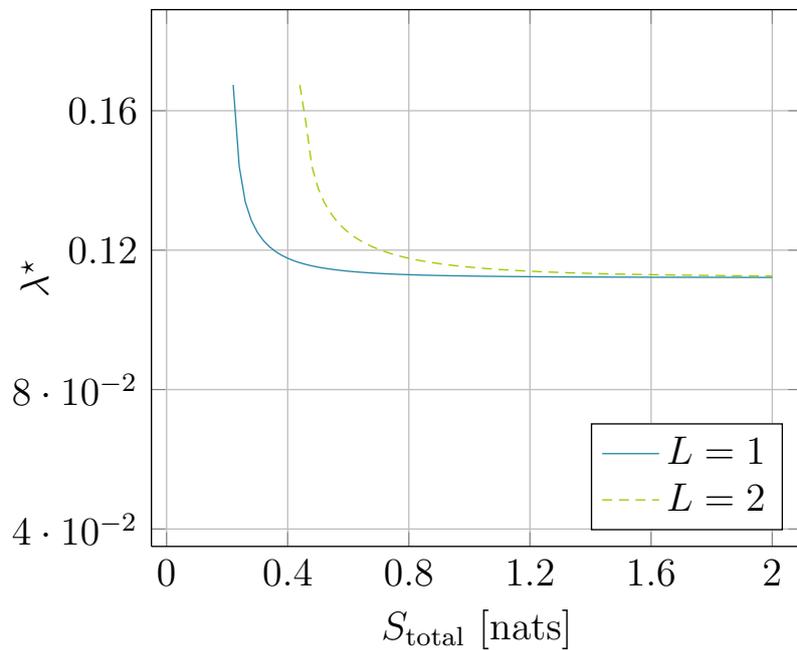
Figure 3.8: The trade-off between SBSs density and total storage size for different file lengths. SNR = 10 dB, $\alpha = 4$, $T = 0.2$ nats/sec/Hz, $\gamma = 3$ and $p^\dagger = 0.3$.

3.6 Closing Remarks

We have studied the caching problem in a scenario where SBSs are stochastically distributed and have finite-rate backhaul links. We derived expressions for the outage probability and average delivery rate, and validated these results via numerical simulations. The results showed that significant gains in terms of outage probability and average delivery rate are possible by having cache-enabled SBSs. We showed that telecom operators can either deploy more base stations or increase the storage size of existing deployment in order to achieve a certain QoE level.

Chapter 4

Multi-Tier Cellular Networks

4.1 Overview

In the previous chapter, we have investigated the gains of caching for a single-tier network. In this chapter, we move to a multi-tier heterogeneous network where base stations in each tier are deployed according to a homogeneous PPP. More precisely, we have a four-tier heterogeneous network consists of mobile terminals (users), cache-enabled small cells, macro cells and central routers. The heterogeneous network experiences delays on the downlink, backhaul and caches. Supposing that small cells are able to cache popular contents proactively, we derive expressions for the average delay of typical users when connected to macro and small cells. Moreover, in order to capture the spatio-temporal content access patterns of users, we suppose *fixed* content popularity, *distance-dependent* and *load-dependent* content popularities. Assuming that the content popularity distribution is perfectly known at the small base stations, we explore three different caching policies based on content-popularity and randomization. In the final part of this chapter, we validate our results via Monte-Carlo simulations and draw our conclusions.

The rest of this chapter is organized as follows. Section 4.2 details the system model under consideration. The performance analysis based on delay and cost is given in Section 4.3. Numerical results for validation are presented in Section 4.4. We finally conclude in Section 4.5.

4.2 System Model

Topology: We consider a multi-tier heterogeneous network in the two-dimensional Euclidean plane $\mathbb{R}^2$ where nodes in each tier k are modeled as a homogeneous PPP $\Phi_k = \{r_i^{(k)}\}_{i \in \mathbb{N}}$ with intensity λ_k, such that $\lambda_1 > ...\lambda_K$ and $r_i^{(k)} \in \mathbb{R}^2$ is referred to as the location of the i-th node at the k-th tier. As a matter of fact, consider a four-tier heterogeneous network consists of mobile terminals (users), small cells, macro cells and central routers with densities $\lambda_{\mathrm{ut}} > \lambda_{\mathrm{sc}} > \lambda_{\mathrm{mc}} > \lambda_{\mathrm{cr}}$ respectively. A *typical* mobile user is located at the

45
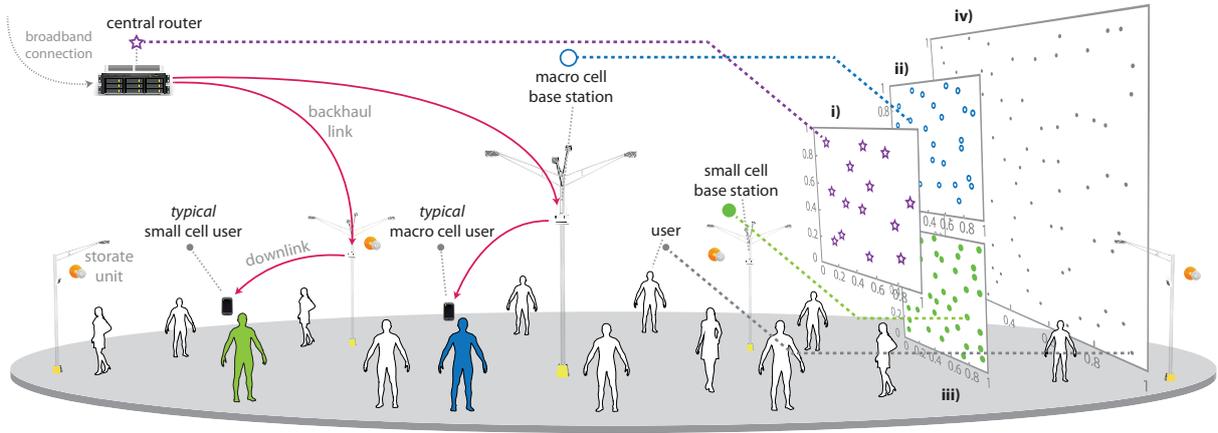
Figure 4.1: An illustration of the considered system model. The snapshots of i) central routers, ii) macro cells, iii) small cells and iv) mobile user terminals are provided on the right side of figure.

Cartesian origin $(0, 0)$ for deriving the performance metrics of the heterogeneous network, whereas the same results for any mobile user hold due to the Slivnyak-Mecke theorem and the stationary and isotropy properties of PPP [195].

Signal Model: We shall consider that the macro and small cells are interfering with each other in the same frequency band. The transmit power is P_{mc} for each macro cell and P_{sc} for each small cell, where we assume that $P_{mc} > P_{sc}$. For notational convenience, let us denote a base station (transmitter) by its position. The received power experienced at a typical user due to a transmitter x is given by $P_x h_x \ell(x)$, where P_x is the transmit power (P_{mc} or P_{sc}), h_x corresponds to the fading power coefficient (square of the fading amplitude) of the channel between transmitter x and typical user, and $\ell(x) = \|x\|^{-\alpha}$ is the singular path-loss function with $\alpha > 2$ [196]. The channel fading power coefficients are independent and identically distributed (i.i.d.) Exponential random variables (Rayleigh fading) with $\mathbb{E}[h_x] = 1$.

Since we assume that the network is interference-limited (i.e., the interference power is dominating the noise power), we simply consider SIR. In case of a typical user connected to a macro cell located at x, the SIR is given as

$$\text{SIR}_{mc}(x) = \frac{P_{mc} h_x \ell(x)}{I_{mm} + I_{sm}} \tag{4.1}$$

where $I_{mm} = \sum_{y \in \Phi_{mc} \backslash \{x\}} P_{mc} h_y \ell(y)$ is the interference experienced from all macro cells except the signalling macro cell at x, and $I_{sm} = \sum_{y \in \Phi_{sc}} P_{sc} h_y \ell(y)$ is the cumulative interference experienced from small cells. In case of a typical user connected to a small cell located at x, the SIR is given as

$$\text{SIR}_{sc}(x) = \frac{P_{sc} h_x \ell(x)}{I_{ss} + I_{ms}} \tag{4.2}$$

where $I_{\mathrm{ss}} = \sum\limits_{y \in \Phi_{\mathrm{sc}} \setminus \{x\}} P_{\mathrm{sc}} h_y \ell(y)$ is the interference experienced from all macro cells except the signalling small cell, and $I_{\mathrm{ms}} = \sum\limits_{y \in \Phi_{\mathrm{mc}}} P_{\mathrm{mc}} h_y \ell(y)$ is the cumulative interference from macro cells. The target SIR in our system model is denoted as γ.

Connectivity and Backhaul: The mobile user terminals are associated with the closest base station, either macro or small cell. As alluded to earlier, each macro or small cell is also connected to its nearest central router. The associated policy will be detailed later on. Each central router has a sufficiently high broadband Internet connection. The wired backhaul is used to provide this broadband connection to macro and small cells via backaul links, such that users' demand can be satisfied. Supposing that a content request of a user is done, the base station is then in charge of starting the delivery immediately.

4.2.1 Caching

When a user has a request, we suppose that the content request is drawn from the distribution f_{pop} which is in decreasing order of content popularities. More formally, the content popularity distribution of a user is a right continuous and monotonically decreasing PDF, given by [192]

$$f_{\mathrm{pop}}(f, \eta) = \begin{cases} (\eta - 1) f^{-\eta}, & f \geq 1, \\ 0, & f < 1, \end{cases} \tag{4.3}$$

where f indicates a point in the support of the corresponding content, and $\eta > 1$ parametrizes the steepness of the popularity distribution curve.

In fact, higher values of η results in steeper distribution, which in turn means that certain contents are highly popular than the rest of contents in $f_{\mathrm{pop}}(f, \eta)$. Conversely, lower values of η yield a more uniform distribution, which in turns say that almost all contents have similar popularities. The content popularity of a user might be evolving over time and space, influenced by the choice of other users and partially known at the base stations. This is somewhat equivalent to say that the parameter η can take different values depending on the scenario. In our case, each base station *perfectly* observes the content popularities according to three different considerations as follows:

- *Fixed*: The content popularity is identical among all users, with fixed steepness factor of $\eta = \eta_0$. Therefore, all small cells observe the same distribution given by $f_{\mathrm{pop}}(f, \eta_0)$.

- *Distance-dependent*: The users have different content popularity distributions, each of them having a distance-dependent steepness factor $\eta = r$, where r is the (random) distance between a user and its signalling small cell. Therefore, we assume that each small cell on average observes a content popularity distribution given by $f_{\mathrm{pop}}(f, \bar{r})$, where $\bar{r}$ is the average distance between the small cell and its users.

- *Load-dependent*: The content popularity of users is load-dependent on average, each

of small cell having parameter $\eta = \lambda_{\mathrm{ut}}/\lambda_{\mathrm{sc}}$. Therefore, all small cells observe the content popularity distribution given by $f_{\mathrm{pop}}(f, \lambda_{\mathrm{ut}}/\lambda_{\mathrm{sc}})$.

Note that the choice of such a continuous content distribution is in fact for ease of analysis. When practice matters or analytical tractability is not a priority, Zipf-like discrete power laws can also be considered for modeling [192]. Indeed, content access statistics in cache-enabled web proxies [197], or more relevantly in base stations [198] are characterized by such discrete power laws (or arguably distributions).

For (some of) caching policies which will be described below, we shall assume that the content popularity distribution $f_{\mathrm{pop}}(f, \eta)$ is perfectly known at the base stations. Practically, in order to have partial knowledge of $f_{\mathrm{pop}}(f, \eta)$ for the caching policies, statistical estimation methods can be employed either at base stations in a distributed manner or alternatively at central routers, by using statistical tools from machine learning (i.e., collaborative filtering [38] and transfer learning [34]).

Given $f_{\mathrm{pop}}(f, \eta)$, the contents in the interval $[1, f_0)$ are *cacheable* contents and called as *catalogue*, whereas the remaining part $[f_0, \infty]$ is considered as non-cacheable contents (i.e., sensor data, voice streaming and online gaming). An interval $[f, f + \Delta f)$ in the support of $f_{\mathrm{pop}}(f, \eta)$ is dedicated to represent the probability of f-th content.

So far, size of a content can start from very few kilo bytes and might go up to hundreds of gigabytes. We restrict ourselves to chunks as contents where we assume that each content/chunk has a fixed length of L bits (for example as in [199]). Indeed, storing/distributing constant-sized chunks of files rather their complete version is one of the key principle of content centric networks [200], as opposite to traditional way of dealing with files on the Internet. Therefore, even though we use "content" due to naming convention, the chunks will be considered from now on, which in turn makes sense to call $f_{\mathrm{pop}}(f, \eta)$ as chunk popularity distribution. This choice makes also analysis simpler as sufficiently small chunks can be transmitted in one time slot in downlink, yielding to avoid time dependence analysis and mixed-time situations in which large-length content delivery and downlink fluctuations would appear in the different time scales.

Each small cell base station has a storage capacity of S, thus caches contents according to a given caching policy. Having such a demand behaviour described above and caching capabilities at the small cells, we then consider the following *offline* caching policies:

- *StdPop* [37]: The most popular contents from the catalogue are stored in the cache of small cells and requires $S_{\mathrm{p}} \geq 0$ amount of storage. We additionally assume that the track of content popularity in a small cell base station requires S_0 amount of storage, defined as a function of number of contents in the catalogue and type of algorithm employed for content popularity estimation, thereby it holds that $S = S_{\mathrm{p}} + S_0 \geq 0$.

- *UniRand* [132]: The S amount of contents are cached uniformly at random. Note that this policy is not aware of the content catalogue, therefore does not require any memory to track the content popularity profile.

- *MixPop*: The S_p amount of storage is used to cache the most popular contents deterministically. The storage overhead is $S_0 \geq 0$ and again defined as a function of number of contents and employed algorithm. On top of this popularity-based policy, we also assume that $S_\mathrm{u} \geq 0$ amount of storage is used to cache contents uniformly at random, thus $S = S_\mathrm{p} + S_0 + S_\mathrm{u} \geq 0$.

In fact, if the catalogue size is sufficiently small, the storage overhead in StdPop and MixPop, due to the track of content popularity can be neglected. However, such an overhead might dominate the total storage space when a large catalogue with low-sized chunks are considered. One can also observe that the StdPop and UniRand policies are special cases of MixPop policy and are given here for the sake of exposition.

The performance of any statistic-aware *online* cache removal policy (i.e., least-recently used (LRU) and least-frequently used (LFU)) would be upper bounded by its offline successor which has perfect content statistics, as such an online approach would require iterative estimation of content popularity in a finite time window, yielding to degrade the overall performance. Such online policies can also be incorporated to our system model after some specific assumptions (see Independent Reference Model [201] for an approximation of LRU policy).

4.2.2 Delay and Quality of Service

QoS is closely related to the delay experienced by users. We consider three different sources of delay which are detailed separately as follows.

Delay in downlink: When macro and small cell base stations have to deliver the contents to their intended mobile users, it is evident that the transmissions throughout the wireless medium of the downlink incur delays mainly due to the interference from concurrent transmissions and channel fading. Consider now a simple retransmission protocol where a packet of requested content is repeatedly transmitted until the success of delivery, up to a pre-defined number of retransmission attempts M. Indeed, inferring whether a delivery is successful or not at the base station essentially relies on the SINR (or SIR in our case) being higher than the predefined threshold γ and feedback. If a packet is delivered successfully, we shall assume that the base station (macro or small cell) receives a one-bit acknowledgement message from the mobile user with negligible delay and error. Otherwise, if the delivery fails, the base station receives a one-bit negative acknowledgement message in the same vein. These attempts take T_0 amount of time. An *outage* event occurs if the packet is not delivered after M attempts. For the rest of the chapter, we denote the downlink delay experienced by the typical macro and small cell users as D_dm and D_ds respectively.

Delay in backhaul: The delay caused in a wired backhaul link is modeled by an Exponentially distributed random variable with mean being proportional to the product of the average link distance (from typical base station to its nearest central router) and the average number of base stations connected to a single central router. In particular, rep-

resenting the delay in macro and small cell backhaul links as $D_{\text{bm}} \sim \text{Exponential}(\bar{\mu}_{\text{bm}})$ and $D_{\text{bs}} \sim \text{Exponential}(\bar{\mu}_{\text{bs}})$ respectively, we (in general) suppose that D_{bs} stochastically dominates D_{bm}.[1] In stochastic sense, this shows that small cell backhaul links are subject to higher delays compared to those of macro cells.

Delay in caches: Serving a user by reading its content from local cache is subject to delay as the storage medium is prone to errors, whereas such a delay may also vary depending on the storage types and underlying mechanisms (i.e., hard disk, solid-state disk (SSD)). In this regard, we model this phenomenon as $D_{\text{ca}} \sim \text{Exponential}(\bar{\mu}_{\text{ca}})$, an Exponentially distributed random variable with mean $\bar{\mu}_{\text{ca}}$ being proportional to the storage type. We also assume that the delay of small cell backhaul links stochastically dominates the delay of reading a content from local caches, meaning that the speed of content reads from caches is stochastically higher than the speed of small cell backhaul links.

4.3 Performance Analysis

We in the following introduce two lemmas [202, 203] which will be used in the delay and cost analysis.

Lemma 3. *The PDF of the length of a link of any node in Φ_{k-1} to the nearest node in Φ_k is given by*

$$f_k(r) = 2\lambda_k \pi r \exp(-\pi \lambda_k r^2). \tag{4.4}$$

If $c(r)$ is a cost function of link length r, then, the expected cost $\bar{c}$ and total expected cost $\bar{C}$ are given by

$$\bar{c} = \mathbb{E}_{\Phi_k}\big[c(r)\big] = \int_0^\infty c(r) f_k(r) \mathrm{d}r \tag{4.5}$$

$$\bar{C} = \mathbb{E}_{\Phi_{k-1}, \Phi_k}\Big[\sum_{\Phi_{k-1}} c(r)\Big] = \lambda_{k-1} \int_0^\infty c(r) f_k(r) \mathrm{d}r. \tag{4.6}$$

When the cost function $c(r)$ has the form of ar^b, the expected cost $\bar{c}$ and total expected cost $\bar{C}$ are expressed as

$$\bar{c} = a \frac{\Gamma\big(\frac{b}{2} + 1\big)}{(\pi \lambda_k)^{b/2}} \tag{4.7}$$

$$\bar{C} = \lambda_{k-1} a \frac{\Gamma\big(\frac{b}{2} + 1\big)}{(\pi \lambda_k)^{b/2}}. \tag{4.8}$$

where $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$ is the Gamma function.

Lemma 4. *Assuming that the nodes in Φ_k are linked to their nearest nodes in Φ_{k+1}, then, the average number of linked nodes from Φ_k to each node in Φ_{k+1} is given by*

$$\lambda_k / \lambda_{k+1} \tag{4.9}$$

---

[1] Given two random variables A and B, we say that A stochastically dominates B if $\mathbb{P}(A > x) \geq \mathbb{P}(B > x)$ for all a, or alternatively, $F_A(x) \leq F_B(x)$ for cumulative distribution functions $F_A(x)$ and $F_B(x)$.

The next lemma and corollary will also be used throughout the chapter.

Lemma 5. *Assume that the noise is negligible (meaning that the communication is interference-limited) and the backhaul is ignored at the macro cells. The probability of successful transmission from the closest macro cell to the typical user is given by*

$$\mathbb{P}_{\mathrm{m}} = \frac{1}{1 + \rho(\gamma, \alpha) + (P_{\mathrm{sc}}/P_{\mathrm{mc}})^{2/\alpha}(\lambda_{\mathrm{sc}}/\lambda_{\mathrm{mc}})\gamma^{2/\alpha}A(\alpha)} \tag{4.10}$$

where $\rho(\gamma, \alpha) = \gamma^{2/\alpha}\int_{\gamma^{-2/\alpha}}^{\infty}\frac{1}{1+u^{\alpha/2}}\mathrm{d}u$ and $A(\alpha) = \frac{2\pi/\alpha}{\sin(2\pi/\alpha)}$. Similarly, the probability of successful transmission from the closest small cell to the typical user is given by

$$\mathbb{P}_{\mathrm{s}} = \frac{1}{1 + \rho(\gamma, \alpha) + (P_{\mathrm{mc}}/P_{\mathrm{sc}})^{2/\alpha}(\lambda_{\mathrm{mc}}/\lambda_{\mathrm{sc}})\gamma^{2/\alpha}A(\alpha)}. \tag{4.11}$$

Proof. See Appendix B.1. $\qquad\square$

Corollary 1. *Consider that the regime is interference-limited and the backhaul effects are ignored. Then, the probability of successful transmission from the closest macro cell located at a distance r is given by*

$$p_{\mathrm{m}}(r) = \exp\left(\pi r^2\left[\lambda_{\mathrm{mc}}\rho(\gamma, \alpha) + (P_{\mathrm{sc}}/P_{\mathrm{mc}})^{2/\alpha}\lambda_{\mathrm{sc}}\gamma^{2/\alpha}A(\alpha)\right]\right). \tag{4.12}$$

Similarly, the probability of successful transmission from the closest small cell located at a distance r is given by

$$p_{\mathrm{s}}(r) = \exp\left(\pi r^2\left[\lambda_{\mathrm{sc}}\rho(\gamma, \alpha) + (P_{\mathrm{mc}}/P_{\mathrm{sc}})^{2/\alpha}\lambda_{\mathrm{mc}}\gamma^{2/\alpha}A(\alpha)\right]\right) \tag{4.13}$$

Proof. The proof is easily obtained by following the same steps in proof of Lemma 5 except averaging over the spatial distribution. $\qquad\square$

4.3.1 Delay Analysis

Considering the aforementioned delay sources (namely downlink, caching and backhaul), the delay experienced by the typical macro and small cell users are respectively defined as

$$D_{\mathrm{m}} = D_{\mathrm{dm}} + D_{\mathrm{bm}}, \tag{4.14}$$

$$D_{\mathrm{s}} = D_{\mathrm{ds}} + \mathbb{1}_{\{f_{\mathrm{s}} \in \Delta_0\}}D_{\mathrm{ca}} + \left(1 - \mathbb{1}_{\{f_{\mathrm{s}} \in \Delta_0\}}\right)D_{\mathrm{bs}} \tag{4.15}$$

where f_{s} is the content requested by the typical small cell user and Δ_0 is the cache of its connected small cell. The indicator function $\mathbb{1}_{\{\ldots\}}$ returns 1 if the statement holds, 0 otherwise. Before proceeding to the next step, let us define functions $B_1(T_0, M, \gamma, \alpha, P_x, P_y, \lambda_x, \lambda_y)$, $B_2(S_{\mathrm{p}}, \eta)$ and $B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \eta)$ given on the top of the next page.

We now state the following results which are related to average delay experienced by typical macro and small cell users.

$$B_1(T_0, M, \gamma, \alpha, P_x, P_y, \lambda_x, \lambda_y) = T_0 \sum_{i=0}^{M-1} (-1)^i \binom{M}{i+1} \frac{1}{1 + i\left[\rho(\gamma, \alpha) + (P_x/P_y)^{2/\alpha}(\lambda_x/\lambda_y)\gamma^{2/\alpha}A(\alpha)\right]} \tag{4.16}$$

$$B_2(S_{\mathrm{p}}, \eta) = 1 - (1 + S_{\mathrm{p}})^{1-\eta} \tag{4.17}$$

$$B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \eta) = \frac{S_{\mathrm{u}}}{f_0 - S_{\mathrm{p}}} \left(1 - (1 + f_0)^{1-\eta} + (1 + S_{\mathrm{p}})^{1-\eta}\right) \tag{4.18}$$

Theorem 6. *The average delay for a typical user connected to a macro cell is given by*

$$\bar{D}_{\mathrm{m}} = B_1(T_0, M, \gamma, \alpha, P_{\mathrm{sc}}, P_{\mathrm{mc}}, \lambda_{\mathrm{sc}}, \lambda_{\mathrm{mc}}) + \frac{1}{2}\beta\lambda_{\mathrm{mc}}\lambda_{\mathrm{cs}}^{-3/2} \tag{4.19}$$

where $B_1(T_0, M, \gamma, \alpha, P_{\mathrm{sc}}, P_{\mathrm{mc}}, \lambda_{\mathrm{sc}}, \lambda_{\mathrm{mc}})$ is given in (4.16).

Proof. See Appendix B.2. □

Corollary 2. *The average delay for a typical user connected to a small cell (with no caching capabilities) is given by*

$$\bar{D}_{\mathrm{m}} = B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}}) + \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2} \tag{4.20}$$

where $B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}})$ is given in (4.16).

Proof. The result is a direct application of Theorem 6, thus is immediately proved by by following similar steps given in Appendix B.2. □

Theorem 7. *When MixPop caching policy is employed at the small cells, the average delay for a typical user connected to a small cell under fixed content popularity distribution is given by*

$$\bar{D}_{\mathrm{fix}}^{(\mathrm{mix})} = B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}}) + \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2} +$$
$$\left(\bar{\mu}_{\mathrm{ca}} - \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2}\right)\left(B_2(S_{\mathrm{p}}, \eta_0) + B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \eta_0)\right) \tag{4.21}$$

where $B_2(S_{\mathrm{p}}, \eta_0)$ and $B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \eta_0)$ are given in (4.17) and (4.18) respectively.

In case of distance-dependent content popularity, the average delay is given by

$$\bar{D}_{\mathrm{dist}}^{(\mathrm{mix})} = B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}}) + \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2} +$$
$$\left(\bar{\mu}_{\mathrm{ca}} - \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2}\right)\left(B_2(S_{\mathrm{p}}, \frac{1}{2\sqrt{\lambda_{\mathrm{sc}}}}) +$$
$$B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \frac{1}{2\sqrt{\lambda_{\mathrm{sc}}}})\right). \tag{4.22}$$

In case of load-dependent content popularity, the average delay is given by

$$\bar{D}_{\text{load}}^{(\text{mix})} = B_1(T_0, M, \gamma, \alpha, P_{\text{mc}}, P_{\text{sc}}, \lambda_{\text{mc}}, \lambda_{\text{sc}}) + \frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2} +$$
$$\left(\bar{\mu}_{\text{ca}} - \frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2}\right)\left(B_2(S_{\text{p}}, \frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}}) +$$
$$B_3(S_{\text{u}}, S_{\text{p}}, f_0, \frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}})\right). \quad (4.23)$$

Proof. See Appendix B.3. □

We have so far stated the results for MixPop caching policy. By slightly modifying the steps in proof of Theorem 7, the following corollaries can be obtained for StdPop and UniRand caching policies respectively.

Corollary 3. *When* StdPop *caching policy is employed at the small cells, the average delay for a typical user connected to a small cell under fixed content popularity distribution is given by*

$$\bar{D}_{\text{fix}}^{(\text{std})} = B_1(T_0, M, \gamma, \alpha, P_{\text{mc}}, P_{\text{sc}}, \lambda_{\text{mc}}, \lambda_{\text{sc}}) +$$
$$\frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2} + \left(\bar{\mu}_{\text{ca}} - \frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2}\right)B_2(S_{\text{p}}, \eta_0). \quad (4.24)$$

In case of distance-dependent content popularity, the average delay is given by

$$\bar{D}_{\text{dist}}^{(\text{std})} = B_1(T_0, M, \gamma, \alpha, P_{\text{mc}}, P_{\text{sc}}, \lambda_{\text{mc}}, \lambda_{\text{sc}}) +$$
$$\frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2} + \left(\bar{\mu}_{\text{ca}} - \frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2}\right)B_2(S_{\text{p}}, \frac{1}{2\sqrt{\lambda_{\text{sc}}}}). \quad (4.25)$$

In case of load-dependent content popularity, the average delay is given by

$$\bar{D}_{\text{load}}^{(\text{std})} = B_1(T_0, M, \gamma, \alpha, P_{\text{mc}}, P_{\text{sc}}, \lambda_{\text{mc}}, \lambda_{\text{sc}}) +$$
$$\frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2} + \left(\bar{\mu}_{\text{ca}} - \frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2}\right)B_2(S_{\text{p}}, \frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}}). \quad (4.26)$$

Proof. Observe that StdPop is a special case of MixPop. Therefore, the proof is done by following the steps in proof of Theorem 7 but only taking into account the first term on the right hand side (RHS) of (B.24), (B.29) and (B.32). □

Corollary 4. *When* UniRand *caching policy is employed at the small cells, the average delay for a typical user connected to a small cell under fixed content popularity distribution is given by*

$$\bar{D}_{\text{fix}}^{(\text{uni})} = B_1(T_0, M, \gamma, \alpha, P_{\text{mc}}, P_{\text{sc}}, \lambda_{\text{mc}}, \lambda_{\text{sc}}) +$$
$$\frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2} + \left(\bar{\mu}_{\text{ca}} - \frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2}\right)B_3(S_{\text{u}}, S_{\text{p}}, f_0, \frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}}). \quad (4.27)$$

In case of distance-dependent content popularity, the average delay is given by

$$\bar{D}_{\text{dist}}^{(\text{uni})} = B_1(T_0, M, \gamma, \alpha, P_{\text{mc}}, P_{\text{sc}}, \lambda_{\text{mc}}, \lambda_{\text{sc}}) +$$
$$\frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2} + \left(\bar{\mu}_{\text{ca}} - \frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2}\right)B_3(S_{\text{u}}, S_{\text{p}}, f_0, \frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}}). \quad (4.28)$$

In case of load-dependent content popularity, the average delay is given by

$$\bar{D}_{\text{load}}^{(\text{uni})} = B_1(T_0, M, \gamma, \alpha, P_{\text{mc}}, P_{\text{sc}}, \lambda_{\text{mc}}, \lambda_{\text{sc}}) +$$
$$\frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2} + \left(\bar{\mu}_{\text{ca}} - \frac{1}{2}\beta\lambda_{\text{sc}}\lambda_{\text{cs}}^{-3/2}\right)B_3(S_{\text{u}}, S_{\text{p}}, f_0, \frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}}). \quad (4.29)$$

Proof. As UniRand is a special case of MixPop, the results are immediate from proof of Theorem 7 by only considering the second term on the RHS of (B.24), (B.29) and (B.32). □

Note that the results above are based on the assumption that typical users are connected to their nearest base stations. Now, in order to move from average delay of typical users to total average delay in network, we use the following distance-based association policy.

Definition 3 (Association Policy). *Let r_{sc} and r_{mc} be the distance from the nearest small and macro cell respectively. A user is associated to the nearest small cell if $r_{\text{sc}} < \kappa r_{\text{mc}}$ with $\kappa \in \mathbb{R}^+$, and the nearest macro cell otherwise.*

Based on this association policy, the association probabilities are given as follows.

Proposition 3 (Association Probability). *A typical user is connected to a small cell with association probability given by*

$$p_{\text{a}} = \frac{\kappa^2\lambda_{\text{sc}}}{\lambda_{\text{mc}} + \kappa^2\lambda_{\text{sc}}}, \quad (4.30)$$

and is connected to a macro cell with association probability $1 - p_{\text{a}}$.

Proof. See Appendix B.4. □

Therefore, after these association probabilities, the total average network delay can be stated as follows.

Proposition 4 (Total Average Network Delay). *Suppose that TDMA is employed among users and the small cells operate under* MixPop *caching policy with fixed content popularity*

distribution. *The total average network delay is then given by*

$$\bar{D}^{(\text{tot})} = \frac{p_{\text{a}}^2 \lambda_{\text{ut}}^2}{\lambda_{\text{sc}}} B_1(T_0, M, \gamma, \alpha, P_{\text{mc}}, P_{\text{sc}}, \lambda_{\text{mc}}, \lambda_{\text{sc}}) +$$

$$\frac{(1 - p_{\text{a}})^2 \lambda_{\text{ut}}^2}{\lambda_{\text{mc}}} B_1(T_0, M, \gamma, \alpha, P_{\text{sc}}, P_{\text{mc}}, \lambda_{\text{sc}}, \lambda_{\text{mc}}) +$$

$$\frac{1}{2} \beta p_{\text{a}} \lambda_{\text{ut}} \lambda_{\text{sc}} \lambda_{\text{cs}}^{-3/2} + \frac{1}{2} \beta (1 - p_{\text{a}}) \lambda_{\text{ut}} \lambda_{\text{mc}} \lambda_{\text{cs}}^{-3/2} +$$

$$p_{\text{a}} \lambda_{\text{ut}} \left(\bar{\mu}_{\text{ca}} - \frac{1}{2} \beta \lambda_{\text{sc}} \lambda_{\text{cs}}^{-3/2} \right) \left(B_2(S_{\text{p}}, \eta_0) + B_3(S_{\text{u}}, S_{\text{p}}, f_0, \eta_0) \right). \quad (4.31)$$

Proof. See Appendix B.5. $\qquad\qquad\square$

Remark 1. *The total average network delay for the other caching policies and content popularities can be derived straightforwardly by using the results in Theorems 6, 7 and Corollaries 3, 4.*

4.3.2 Cost Analysis

In a similar fashion to [203, 204], we define our total network cost by taking into account the deployment and operational costs, such as

$$C = \underbrace{c_1 \lambda_{\text{cs}} + c_2 \lambda_{\text{mc}} + c_3 \lambda_{\text{sc}} + \bar{L}_{\text{mc}} + \bar{L}_{\text{sc}}}_{\text{(i)}} + \underbrace{\varphi \bar{D}^{(\text{tot})}}_{\text{(ii)}} \quad (4.32)$$

where the term (i) captures deployment and load-independent operational costs, and (ii) takes into account the load-dependent operational costs. The coefficients c_1, c_2 and c_3 reflects the unit cost of central routers, macro cells and small cells respectively. The parameters $\bar{L}_{\text{mc}}$ and $\bar{L}_{\text{sc}}$ are the average costs of constructing/deploying backhaul links from the nearest central routers to macro and small cells respectively. The load-dependent cost is captured by the total average network delay $\bar{D}^{(\text{tot})}$ times the unit cost φ. Based on this definition, the total average network cost is stated as follows.

Proposition 5 (Total Average Network Cost). *When the cost of link construction between a base station (macro or small cell) and its nearest central router has the form of $L(r) = ar^b$, the total average cost of the network is given by*

$$\bar{C}^{(\text{tot})} = c_1 \lambda_{\text{cs}} + c_2 \lambda_{\text{mc}} + c_3 \lambda_{\text{sc}} + \lambda_{\text{mc}} a_{\text{mc}} \frac{\Gamma\left(\frac{b_{\text{mc}}}{2} + 1\right)}{\left(\pi \lambda_{\text{cs}}\right)^{b_{\text{mc}}/2}} + \lambda_{\text{sc}} a_{\text{sc}} \frac{\Gamma\left(\frac{b_{\text{sc}}}{2} + 1\right)}{\left(\pi \lambda_{\text{cs}}\right)^{b_{\text{sc}}/2}} + \varphi \bar{D}^{(\text{tot})} \quad (4.33)$$

where a_{mc} and b_{mc} are the coefficients for macro cells, and a_{sc} and b_{sc} are for the small cells.

Proof. See Appendix B.6. $\qquad\qquad\square$

In the above, we have provided the expressions for total average network delay and total network cost. The optimization of these metrics with respect to system design parameters are left for future work.

4.4   Numerical Results

In this section, we conduct a numerical study to validate our results in the previous section. The list of simulation parameters are given in Table 4.1 and will be used throughout this section unless otherwise stated. In the following, we discuss impact of key design parameters on the average delay, namely 1) macro cell density, 2) small cell density, 3) target SIR, and 4) storage size.

Table 4.1: Simulation Parameters for Multi-Tier Network.

Parameters	Values
λ_{cr}	1.4×10^{-6} unit/m^2
λ_{mc}	2.8×10^{-6} unit/m^2
λ_{sc}	3.6×10^{-6} unit/m^2
λ_{ut}	7.2×10^{-6} unit/m^2
P_{mc}	20 Watts
P_{sc}	2 Watts
α	4
γ	3 dB
M	4
T_0	0.1 ms
μ_{ca}	0.01 ms
f_0	500 GByte
η_0	1.45
S, S_{p}, S_0 and S_{u}	100, 9.5, 0.5 and 90 GByte

Impact of macro cell density λ_{mc}: The change of average delay with respect to the macro cell density is given in Fig. 4.2a. Therein, as the number of macro cells increases, we observe an increment in average delay. This is mainly due to the backhaul as the delay in backhaul is proportional to the distance and average number of connected macro cells. In this setup, even though the average distance from a macro cell to its central central router decreases (thus less delay in the backhaul), the increasing number of base stations contributes more to the average delay, thus yielding such a behaviour. On the other

56

hand, the average delay in small cells remains static in this setup. However, we note that the average delay experienced by a typical small cell user is reduced by adding caching capabilities at the base stations. For instance, when content popularity is load-dependent and caching policy is MixPop, the average delay is reasonably less than other candidates (including typical user with no caching at small cells).

Impact of small cell density λ_{sc}: The change of average delay with respect to the small cell density is depicted in Fig. 4.2b. As similar to the previous figure for macro cell density, we see that the average delay increases for all kind of small cell users. However, in this numerical setup, the rate of increment in delay with no-caching capabilities at the small cells is higher than the delay experienced by typical users with cache-enabled small cells. Compared to the fixed and load-dependent content popularities, typical user under load-dependent content popularity experiences less delays when the number of small cells increases.

Impact of target SIR γ: In our setup, yet another important design parameter is the target SIR. In this regard, the variation of average delay with respect to the target SIR is illustrated in Fig. 4.2c. As observed in the figure, the average delay increases by imposing higher target SIR values. This change is only visible in low values of target SIR whereas variation of delay in higher values of target SIR is negligible. This might stem from the fact that downlink delay is not a dominating factor in our scenario compared to the backhaul delay. A typical user connected to the small cell with no caching capabilities experiences the highest delay, whereas the minimum delay is achieved by using MixPop policy under load-dependent content popularity. The delay of typical macro cell user remains between small cell user with no-caching and caching capabilities at the base stations.

Impact of storage size S: Yet another crucial design parameter in our setup is the storage size. The impact of storage size on the average delay is shown in Fig. 4.2d. Indeed, as observed from the figure, dramatical decrease in delay is observed by increasing the storage size of small base stations. As similar to previous observations, the most sensitive content popularity for the average delay is the load-dependent content popularity.

4.5   Closing Remarks

In this chapter, we have characterized the average delay of small and macro cells users under backhaul constraints and caching capabilities at the small base stations. Several content popularity distributions and caching policies have been considered. The main conclusion from this chapter is that caching at the small base stations allows telecom operators to balance average access delay to the contents, especially if heterogeneous network densification under limited backhaul is considered for the deployments.
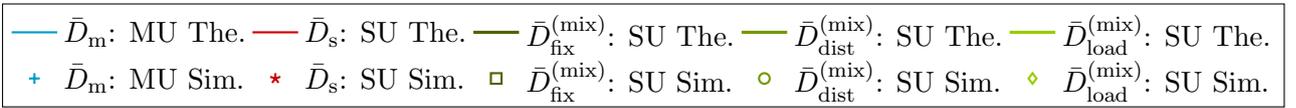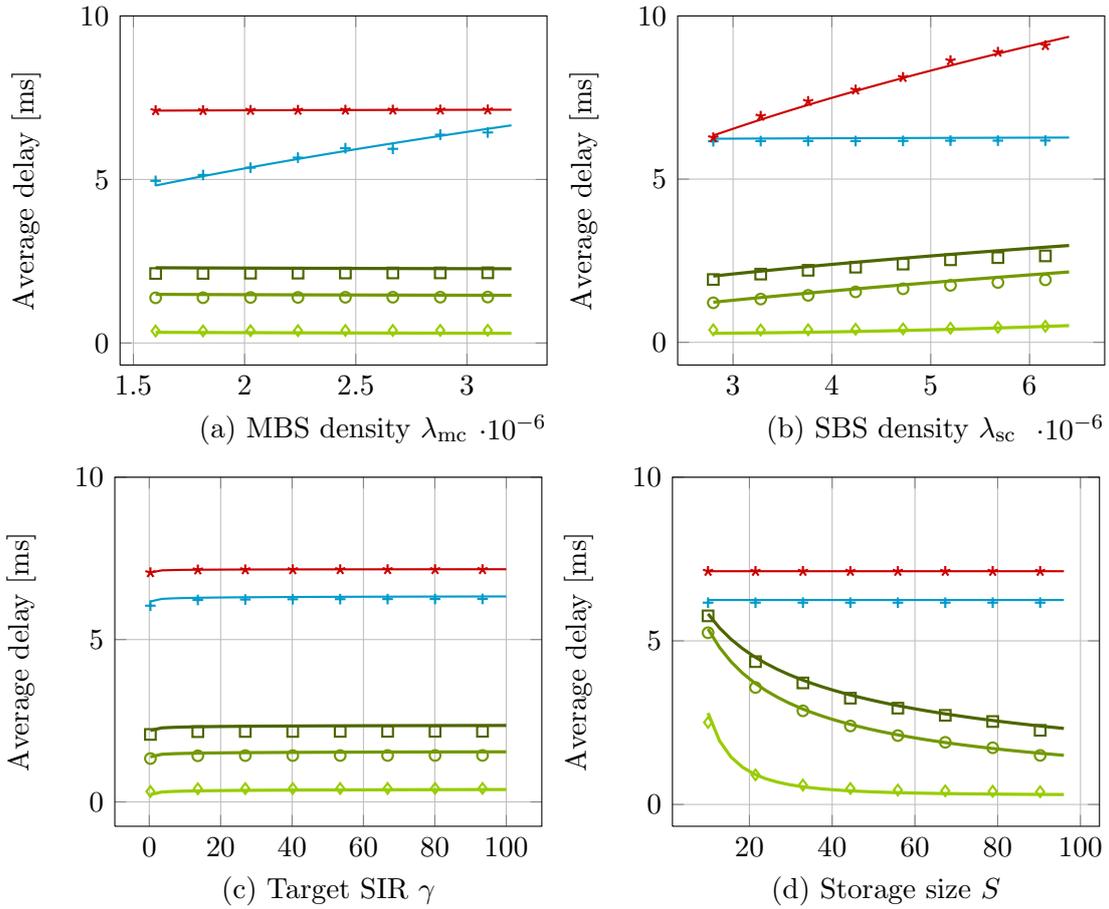
Figure 4.2: Evolution of average delay with respect to the a) macro cell density, b) small cell density, c) target SIR and d) storage size.

Chapter 5

Clustered Cellular Networks

5.1 Overview

In the previous chapter, we have considered a multi-tier network where SBS are distributed according to a PPP. In order to move forward to a more realistic deployment, in this chapter, we focus on clustering of cache-enabled small base stations. In particular, the heterogeneous network consists of mobile users, clustered cache-enabled small cells, macro cells and central routers. For small cells, we shall consider two different topologies, namely 1) coverage and 2) capacity-aided deployments. The coverage-aided deployment is based on a Poisson hole process and capacity-aided deployment is modeled by a Matérn cluster process. Due to the non-tractability nature of point processes, we restrict ourselves to approximations and validate these results via numerical simulations. A hierarchical model based on random trees and a geographical caching problem are also described, whereas further investigations are left for future work.

The rest of this chapter is organized as follows. Section 5.2 provides the details of system model including cache-enabled clustered small cells. Section 5.3 conducts a performance analysis based on delivery rate metric. Numerical results are presented in Section 5.3.1 and validate our approximations for average delivery rate. Section 5.4 focuses on formulation of geographical caching methods which makes use of a hierarchical model. Finally, conclusions are given in Section 5.5.

5.2 System Model

The system model consists of macro cells, cache-enabled clustered small cells, central routers and mobile users. We focus on two different topologies which are similar to the one studied in [205].

5.2.1 Coverage-Aided Topology

We model a heterogeneous cellular network which consists of macro and cache-enabled small cell base stations. The macro cells are modeled by an independent homogeneous PPP of intensity λ_{mc}, denoted by $\Phi_{\mathrm{mc}} = \{y_i\}_{i\in\mathbb{N}}$ where y_i denotes the location of the i-th macro cell base station. Additionally, the *potential* small cells are given by another independent homogeneous PPP of intensity $\lambda_{\mathrm{sc}'}$, denoted by $\Phi_{\mathrm{sc}'} = \{x_i'\}_{i\in\mathbb{N}}$ where x_i' represents the location of the i-th small cell base station. We suppose that each macro cell has an exclusion region which is made of a disk with radius R_{c} centered at the position of macro cell. Assuming that the aim of small cells is to fill the coverage holes of macro cells to provide a better service to users, these small cells are deployed outside of the exclusion regions of macro cells. Therefore, the deployed small cells form clusters according to a *Poisson hole process* (a Cox process) as follows [206].

Definition 4 (Clustering process of coverage-aided small cells). *Let Φ_{mc} be a homogeneous PPP of intensity λ_{mc} for macro cells and $\Phi_{\mathrm{sc}'}$ be an independent and homogeneous PPP of intensity $\lambda_{\mathrm{sc}'}$ for potential small cells, with $\lambda_{\mathrm{sc}'} > \lambda_{\mathrm{mc}}$. For each $y \in \Phi_{\mathrm{mc}}$, remove all the points in*

$$\Phi_{\mathrm{sc}'} \cap \mathcal{B}(y, R_{\mathrm{c}}) \tag{5.1}$$

where $\mathcal{B}(y, R_{\mathrm{c}})$ is the ball of radius R_{c} centered at y. Then, the remaining points of $\Phi_{\mathrm{sc}'}$ form clusters, known as the Poisson hole process Φ_{sc} and represents the deployed small cells. Moreover, this process has the intensity of

$$\lambda_{\mathrm{sc}} = \lambda_{\mathrm{sc}'}\exp(-\lambda_{\mathrm{mc}}\pi R_{\mathrm{c}}^2). \tag{5.2}$$

On the other hand, central routers are distributed in the plane according to an independent homogeneous PPP of intensity λ_{cr}, denoted by $\Phi_{\mathrm{cr}} = \{u_i\}_{i\in\mathbb{N}}$. These routers are in charge of providing broadband Internet connection to macro and small cells via backhaul links. Mobile user terminals are also positioned in the whole plane according to an independent homogeneous PPP of intensity λ_{ut}, denoted by $\Phi_{\mathrm{ut}} = \{z_i\}_{i\in\mathbb{N}}$. The visualization of the system model, together with the coverage-aided deployment of base stations and realizations/snapshots of point processes are given in Fig. 5.1.

5.2.2 Capacity-Aided Topology

Let us consider a two-tier heterogeneous cellular network consists of macro and cache-enabled small cell base stations. The macro cells are distributed on the two-dimensional Euclidean plane according to an independent homogeneous PPP of intensity λ_{mc}, denoted by $\Phi_{\mathrm{mc}} = \{y_i\}_{i\in\mathbb{N}}$ where y_i denotes the position of the i-th macro cell base station. On the other hand, the small cells are placed in hot-spots to sustain the demand of highly concentrated users, according to an independent *Matérn cluster process* $\Phi_{\mathrm{sc}} = \{x_i\}_{i\in\mathbb{N}}$ whose parent PPP $\Phi_{\mathrm{sc}'}$ has intensity $\lambda_{\mathrm{sc}'}$. The process is given as follows [206].
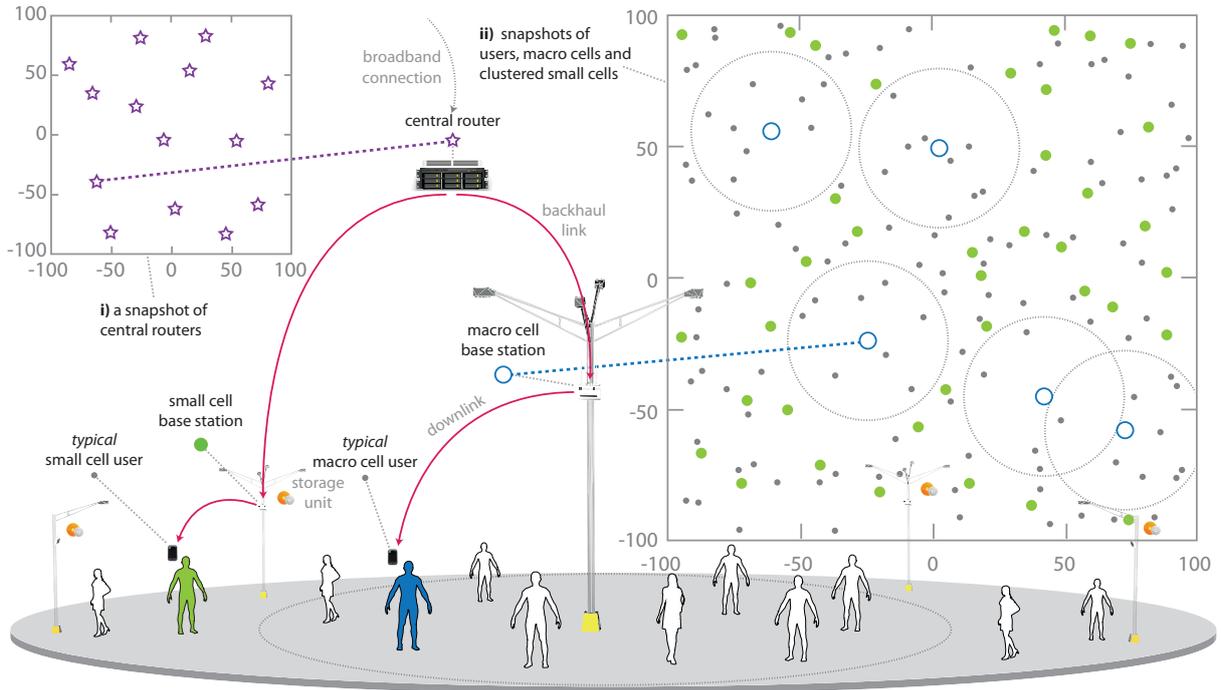
Figure 5.1: An illustration of the coverage-aided deployment.

Definition 5 (Clustering process of capacity-aided small cells)**.** *Let $\Phi_{\mathrm{sc}'}$ be a parent process modeled by a homogeneous PPP of intensity $\lambda_{\mathrm{sc}'}$. Then, the clustering process of small cells is given by*

$$\Phi_{\mathrm{sc}} = \bigcup_{x' \in \Phi_{\mathrm{sc}'}} N^{x'} \tag{5.3}$$

where $N^{x'}$ is a Poisson number of i.i.d. points with mean $\bar{c}$, distributed uniformly in the ball $\mathcal{B}(x', R_{\mathrm{c}})$. Then, the process Φ_{sc} is called Matérn cluster process Φ_{sc} and has intensity of

$$\lambda_{\mathrm{sc}} = \lambda_{\mathrm{sc}'} \bar{c}. \tag{5.4}$$

For the users, we suppose that mobile users (both macro and small cell users) are distributed on the two-dimensional Euclidean plane, however small cell users are highly concentrated on hot-spot regions served by small cells. From this motivation, we assume that all mobile users are distributed according to a Cox process $\Phi_{\mathrm{ut}} = \{z_i\}_{i \in \mathbb{N}}$ [206]. In particular, the population centers of radius R_{c} are drawn from the parent PPP $\Phi_{\mathrm{sc}'}$, where *small cell users* in these clusters are distributed uniformly at random and are covered by small cells realized from the Matern cluster process Φ_{sc}. By doing so, these mobile users are (on average) covered by their small cells deployed in these hot-spot areas.

Recalling $\bar{c}$ as the average number of small cells per cluster, the density of active small cell users per cluster is then $\lambda_{\mathrm{ut-s}} = \frac{\bar{c}}{\pi R_{\mathrm{c}}^2}$. The *macro cell users* distributed in the rest of the network follow a PPP with density $\lambda_{\mathrm{ut-m}}$ and are served by their own macro cells. In order to ease the calculations, we consider that each macro serves only one macro user on

average and same holds for each macro cell and its user. With this consideration in mind, the densities of macro and small call users are equal to that of the macro and small cells, respectively. Under this setting, the macro and small cell users form a Cox process with density $\lambda_{\mathrm{ut}} = \lambda_{\mathrm{mc}} + \lambda_{\mathrm{sc}}$, clustered in hot-spots and uniformly distributed in the rest.

On the other hand, we consider that central routers are modeled by an independent homogeneous PPP of intensity λ_{cr}, denoted by $\Phi_{\mathrm{cr}} = \{u_i\}_{i \in \mathbb{N}}$, aiming to provide broadband Internet connection to macro and small cells via backhaul links. The visualization of the system model, together with the capacity-aided deployment of base stations and snapshots of point processes are given in Fig. 5.2. The next subsections detail signal, connectivity, backhaul and caching model of these coverage and capacity-aided deployments.
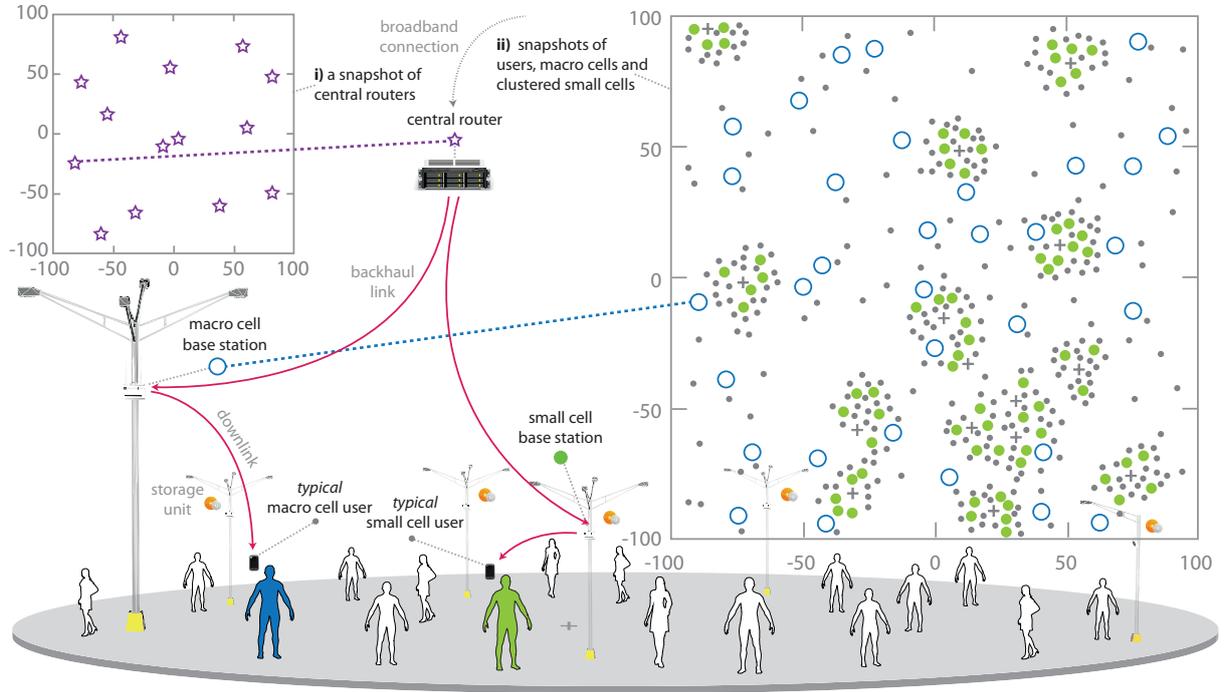
Figure 5.2: An illustration of the capacity-aided deployment.

5.2.3 Signal Model, Connectivity and Backhaul

The transmissions of macro and small cells occur in the same frequency with frequency reuse factor 1. The transmit power is P_{mc} for each macro cell and P_{sc} for each small cell. Macro cells, all users and small cells have single antennas. Having a macro cell positioned at y and receiver at z (or simply call as transmitter y and user z), the channel coefficient is denoted by $h_{y,z} \in \mathbb{C}$. In case of small cell as a transmitter, the channel coefficient between transmitter y and user z is given by $g_{y,z} \in \mathbb{C}$. All the channel power coefficients are i.i.d. Exponential random variables (Rayleigh fading) with $\mathbb{E}[|h_{y,z}|^2] = 1$ and $\mathbb{E}[|g_{x,z}|^2] = 1$. Supposing that the downlink rate of the typical user is a function of received SIR, the

target rate of signalling macro and small cell are given by τ_{mc} and τ_{sc} respectively. The details will be given later on.

Each user is either associated to the nearest macro cell or nearest small cell. The backhaul connection of each base station is provided from its nearest central router. A content request by a mobile user is done via uplink. Having content requests of connected users, the base stations start delivery immediately via the downlink.

Supposing that each central router has a sufficiently high-capacity broadband Internet connection, macro and small cells are connected to their nearest central routers via error-free *wired* backhaul links. In particular, each central router provides broadband internet connection to its connected base stations via backhaul links and has a total capacity of C_{cr}. This total capacity C_{cr} is an Exponentially distributed random variable with mean μ.

5.2.4   Caching

We shall assume that the global content popularity distribution of users follow a power law defined as [192]

$$f_{\mathrm{pop}}\left(f, \eta\right) = \begin{cases} (\eta - 1) f^{-\eta}, & f \geq 1, \\ 0, & f < 1, \end{cases} \tag{5.5}$$

where f indicates a point in the support of the corresponding content and the parameter η models the steepness of the distribution. Such laws are observed in many real-world phenomena (see [192] for example) and its discrete version (i.e., Zipf) is commonly used in the caching literature [53, 132]. Higher values of η yields a stepper distribution, meaning that a few amount of contents in the catalogue is highly popular than the rest, whereas low values of η corresponds to a more uniform behaviour of content popularities.

In practice, the global content popularity distribution might evolve over the space and time, is influenced by the users' local and global preferences and partially known at the small cells. In this work, we assume that this global popularity distribution is static and perfectly known at the small cells. In $f_{\mathrm{pop}}\left(f, \eta\right)$, the contents in the interval $[1, F)$ are *cacheable* and called as *catalogue*, whereas the remaining part $[F, \infty]$ is called as non-cacheable contents (i.e., voice traffic, online gaming and sensor information). An interval $[f, f + \Delta f)$ in the support of $f_{\mathrm{pop}}\left(f, \eta\right)$ is considered as the probability of f-th content.

We assume that each content in the catalogue has a fixed length (see [199] for instance) and called as *chunk*. Each chunk can belong to a part of cacheable video file, audio or picture and so on. In fact, storing/distributing fixed-length chunks is one of the key principle in content centric networks [200] as opposed to classical way of storing contents on the Internet. Therefore, even though we use the term "content" in the paper, the function $f_{\mathrm{pop}}\left(f, \eta\right)$ is actually a chunk popularity distribution. Each small cell has a storage capacity of F_{sc} contents/chunks, with $1 \leq F_{\mathrm{sc}} \leq F$.

5.2.5 Hierarchical Model

The coverage and capacity-aided deployments of small cells together with macro cells, users and central routers can be modeled as random stationary graphs (hierarchical trees in particular) [202]. In particular, a random hierarchical tree whose root node is a central router located at u is given by $\Psi = \{(u, \mathbf{v}_u)\}$, where $\mathbf{v}_u$ is a mark vector containing all random variables associated with the central router. In particular, the mark vector $\mathbf{v}_u$ contains information of macro and small cells (with their users) which are associated to the central router at u (see [53] for a similar treatment), such as:

- N_{mc}: The number of macro cells connected to the central router located at u.

- $\mathbf{r}_{\mathrm{mc}} \in \mathbb{C}^{1 \times N_{\mathrm{mc}}}$: The relative position vector of those macro cells connected to the central router at u. Therein, each element $r_{u,y}$ represents the distance from central router at u to macro cell at y. These positions are conditioned on N_{mc}.

- N_{mu}: The number of users connected to a macro cell located at $y \in \{\mathbf{r}_{\mathrm{mc}}\}$.

- $\mathbf{r}_{\mathrm{mu}} \in \mathbb{C}^{1 \times N_{\mathrm{mu}}}$: The relative positions of those macro cell users which are conditioned on N_{mu}. Each element $r_{y,z}$ represents distance from macro cell at y to its user z.

- $\mathbf{d}_{y,z} \in [0,1]^{1 \times F+1}$: The content demand from macro cell user located at $z \in \mathbf{r}_z$ to macro cell at $y \in \{\mathbf{r}_{\mathrm{mc}}\}$. We consider that each user demands only one content and is drawn from $f_{\mathrm{pop}}(f, \eta)$. Observe that the catalogue has F number of contents and the last entry in $\mathbf{d}_{y,z}$ is dedicated to the non-cacheable content demand. Now, since each macro cell user is connected to only one cell, the total content demands observed at the macro cell y is given by $\mathbf{d}_y = \sum_{z \in \{\mathbf{r}_{\mathrm{mu}}\}} \mathbf{d}_{y,z}$. Here, each element $d_{y,f}$ represents the cumulative number of observed demands at macro cell y for content f.

- $h_{y,z}$: The channel power coefficient representing the channel propagation between the macro cell $y \in \{\mathbf{r}_{\mathrm{mc}}\}$ and user $z \in \{\mathbf{r}_{\mathrm{mu}}\}$.

And also:

- N_{sc}: The number of small cells connected to the central router located at u.

- $\mathbf{r}_{\mathrm{sc}} \in \mathbb{C}^{1 \times N_{\mathrm{sc}}}$: The relative positions of those small cells connected to the central router at u. Here, each element $r_{u,x}$ represents the distance from central router at u to small cell at y. These positions are conditioned on N_{sc}.

- $\mathbf{s}_x \in [0,1]^{1 \times F}$: The caching vector of the small cell located at $x \in \{\mathbf{r}_{\mathrm{sc}}\}$.

- N_{su}: The number of users connected to a small cell located at $x \in \{\mathbf{r}_{\mathrm{sc}}\}$.

- $\mathbf{r}_{\mathrm{su}} \in \mathbb{C}^{1 \times N_{\mathrm{su}}}$: The relative positions of users connected to the small cell at $x \in \{\mathbf{r}_{\mathrm{sc}}\}$. They are conditioned on N_{su}. Each element $r_{y,z}$ represents distance from small cell at x to its user at z.

- $\mathbf{d}_{x,z} \in [0,1]^{1 \times F+1}$: The content demand vector of the small cell user located at $x \in \{\mathbf{r}_{\mathrm{su}}\}$. Similar to macro cell user case, we consider that each user can requests only one content and is sampled from $f_{\mathrm{pop}}(f, \eta)$. Since each small cell user is associated to only one cell, the total content demands observed at a small cell x is given by $\mathbf{d}_x = \sum_{z \in \{\mathbf{r}_{\mathrm{su}}\}} \mathbf{d}_{x,z}$. Therein, each element $d_{x,f}$ represents the cumulative number of observed demands at small cell y for content f.

- $g_{x,z}$: The channel coefficient that describes the channel propagation between the small cell $x \in \{\mathbf{r}_{\mathrm{sc}}\}$ and its connected user $z \in \{\mathbf{r}_{\mathrm{su}}\}$.

5.3 Performance Analysis

In order to conduct the performance analysis of both coverage and capacity-aided deployments, we first start by defining SIR.

Definition 6 (Signal to noise ratio)**.** *The SIR of a typical user connected to a macro cell (namely typical macro cell user) located at random position y is defined as*

$$\mathrm{SIR}_{\mathrm{mu}} \triangleq \frac{P_{\mathrm{mc}} h_y \ell(y)}{I_{\mathrm{mm}} + I_{\mathrm{sm}}} \tag{5.6}$$

where $\ell(y) = \|y\|^{-\alpha}$ is the path loss function (unless otherwise stated) with exponent α, $I_{\mathrm{mm}} = \sum_{y_i \in \Phi_{\mathrm{mc}} \backslash \{y\}} P_{\mathrm{mc}} h_{y_i} \ell(y_i)$ is the cumulative interference from other macro cells except the signalling cell at y, and $I_{\mathrm{sm}} = \sum_{x_i \in \Phi_{\mathrm{sc}}} P_{\mathrm{sc}} g_{x_i} \ell(x_i)$ is the total interference from clustered small cells. Similarly, the SIR of a typical user connected to a small cell (namely typical small cell user) located at random position x is given by

$$\mathrm{SIR}_{\mathrm{su}} \triangleq \frac{P_{\mathrm{mc}} g_x \ell(x)}{I_{\mathrm{ss}} + I_{\mathrm{ms}}} \tag{5.7}$$

where $I_{\mathrm{ss}} = \sum_{x_i \in \Phi_{\mathrm{sc}} \backslash \{x\}} P_{\mathrm{sc}} g_{x_i} \ell(x_i)$ is the cumulative interference from other clustered small cells except the signalling cell at x, and $I_{\mathrm{sm}} = \sum_{y_i \in \Phi_{\mathrm{sc}}} P_{\mathrm{mc}} h_{y_i} \ell(y)$ is the total interference from macro cells.

The amount of backhaul rate allocated to typical users are defined by the following policy.

Definition 7 (Backhaul Rate Splitting Policy)**.** *Following the hierarchical model, suppose that a typical user located at $z \in \mathbf{r}_z$ (typical macro cell user) is connected to the macro cell at $y \in \{\mathbf{r}_{\mathrm{mc}}\}$, and this macro cell is connected to the nearest central router at u. Then, the rate of backhaul link to the macro cell at y is given as*

$$R'_{\mathrm{mu}} \triangleq \frac{\gamma C_{\mathrm{cr}}}{\mathbb{E}\left[N_{\mathrm{mc}} N_{\mathrm{mu}}\right]}, \tag{5.8}$$

where $\gamma \in [0, 1]$ is a fraction of capacity allocated to the macro cells. In case of small cell, in a similar vein, a typical user located at $z \in \mathbf{s}_z$ is connected to the small cell at $x \in \{\mathbf{r}_{\mathrm{sc}}\}$ whose central router is at u. Then, the rate of backhaul link to the small cell is given as

$$R'_{\mathrm{su}} \triangleq \frac{(1 - \gamma)C_{\mathrm{cr}}}{\mathbb{E}\big[N_{\mathrm{sc}}N_{\mathrm{su}}\big]}. \tag{5.9}$$

By using the definitions of SIR and backhaul rates of typical users, we then define our main performance metric, namely delivery rate, as follows.

Definition 8 (Delivery Rate). *The delivery rate of a typical user connected to a macro cell is defined as*

$$R_{\mathrm{mu}} \triangleq \begin{cases} \tau_{\mathrm{mc}}, & if \ \log(1 + \mathrm{SIR}_{\mathrm{mu}}) > \tau_{\mathrm{mc}} \ and \ R'_{\mathrm{mu}} > \tau_{\mathrm{mc}}, \\ 0, & otherwise. \end{cases} \tag{5.10}$$

Similarly, the delivery rate of a typical user connected to a small cell is defined as

$$R_{\mathrm{su}} \triangleq \begin{cases} \tau_{\mathrm{sc}}, & if \ \log(1 + \mathrm{SIR}_{\mathrm{su}}) > \tau_{\mathrm{sc}} \ and \ R'_{\mathrm{su}} > \tau_{\mathrm{sc}}, \\ \tau_{\mathrm{sc}}, & if \ \log(1 + \mathrm{SIR}_{\mathrm{su}}) > \tau_{\mathrm{sc}} \ and \ f_z \in \Delta_x, \\ 0, & otherwise, \end{cases} \tag{5.11}$$

where f_z represents the content requested by the typical small cell user and Δ_x is the cache of the small cell.

Before stating our results for the average delivery rate, the following two lemmas are compiled from [202, 203] and will be used to derive expressions for average delivery rate.

Lemma 8. *Considering that Φ_{k-1} and Φ_k are two independent PPPs, the PDF of the distance of a link of any node in Φ_{k-1} to the nearest node in Φ_k is given by*

$$f_k(r) = 2\lambda_k \pi r \exp(-\pi\lambda_k r^2). \tag{5.12}$$

If $c(r)$ is a cost function of link distance r, then, the expected cost $\bar{c}$ and total expected cost $\bar{C}$ are expressed as

$$\bar{c} = \mathbb{E}_{\Phi_k}\big[c(r)\big] = \int_0^\infty c(r)f_k(r)\mathrm{d}r \tag{5.13}$$

$$\bar{C} = \mathbb{E}_{\Phi_{k-1}, \Phi_k}\bigg[\sum_{\Phi_{k-1}} c(r)\bigg] = \lambda_{k-1}\int_0^\infty c(r)f_k(r)\mathrm{d}r. \tag{5.14}$$

When the cost function $c(r)$ is expressed in the the form of ar^b, the expected cost $\bar{c}$ and total expected cost $\bar{C}$ are given by

$$\bar{c} = a\frac{\Gamma\big(\frac{b}{2} + 1\big)}{(\pi\lambda_k)^{b/2}} \tag{5.15}$$

$$\bar{C} = \lambda_{k-1}a\frac{\Gamma\big(\frac{b}{2} + 1\big)}{(\pi\lambda_k)^{b/2}}. \tag{5.16}$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ is the Gamma function.

$$B_1^{(\text{cov})} = \int_0^{R_\text{c}} e^{-\frac{(e^{\tau_\text{mc}}-1)}{P_\text{mc}r_\text{mc}^{-\alpha}}} \mathcal{L}_{I_\text{mm}}\Big(\frac{e^{\tau_\text{mc}}-1}{P_\text{mc}r_\text{mc}^{-\alpha}}\Big)\mathcal{L}_{I_\text{sm}}\Big(\frac{e^{\tau_\text{mc}}-1}{P_\text{mc}r_\text{mc}^{-\alpha}}\Big)\frac{k}{\nu}\Big(\frac{r_\text{mc}}{\nu}\Big)^{k-1}e^{-(r_\text{mc}/\nu)^k}\mathrm{d}r_\text{mc}$$

$$(5.20)$$

$$\mathcal{L}_{I_\text{mm}}(s) = \exp\Big(\frac{-s\pi\lambda_\text{mc}P_\text{mc}(2/\alpha)}{1-2/\alpha}r_\text{mc}^{2-\alpha}F\big(1,1-2/\alpha;2-2/\alpha;-sP_\text{mc}r_\text{mc}^{-\alpha}\big)\Big) \qquad (5.21)$$

$$\mathcal{L}_{I_\text{sm}}(s) = \exp\Big\{-\lambda_{\text{sc}'}\Big(\frac{(sP_\text{sc})^{2/\alpha}\pi^2(2/\alpha)}{\sin(\pi\frac{2}{\alpha})} - \pi R_\text{c}^2 A_\text{mc}(s,R_\text{c})\Big)\Big\} \qquad (5.22)$$

$$A_\text{mc}(s,R_\text{c}) = \frac{1}{\pi R_\text{c}^2}\int_0^{2\pi}\int_0^{r_\text{mc}\cos\varphi+\sqrt{R_\text{c}^2-r_\text{mc}^2\sin^2\varphi}}\frac{r\mathrm{d}r\mathrm{d}\varphi}{1+s^{-1}P_\text{sc}^{-1}r^\alpha} \qquad (5.23)$$

$$B_2^{(\text{cov})} = 1 - \exp\Big(-\frac{\tau_\text{mc}\lambda_\text{cr}\big(\lambda_\text{mc}+\lambda_{\text{sc}'}\exp(-\lambda_\text{mc}\pi R_\text{c}^2)\big)}{\mu\gamma\lambda_\text{mc}^2\lambda_\text{ut}}\Big) \qquad (5.24)$$

Lemma 9. *Considering that Φ_k and Φ_{k+1} are two independent PPPs and the nodes in Φ_k are linked to their nearest nodes in Φ_{k+1}, the average number of linked nodes from Φ_k to each node in Φ_{k+1} is then expressed as*

$$\lambda_k/\lambda_{k+1} \qquad (5.17)$$

We are now ready to give the expressions for average delivery rate of typical macro and small cell users.

Theorem 10 (Average Delivery Rate of Typical Macro Cell User). *The average delivery rate of a typical user connected to the nearest macro cell cell in coverage-aided deployment is approximately given by*

$$\bar{R}_\text{mu}^{(\text{cov})} \approx \tau_\text{mc}B_1^{(\text{cov})}B_2^{(\text{cov})} \qquad (5.18)$$

where $B_1^{(\text{cov})}$ and $B_2^{(\text{cov})}$ are given in (5.20) and (5.24) respectively. Therein, Laplace transforms and other related function definitions are given below $B_1^{(\text{cov})}$, and $F(x,y;z;w)$ is the hypergeometric function [207].

In case of capacity-aided deployment, the average delivery rate of this user is approximately given by

$$\bar{R}_\text{mu}^{(\text{cap})} \approx \tau_\text{mc}B_1^{(\text{cap})}B_2^{(\text{cap})} \qquad (5.19)$$

where $B_1^{(\text{cap})}$ and $B_2^{(\text{cap})}$ are given in (5.25) and (5.30) respectively.

Proof. See Appendix C.1. □

67

$$B_1^{(\text{cap})} = \int_0^{R_c} e^{-\frac{(e^{\tau_{mc}}-1)}{P_{mc}r_{mc}^{-\alpha}}} \mathcal{L}_{I_{mm}}\Big(\frac{e^{\tau_{mc}}-1}{P_{mc}r_{mc}^{-\alpha}}\Big) \mathcal{L}_{I_{sm}}\Big(\frac{e^{\tau_{mc}}-1}{P_{mc}r_{mc}^{-\alpha}}\Big)\frac{k}{\nu}\Big(\frac{r_{mc}}{\nu}\Big)^{k-1} e^{-(r_{mc}/\nu)^k}\mathrm{d}r_{mc} \quad (5.25)$$

$$\mathcal{L}_{I_{mm}}(s) = \exp\Big(\frac{-s\pi\lambda_{mc}P_{mc}(2/\alpha)}{1-2/\alpha}r_{mc}^{2-\alpha}F\big(1, 1-2/\alpha; 2-2/\alpha; -sP_{mc}r_{mc}^{-\alpha}\big)\Big) \quad (5.26)$$

$$\mathcal{L}_{I_{sm}}(s) = \exp\Big(-\lambda_{sc'}\int_{\mathbb{R}^2}\big(1-\exp(-\bar{c}\nu(s,y))\big)\mathrm{d}y\Big) \quad (5.27)$$

$$\nu(s,y) = \int_{\mathbb{R}^2}\frac{f(x)}{1+(sP_{sc}\ell(x-y))^{-1}}\mathrm{d}x \quad (5.28)$$

$$f(x) = \begin{cases} \frac{1}{\pi R_c^2}, & \text{if } \|x\| < R_c, \\ 0, & \text{otherwise.} \end{cases} \quad (5.29)$$

$$B_2^{(\text{cap})} = 1 - \exp\Big(-\frac{\tau_{mc}\lambda_{cr}}{\mu\gamma\lambda_{ut-m}}\Big) \quad (5.30)$$

Theorem 11 (Average Delivery Rate of Typical Small Cell User). *The average delivery rate of the typical user connected to the nearest small cell in coverage-aided deployment is approximately given by*

$$\bar{R}_{su}^{(\text{cov})} \approx \tau_{sc}C_1^{(\text{cov})}C_2^{(\text{cov})} + \tau_{sc}C_1^{(\text{cov})}C_3^{(\text{cov})} - \tau_{sc}C_1^{(\text{cov})}C_2^{(\text{cov})}C_3^{(\text{cov})} \quad (5.31)$$

where $C_1^{(\text{cov})}$, $C_2^{(\text{cov})}$ and $C_3^{(\text{cov})}$ are given in (5.33), (5.37) and (5.38) respectively. Therein, Laplace transforms and other related function definitions are given below $C_1^{(\text{cov})}$, and $F(x, y; z; w)$ is the hypergeometric function [207].

In case of capacity-aided deployment, the average delivery rate of this user is approximately given by

$$\bar{R}_{su}^{(\text{cap})} \approx \tau_{sc}C_1^{(\text{cap})}C_2^{(\text{cap})} + \tau_{sc}C_1^{(\text{cap})}C_3^{(\text{cov})} - \tau_{sc}C_1^{(\text{cap})}C_2^{(\text{cap})}C_3^{(\text{cap})} \quad (5.32)$$

where $C_1^{(\text{cap})}$, $C_2^{(\text{cap})}$ and $C_3^{(\text{cap})}$ are given in (5.39), (5.43) and (5.44) respectively.

Proof. See Appendix C.2. □

5.3.1 Validation of the Proposed Model

In this subsection, we validate our expressions for average delivery rate via Monte-Carlo simulations. The list of simulation parameters for coverage and capacity-aided deployments are given in Tables 5.1 and 5.2 respectively. These parameter values shall be used throughout this section unless otherwise stated. In the following, we investigate the impact of important parameters on the delivery rate, namely 1) target bitrate, 2) rate splitting ratio, 3) storage size.

$$C_1^{(\mathrm{cov})} = \int_0^{R_{\mathrm{c}}} e^{-\frac{(e^{\tau_{\mathrm{sc}}}-1)}{P_{\mathrm{sc}} r_{\mathrm{sc}}^{-\alpha}}} \mathcal{L}_{I_{\mathrm{ss}}}\Big(\frac{e^{\tau_{\mathrm{sc}}}-1}{P_{\mathrm{sc}} r_{\mathrm{sc}}^{-\alpha}}\Big) \mathcal{L}_{I_{\mathrm{ms}}}\Big(\frac{e^{\tau_{\mathrm{sc}}}-1}{P_{\mathrm{sc}} r_{\mathrm{sc}}^{-\alpha}}\Big) \frac{k}{\nu}\Big(\frac{r_{\mathrm{sc}}}{\nu}\Big)^{k-1} e^{-(r_{\mathrm{sc}}/\nu)^k} \mathrm{d}r_{\mathrm{sc}} \tag{5.33}$$

$$\mathcal{L}_{I_{\mathrm{ss}}}(s) = \exp\Big(\frac{-s\pi\lambda_{\mathrm{sc}'} P_{\mathrm{sc}}(2/\alpha)}{1-2/\alpha} r_{\mathrm{sc}}^{2-\alpha} F\big(1, 1-2/\alpha; 2-2/\alpha; -sP_{\mathrm{sc}} r_{\mathrm{sc}}^{-\alpha}\big)\Big) \tag{5.34}$$

$$\mathcal{L}_{I_{\mathrm{ms}}}(s) = \exp\Big\{-\lambda_{\mathrm{mc}}\Big(\frac{(sP_{\mathrm{mc}})^{2/\alpha}\pi^2(2/\alpha)}{\sin(\pi\frac{2}{\alpha})} - \pi R_{\mathrm{c}}^2 A_{\mathrm{sc}}(s, R_{\mathrm{c}})\Big)\Big\} \tag{5.35}$$

$$A_{\mathrm{sc}}(s, R_{\mathrm{c}}) = \frac{1}{\pi R_{\mathrm{c}}^2} \int_0^{2\pi} \int_0^{r_{\mathrm{sc}}\cos\varphi + \sqrt{R_{\mathrm{c}}^2 - r_{\mathrm{sc}}^2 \sin^2\varphi}} \frac{r\mathrm{d}r\mathrm{d}\varphi}{1 + s^{-1} P_{\mathrm{mc}}^{-1} r^\alpha} \tag{5.36}$$

$$C_2^{(\mathrm{cov})} = 1 - \exp\Big(-\frac{\tau_{\mathrm{sc}}\lambda_{\mathrm{cr}}(\lambda_{\mathrm{mr}} + \lambda_{\mathrm{sc}})}{\mu\gamma\lambda_{\mathrm{sc}}^2\lambda_{\mathrm{ut}}}\Big) \tag{5.37}$$

$$C_3^{(\mathrm{cov})} = 1 - \big(1 + F_{\mathrm{sc}}\big)^{1-\eta} \tag{5.38}$$

$$C_1^{(\mathrm{cap})} = \int_0^{R_{\mathrm{c}}} e^{-\frac{(e^{\tau_{\mathrm{sc}}}-1)}{P_{\mathrm{sc}} r_{\mathrm{sc}}^{-\alpha}}} \mathcal{L}_{I_{\mathrm{ss}}}\Big(\frac{e^{\tau_{\mathrm{sc}}}-1}{P_{\mathrm{sc}} r_{\mathrm{sc}}^{-\alpha}}\Big) \mathcal{L}_{I_{\mathrm{ms}}}\Big(\frac{e^{\tau_{\mathrm{sc}}}-1}{P_{\mathrm{sc}} r_{\mathrm{sc}}^{-\alpha}}\Big) \frac{k}{\nu}\Big(\frac{r_{\mathrm{sc}}}{\nu}\Big)^{k-1} e^{-(r_{\mathrm{sc}}/\nu)^k} \mathrm{d}r_{\mathrm{sc}} \tag{5.39}$$

$$\mathcal{L}_{I_{\mathrm{ss}}}(s) = \exp\Big(-\lambda_{\mathrm{sc}'} \int_{\mathbb{R}^2} \big(1 - \exp(-\bar{c}\nu(s, x))\big)\mathrm{d}x\Big) \int_{\mathbb{R}^2} \big(\exp(-\bar{c}\nu(s, x))\big) f(x)\mathrm{d}x \tag{5.40}$$

$$\nu(s, x) = \int_{\mathbb{R}^2} \frac{f(y)}{1 + (sP_{\mathrm{sc}}\tilde{\ell}(y-x))^{-1}}\mathrm{d}y \tag{5.41}$$

$$\mathcal{L}_{I_{\mathrm{ms}}}(s) = \exp\Big(-\lambda_{\mathrm{mc}}\frac{(sP_{\mathrm{mc}})^{2/\alpha}\pi^2(2/\alpha)}{\sin(\pi\frac{2}{\alpha})}\Big) \tag{5.42}$$

$$C_2^{(\mathrm{cap})} = 1 - \exp\Big(-\frac{\tau_{\mathrm{mc}}\lambda_{\mathrm{cr}}}{\mu\gamma\lambda_{\mathrm{ut}-\mathrm{s}}}\Big) \tag{5.43}$$

$$C_3^{(\mathrm{cap})} = 1 - \big(1 + F_{\mathrm{sc}}\big)^{1-\eta} \tag{5.44}$$

Table 5.1: Simulation Parameters for Coverage-aided Deployment.

Parameters	Values
λ_{cr}, λ_{mc}, $\lambda_{sc'}$, λ_{ut}	1.0×10^{-5}, 1.5×10^{-5}, 5.5×10^{-5} and 12.8×10^{-5} unit/m^2
P_{mc}, P_{sc}	16 and 3 Watt
τ_{mc}, τ_{sc}	4 bits/s/Hz
α	4
R_c	80 meters
μ, γ	30 bits/s/Hz, 0.6
f_0, F_{sc}, η	500 Gbye, 4 GByte and 1.45

Table 5.2: Simulation Parameters for Capacity-aided Deployment.

Parameters	Values
λ_{cr}, λ_{mc}, $\lambda_{sc'}$, λ_{ut-m}	1.0×10^{-5}, 1.5×10^{-5}, 1.5×10^{-5}, 3.0×10^{-5} unit/m^2
$\hat{c}$	3 units
P_{mc}, P_{sc}	16, 3 Watt
τ_{mc}, τ_{sc}	4 bits/s/Hz
α	4
R_c	80 meters
μ, γ	30 bits/s/Hz, 0.6
f_0, F_{sc}, η	500 Gbye, 4 GByte and 1.45

Impact of target bitrate τ: The target bitrate τ (namely τ_{mc} or τ_{sc} depending on the cell type) is a crucial parameter for QoS. The impact of this parameter on the average delivery rate is given in Figures 5.3a and 5.4b for coverage and capacity-aided deployments respectively. In these figures, one can see that theoretical and simulation curves are following similar trends. In fact, the theoretical curves based on approximations are pretty loose in low bitrates. However, in high bitrate values, approximation and simulation curves match reasonably well. Note that the target bitrate is dictated both for downlink and backhaul links in our system model. Therefore, a concave behaviour of average delivery rate appears in these plots, which in turn induces target bitrate to be

70

set carefully for maximum average delivery rate. In this parameter setting, the average delivery rate of typical macro cell user is generally outperforming rate of typical small cell user with/without caching capabilities at small cells. However, the decreasing trend of average delivery rate for macro cell user is higher than the small cell user case, thus, yielding small cell users to perform better in higher target bitrates. The typical small cell user with no caching capabilities has the lowest performance.

Impact of rate splitting ratio γ: In fact, higher value of backhaul rate splitting ratio in central routers corresponds to a higher backhaul capacity dedication for macro cells. To show this, the change of average delivery rate with respect to the backhaul rate splitting ratio is given in Figures 5.3b and 5.4b for coverage and capacity-aided deployments respectively. Indeed, as seen from figures, a dramatical increase in average delivery rate occurs which in turn confirms our intuitions. The rate of decrement for the typical small cell user is relatively slow compared to the typical macro cell user. In order to find a balance between average delivery rate of typical macro and small cell users, for fairness, the plots show that one has to set the rate splitting ratio carefully. In all of these cases, we observe that having caching capabilities at small cells improves system performance in terms of average delivery rate. In other words, a heterogeneous network consists of macro cells and cache-enabled allows higher average delivery rates while having fairness between users at different tiers.

Impact of storage size F_{sc}: The variation of average delivery rate with respect to the storage size is given in Figures 5.3c and 5.4c for coverage and capacity-aided deployments respectively. In this parameter settings, increasing storage size of small cells both in coverage and capacity-aided deployments yields higher average delivery rates. On the other hand, the increment of storage size in coverage-aided deployment is more visible compared to capacity-aided deployment and allows typical small cell users to achieve higher rates than typical macro cell users. Given the fact that caching is monotonically improving overall system performance of small cell users, high values of storage size does not seem necessary if one considers a linear cost for storage size, even though more storage is desirable for improving average delivery rate.
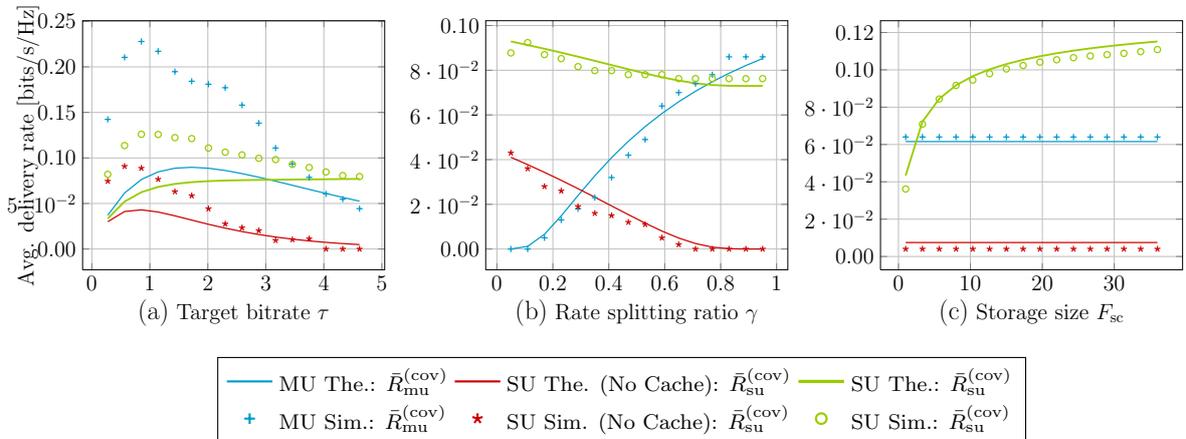
Figure 5.3: Evolution of average delivery rate in coverage-aided deployment.

Figure 5.4: Evolution of average delivery rate in capacity-aided deployment.

5.4 Geographical Caching Methods

This section covers geographical caching methods from a global cost perspective. We define required structures and pose our optimization problem. Investigations on optimal cache allocation in our scenario is left for future work.

We start by constructing cost function which captures the cost of deployment as well as cost of operating macro cell users and small cell users from caches or backhaul. Recall the hierarchical model in Section 5.2.5 and assume that an interference-avoidance scheme is employed among cells (thus we consider SNR). Then, we suppose that the following general form for the cost function, that is

$$\mathsf{C} = \sum_{u \in \Phi_{\mathrm{cr}}} C(u, \mathbf{v}_u) \tag{5.45}$$

where $C(u, \mathbf{v}_u)$ is given by

$$\underbrace{c_{\mathrm{cr}} + c_{\mathrm{mc}} N_{\mathrm{mc}} + c_{\mathrm{sc}} N_{\mathrm{sc}}}_{(i)} + \underbrace{C_{\mathrm{mc}} + C_{\mathrm{sc}}}_{(ii)}. \tag{5.46}$$

Therein, the term (i) accounts for deployment costs with c_{cr} representing the unit cost of deploying central router, c_{mc} and c_{sc} are the unit costs of deploying macro and small cells with their backhaul links respectively. The term (ii) is for operational costs of macro and small cells. We now focus on the terms C_{mc} and C_{sc} respectively.

For macro cells, assuming that the unit cost of downlink transmission is inversely proportional to the received SNR of users (for a given fixed transmit power), we define a vector $\mathbf{w}_y \in \mathbb{R}^{1 \times N_{\mathrm{mu}}}$ which represents the unit cost of wireless transmission for each macro cell at y. Therein, each element is defined as $w_{y,z} = \frac{\sigma^2}{h_{y,z} \|r_{y,z}\|^{-\alpha}}$ where σ^2 is the noise power. On the other hand, the unit backhaul cost for macro cell is denoted by $\mathbf{b}_y \in \mathbb{R}^{1 \times N_{\mathrm{mu}}}$ where

each element $b_{u,y}$ is proportional to the distance such as $b_{u,y} = a_{\mathrm{mc}} \times (r_{u,y})^{b_{\mathrm{mc}}}$ with arbitrary design coefficients a_{mc} and b_{mc}. Therefore, combining the unit cost of downlink usage $\mathbf{w}_y$, unit backhaul cost $\mathbf{b}_y$ and the observed demand $\mathbf{d}_y$, we have the following total cost for macro cells, that is

$$C_{\mathrm{mc}} = \sum_{y \in \{\mathbf{r}_{\mathrm{mc}}\}} \left(\mathbf{w}_y \mathbf{d}_y^T + \mathbf{b}_y \mathbf{d}_y^T \right). \tag{5.47}$$

Now, we construct the cost function for small cells, which captures the cost of serving all small cell users in addition to the cost of taking caching decisions at the small cells. Depending on which content is stored and which one is going to be demanded, four different cases might be integrated into the cost function as follows:

1. *Part of the content in the catalogue is* **cached** *and is going to be* **demanded***:* The content has to be prefetched via the backhaul only one time, thus the induced cost is expressed by

$$\mathbf{s}_x \mathbf{e}_x^T \tag{5.48}$$

where $\mathbf{s}_x$ is the cache indicator (allocation) vector for small cell $x \in \{\mathbf{r}_{\mathrm{sc}}\}$ and $\mathbf{e}_x \in \{0,1\}^{1 \times F}$ is the prefetching cost vector for the same small cell. The entries of $\mathbf{e}_x$ are evaluated as follows

$$e_{x,f} = \begin{cases} a_{\mathrm{sc}} \times (r_{u,x})^{b_{\mathrm{sc}}} & \text{if } d_{x,f} > 0 \\ \\ 0 & \text{otherwise} \end{cases} \tag{5.49}$$

where a_{sc} and b_{sc} are the arbitrary design coefficients.

2. *Part of the content in the catalogue is* **cached** *but is* **not** *going to be* **demanded***:* Prefetching the content induces unnecessary usage of backhaul, thus the cost is given by

$$\mathbf{s}_x \bar{\mathbf{e}}_x^T \tag{5.50}$$

where $\bar{\mathbf{e}}_x$ is the complementary of $\mathbf{e}_x$ with the entries defined such as

$$\bar{e}_{x,f} = \begin{cases} 0 & \text{if } d_{x,f} > 0 \\ \\ a_{\mathrm{sc}} \times (r_{u,x})^{b_{\mathrm{sc}}} & \text{otherwise.} \end{cases} \tag{5.51}$$

3. *Part of the content is* **not cached** *but is going to be* **demanded***:* The content demand is going to be satisfied each time via the backhaul due to its non-availability in the cache, that is to say

$$(\mathbf{1} - \bar{\mathbf{s}}_x) \tilde{\mathbf{e}}_x^T, \tag{5.52}$$

where $\mathbf{1}$ is the all one vector with F elements. The vector $\tilde{\mathbf{g}}_x$ is the cost of serving the contents from backhaul with entries defined as

$$
\tilde{e}_{x,f} = \begin{cases} a_{\mathrm{sc}} \times (r_{u,x})^{b_{\mathrm{sc}}} \times d_{x,f} & \text{if } d_{x,f} > 0 \\[1.5em] 0 & \text{otherwise.} \end{cases} \tag{5.53}
$$

4. *Part of the content in the catalog is* **not cached** *and is* **not** *going to be* **demanded**: The cost of using backhaul in this case is zero.

Now, again assuming that the unit cost of downlink transmission is proportional to the received SNR of small cell users (similar to macro cell case), we denote this cost as $\mathbf{w}_x$ where each entry $w_{x,z} = \frac{\sigma^2}{g_{x,z} \| r_{x,z} \|^{-\alpha}}$. Combining cost of backhaul usage and wireless transmission, we have

$$
C_{\mathrm{sc}} = \sum_{x \in \{\mathbf{r}_{\mathrm{sc}}\}} \left(\mathbf{s}_x \mathbf{e}_x^T + \mathbf{s}_x \bar{\mathbf{e}}_x^T + (\mathbf{1} - \mathbf{s}_x) \tilde{\mathbf{e}}_x^T + \mathbf{w}_x \mathbf{d}_x^T \right). \tag{5.54}
$$

Therefore, having defined C_{mc} and C_{sc}, the expression $C(u, \mathbf{v}_u)$ is given by

$$
C(u, \mathbf{v}_u) = c_{\mathrm{cr}} + c_{\mathrm{mc}} N_{\mathrm{mc}} + c_{\mathrm{sc}} N_{\mathrm{sc}} + \sum_{y \in \{\mathbf{r}_{\mathrm{mc}}\}} \left(\mathbf{w}_y \mathbf{d}_y^T + \mathbf{b}_y \mathbf{d}_y^T \right)
$$
$$
+ \sum_{x \in \{\mathbf{r}_{\mathrm{sc}}\}} \left(\mathbf{s}_x \mathbf{e}_x^T + \mathbf{s}_x \bar{\mathbf{e}}_x^T + (\mathbf{1} - \mathbf{s}_x) \tilde{\mathbf{e}}_x^T + \mathbf{w}_x \mathbf{d}_x^T \right). \tag{5.55}
$$

Finally, the optimal cache allocation which minimizes average total cost is formulated as

$$
\begin{aligned}
\underset{\forall \mathbf{s}_x}{\text{minimize}} \quad & \mathbb{E}\left[\sum_{u \in \Phi_{\mathrm{cr}}} C(u, \mathbf{v}_u) \right] \\
\text{subject to} \quad & \|\mathbf{s}_x\|_1 \leq F_{\mathrm{sc}}.
\end{aligned} \tag{5.56}
$$

where F_{sc} is the storage size of small cell x. The problem above is mixed-integer problem as each entry of $\mathbf{s}_x$ takes binary values (namely 0 or 1). We use three different distributed caching policies to solve the problem for each given realization of the topology (namely random hierarchical trees), fading and users' demand. In particular:

- *AppCache*: Each small cell solves the convex relaxation of the problem under the knowledge of demand drawn from content popularity distribution. In other words, a small cell at x fills $\mathbf{s}_x$ for given $\mathbf{d}_x$ and storage size F_{sc} by solving relaxed problem in which the entries of $\mathbf{s}_x$ take values in the interval $[0 - 1]$ and apply rounding. In general, as long as the cache allocation problem is formulated (either convex or non-convex), various approximation methods can be applied for the solution.

- *PopCache*: Each small cell caches the most popular contents under the knowledge of its local demand and storage size. Therefore, the vector $\mathbf{s}_x$ is filled according to the knowledge of $\mathbf{d}_x$ and storage constraint F_{sc}. Similar policies can be found in [37].

- *UniCache*: Contents are cached uniformly at random until the storage size F_{sc} is fulfilled. This policy does not require the knowledge of users' demand. Similar policies can be found in [132].

Given these caching policies, following two storage deployment policies can be applied:

- *Scaled*: Storage size of each small cell is fixed to a constant value, thus the total storage size in the deployment increases with the number of deployed small cells.

- *Fixed*: For a given fixed total storage budget F_{tot} in each hierarchical random tree $\{(u, \mathbf{v}_x)\}$, the storage size of each small cell is inversely proportional to the total number of deployed small cells, meaning that $F_{sc} = \frac{F_{tot}}{N_{sc}}$.

5.5   Closing Remarks

In this chapter, we have modeled a heterogeneous network which consists of macro cells and clustered small base stations. The clustering processes of small cells allowed us to model the deployments of these small cells in a more realistic manner. Approximations of average delivery have been provided and validated via numerical results. Given the fact that adding more storage to the small cell is more desirable from average delivery rate point of view, we have showed that the critical system parameters needs to be adjusted carefully. In other words, target bitrate, backhaul rate splitting rate and storage size can be chosen in a level so that fairness between small/macro users and deployment cost in this heterogeneous network can be sustained.

Part II

Content Popularity Learning and Algorithmic Aspects

Chapter 6

Proactive Caching

6.1 Overview

The fact that the huge amount of users' information is often available and the human behaviour has a certain predictability [6], users' future events can be inferred. Therefore, in this chapter, we explore such a proactive caching framework by leveraging context-awareness and storage capabilities at the edge of the network in order to sustain peak data demands and offload the backhaul. More precisely, estimating users' future demands and content popularity can be used to proactively store the content before the actual requests take place. In addition, whenever a D2D communication is available, the proactive caching approach exploits users' social relationships (and their influence within the social community), as well as users' storage for content dissemination and physical proximity.

Rest of this chapter is organized as follows. Our system model and corresponding problem formulation is presented in Section 6.2. The details of proactive caching at the SBSs and UTs are given in Sections 6.3 and 6.4 respectively and discussions of numerical results are carried out in the same sections. Finally, Section 6.5 draws some conclusions.

6.2 System Model

Let us consider a scenario that consists of M SBSs $\mathcal{M} = \{1, ..., M\}$ and N UTs $\mathcal{N} = \{1, ..., N\}$. The broadband connection of every SBS $m \in \mathcal{M}$ is provided by a CS via a limited backhaul link with capacity c_m.[1] We suppose that the capacity of the wireless small cell link between SBS m and UT n is given by $c_{m,n}$. Depending on the content availability and users' proximity, the SBSs can establish D2D communications between users n and n', whereas the corresponding D2D link capacity is denoted by $\ddot{c}_{n,n'}$. This scenario is illustrated in Fig. 6.1. Suppose that user n requests a content from a library of

---

[1]This controller is typically a network entity located at the evolved packet core (EPC) or at the network edge (small cell gateway).
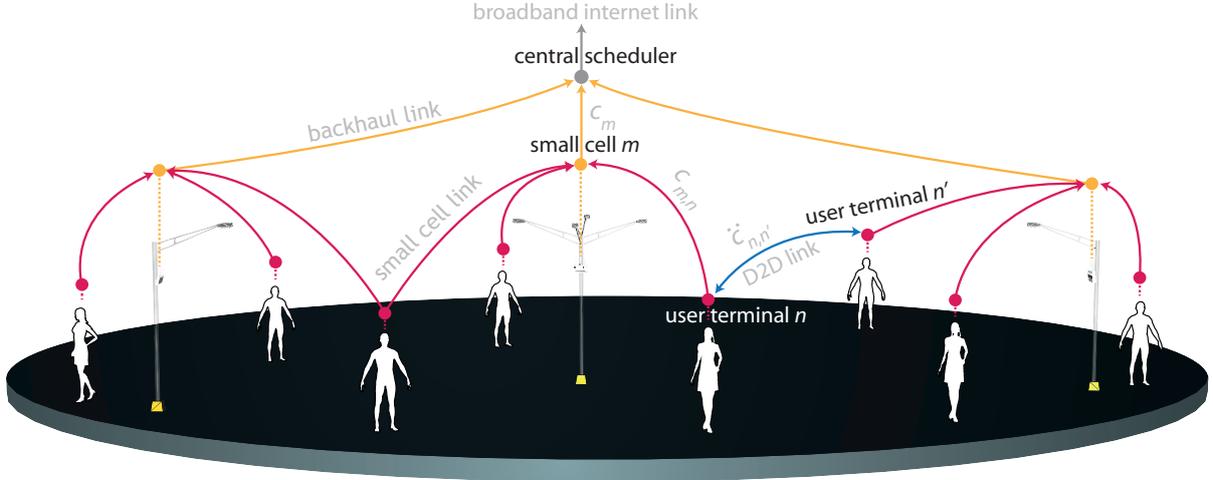
Figure 6.1: A sketch of the scenario given in the system model. A central scheduler is in charge of providing broadband connection to M SBSs via backhaul links. Depending on the users' contents availability in the caches of SBS and UTs, the SBSs serve their user either via wireless small cell links or D2D communications.

F contents, represented by $\mathcal{F} = \{1, ..., F\}$, according to probabilities $\mathcal{P}_n = \{p_{n,1}, ..., p_{n,F}\}$. In this library, the length of contents are $\mathcal{L} = \{l_1, ..., l_F\}$ and the bitrate are given by set of $\mathcal{B} = \{b_1, ..., b_F\}$. Now, suppose that R number of content requests are drawn by users randomly during T time slots. Then, we say that a request $r \in \mathcal{R} = \{1, ..., R\}$ is *satisfied* if the rate of delivery is equal or greater than the bitrate of the requested content as follows:

$$\frac{l_r}{t'_r - t_r} \geq b_r, \tag{6.1}$$

where $l_r \in \mathcal{L}$ represent the length of the requested content, t_r (t'_r) is the start (end) time of the delivery, and $b_r \in \mathcal{B}$ is the bitrate of the content $f_r \in \mathcal{F}$. Given this definition, the *satisfaction ratio* can be expressed as:

$$\eta(\mathcal{R}) = \frac{1}{R} \sum_{r \in \mathcal{R}} \mathbb{1} \left\{ \frac{l_r}{t'_r - t_r} \geq b_r \right\}, \tag{6.2}$$

where $\mathbb{1}\{...\}$ is the indicator function which yields 1 when the condition holds and 0 otherwise.

Our target as the network operator is to keep the satisfaction ratio above a threshold, while minimizing the usage of the backhaul. As stated before, this can be done via proactive caching in SBSs and UTs, in which we detail these two case studies separately in the following sections.

6.3 Proactive Caching at Base Stations

Results have shown that the backhaul constitutes one of the most important challenges for SCN deployments and this is going to increase dramatically due to the densely deployed SBSs. From this observation, suppose that the total capacity of the backhaul is lower than the available wireless link capacity between SBSs and UTs, such as $\sum_{m \in \mathcal{M}} c_m \ll \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} c_{m,n}$. Since in this case we suppose that the backhaul is the bottleneck, one reasonable option is to avoid its usage by storing the users' contents proactively at the SBSs, during peak-off hours. In other words, if the users' contents can be stored at SBSs before the users' actual contents arrives, the backhaul will not be used for a certain level, depending on on how smartly the content is placed.

Let us consider that the rate of the backhaul link during the content delivery for request r at time t is $\lambda_r(t)$. Then, the *backhaul load* under given these definitions can be expressed as follows:

$$\rho(\mathcal{R}) = \frac{1}{R} \sum_{r \in \mathcal{R}} \frac{1}{l_r} \sum_{t=t_r}^{t=t_r'} \lambda_r(t). \tag{6.3}$$

Additionally, suppose that the storage capacity of SBS m is given by s_m and the amount of its consumption at time t is denoted by $\kappa_m(t)$. Hence, the backhaul minimization problem subject to the link capacities, storage and QoS constraints can be formulated as follows:

$$\begin{aligned}
\underset{t_r', r \in \mathcal{R}}{\text{minimize}} \qquad & \rho(\mathcal{R}) & \tag{6.4} \\
\text{subject to} \qquad & \lambda_r(t) \leq c_m, & \forall m \in \mathcal{M}, \\
& \kappa_m(t) \leq s_m, & \forall m \in \mathcal{M}, \\
& \eta(\mathcal{R}) \geq \eta_{min}, & \forall r \in \mathcal{R},
\end{aligned}$$

where η_{min} is the target satisfaction ratio. Since dealing with (6.4) is computationally intractable, a heuristic approach similar to the one in [39] can be performed by storing users' popular content in the cache of SBSs. Before such a caching procedure is applied, we suppose that each SBS m has to track, learn and build its user' content profile to infer their future demands. Assume that $\mathbf{P}_m$ is the discrete content probabilities of users in SBS m in which we refer as *popularity matrix*, each row representing the users and columns are content popularities/ratings. Indeed, a perfectly known $\mathbf{P}_m$ could easily allow us to store the content according to this caching procedure. Unfortunately, this situation in practice is not the case, in which the matrix is not perfectly known, large and indeed sparse. Given these observations and inspired from the *Netflix paradigm* [208], supervised machine learning tools can be used to exploit users-content correlations. Inferring the probability that user n requests content f (namely estimating the popularity matrix), and storing the predicted content accordingly can clearly offload the backhaul.

The proposed proactive caching procedure is composed of training and placement steps. The first step is the training step in which each SBS m builds a model for the

popularity matrix $\mathbf{P}_m$ based on the available information. The estimation of $\mathbf{P}_m$ boils down to solving a least square minimization problem as follows:

$$\min_{\{b_n, b_f\}} \sum_{n,f} \left(r_{nf} - \hat{r}_{nf} \right)^2 + \lambda \left(\sum_n b_n^2 + \sum_f b_f^2 \right), \tag{6.5}$$

where the sum is over the (n, f) user/content pairs in the training set, containing how user n rated content f (i.e., r_{nf}). The total number of users in the training set is N and F is the total number of contents, thus, the minimization is done over all the $N + F$ parameters. In this formulation, $\hat{r}_{nf} = \bar{r} + b_n + b_f$ is the baseline estimator where b_f is the relative quality of each content f compared to the average $\bar{r}$. The bias of each user n relative to b_n is given by $\bar{r}$. Additionally, the parameter λ is used for balancing the regularization and fitting the training data.

In the numerical setup, we use the regularized singular value decomposition (SVD) due to its numerical accuracy (see [209] for comprehensive study of CF methods). Roughly speaking, since the entries of $\mathbf{P}_m$ are not fully known, the model construction is done via gradient descent by using the least-squares property of the SVD. Thus, $\hat{\mathbf{P}}_m$ is constructed as the low rank version of $\mathbf{P}_m$.

So far, we have described the first step. In the last step (namely, the placement step of the caching procedure), the content is cached proactively by storing the most popular content based on the estimation of $\hat{\mathbf{P}}_m$, until the storage capacity is fulfilled. In the following, we show the gains of proactive caching in a numerical setup and discuss the impact of various parameters of interest. A sketch of the proactive caching procedure at the base stations is summarized in Fig. 6.2.

6.3.1 Numerical Results and Discussions

The list of parameters used in the numerical study is provided in Table 6.1. In order to see the impact of the parameters of interest, the length and bitrate of the content, wireless small cell links and storage capacities are set to the identical values. We consider three regimes of interest: (i) low load, (ii) medium load, and (iii) high load.

In the numerical study, R number of requests are drawn over a time duration T, given the fact that the arrival times of these requests are sampled uniformly at random. The users' content requests are drawn from the ZipF(α) distribution. Given that knowledge, at $t = 0$, the perfect popularity matrix $\mathbf{P}_m$ is constructed for each SBS m. Removing 20% of the entries of this matrix uniformly at random, the remaining entries are used for the model construction in CF. The prediction of missing entries are then carried out by the regularized SVD [210]. Once the popularity matrix is estimated, the proactive caching is applied by greedily storing the most popular content subject to the storage size of the SBS. In the numerical setup, afer completing the training and placement steps of the proactive procedure at $t = 0$, the users' are served depending on their request arrival time until all content delivery processes finish. We use random caching as a baseline referred to as *reactive*.
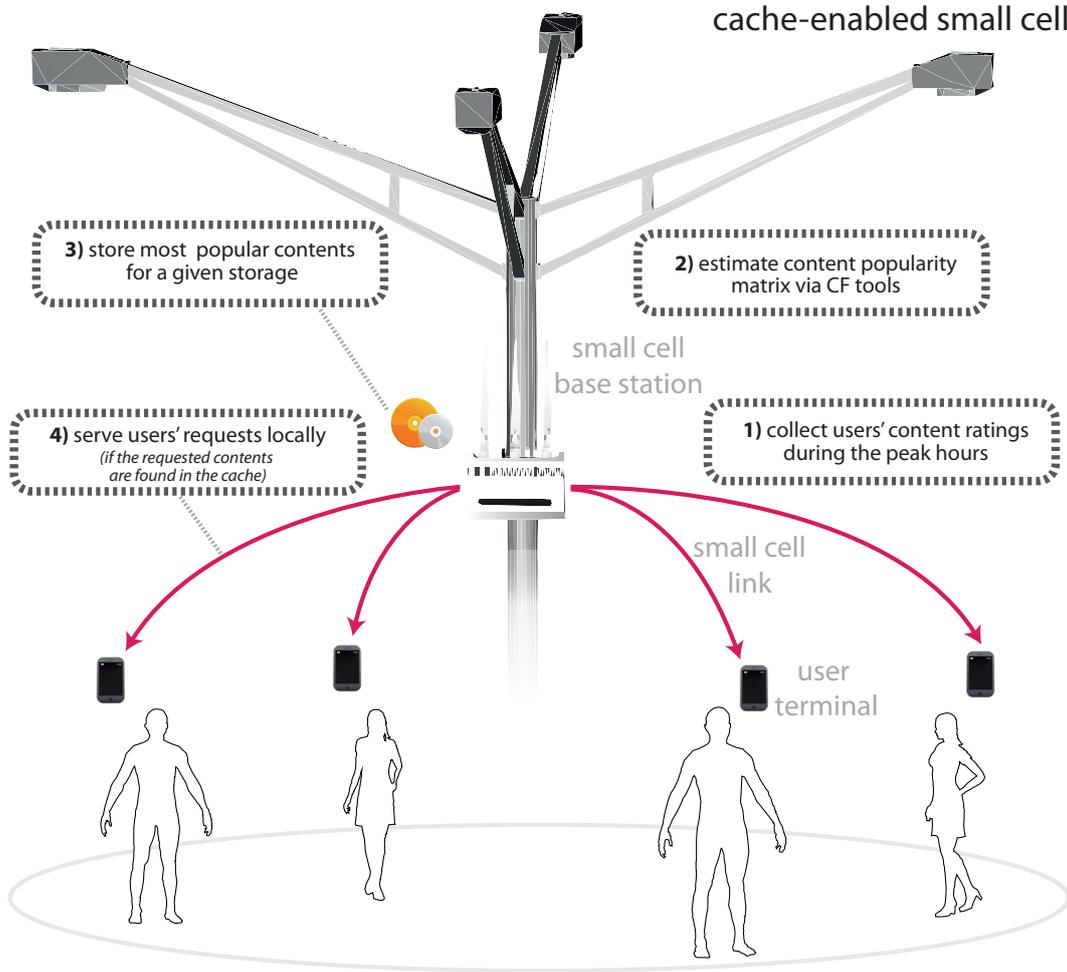
Figure 6.2: A practical procedure for proactive caching at the base stations.

In order to compare the benefits of caching both for proactive and reactive cases, three parameters of interest are detailed: (i) number of requests R, (ii) total cache size S, and (iii) ZipF distribution parameter α. The gains in the plots are normalized for ease of understanding. The evolution of the satisfaction ratios and the backhaul loads with respect to the variation of these parameters are given in Fig. 6.3.

In the figures, we see that the satisfaction ratio decreases as the number of users' content requests increases. The reason is somewhat obvious as the capacity constraints starts to be limiting factor for the delivery of high amount of requests. Concerning the backhaul load in very small number of requests, the reactive approach is generating less load compared to the proactive case which can be explained by *cold start* phenomenon of the CF used in the proactive case. However, as the number of request increases, the amount of information given to the CF for training step increases. Therefore, in the end, the proactive approach with sufficient amount of information outperforms the reactive approach with an almost constant gain.

Table 6.1: The numerical setup parameters for proactive caching at the SBSs.

Parameter	Description	Value
T	Time slots	1024 seconds
M	Number of SBSs	4
N	Number of UTs	32
F	Number of contents	128
l_f	Length of content f	1 Mbit
b_f	Bitrate of content f	1 Mbit/s
$\sum_m c_m$	Total backhaul link capacity	2 Mbit/s
$\sum_m \sum_n c_{m,n}$	Total wireless small cell link capacity	64 Mbit/s
R	Number of requests	$0 \sim 2048$
S	Total cache size	$0 \sim l_f \times F$
α	ZipF parameter	$0 \sim 2$

One important parameter of interest in our scenario is the total storage size of SBSs. As we increase the storage, the SBSs gain more capability to store the content from the catalog, yielding the satisfaction ratio up to 1 and backhal load up to 0 in the extreme values of the storage size. Looking at more practical situations in which the storage size is somewhere between 0 and 1, we see that the proactive approach outperforms the reactive case in terms of the satisfaction ratio as well as the backhaul load.

The content popularity parameter α indeed has an impact on the performance metrics. In the low values of α where the distribution follows a uniform behaviour, the proactive approach outperforms the reactive case with a relatively low difference. However, as the α increases, a few amount of content become highly popular than the rest of the content in the catalog. Thus, the difference between the gain of proactive and reactive approaches become quite visible in terms of the satisfaction ratio and the backhaul load.

6.4 Proactive Caching at User Terminals

Yet another mean of offloading the traffic at SBSs (thus, offloading the backhaul as a consequence) can be achieved by caching users' contents at the UTs and exploiting D2D communications for content dissemination. For this purpose, the interplay between users' social ties and physical proximity can be taken into account for proactive caching decision.
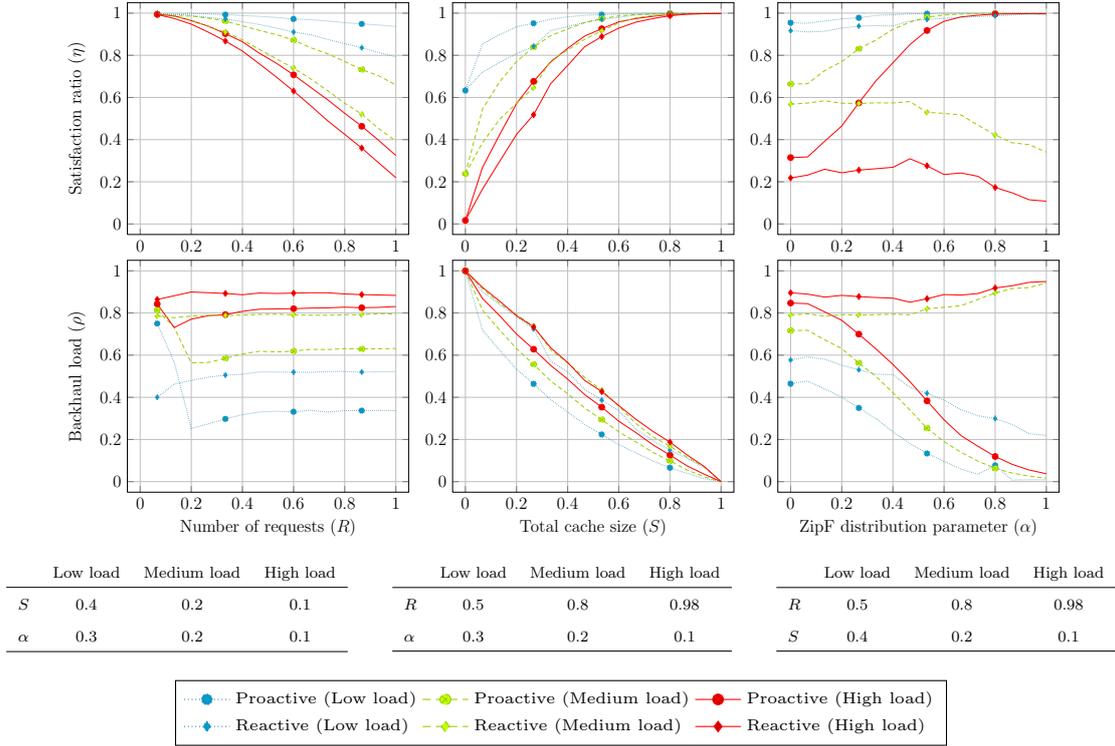
Figure 6.3: Backhaul Offloading via Proactive Caching: Dynamics of the satisfied requests and backhaul load with respect to the number of requests, total cache size and ZipF parameter.

In particular, when a content request arrives to the network, the SBS can take benefit of the influential users who have the content, requesting them to join the content delivery via D2D opportunities. If such a opportunity does not exist and the requested content is not available, as a last resort, the content can be delivered by the SBS but with the cost of using the backhaul.

Let us consider that the storage capacity of UT n is $\ddot{s}_n$ and its usage at time t is given by $\ddot{\kappa}(t)$. Also suppose that $\dot{\lambda}_r(t)$ is the total rate of the SBSs during the content delivery of request r at time t and the D2D link rate is $\ddot{\lambda}_r(t)$. Then, *small cell load* can be expressed as follows:

$$\ddot{\rho}(\mathcal{R}) = \frac{1}{R} \sum_{r \in \mathcal{R}} \sum_{t=t_r}^{t=t_r'} \frac{\dot{\lambda}_r(t)}{\dot{\lambda}_r(t) + \ddot{\lambda}_r(t)}. \tag{6.6}$$

Given that definition and using a formulation similar to (6.4), the D2D caching opti-

mization problem can be written as:

$$\underset{t'_r, r \in \mathcal{R}}{\text{minimize}} \qquad \ddot{\rho}(\mathcal{R}) \qquad\qquad\qquad (6.7)$$

$$\text{subject to} \qquad \dot{\lambda}_r(t) \leq c_{m,n}, \qquad\qquad \forall m \in \mathcal{M}, \forall n \in \mathcal{N},$$

$$\ddot{\lambda}_r(t) \leq \ddot{c}_{n,n'}, \qquad\qquad \forall (n, n') \in \mathcal{N},$$

$$\ddot{\kappa}_n(t) \leq \ddot{s}_n, \qquad\qquad\quad \forall n \in \mathcal{N},$$

$$\eta(\mathcal{R}) \geq \eta_{min} \qquad\qquad\quad \forall r \in \mathcal{R}.$$

According to our scenario, the first step for solving (6.7) is to infer the set of influential users. This, as mentioned before, is done via the notion of *centrality* metric [211]. In general, the centrality measures is used to quantify the social influence of a node in the network and also related to how the node is well connected. A node with higher value of this measure in turns means that such a node is more central (thus influential) than the nodes who have lower values of this measure. Several definition of centrality metrics exist on literature [211], whereas we only focus on the eigenvector centrality for exposition. Let $G = (\mathcal{N}, \mathcal{E})$ be the social graph which consists of N nodes/users, where $\mathcal{N}$ represents the set of nodes and $\mathcal{E}$ is set of the links between them. We know that, the graph G can be represented by its adjacency (or D2D connectivity) matrix $\mathbf{A}_{N \times N}$, where the entry $a_{n,n'}$, $n, n' = 1, ..., N$ is 1 if link (or edge) $\ddot{c}_{n,n'}$ exists, or 0 otherwise. For this matrix, let the eigenvalues to be represented by $\lambda_1 \geq ... \geq \lambda_N$ in decreasing order, and the corresponding eigenvectors of these eigenvalues be given by $\mathbf{v}_1, ... \mathbf{v}_N$. The eigenvector-centrality in this case is basically the eigenvector $\mathbf{v}_1$ that has the largest eigenvalue λ_1. Knowing K-most influential users of the social network via notion of centrality, a clustering method (i.e., K-means [212]) can be then formed around the users for community formation.

Once the set of influential users is identified and their communities are formed, the next step is to analyze the content dissemination within each social community. By doing so, the critical content of each community can be stored in the cache of influential users. To show this, suppose that there is a set number of available contents, denoted by $\mathcal{F} = \mathcal{F}_0 + \mathcal{F}_h$, where $\mathcal{F}_h$ is the set of contents with viewing history and $\mathcal{F}_0$ represents the set of contents without history. We further assume that each user is interested to only one type of available contents $\mathcal{F}$. Let π_f be the probability that content f is chosen by a given user, and as a prior [213], assume that the distribution follows a Beta distribution [213]. Then, the selection of user n given as the conjugate probability of the Beta distribution has a Bernoulli distribution. This in turn shows that the resulting user-content partition is analogous to that of the Chinese restaurant process (CRP) [213]. The CRP is a metaphor in which the objects are customers in a restaurant, and the classes are represented by the tables which the customers sit. More precisely, in CRP, there exists a restaurant with a large number of tables, each with infinite number of sets, and customers arrive sequentially each of them choosing a table at random.

In the CRP with concentration parameter β, each customer decides to occupy a table with a probability proportional to the number of occupiers of that table, and chooses the next available table with proportional to the parameter β. Being more specific, the

first customer selects the first table with probability $\frac{\beta}{\beta} = 1$. The second customer selects the first table with probability $\frac{1}{1+\beta}$, and the second table with probability $\frac{\beta}{1+\beta}$. Once the second customer selects the table, in the next, the third customer selects the first table with probability $\frac{1}{2+\beta}$, the second table with probability $\frac{1}{2+\beta}$ and the third table with probability $\frac{\beta}{2+\beta}$. This selection process continues until all customers have seats, yielding a distribution over allocation of customers to tables. In this process, the decision of subsequent customers are affected by the feedback of previous customers, where customers learn the previous selections to update their beliefs and probabilities in which they select the tables.

From this point, the behaviour of the content dissemination in the social network is similar to the table selection in an CRP. Looking to the social network as a Chinese restaurant, the contents as the large number of tables and the users as the customers, we can model the content dissemination process by an CRP. This means that, within each social community, users intend to request the sought-after content sequentially, and once a content is downloaded, a hit is recorded (i.e., history). This, in turn, changes the probability that this content will be requested by others within the same social community, where popular contents will be requested more frequently and new contents less frequently. Suppose a random binary matrix $\mathbf{Z}_{N \times F}$, indicating the selection of contents by users, where $z_{nf} = 1$ if user n chooses content f and 0 otherwise. Then, we can show that [213]:

$$P(\mathbf{Z}) = \frac{\beta^{F'}\Gamma(\beta)}{\Gamma(\beta + N)} \prod_{f=1}^{F'} (m_f - 1)! \tag{6.8}$$

where $\Gamma(.)$ is the Gamma function [214], m_f is the number of users already assigned to content f (i.e., viewing history) and F' is the number of partitions with $m_f > 0$. Therefore, for a given $P(\mathbf{Z})$, the popular contents of each community can be stored inside the cache of influential users. A sketch of the proactive caching procedure at the user terminals is summarized in Fig. 6.4.

6.4.1 Numerical Results and Discussions

In the numerical setup, for similar purposes as in the previous section, the wireless link capacities are assumed to be equal among the users. The total D2D link capacity of each user is shared among the number of social links. The list of parameters are given in Table 6.2.

Starting from $t = 0$, request arrival times are drawn uniformly at random until the time T. The social network is constructed by using the preferential attachment model [215]. As states before, the eigenvector centrality is used to quantize the influential users in the social network, then, K-most influential are formed into K communities via K-means clustering [212]. In each community, the content popularity distribution is sampled from the CRP(β). Given the content popularity, the proactive caching is done by storing the popular files greedily inside the influential users until no storage space remains. Similar to the case study in previous section, random caching is used as a baseline.

cache-enabled D2D communications

3) estimate content popularities within each community

2) form communities/clusters

small cell base station

4) store popular contents at influential users' devices

1) infer influential users based on centrality measures

social network overlay

D2D link

5) join to content delivery via D2D *(if the requested content is found nearby)*
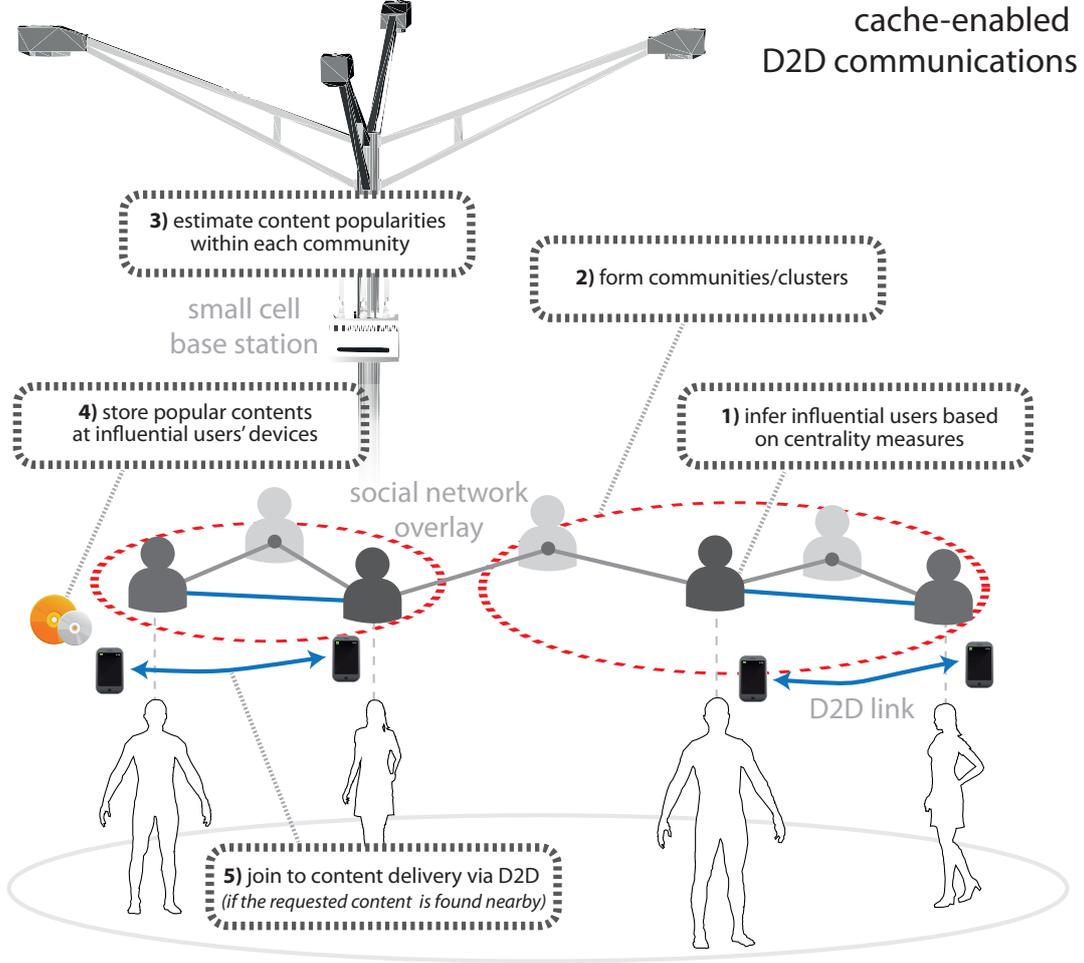
Figure 6.4: A practical procedure for proactive caching at the user terminals.

Parameters of interests in this case are: (i) number of requests R, (ii) total D2D cache size S, and (iii) CRP concentration parameter β. The results are normalized for ease of understanding. The impact of parameter of interests on the satisfaction ratio and small cell load are given in Fig. 6.5.

In the figure, increasing the number of requests, we see that the satisfaction ratio decreases rapidly and the small cell load decreases at a low pace. The gains of proactive caching approach are higher than the reactive approach in all regimes.

When an increment of D2D size is the case, we observe an increment in the satisfaction ratio and decrement in the small cell load. Even though both proactive and reactive cases have the gains, the proactive approach has more desirable performance compared to the reactive approach.

The concentration parameter β has also an impact on the performance. When β increases (i.e., the number of distinct contents grows), the satisfaction ratio and the small cell loads tends to be almost constant in the reactive approach. On the other hand, as

Table 6.2: The numerical setup parameters for proactive caching at the UTs.

Parameter	Description	Value
T	Time slots	1024 seconds
M	Number of SBSs	4
K	Number of communities	3
N	Number of UTs	32
F	Number of contents	128
l_f	Length of content f	1 Mbit
b_f	Bitrate of content f	1 Mbit/s
$\sum_m \sum_n c_{m,n}$	Total SBSs link capacity	32 Mbit/s
$\sum_n \sum_{n',n' \neq n} \ddot{c}_{n,n'}$	Total D2D link capacity	64 Mbit/s
R	Number of requests	$0 \sim 9464$
S	Total D2D cache size	$0 \sim l_f \times F$
β	CRP concentration parameter	$0 \sim 100$

β increases, the satisfaction ratio in the proactive approach decreases and the small cell increases. The performance gap between the proactive and reactive approaches gets closer and closer as β increases. This is due to the facts that the contents catalog size is growing while UTs having a limited cache size.

6.5   Closing Remarks

In this chapter, we have proposed a novel proactive network paradigm based on caching at the edge of the network. Using tools from machine learning, we exploited users' predictable behaviour and their social relationships for caching at the edge of the network. Our approach showed that peak mobile traffic demands can be significantly minimized, yielding backhaul offloadings and resource savings.
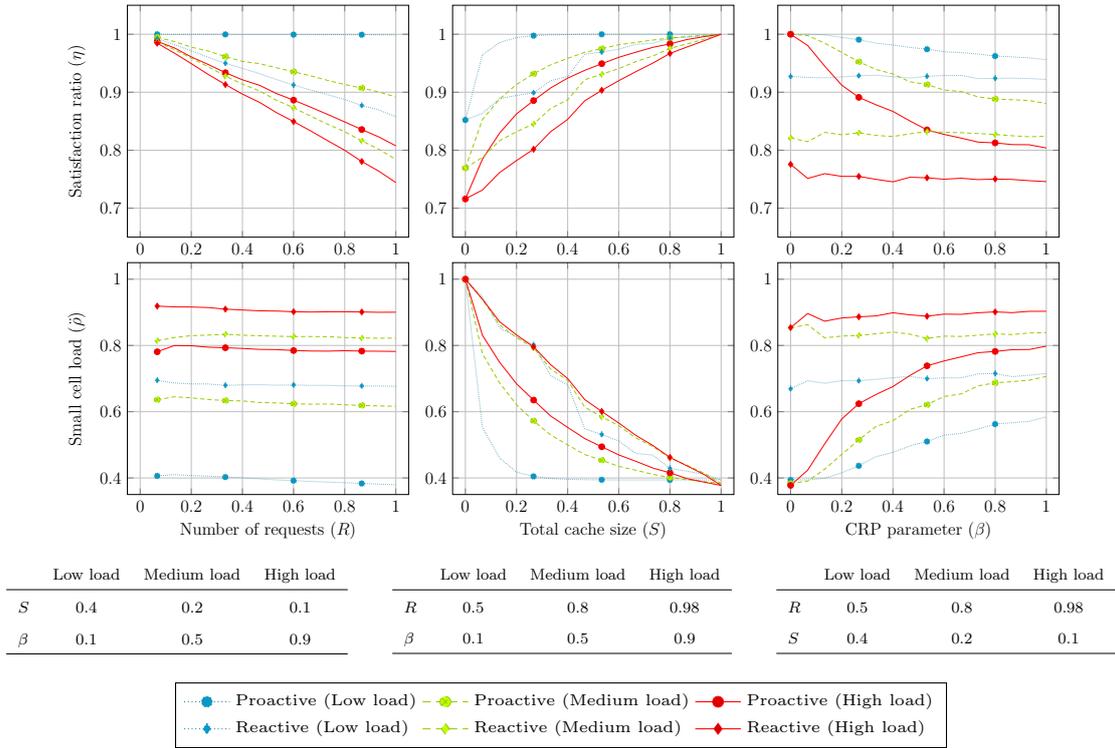
Figure 6.5: Social-Aware Caching via D2D: Dynamics of the satisfied requests and small cell load with respect to the number of requests, total cache size and CRP concentration parameter β.

Chapter 7

Transfer Learning

7.1 Overview

In the previous chapter, by exploiting spatio-social caching coupled with D2D communication, we proposed a novel proactive networking paradigm in which SBSs and UTs proactively cache contents at the network edge. As a result, the overall performance of the network in terms of users' satisfaction and backhaul offloading was improved. Therein, the *proactive* caching problem assumed non-perfect knowledge of the content popularity matrix, and supervised machine learning and CF techniques were used to estimate the popularity matrix leveraging user-content correlations. Nevertheless, the content popularity matrix remains typically large and sparse with very few users ratings, rendering CF learning methods inefficient mainly due to *data sparseness* and *cold-start* problems [209].

In this chapter, given the fact that data sparsity and cold-start problems degrade the performance of proactive caching, we leverage the framework of *transfer learning (TL)* and recent advances in machine learning [177]. TL is motivated by the fact that in many real-world applications, it is hard or even impossible to collect and label training data to build suitable prediction models. Exploiting available data from other rich information sources such as D2D interactions (called as *source domain*), allows TL to substantially improve the prediction task in the so-called *target domain*. TL has been applied to various data mining problems such as classification and regression [177]. TL methods can be mainly grouped into *inductive*, *transductive* and *unsupervised* TL methods depending on the availability of labels in the source and target domains. All these approaches boil down to answering the following fundamental questions: 1) *what* information to transfer? 2) *how* to transfer it? and 3) *when* to transfer it? While "what to transfer" deals with which part of the knowledge should be transferred between domains and tasks, "when to transfer" focuses on the timing of the operations in order to avoid negative transfer, especially when the source and target domains are uncorrelated. On the other hand, "how to transfer" deals with what kind of information should be transferred between domains and tasks.

The main contribution of this chapter is to propose a TL-based content caching mechanism to maximize the backhaul offloading gains as a function of storage constraints and users' content popularity matrix. This is done by learning and transferring hidden latent features extracted from the source domain to the target domain. In the source domain, we take into account users' D2D interactions while accessing/sharing statistics of contents within their social community as prior information in the knowledge transfer. It is shown that the content popularity matrix estimation in the target domain can be significantly improved instead of *learning from scratch* with unknown users' ratings. To the best of our knowledge, this is perhaps the first contribution of unsupervised transfer learning in cache-enabled small cells.

The rest of the chapter is organized as follows. The network model under consideration is provided in Section 7.2, accompanied with the caching problem formulation in both source and target domains. Section 7.3 presents the classical CF-based caching and that of the proposed transfer learning. The numerical results capturing the impact of various parameters on the users' satisfaction and backhaul offloading gains are given in Section 7.4. We finally conclude in Section 7.5.

7.2 Network Model

Let us assume an information system denoted by $S^{(\mathcal{S})}$ in the source domain and an information system denoted by $S^{(\mathcal{T})}$ in the target domain. A sketch of the network model is shown in Fig. 7.1.

7.2.1 Target Domain

Let us consider a network deployment consisting of M_{tar} SBSs from the set $\mathcal{M}_{tar} = \{1, \ldots, M_{tar}\}$ and N_{tar} UTs from the set $\mathcal{N}_{tar} = \{1, \ldots, N_{tar}\}$. Each SBS m is connected to the core network via a limited backhaul link with capacity $0 < C_m < \infty$ and each SBS has a total wireless link capacity C'_m for serving its UTs in the downlink. We further assume that $\mathbb{E}[C_m] < \mathbb{E}[C'_m]$. UTs request contents from a library $\mathcal{F}_{tar} = \{1, \ldots, F_{tar}\}$, where each content f has a size of $L(f)$ and a bitrate requirement of $B(f)$. Moreover, we suppose that users' content requests follow a Zipf-like distribution $P_{\mathcal{F}_{tar}}(f), \forall f \in \mathcal{F}_{tar}$ defined as [197]:

$$P_{\mathcal{F}_{tar}}(f) = \frac{\Omega}{f^\alpha} \tag{7.1}$$

where $\Omega = \left(\sum_{i=1}^{F_{tar}} \frac{1}{i^\alpha} \right)^{-1}$ and α characterizes the steepness of the distribution, reflecting different content popularities. Having such a content popularity in the ordered case, the content popularity matrix for the m-th SBS at time t is given by $\mathbf{P}^m(t) \in \mathbb{R}^{N_{tar} \times F_{tar}}$ where each entry $P_{n,f}^m(t)$ represent the probability that the n-th user requests the f-th content.

In order to avoid any kind of bottleneck during the delivery of users' content requests,
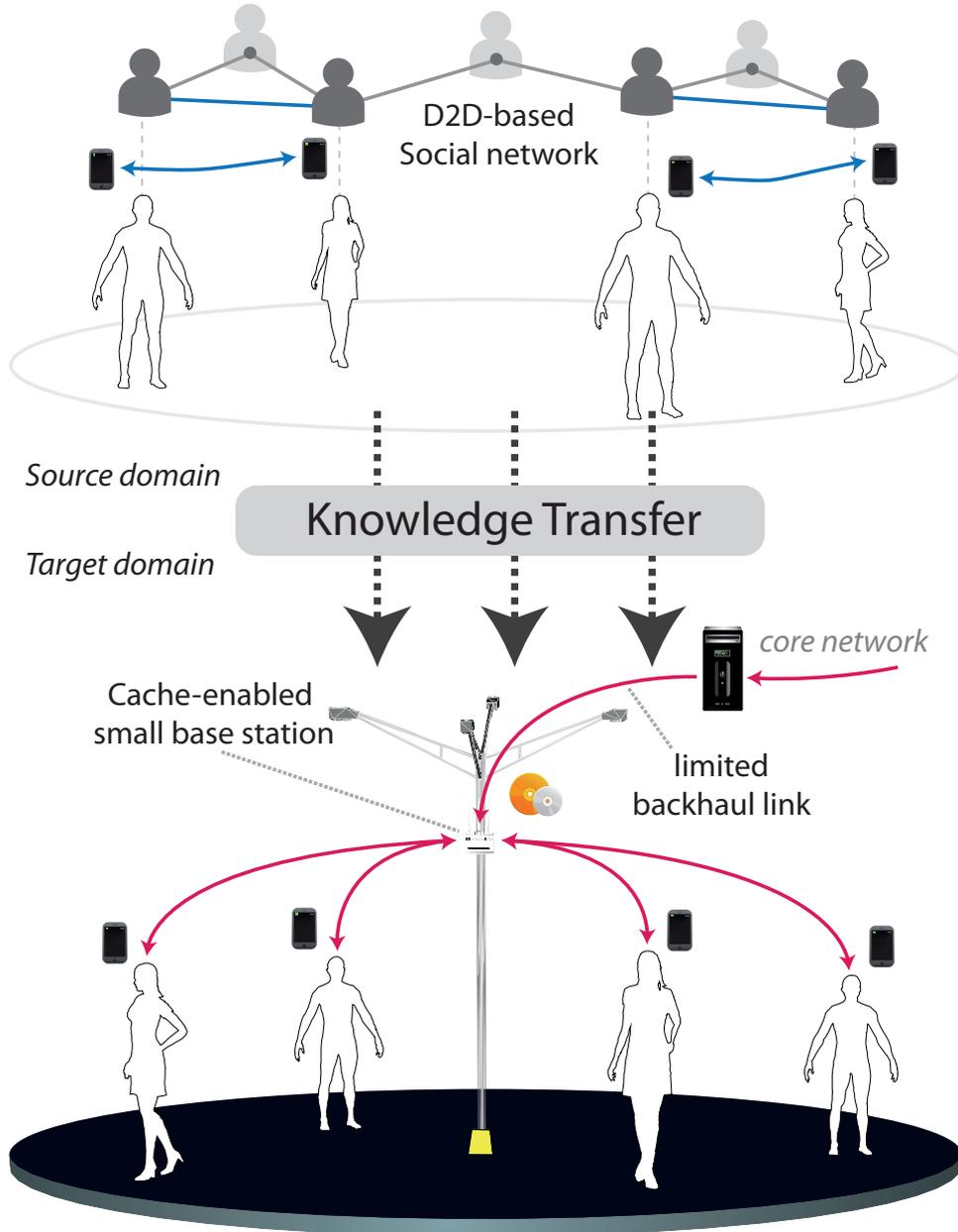
Figure 7.1: An illustration of the network model which consists of two information systems $S^{(\mathcal{S})}$ and $S^{(\mathcal{T})}$. Due to the lack of prior information in the target domain, the information extracted from users' social interactions and their ratings in the source domain is transferred to the target domain.

we assume that each SBS has a finite storage capacity of S_m and caches selected contents from the library $\mathcal{F}_{tar}$. Thus, the amount of requests SBSs satisfy from their local caches is of high importance to avoid peak demands and minimize the latency of content delivery. Our goal is to offload the backhaul while satisfying users' content requests, by pre-fetching strategic contents from the core network (CN) at suitable times and cache them at the

SBSs, subject to their storage constraints. To formalize this, suppose that D number of requests from the set $\mathcal{D} = \{1, ..., D\}$ are made by users during T time-slots. Then, a request $d \in \mathcal{D}$ within time window T is served immediately and is said to be *satisfied*, if the rate of delivery is equal or greater than the content bitrate, such that:

$$\frac{L(f_d)}{\tau'(d) - \tau(d)} \geq B(f_d) \tag{7.2}$$

where f_d is the requested content, $L(f_d)$ and $B(f_d)$ are the size and bitrate of the content, $\tau(d)$ is the arrival time of the request and $\tau'(d)$ the end time delivery. Given these definitions, the users' average *satisfaction ratio* can be expressed as:

$$\eta(\mathcal{D}) = \frac{1}{D} \sum_{d \in \mathcal{D}} \mathbb{1}\left\{ \frac{L(f_d)}{\tau'(d) - \tau(d)} \geq B(f_d) \right\} \tag{7.3}$$

where $\mathbb{1}\{...\}$ is the indicator function which returns 1 if the statement holds and 0 otherwise. Suppose that the instantaneous backhaul rate for the content delivery of request d at time t is given by $R_d(t) \leq C_m, \forall m \in \mathcal{M}_{tar}$. Then, the average *backhaul load* is defined as:

$$\rho(\mathcal{D}) = \frac{1}{D} \sum_{d \in \mathcal{D}} \frac{1}{L(f_d)} \sum_{t=\tau(f_d)}^{\tau'(f_d)} R_d(t). \tag{7.4}$$

Now, denote $\mathbf{X}(t) \in \{0,1\}^{M_{tar} \times F_{tar}}$ as the cache decision matrix of SBSs, where $x_{m,f}(t)$ equals 1 if the f-th content is cached at the m-th SBS at time t, and 0 otherwise. Therefore, the backhaul offloading problem can be formally expressed as:

$$\begin{aligned}
\underset{\mathbf{X}(t), \mathbf{P}^m(t)}{\text{minimize}} \quad & \rho(\mathcal{D}) \tag{7.5} \\
\text{subject to} \quad & L_{\min} \leq L(f_d) \leq L_{\max}, && \forall d \in \mathcal{D}, \\
& B_{\min} \leq B(f_d) \leq B_{\max}, && \forall d \in \mathcal{D}, \\
& R_d(t) \leq C_m, && \forall t, \forall d \in \mathcal{D}, \forall m \in \mathcal{M}_{tar}, \\
& R'_d(t) \leq C'_m, && \forall t, \forall d \in \mathcal{D}, \forall m \in \mathcal{M}_{tar}, \\
& \sum_{f \in \mathcal{F}_{tar}} L(f) x_{m,f}(t) \leq S_m, && \forall t, \forall m \in \mathcal{M}_{tar}, \\
& \sum_{n \in \mathcal{N}_{tar}} \sum_{f \in \mathcal{F}_{tar}} P^m_{n,f}(t) = 1, && \forall t, \forall m \in \mathcal{M}_{tar}, \\
& x_{m,f}(t) \in \{0,1\}, && \forall t, \forall f \in \mathcal{F}_{tar}, \forall m \in \mathcal{M}_{tar}, \\
& \eta_{\min} \leq \eta(\mathcal{D})
\end{aligned}$$

where $R'_d(t)$ is the instantaneous wireless link rate for request d and $\eta_{\min}$ is the minimum target satisfaction ratio respectively. In order to solve this problem, a joint optimization of the cache decision $\mathbf{X}(t)$ and the content popularity matrix estimation $\mathbf{P}^m(t)$ is needed. Moreover, solving (7.5) is very challenging due to:

i) limited backhaul and wireless link capacity as well as the limited storage capacity of SBSs,

ii) large number of users with unknown ratings and library size,

iii) SBSs need to track, learn and estimate users' content popularity/rating matrix $\mathbf{P}^m(t)$ for cache decision while dealing with data sparsity.

For simplicity, we drop now the index of the SBSs and assume that the content popularity is stationary during T time slots, thus $\mathbf{P}^m(t)$ is denoted as $\mathbf{P}_{tar}$. Moreover, for sake of exposition, we restrict ourselves to caching policies in which the contents are stored during the peak-off hours, thus $\mathbf{X}(t)$ remains fixed during the content delivery and represented as $\mathbf{X}$. In the following, we examine the source domain which we exploit when dealing with the sparsity of $\mathbf{P}_{tar}$ in the target domain.

7.2.2 Source Domain

As advocated in [18], we leverage the existence of a D2D-based social network overlay made of users' interactions within their social communities, referred as the *source domain* in the sequel. Specifically, this source domain contains the behaviour of users' interactions within their social communities, modeled as a CRP [213]. This constitutes the prior information used in the transfer learning procedure.

In the CRP with parameter β, every customer selects an occupied table with a probability proportional to the number of occupants, and selects the next vacant table with probability proportional to β. More precisely, the first customer selects the first table with probability $\frac{\beta}{\beta} = 1$. The second customer selects the first table with probability $\frac{1}{1+\beta}$, and the second table with probability $\frac{\beta}{1+\beta}$. After the second customer selects the second table, the third customer chooses the first table with probability $\frac{1}{2+\beta}$, the second table with probability $\frac{1}{2+\beta}$ and the third table with probability $\frac{\beta}{2+\beta}$. This stochastic Dirichlet process continues until all customers select their seats, defining a distribution over allocation of customers to tables.

In this regard, the content dissemination in the social network is analogous to the table selection in a CRP. If we view this network as a CRP, the contents as the large number of tables, and users as the customers, we can make an analogy between the content dissemination and the CRP. First, suppose that there exist N_{D2D} users in this network. Let $F_{D2D} = F_0 + F_h$ be the total number of contents in which F_h represents the number of contents with viewing histories and F_0 is the number of contents without history. Denote also $\mathbf{Z}_{D2D} \in \{0,1\}^{N_{D2D} \times F_{D2D}}$ as a random binary matrix indicating which contents are selected by each user, where $z_{n,f} = 1$ if the n-th user selects the f-th content and 0 otherwise. Then, it can be shown that [213]:

$$P(\mathbf{Z}_{D2D}) = \frac{\beta^{F_h} \Gamma(\beta)}{\Gamma(\beta + N_{D2D})} \prod_{f=1}^{F_h} (m_f - 1)! \tag{7.6}$$

where $\Gamma(.)$ is the Gamma function, m_f is the number of users assigned to content f (i.e., viewing history) and F_h is the number of contents with viewing histories with $m_f > 0$.

In the target domain, the caching problem boils down to estimating the content popularity matrix which is assumed to be largely unknown, yielding degraded performance (i.e., very low cache hit ratios, slow convergence, etc.). Moreover, this degradation can be more severe in cases where the number of users and library size is extremely large. Therefore, in order to handle these issues and cache contents more efficiently, we propose a novel proactive caching procedure using transfer learning which exploits the rich contextual information extracted from users' social interactions. This caching procedure is shown to yield more backhaul offloading gains compared to a number of baselines, including random caching and the classical CF-based estimation methods [18].

7.3 Transfer Learning: Boosting Content Popularity Matrix Estimation

First, we start by explaining the classical CF-based learning, then detail our proposed TL solution.

7.3.1 Classical CF-based Learning

The classical CF-based estimation procedure is composed of a training and prediction phase. In the training part, the goal is to estimate the content popularity matrix $\mathbf{P}_{tar} \in \mathbb{R}^{N_{tar} \times F_{tar}}$, where each SBS constructs a model based on the already available information (i.e., users' content ratings). Let $\mathcal{N}_{tar}$ and $\mathcal{F}_{tar}$ represent the set of users and contents associated with N_{tar} users and F_{tar} contents. In particular, $\mathbf{P}_{tar}$ with entries $P_{tar,ij}$ is the (sparse) content popularity matrix in the target domain. $\mathcal{R}_{tar} = \{(i, j, r) : r = P_{tar,ij}, P_{tar,ij} \neq 0\}$ denotes the set of known user ratings. In the prediction phase, in order to predict the unobserved ratings in $\mathcal{N}_{tar}$, low-rank matrix factorization techniques are used to estimate the unknown entries of $\mathbf{P}_{tar}$. The objective here is to construct a k-rank approximate popularity matrix $\mathbf{P}_{tar} \approx \mathbf{N}_{tar}^T \mathbf{F}_{tar}$, where the factor matrices $\mathbf{N}_{tar} \in \mathbb{R}^{k \times N_{tar}}$ and $\mathbf{F}_{tar} \in \mathbb{R}^{k \times Ftar}$ are learned by minimizing the following cost function:

$$\underset{(i,j)\in\mathbf{P}_{tar}}{\text{minimize}} \quad \sum_{(i,j)\in\mathbf{P}_{tar}} \left(\mathbf{n}_i^T \mathbf{f}_j - P_{tar,ij} \right)^2 + \mu \left(||\mathbf{N}_{tar}||_F^2 + ||\mathbf{F}_{tar}||_F^2 \right) \tag{7.7}$$

where the sum is over the (i,j) user/content pairs in the training set. In addition, $\mathbf{n}_i$ and $\mathbf{f}_j$ represent the i-th and j-th columns of $\mathbf{N}_{tar}$ and $\mathbf{F}_{tar}$ respectively, and $||.||_F^2$ denotes the Frobenius norm. In (7.7), the parameter μ provides a balance between regularization and fitting training data. Unfortunately, users may rate very few contents, causing $\mathbf{P}_{tar}$ to be extremely sparse, and thus (7.7) suffers from severe over-fitting issues and engenders poor performance.

7.3.2 TL-based Content Caching

To alleviate data sparsity, solving (7.7) can be done more efficiently by exploiting and transferring the vast amount of available user-content ratings (i.e., prior information) from a different-yet-related source domain. Formally speaking, let us denote the source domain as $S^{(\mathcal{S})}$, and assume that this domain is associated with a set of N_{D2D} users and F_{D2D} contents denoted by $\mathcal{N}_{D2D}$ and $\mathcal{F}_{D2D}$ respectively. Additionally, the user-content popularity matrix in the source domain is given by matrix $\mathbf{P}_{D2D} \in \mathbb{R}^{N_{D2D} \times F_{D2D}}$ and likewise let $\mathcal{R}_{D2D} = \{(i, j, r) : r = P_{D2D,ij}, P_{D2D,ij} \neq 0\}$ represent the set of observed user ratings in the source domain. The underlying principle of the proposed approach is to smartly "borrow" carefully-chosen user social behavior information from $S^{(\mathcal{S})}$ to better learn $S^{(\mathcal{T})}$.

The transfer learning procedure from $S^{(\mathcal{S})}$ to $S^{(\mathcal{T})}$ is composed of two interrelated phases. In the first phase, a content *correspondence* is established in order to identify similarly-rated contents in both source and target domains. In the second phase, an optimization problem is formulated by combining the source and target domains for *knowledge transfer*, to jointly learn the popularity matrix $\mathbf{P}_{tar}$ in the target domain. In this regard, we suppose that both source and target domains correspond to one information system $s \in \{S^{(\mathcal{S})}, S^{(\mathcal{T})}\}$, that is made of N_s users and F_s contents given by $\mathcal{N}_s$ and $\mathcal{F}_s$ respectively. In each system s, we observe $\mathbf{P}_s$ with entries $P_{s,ij}$. Let $\mathcal{R}_s = \{(i, j, r) : r = P_{s,ij}, P_{s,ij} \neq 0\}$ represent the set of observed user ratings in each system and the set of *shared contents* is given by $\tilde{\mathcal{F}}$. Moreover, let $\mathcal{N}^* = \mathcal{N}_{D2D} \cup \mathcal{N}_{tar}$ and $\mathcal{F}^* = \mathcal{F}_{D2D} \cup \mathcal{F}_{tar}$ be the union of the collections of users and contents, respectively, where $N^* = |\mathcal{N}^*|$ and $F^* = |\mathcal{F}^*|$ represent the total number of unique users and contents in the union of both systems.

In the proposed TL approach, we model the users $\mathcal{N}^*$ and contents $\mathcal{F}^*$ by a user factor matrix $\mathbf{N} \in \mathbb{R}^{k \times N^*}$ and a content factor matrix $\mathbf{F} \in \mathbb{R}^{k \times F^*}$, where the i-th and j-th columns of these matrices are given by $\mathbf{n}_i$ and $\mathbf{f}_j$, respectively. The aim is to approximate the popularity matrix $\mathbf{P}_s \approx \mathbf{N}_s^T \mathbf{F}_s$ by jointly learning the factor matrices $\mathbf{N}$ and $\mathbf{F}$. This is formally done by minimizing the following cost function:

$$\underset{(i,j) \in \mathbf{P}_s}{\text{minimize}} \quad \sum_s \left(\alpha_s \sum_{(i,j) \in \mathbf{P}_s} \left(\mathbf{n}_i^T \mathbf{f}_j - P_{s,ij} \right)^2 \right) + \mu \left(||\mathbf{N}||_F^2 + ||\mathbf{F}||_F^2 \right) \qquad (7.8)$$

where the parameter α_s is the weight of each system. By doing so, $\mathbf{P}_{D2D}$ and $\mathbf{P}_{tar}$ are jointly factorized, and thus the set of factor matrices $\mathbf{F}_{D2D}$ and $\mathbf{F}_{tar}$ become interdependent as the features of a shared content are similar for knowledge sharing. A practical TL-based caching procedure is sketched in Fig. 7.2.

7.4 Numerical Results and Discussion

The objective of this section is to validate the effectiveness of the proposed TL caching procedure and draw key insights. In particular, we consider the following caching policies for comparison:
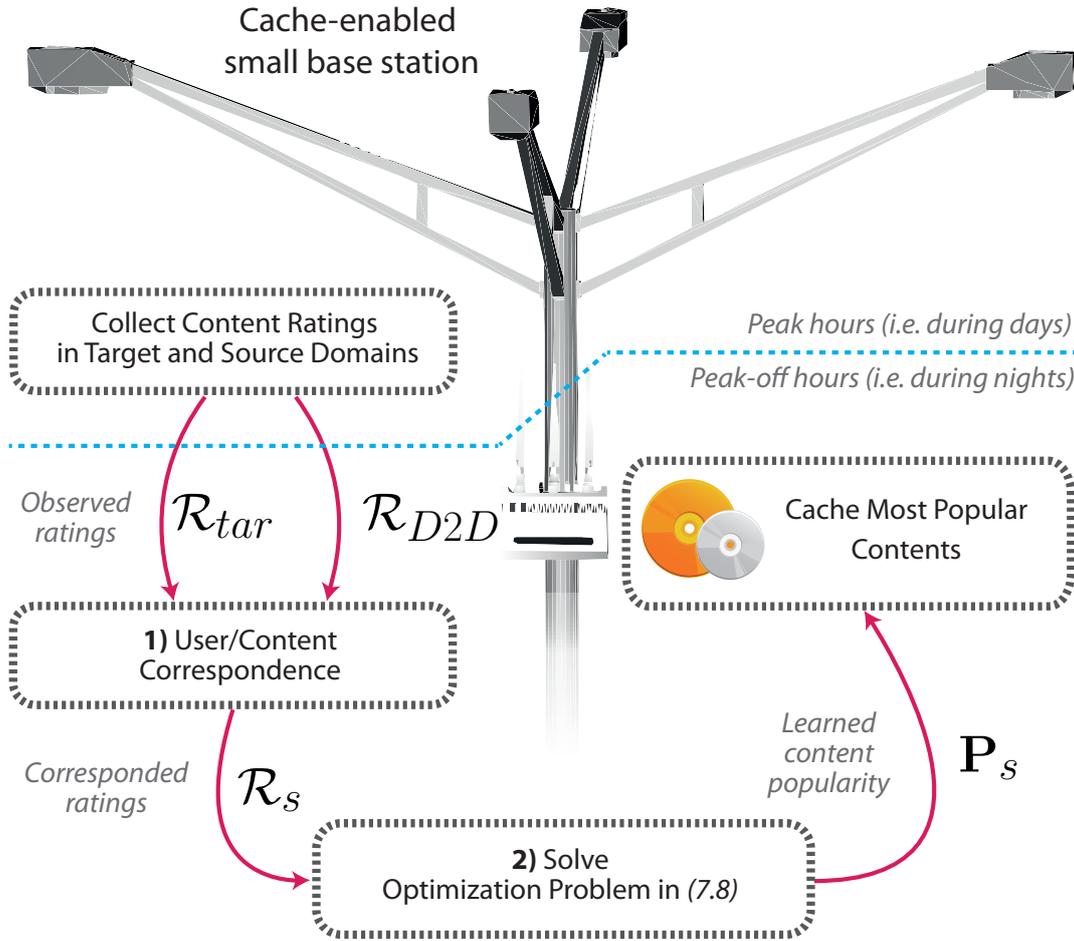
Figure 7.2: An illustration of the proposed TL-based caching procedure.

1) *Ground Truth*: Given the perfect rating matrix $\mathbf{P}_{tar}$, the most popular contents are stored greedily.

2) *Random caching* [18]: Contents are cached uniformly at random.

3) *Collaborative Filtering* [209]: The content popularity matrix $\mathbf{P}_{tar}$ is estimated via CF from a training set with 4% of ratings. Then, the most popular contents are stored accordingly.

4) *Transfer Learning*: $\mathbf{P}_{tar}$ and $\mathbf{P}_{D2D}$ matrices are jointly factorized via TL by using a training set with 12% of ratings and perfect user-content correspondence. Then the most popular contents are stored accordingly.

In the numerical setup, having contents cached according to these policies, the SBSs serve their users according to a traffic arrival process. This process is drawn from a Poisson process with intensity λ. The storage size of SBSs, content lengths, capacities of non-interfering wireless and backhaul links are assumed to have same constant values
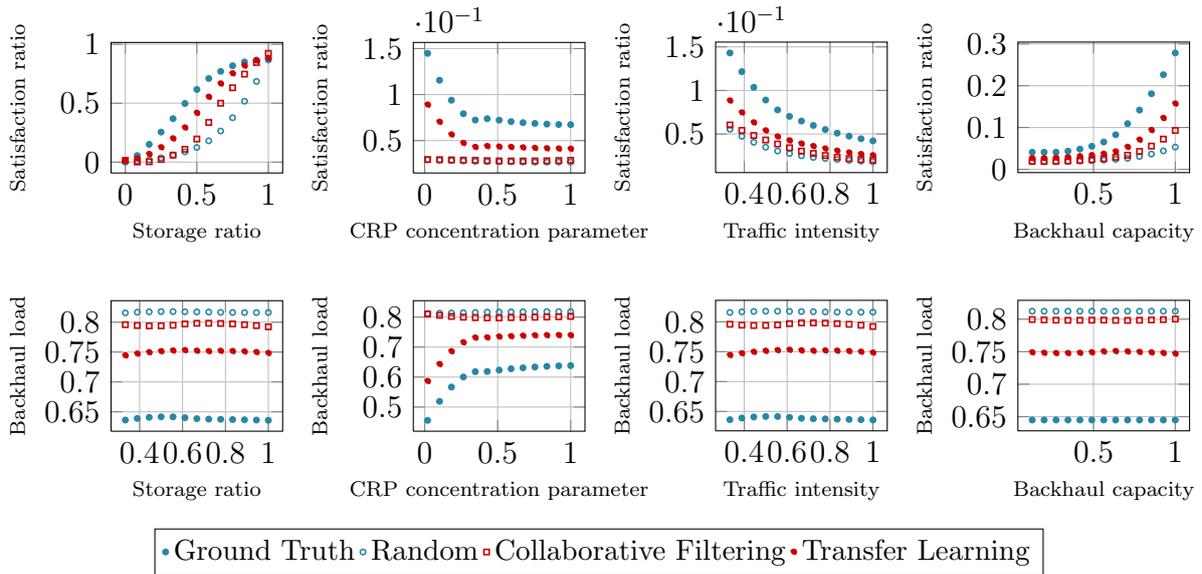
Figure 7.3: Evolution of the aggregate backhaul load and users' satisfaction ratio.

individually, in order to showcase the performance of the caching policies. The numerical results of users' satisfaction ratio and backhaul load are obtained by averaging out 1000 Monte-Carlo realizations. The simulation parameters are summarized in Table 7.1, unless stated otherwise.

The dynamics of users' satisfaction ratio and backhaul load with respect to the storage size, demand shape in the source domain, traffic intensity and backhaul capacity are given in Fig. 7.3. The results are normalized to show the various percentage gains, whereas the actual values are shown in Table 7.1. In the following, we discuss in detail the impact of these parameters.

Impact of the storage size (S_m)

The storage size is indeed one of the crucial parameter in cache-enabled SBSs, and it is expected that higher storage sizes result in better performance in terms of satisfaction ratio and backhaul offloading. According to this setup, we would like to note that the biggest improvement in satisfaction ratio and decrement in the backhaul load is achieved by the ground truth baseline where the content popularity is perfectly known. The random approach on the other hand has the worst-case performance. The CF approach exhibits similar performance as the random approach due to the cold-start problem, whereas the satisfaction ratio and backhaul offloading gains of TL are close to the ground truth base-line. In particular, it is shown that the TL policy outperforms its CF counterpart, with satisfaction and backhaul offloading gains up to 22% and 5% respectively.

Table 7.1: List of simulation parameters for TL-based approach.

Parameter	Description	Default-Varied Values
M_{tar}	Number of SBSs	4
N_{tar}	Number of UTs	32
F_{tar}	Library size	32 contents
L	Content length	1 MBit
B	Bitrate requirement	1 MBit
$\sum C'_m$	Total wireless capacity	32 MBit/s
T	Time slots	128 seconds
α	Zipf parameter	2
β	CRP concentration parameter	2 - [2 $\sim$ 100]
$\sum S_m$	Total storage size	6 - [0 $\sim$ 32] MBit
$\sum C_m$	Total backhaul capacity	1 - [1 $\sim$ 8] MBit/s
λ	Traffic intensity	1 - [1 $\sim$ 3] demand/s

Impact of the demand shape in the source domain (β)

The demand shape in the source domain, characterized by the CRP concentration parameter β provides meaningful insights to our problem. In fact, as β increases, the demand shape tends to be more uniform, requiring higher storage sizes at the SBSs to sustain the same performance. In a storage limited case, we see that the satisfaction ratio decreases and the backhaul load increases with the increment of β. Compared to the CF approach, the gains of TL are around 6% for the satisfaction gains and 22% for the backhaul offloading. However, the gap between TL and CF becomes smaller as β increases.

Impact of the traffic intensity (λ)

As the average number of request arrivals per time slot increases, bottlenecks in the network are expected to occur due to the limited resources of SBSs, resulting in less satisfaction ratios. This is visible in the high arrival rate regime, whereas the relative backhaul load remains constant. It can be shown that the ground truth caching with perfect knowledge of content popularity outperforms the other policies while the random approach has the worst performance. On the other hand, the performance of TL is

in between these approaches and has up to 3% satisfaction gains and 18% of backhaul offloading gain compared to the CF.

Impact of the backhaul capacity (C_m)

The total backhaul capacity is assumed to be sufficiently smaller than the capacity of wireless links. The increment of this capacity clearly results in higher satisfaction ratios in all cases. Note that any content not available in the caches of SBSs is delivered via the backhaul. Therefore, increasing the backhaul capacity avoids the bottlenecks during the delivery, thus yielding higher users' satisfaction. On the other hand, the backhaul load remains constant in this setting. It can be seen that TL approach has satisfaction ratio gains of up to 6% and backhaul offloading of up to 5% compared to the CF approach.

Impact of source-target correspondence
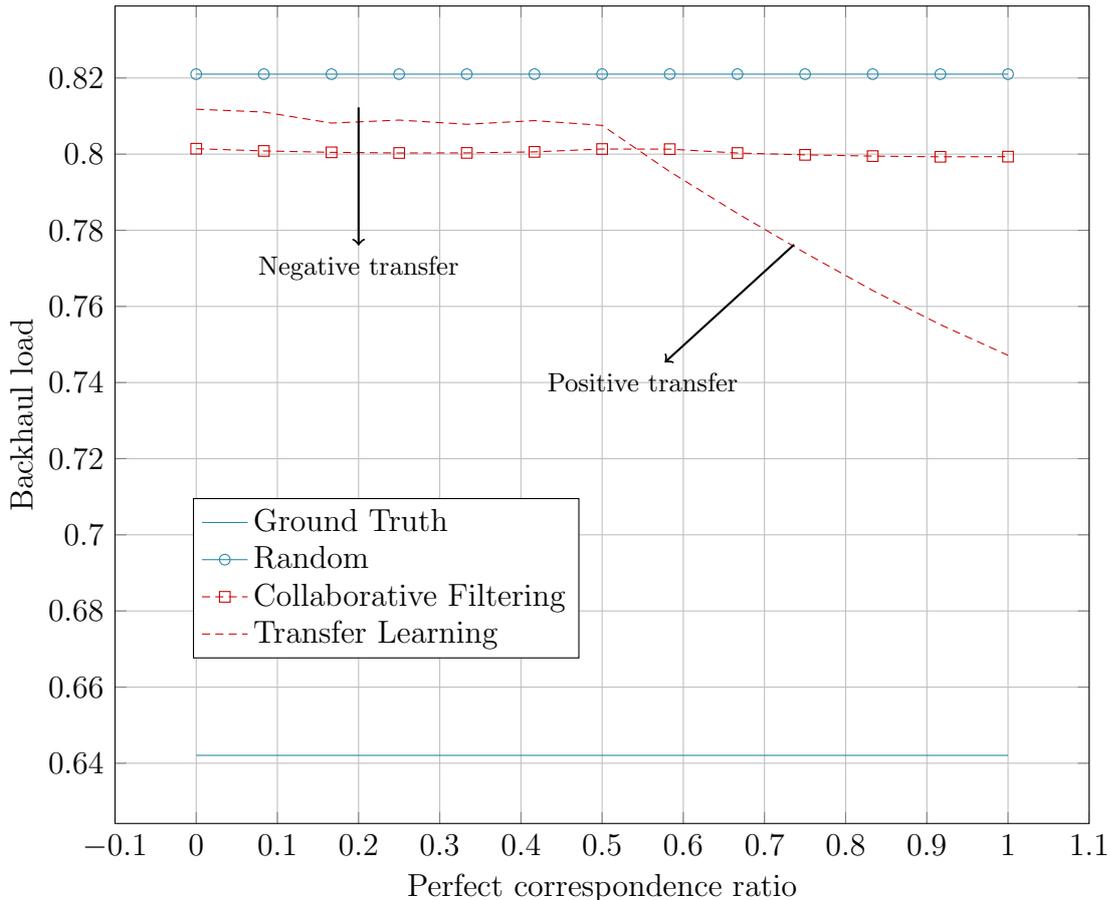
Figure 7.4: Evolution of the backhaul load with respect to the perfect correspondence ratio.

We have so far assumed that the user/content correspondence between the target and source domains is perfect. This is a strong assumption and such an operation requires a more careful treatment to avoid negative transfer. Here, we relax this assumption by introducing a perfect correspondence ratio. This ratio represents the amount of perfect user/content matching between both source and target domains. A ratio of 0 means that 100% of correspondence is done uniformly at random and 1 is equivalent to the perfect case. It is shown in Fig. 7.4 that TL has a poor performance in the low values of this ratio, with similar performance as the random caching due to the negative transfer. However, as this ratio increases, the performance of TL improves, outperforming the CF with a ratio of 0.58. This underscores the importance of such an operation for the positive transfer and is left for future work.

7.5 Closing Remarks

We proposed a novel transfer learning-based caching procedure which was shown to yield higher users' satisfaction and backhaul offloading gains overcoming the data sparsity and cold start problems. Numerical results confirmed that the overall performance can be improved by transferring a judiciously-extracted knowledge from a source domain to a target domain via TL.

Chapter 8

Big Data for Caching

8.1 Overview

The unprecedented increase in data traffic demand driven by mobile video, online social media and over-the-top (OTT) applications are compelling mobile operators to look for innovative ways to manage their increasingly complex networks. This explosion of traffic stemming from diverse domain (e.g., healthcare, machine-to-machine communication, connected cars, user-generated content, smart metering, to mention a few) have different characteristics (e.g., structured/non-structured) and is commonly referred to as *Big Data* [216]. While big data comes with "big blessings" there are formidable challenges in dealing with large-scale data sets due to the sheer volume and dimensionality of the data. A fundamental challenge of big data analytics is to shift through large volumes of data in order to discover hidden patterns for actionable decision making. Indeed, the era of collecting and storing data in remote standalone servers where decision making is done offline has dawned. Rather, telecom operators are exploring decentralized and flexible network architectures whereby predictive resource management play a crucial role leveraging recent advances in storage/memory, context-awareness and edge/cloud computing [18, 217, 218]. In the realm of wireless, big data brings to network planning a variety of new information sets that can be inter-connected to achieve a better understanding of users and networks (e.g., location, user velocity, social geodata, etc.). Moreover, public data from social networks such as Twitter and Facebook provides additional side information about the life of the network, which can be further exploited. The associated benefits are a higher accuracy of user location information or the ability to easily identify and predict user clustering, for example for special events. Undoubtedly, the huge potential associated with big data has sparked a flurry of research interest from industry, government and academics (see [219] for a recent survey), and will continue to do so in the coming years.

In this chapter, we are intent to propose a proactive caching architecture for optimization of 5G wireless networks where we exploit large amount of available data with the help of big data analytics and machine learning tools. In other words, we investigate the gains of proactive caching both in terms of backhaul offloadings and request satisfac-

tions, where machine learning tools are used to model and predict the spatio-temporal user behaviour for proactive cache decision. By caching strategic contents at the edge of network, namely at the base stations, network resources are utilized more efficiently and users' experience is further improved. However, the estimation of content popularity tied with spatio-temporal behaviour of users is a very complex problem due to the high dimensional aspects of data, data sparsity and lack of measurements. In this regard, we present a platform to parallelize the computation and execution of the content prediction algorithms for cache decision at the base stations. As a real-world case study, a large amount of data collected from a Turkish telecom operator, one of the largest mobile operator in Turkey with 16.2 million of active subscribers, is examined for various caching scenarios. Particularly, the traces of mobile users' activities are collected from several base stations in hours of time interval and are analysed inside the network under the privacy concerns and regulations. The analysis is carried out on a big data platform and caching at the base stations has been investigated for further improvements of users' experience and backhaul offloadings.

Our main contribution in this chapter is to make tighter connections of big data phenomena with caching in $5G$ wireless networks, by proposing a proactive caching architecture where statistical machine learning tools are exploited for content popularity estimation. Combined with a large-scale real-world case study, this is perhaps the first attempt on this direction and highlights a huge potential of big data for 5G wireless networks.

The rest of chapter is organized as follows. Our network model for proactive caching is detailed in Section 8.2. A practical case study of content popularity estimation on a big data platform is presented in Section 8.3, including a characterization of users' traffic pattern. Subsequently, numerical results for cache-enabled base stations and relevant discussions are carried out in Section 8.4. We finally conclude in Section 8.5.

8.2 Network Model

Suppose a network deployment of M SBSs from the set $\mathcal{M} = \{1, \ldots, M\}$ and N UTs from the set $\mathcal{N} = \{1, \ldots, N\}$. Each SBS m has access to the broadband Internet connection via a wired backhaul link with capacity C_m Mbyte/s, and is able to provide this broadband service to its users via a wireless link with total capacity of C'_m Mbyte/s. Due to the motivation that the backhaul capacity is generally limited in densely deployed SBSs scenarios [220], we further consider that $C_m < C'_m$. Also, assume that each user $n \in \mathcal{N}$ is connected to only one SBS and is served via unicast sessions[1]. In particular, we assume that UTs request contents (i.e., videos, files, news, etc.) from a library $\mathcal{F} = \{1, \ldots, F\}$, where each content f in this library has a size of $L(f)$ Mbyte and bitrate requirement of

---

[1]The unicast service model can also be extended to the multicast case. See [59,101] for studies in this direction.

$B(f)$ Mbyte/s, with

$$L_{\min} = \min_{f \in \mathcal{F}}\{L(f)\} > 0 \qquad (8.1)$$

$$L_{\max} = \max_{f \in \mathcal{F}}\{L(f)\} < \infty \qquad (8.2)$$

and

$$B_{\min} = \min_{f \in \mathcal{F}}\{B(f)\} > 0 \qquad (8.3)$$

$$B_{\max} = \max_{f \in \mathcal{F}}\{B(f)\} < \infty. \qquad (8.4)$$

The users' content requests in fact follow a Zipf-like distribution $P_{\mathcal{F}}(f), \forall f \in \mathcal{F}$ given as [197]:

$$P_{\mathcal{F}}(f) = \frac{\Omega}{f^\alpha} \qquad (8.5)$$

where

$$\Omega = \Big(\sum_{i=1}^{F} \frac{1}{i^\alpha} \Big)^{-1}.$$

The parameter α in (8.5) describes the steepness of the distribution. This kind of power laws is used to characterize many real-world phenomena, such as the distribution of files in the web-proxies [197] and the traffic dynamics of cellular devices [198]. Higher values of α corresponds to a steeper distribution, meaning that a small subset of contents are highly popular than the rest of the catalog (namely users have very similar interests). On the other hand, the lower values describe a more uniform behaviour with almost equal popularity of contents (namely users have more distinct interests). The parameter α can take different values depending on users' behaviour and SBSs deployment strategies (i.e., home, enterprise, urban and rural environments), and its practical value in our experimental setup will be given in the subsequent sections.

Given such a global content popularity in the decreasing ordered case, the content popularity matrix of the m-th SBS at time t is specifically described by $\mathbf{P}^m(t) \in \mathbb{R}^{N \times F}$ where each entry $P_{n,f}^m(t)$ corresponds to the probability that the n-th user requests the f-th content. In fact, the matrix $\mathbf{P}^m(t)$ is the local content popularity distribution observed at the base station m at time t, whereas the Zipf distribution $P_{\mathcal{F}}(f), \forall f \in \mathcal{F}$ is used to characterize the global content popularity distribution of all contents in (decreasing) sorted order.

In this scenario, we consider that each SBS has a finite storage capacity of S_m and proactively caches selected contents from the library $\mathcal{F}$ during peak-off hours. By doing so, the bottlenecks caused by the limited-backhaul are avoided during the delivery of users' content requests in peak hours. The amount of satisfied requests and backhaul load are of paramount importance and are defined as follows. Suppose that D number of contents are requested during the duration of T seconds, and are represented by the set $\mathcal{D} = \{1, ..., D\}$. Assume that the delivery of content is started immediately when the

request $d \in \mathcal{D}$ arrives to the SBS. Then, the request d is called *satisfied* if the rate of content delivery is equal or higher than the bitrate of the content in the end of service, such as:

$$\frac{L(f_d)}{\tau'(f_d) - \tau(f_d)} \geq B(f_d) \tag{8.6}$$

where f_d describes the requested content, $L(f_d)$ and $B(f_d)$ are the size and bitrate of the content, $\tau(f_d)$ is the arrival time of the content request and $\tau'(f_d)$ the end time delivery.[2] Defining the condition in (8.6) stems from the fact that, if the delivery rate is not equal nor higher than the bitrate of the requested content, the interruption during the playback (or download) occurs thus users would have less QoE[3]. Therefore, the situations where this condition holds are more desirable for better QoE. In (8.6), note also that the end time of delivery for request d, denoted by $\tau'(d)$, highly depends on the load of the system, capacities of the backhaul and wireless links as well as availability of contents at the base stations. Given this definition of satisfied requests and related explanations, the users' average request *satisfaction ratio* is then defined for the set of all requests, that is:

$$\eta(\mathcal{D}) = \frac{1}{D} \sum_{d \in \mathcal{D}} \mathbb{1} \left\{ \frac{L(f_d)}{\tau'(f_d) - \tau(f_d)} \geq B(f_d) \right\} \tag{8.7}$$

where $\mathbb{1}\{...\}$ is the indicator function which takes 1 if the statement holds and 0 otherwise. Now, denoting $R_d(t)$ Mbyte/s as the instantaneous rate of backhaul for the request d at time t, with $R_d(t) \leq C_m$, $\forall m \in \mathcal{M}$, the average *backhaul load* is then expressed as:

$$\rho(\mathcal{D}) = \frac{1}{D} \sum_{d \in \mathcal{D}} \frac{1}{L(f_d)} \sum_{t=\tau(f_d)}^{\tau'(f_d)} R_d(t). \tag{8.8}$$

Here, the outer sum is over the set of all requests whereas the inner sum gives the total amount of information passed over the backhul for request d which is at most equal to the length of requested file $L(f_d)$. The instantaneous rate of backhul for request d, denoted by $R_d(t)$, heavily depends on the load of the system, capacity of the backhaul link and cached contents at the base stations.

In fact, by pre-fetching the contents at the SBSs, the access delays to the contents are minimized especially during the peak hours, thus yielding higher satisfaction ratio and less backhaul load. To elaborate this, now consider the cache decision matrix of SBSs as $\mathbf{X}(t) \in \{0,1\}^{M \times F}$, where the entry $x_{m,f}(t)$ takes 1 if the f-th content is cached at the m-th SBS at time t, and 0 otherwise. Then, the backhaul offloading problem under a

---

[2]One can also consider/exploit future information (i.e., start time of requests, end time of content delivery) in the context of proactive resource allocation (see [221] for instance).

[3]In practice, a video content has typically a bitrate requirement ranging from 1.5 to 68 Mbit/s [222].

specific request satisfaction constraint is formally given as follows:

$$\underset{\mathbf{X}(t),\mathbf{P}^m(t)}{\text{minimize}} \qquad \rho(\mathcal{D}) \tag{8.9}$$

$$\text{subject to} \qquad L_{\min} \leq L(f_d) \leq L_{\max}, \qquad \forall d \in \mathcal{D}, \tag{8.9a}$$

$$B_{\min} \leq B(f_d) \leq B_{\max}, \qquad \forall d \in \mathcal{D}, \tag{8.9b}$$

$$R_d(t) \leq C_m, \qquad \forall t, \forall d \in \mathcal{D}, \forall m \in \mathcal{M}, \tag{8.9c}$$

$$R'_d(t) \leq C'_m, \qquad \forall t, \forall d \in \mathcal{D}, \forall m \in \mathcal{M}, \tag{8.9d}$$

$$\sum_{f \in \mathcal{F}} L(f)x_{m,f}(t) \leq S_m, \quad \forall t, \forall m \in \mathcal{M}, \tag{8.9e}$$

$$\sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} P^m_{n,f}(t) = 1, \quad \forall t, \forall m \in \mathcal{M}, \tag{8.9f}$$

$$x_{m,f}(t) \in \{0,1\}, \quad \forall t, \forall f \in \mathcal{F}, \forall m \in \mathcal{M}, \tag{8.9g}$$

$$\eta_{\min} \leq \eta(\mathcal{D}), \tag{8.9h}$$

where $R'_d(t)$ Mbyte/s describes the instantaneous rate of wireless link for request d and $\eta_{\min}$ represents the minimum target satisfaction ratio. In particular, the constraints (8.9a) and (8.9b) are to bound the length and bitrate of contents in the catalog for feasible solution, the constraints (8.9c) and (8.9d) are the backhaul and wireless link capacity constraints, (8.9e) holds for storage capacity for caching, (8.9f) is to ensure the content popularity matrix as a probability measure, (8.9g) denotes the binary decision variables of caching, and finally the expression in (8.9h) is the satisfaction ratio constraint for QoE.

In order to tackle this problem, the cache decision matrix $\mathbf{X}(t)$ and the content popularity matrix estimation $\mathbf{P}^m(t)$ have to be optimized jointly. However, solving the problem (8.9) is very challenging as:

i) the storage capacity of SBSs, the backhaul and wireless link capacities are limited.

ii) the catalog size and number of users with unknown ratings[4] are very large in practice.

iii) the optimal uncoded[5] cache decision for a given demand is non-tractable [39,85,102].

iv) the SBSs have to track, learn and estimate the sparse content popularity/rating matrix SBSs $\mathbf{P}^m(t)$ while making the cache decision.

In order to overcome these issues, we restrict ourselves to the fact that cache decision is made during peak-off hours, thus $\mathbf{X}(t)$ remains static during the content delivery in peak hours and is represented by $\mathbf{X}$. Additionally, the content popularity matrix is stationary during T time slots and identical among the base stations, thus $\mathbf{P}^m(t)$ is represented by $\mathbf{P}$.

---

[4]The term "rating" refers to the empirical value of content popularity/probability and is interchangeable throughout the chapter.

[5]In the information theoretical sense, the caching decision can be categorized into "coding" and "uncoded" groups (see [223] for example).

After these considerations, we now suppose that the problem can be decomposed into two parts in which the content popularity matrix $\mathbf{P}$ is first estimated, then is used in the caching decision $\mathbf{X}$ accordingly. In fact, if sufficient amount of users' ratings are available at the SBSs, we can construct a k-rank approximate popularity matrix $\mathbf{P} \approx \mathbf{N}^T\mathbf{F}$, by jointly learning the factor matrices $\mathbf{N} \in \mathbb{R}^{k \times N}$ and $\mathbf{F} \in \mathbb{R}^{k \times F}$ that minimizes the following cost function:

$$\underset{\mathbf{P}}{\text{minimize}} \sum_{(i,j) \in \mathbf{P}} \left(\mathbf{n}_i^T\mathbf{f}_j - P_{ij}\right)^2 + \mu\left(||\mathbf{N}||_F^2 + ||\mathbf{F}||_F^2\right) \tag{8.10}$$

where the summation is done over the user/content rating pairs (i,j) in the training set. The vectors $\mathbf{n}_i$ and $\mathbf{f}_j$ here describe the i-th and j-th columns of $\mathbf{N}$ and $\mathbf{F}$ matrices respectively, and $||.||_F^2$ represents the Frobenius norm. The parameter μ is used to provide a balance between the regularization and fitting the training data. Therein, high correspondence between the user factor matrix $\mathbf{N}$ and content factor matrix $\mathbf{F}$ leads to a better estimate of $\mathbf{P}$. In fact, the problem (8.10) is a regularized least square problem where the matrix factorization is embedded in the formulation. Despite various approaches, the matrix factorization methods are commonly used to solve this kind of problems and has many applications such as in recommendation systems (i.e., Netflix video recommendation). In our case detailed in the following sections, we have used regularized sparse SVD to solve the problem algorithmically which exploits the least square nature of the problem. The overview of these approaches, sometimes called CF tools, can be found in [209, 224]. When the estimation of content popularity matrix $\mathbf{P}$ is obtained, the caching decision $\mathbf{X}$ can be made in this scenario accordingly.

In practice, the estimation of $\mathbf{P}$ in (8.10) can be done by collecting/analysing large amount of available data on a *big-data platform* of the network operator, and strategic/popular contents from this estimation can be stored at the *cache-enabled base stations* whose cache decisions are represented by $\mathbf{X}$. By doing this, the backhaul offloading problem in (8.9) is minimized and higher satisfactions are achieved. Our network model including such an infrastructure is illustrated in Fig. 8.1. In the following, as a case study, we detail our big data platform and present users' traffic characteristics by analysing large amount of data on this platform. The processed data will be used to estimate the content popularity matrix $\mathbf{P}$ which is essentially required for the cache decision $\mathbf{X}$ and will be detailed in the upcoming sections.

8.3 Big Data Platform

The big data platform used in this work runs in the operator's core network. As mentioned before, the purpose of this platform is to store users' traffic data and extract useful information which are going to be used for content popularity estimation. In a nutshell, the operator's network consists of several districts with more than 10 regional core areas throughout Turkey. The average total traffic over all regional areas consists of approximately over 15 billion packets in uplink direction and over 20 billion packets in
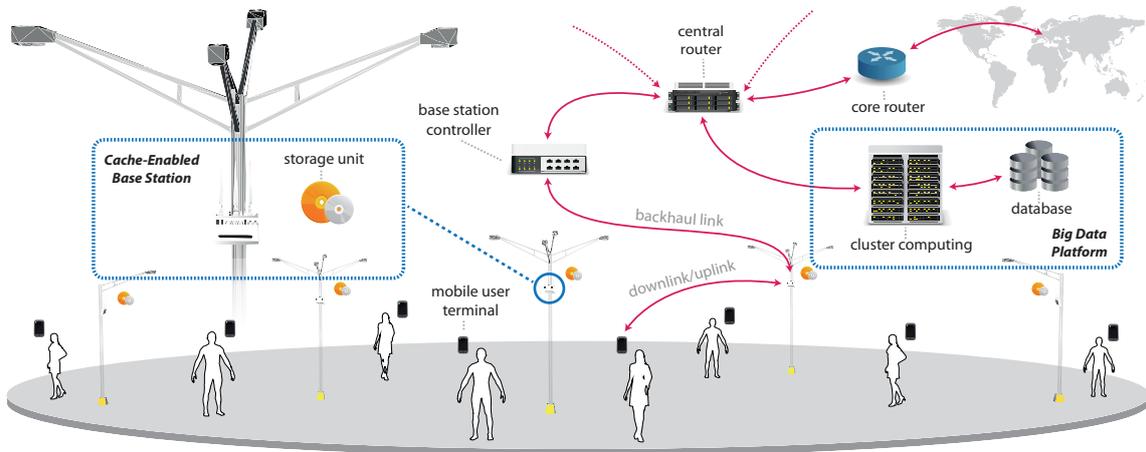
Figure 8.1: An illustration of the network model. A *big data platform* is in charge of tracking/predicting users' demand, whereas *cache-enabled base stations* store the strategic contents predicted on the big data platform.

the downlink direction daily. This corresponds to approximately over 80 TByte of total data flowing in uplink and downlink daily in a mobile operator's core network. The data usage behaviour results in exponential increase in data traffic of a mobile operator. For example, in 2012, the approximate total data traffic was over 7 TByte in both uplink and downlink daily traffic.

The streaming traces which will be detailed in the sequel, are obtained from one of the operator's core network region, includes the mobile traffic from many base stations, and are captured by a server on a high speed link of 200 Mbit/sec at peak hours. In order to capture Internet traffic data by the server in this platform, a procedure is initialized by mirroring real-world Gn interface data.[6] After mirroring stage of Gn interface, network traffic is transferred into the server on the platform. For our analysis, we have collected traffic of approximately 7 hours starting from 12 pm to 7 pm on Saturday 21'st of March 2015. This traffic is processed on the big data platform which is essentially based on Hadoop.

8.3.1 Hadoop platform

Among the available platforms, Hadoop stands out as the most notable one as it is an open source solution [225]. It is made up of a storage module, namely Hadoop Distributed File System (HDFS) and a computation module, namely MapReduce. Whereas HDFS can have centralized or distributed implementations, MapReduce inherently has a distributed structure that enables it to execute jobs in parallel on multiple nodes.

---

[6]Gn is an interface between Serving GPRS Support Node (SGCN) and Gateway GPRS Support Node (GGSN). Network packets sent from a user terminal to the packet data network (PDN), e.g. internet, pass through SGCN and GGSN where GPRS Tunneling Protocol (GTP) constitutes the main protocol in network packets flowing through Gn interface.

As stated in previous subsection, the accuracy and precision of the proposed mechanism was tested in operator's network. A data processing platform was implemented through using Cloudera's Distribution Including Apache Hadoop (CDH4) [226] version on four nodes including one cluster name node, with computations powers corresponding to each node with INTEL Xeon CPU E5-2670 running @2.6 GHz, 32 Core CPU, 132 GByte RAM, 20 TByte hard disk. This platform is used to extract the useful information from raw data which is described as follows.

8.3.2 Data extraction process

First, the raw data is parsed using Wireshark command line utility *tshark* [227] in order to extract the relevant fields of CELL-ID (or service area code (SAC) in our case, in order to uniquely identify a *service area* within a *location area*[7]), LAC, Hypertext Transfer Protocol (HTTP) request-uniform resource identifier (URI), tunnel endpoint identifier (TEID)[8] and TEID-DATA for data and control plane packets respectively, and FRAME TIME indicating arrival time of packets. The HTTP Request-URI is a Uniform Resource Identifier that identifies the resource upon which to apply the request. The *control* packets contain the information elements that carry the information required for future data packets. It contains cell identification ID (CELL-ID), LAC and TEID-DATA fields. The *data* packets contain HTTP-URI and TEID fields.

In the next step, after obtaining those relevant fields from both control and data packets, the extracted data is transferred into HDFS for further analysis. In HDFS, there can be done many data analytics performed over the collected data using Hive Query language (QL) [228]. For example, in order to calculate the HTTP Request-URIs at specific location, the HTTP-URI can be joined with CELL-ID-LAC fields over the same TEID and TEID-DATA fields for data and control packets respectively. In our analysis, due to the limitations on observable number of rows of HTTP-URI fields with a corresponding CELL-ID-LAC fields after mapping, we have proceeded with HTTP Request-URIs and TEID mappings.

From HDFS, a temporary table named *traces-table-temp* is constructed using Hive QL. The *traces-table-temp* has HTTP Request-URI, FRAME TIME and TEID fields. After constructing this table, the sizes of each HTTP Request-URI request is calculated using a separate URI-*size calculator* program that uses HTTPClient API [229] in order to obtain the final table called *traces-table* with fields of SIZE, HTTP Request-URIs, FRAME TIME and TEID. This table has approximately over 420.000 of 4 millions HTTP Request-URI's with SIZE field returned as not zero or null due to unavailability of HTTP response for some requests. Note that in a given session with a specific TEID, there can be multiple HTTP Request-URIs. Each TEID belongs to specific user. Each user can also have

---

[7]The service area identified by SAC is an area of one or more base stations, and belongs to a location area which is uniquely identified by location area code (LAC). Typically, tens or even hundreds of base stations operates in a given location area.

[8]A TEID uniquely identifies a tunnel endpoint on the receiving end of the GTP tunnel. A local TEID value is assigned at the receiving end of a GTP tunnel in order to send messages through the tunnel.

multiple TEIDs with multiple HTTP Request-URIs. The steps of data extraction process on the platform is summarized in Fig. 8.2. Note that the data extraction process is specific to our scenario for proactive caching. However, similar studies in terms of usage of big data platform and exploitation of big data analytics for telecom operators can be found in [230–235].
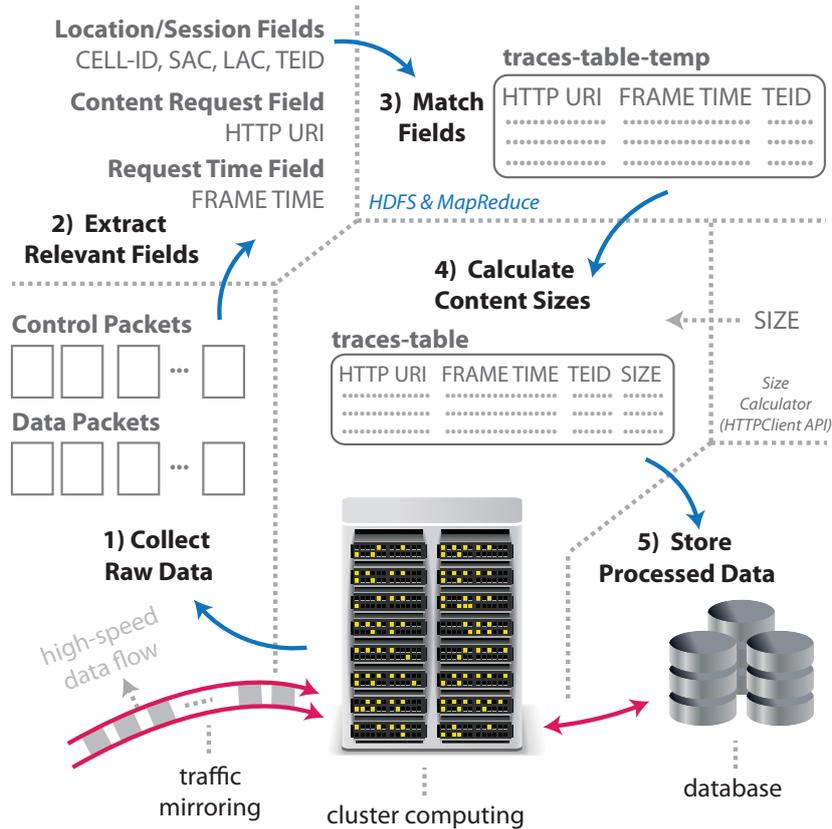
Figure 8.2: An overview of the data extraction process on the big data platform.

8.3.3 Traffic Characteristics

Based on information available in *traces-table*, the global content popularity distribution (namely HTTP-URI popularity distribution) in a decreasing ranked order is plotted in Fig. 8.3a. According to this available experimental data, we observe that the popularity behaviour of contents follows a Zipf law with steepness parameter $\alpha = 1.36$.[9] Therein, the Zipf curve is calculated in the least square sense from the collected traces and the parameter α is then found by evaluating the slope of the curve. On the other hand, cumulative size of ranked contents is given in Fig. 8.3b. The cumulative size up to 41-th most-popular contents has 0.1 GByte of size, whereas a dramatical increase appears

---

[9] The value of steepness parameter α can change depending on the scenario. For instance, the steepness parameter of content popularities in YouTube catalog varies from 1.5 to 2.5 [236, 237].
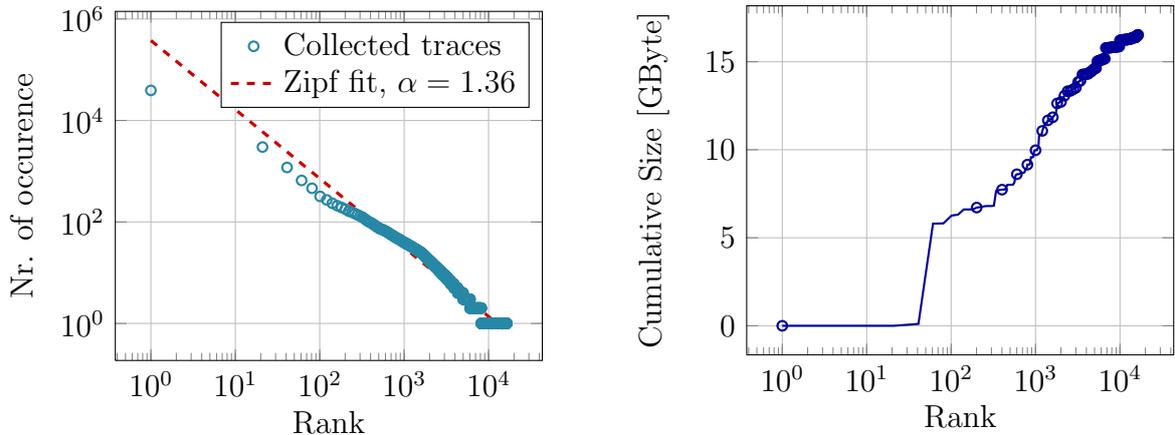
(a) Global content popularity distribution.

(b) Cumulative size distribution.

Figure 8.3: Behaviour of content popularity distribution.

afterwards. This basically shows that most of the requested contents in our traces has low content sizes and contents with larger sizes are relatively less requested.

We would like to note that a detailed characterization of the traffic for caching is left for future work. Indeed, characterization of the traffic in web proxies which are placed in the intermediate level of network [197], a specific video content catalog in a campus network [238], mobile traffic of users in Mexico [239] can be found in the literature. Compared to these works, we focus on the characterization traffic of mobile users collected from base stations in a large regional area and exploit this information for proactive caching (i.e., content popularity distribution, cumulative size distribution). Based on information available in *traces-table*, we in the following simulate a scenario of cache-enabled base stations.

8.4 Numerical Results and Discussions

The list of parameters for numerical setup is given in Table 8.1. For ease of analysis, the storage, backhaul, and wireless link capacities of small cells are assumed to be identical within each other.

In the simulations, all of D number of requests are taken from the processed data (namely *traces-table*), spanning over a time duration of 6 hours 47 minutes. The arrival times of each request (FRAME TIME), requested content (HTTP-URI) and content size (SIZE) are taken from the same table. Then, these requests are associated to M base stations pseudo-randomly. In order to solve the backhaul offloading problem in (8.9), the content popularity matrix $\mathbf{P}$ and caching strategy $\mathbf{X}$ are evaluated separately. In particular, the following two methods are used for constructing the content popularity matrix $\mathbf{P}$:

Table 8.1: List of simulation parameters.

Parameter	Description	Value
T	Time slots	6 hours 47 minutes
D	Number of requests	422529
F	Number of contents	16419
M	Number of small cells	16
$L_{\min}$	Min. size of a content	1 Byte
$L_{\max}$	Max. size of a content	6.024 GByte
$B(f)$	Bitrate of content f	4 Mbyte/s
$\sum_m C_m$	Total backhaul link capacity	3.8 Mbyte/s
$\sum_m \sum_n C'_m$	Total wireless link capacity	120 Mbyte/s

- *Ground Truth*: The content popularity matrix $\mathbf{P}$ is constructed from all available information in *traces-table* instead of solving the problem in (8.10). Note that the rows of $\mathbf{P}$ represent base stations and columns are contents. The rating density of this matrix is 6.42%.

- *Collaborative Filtering*: For the estimation of content popularity matrix $\mathbf{P}$, the problem in (8.10) is attempted by first choosing 10% of ratings from *traces-table* uniformly at random. Then, these ratings are used in the training stage of the algorithm and missing entries/ratings of $\mathbf{P}$ are estimated. Particularly, the regularized SVD from the CF methods [209, 240] is used in the algorithmic part.

After constructing the content popularity matrix $\mathbf{P}$ based on these above methods, the cache decision (modelled by the matrix $\mathbf{X}$) is made by storing the most-popular contents greedily at the SBSs until no storage space remains (see [39] for the details). Having these contents cached proactively at the SBSs at $t = 0$, the requests are then served until all of the contents are delivered. The performance metrics request satisfaction and backhaul load are calculated accordingly.

The evolution of users' request satisfaction with respect to the storage size is given Fig. 8.4a. The storage size is given in terms of percentage where 100% of storage size represents the sum of all size of contents in the catalog (17.7 GByte). From zero storage (0%) to full storage (100%), we can seen that the users' request satisfaction increases monotonically and goes up to 100%, both in ground truth and collaborative filtering approaches. However, there is a performance gap between the ground truth and CF until 87% of storage size, which is due to the estimation errors. For instance, with 40% of

storage size, the ground truth achieves 92% of satisfaction whereas the CF has value of 69%.

The evolution of backhaul load/usage with respect to the storage size of SBSs is given in Fig. 8.4b. As the storage size of SBSs increases, we see that both approaches reduces backhaul usage (namely higher offloading gains). For example, with 87% of storage size for caching, both approaches offload 98% of backhaul usage. The performance of ground truth is evidently higher than the CF as all of the available information is taken into consideration for caching. We also note that there is a dramatical decrease of backhaul usage in both approaches after a specific storage size. In fact, most of the previous works on caching assume a content catalog with identical content sizes. In our case, we are dealing with real traces in the numerical setup where the size of contents differs from content to content, as discussed in the previous section (see Fig. 8.3b). According to this scenario, on the one hand, caching a highly popular content with very small size might not reduce the backhaul usage dramatically. On the other hand, caching a popular content with very high size can dramatically reduce the backhaul usage. Therefore, as the CF approach used here is solely based on content popularity, it fails to capture these content size aspects on the backhaul usage, which in turn results in higher storage requirements to achieve the same performance as in the ground truth. This shows the importance of size distribution of popular contents.
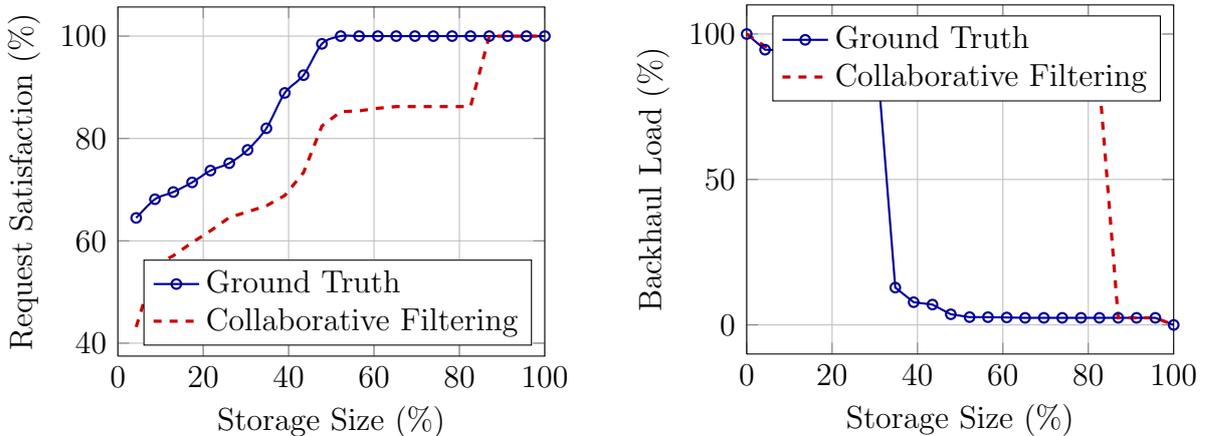
(a) Evolution of satisfaction with respect to the storage size.

(b) Evolution of backhaul usage with respect to the storage size.

Figure 8.4: Simulation results of proactive caching at the base stations.

We have so far compared the performance gains of these approaches with 10% of rating density in CF. In fact, as the rating density of CF for training increases, we expect to have less estimation error, thus resulting closer satisfaction gains to the ground truth. To show this, the change of root-mean-square error (RMSE) with respect to the training rating density is given in Fig. 8.5. Therein, we define the error as the root-mean-square of difference between users' content satisfaction of the ground truth and CF approaches over all possible storage sizes. Clearly, as observed in Fig. 8.5, the performance of CF is

improved by increasing the rating density, thus confirming our intuitions.
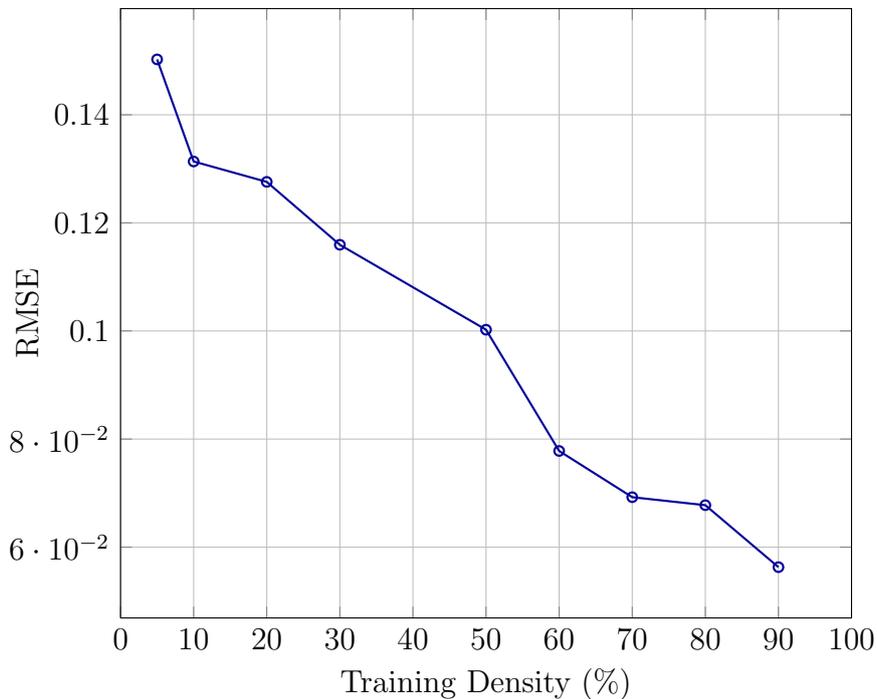
Figure 8.5: Evolution of RMSE with respect to the training density.

8.5 Closing Remarks

In this chapter, we have studied a proactive caching approach for 5G wireless networks by exploiting large amount of available data and employing machine learning tools. In particular, an experimental setup for data collection/extraction process has been demonstrated on a big data platform and machine learning tools (CF in particular) have been applied to predict the content popularity distribution. Depending on the rating density and storage size, the numerical results showed that several caching gains are possible in terms of users' request satisfactions and backhaul offloadings.

Part III

Conclusions, Outlook and Appendices

Chapter 9

Conclusions and Outlook

In this thesis, we have focused on proactive caching paradigm by leveraging small cell network deployments and caching capabilities at the edge of network, namely at small cells and user terminals. In the first part, we have characterized the gains of caching for different topologies, content popularity distributions and caching policies. The modeling has been carried out by using recent tools from stochastic geometry, and our expressions for average delivery rate and delay have been validated via numerical simulations. In the second part of the thesis, we have approached to the problem from a practical point of view, and conducted several studies for content popularity estimation and algorithmic aspects. The tools from machine learning and enabling big data allowed us to show the benefits of caching in practical scenarios, where we have drawn several conclusions based on storage size, content rating density, behaviour of content popularity and caching policy. Despite the fact that caching is gainful especially in limited-backhaul scenarios, there exist still several challenges which needs to be investigated in the future.

In particular, in Part I of the thesis where we have focused on modeling and performance analysis, we have the following future directions.

In Chapter 3 (Single-Tier Cellular Networks), the model and analysis can be extended by considering

- *Average delivery rate*: The performance metric we have defined is based on fixed rate transmission and does not exploit the full potential of instantaneous SINR to achieve higher rates in downlink. Even though this is done for tractability and we expect that new insights might be somewhat similar, additional effort for detailing of this metric is of high interest, in order to have a more realistic view to the system. Recent works in [52,58] take different performance metrics in this regard.

- *Coupling with physical layer parameters*: The coupling between physical layer parameters (i.e., SINR, target bitrate) and caching parameters (storage size, content popularity shape) are separable in the sense that, one can decompose these components and investigate the behaviour of them independently. Even though what we have done is useful for providing a holistic view to the system, a better under-

standing/coupling of caching problem with physical layer parameters is needed in the context of stochastic geometry modeling. The works in [52, 58] have proposed new models for that purpose.

- *Operational caching regime*: It seems evident from our investigations that adding more storage units to the small base stations results in a better system performance and the gains are not linear with respect to the storage size. Therefore, if one can introduce cost of installing and operating storage units, one can also wish to characterize in which regime caching can be useful/operational from cost point of view, while considering storage size and content popularity behaviour. In other words, putting very big storage units might be very costly and might only improve system performance gradually. Therefore, a characterization of this regime in stochastic geometry framework is needed.

- *Imperfect content popularity*: It is now clear that the gains of caching depends on content popularity behaviour/shape, where in our scenario we have assumed that the popularity distribution is known perfectly at the small base stations. An interesting venue in this direction is to introduce an imperfect content popularity model where the shape parameter or exact distribution is partially known. This could reveal potential gains of caching in a more realistic setup, and in some cases, one might not wish to cache contents if the uncertainty is above a threshold. The recent efforts for learning unknown content popularity in the context of stochastic geometry modeling can be found in [141, 142].

- *Demand locality*: The system model we have considered is based on the assumption that content popularity distribution f_{pop} is identical among users and is known at the base stations. If one can come up with a mathematical structure to capture local demand of users and make connection of these local demands with global content popularity distribution, a better understanding of caching problem for small cell deployments would be pointed out. In other sense, some deployment areas with few users' content demands or demand with different characteristics might lead telecoms operators to make choice on whether they should deploy storage units or not in such areas. The recent work in [54] investigates such a demand locality aspect.

- *Locally vs globally storing most popular contents*: Very relevant to demand locality as stated above, yet interesting question to answer is to whether store globally or locally popular contents at the small base stations. As most of stochastic geometry analysis in the literature is so far based on typical user assumption and exploitation of motion-invariant properties of point processes, which is in fact due toe tractability reasons, one may need to check out in details how to introduce such a structure in order to capture these aspects of caching. In this regard, the recent work in [132] compares the performance of globally storing popular contents with a random caching policy, and shows a regime where storing popular contents is not dramatically gainful. In any case, whether storing locally or globally popular contents requires investigations.

- *Backhaul rate splitting strategies*: The limited-backhaul we have considered so far is based on the idea of splitting the total backhaul link capacity among small base stations. This does not capture dynamic rate splitting strategies that rely on active base stations and load of the system. An interesting path to extend this model is to introduce a more comprehensive rate splitting strategy by considering dynamic load aspects of demand and users. Even though we have introduced novel backhaul delay/rate splitting strategies in Chapters 4 and 5, further investigations are still needed for a better understanding of gains of caching under limited-backhaul.

- *Heterogeneous networks*: In fact, we have considered multi-tier heterogeneous networks in Chapters 4 and 5. However, generalization of the single tier model in this chapter might be still extended to a more general setting with K-tiers, where some base stations in some tiers might have open/closed access, caching/non-caching capabilities and different path-loss exponents, transmit powers and target SINR requirements. In fact, several relevant works have appeared recently in this direction (see [58, 141, 142, 162, 166] for instance).

- *D2D communications*: Cache-enabled D2D communications on top of single or multi-tier heterogeneous network can also bring additional insights to the network designers. The inband/outband D2D communications supported with caching capabilities requires a better understanding of clustering of users, similarity of content access patterns, proximity, cache availability and connectivity conditions to base stations. In this regard, the work in [53] contains a scenario of D2D communications with users being clustered.

- *Wireless backhaul*: The backhaul we have considered so far relies on wired error-free links. When wireless backhaul is introduced into the scenario, one can take benefit of such a backhaul by prefeching/broadcasting/delivering contents in a more flexible manner, thus some of capacity limitations might not hold depending on the load/network conditions. This might lead to some scenarios where caching is not very useful in terms of delivery rate.

- *Green aspects*: Area power consumption and energy efficiency of caching is of high interest, as the unit energy consumed by the caches prefetching/delivery is in general assumed to be lower than the unit energy in the backhul prefecthing/delivery. Some recent works (see [33, 161] for instance) have started to focus on these aspects by using tools from stochastic geometry, however, more investigations are clearly needed to gain more insights about the network deployment.

- *Physical layer caching and multicast*: In fact, one of the benefit of caching is to provide more multi-casting opportunities at the base stations, thus, providing higher satisfactions at the mobile user terminals. Even though there exist some works in this regard in the literature, the characterization of gains in the stochastic geometry framework is still missing. In a relevant context, a scenario based on joint transmissions of cache-enabled base stations is considered in [58].

- *More curves/tradeoffs*: Characterization of caching gains with different parameter settings is of high relevance, and more tradeoffs related to caching need to be discovered.

In Chapter 4 (Multi-Tier Cellular Networks), the model and analysis can be further improved by considering

- *Heterogeneity in storage*: We have so far considered that storage size of small base stations are identical among each other. One may wish to relax this assumption by introducing heterogeneity in the storage sizes and characterize average delays in this system model. In fact, one can expect that the more heterogeneous storage distribution in the network result in more degraded system performance. However, exact answer to this question requires a clean-state thinking and more rigorous claims with mathematical proofs.

- *Heterogeneity in content lengths*: Most of previous works related to caching are based on the assumption that lengths of contents in the catalogue are identical. This is in fact for ease of analysis and does not induce loss of generality. As similar to heterogeneity in storage sizes, the contents/chunks with different lengths can be investigated in details and exact performance analysis of average delay can be revealed as an extension. The content length heterogeneity combined with an elegant mathematical model might lead to have better insights about the network performance.

- *Effect of content chunking*: Even though we have briefly mentioned about the chunks in the system model, so far, the cost/impact of having relatively small chunks on the performance of caching policies have not been considered in details. For example, if the chunks are infinitely small, caching based on StdPop or MixPop policies might not lead to a good performance due to more allocation of storage for chunk/content popularity tracking. One in this situation can characterize this phenomena and see the optimal chunking regime. Additionally, for a given chunk size, one can also find a balance between storage size allocated to caching popular files and caching uniformly at random in MixPop policy.

- *Traffic arrival/departure process*: All the analysis and calculations conducted in this chapter were based on the snapshots/realizations of topology together with system related random variables. The notion of time together with traffic arrival/departure process can lead to interesting insights about the overall performance system and caching policies. However, introducing time dynamics in the stochastic geometry setup requires careful technical treatment, as for instance the interference becomes temporally dependent. The traffic process together with some approximations and mild assumptions might bring fruitful design insights to the network designers. Recently, the works in [162, 163] consider a multiclass processor-sharing queue model to investigate the caching gains in a heterogeneous network.

- *Online caching policies*: The average delay characterization of online caching policies is yet another venue of work to investigate. In fact, together with time dynamics, the performance of online caching policies in our setup may result in a better understanding of caching policies, as one has to track/learn content popularity online and cache accordingly. The exploration v.s. exploitation trade-off of content popularity in caching decisions seems interesting to characterize, and the cost of unnecessary backhaul usage, especially in very limited backhaul scenarios is of interest. Recently, the work in [166] considers such an online caching policy.

- *Request overhead in uplink*: The content requests done by users via uplink have been neglected in this model. In fact, if the user is in outage in uplink, the content request might not reach to the base station which in turn changes behaviour of content popularity observed at the small cells and causes outages in delivery via downlink. These aspects can be considered in the system model and performance analysis in order to provide a holistic view to the problem.

- *Association policies*: Note that we have taken into account is the closest base station association policy. On top of this policy, SIR/SINR based caching-aware association policies together with multiple connectivity can lead to interesting results.

- *Total average network delay and cost optimization*: In the performance analysis, we have provided the expressions for total average network delay and cost. However, minimization of these metrics with respect to the storage size, number of small cells and macro cells are left for future work. Interested readers may wish to perform a global network optimization based on these metrics. Clearly, this would provide answers to the optimal deployment of cache-enabled small cells.

- *Wireless backhaul*: In addition to the wired backhaul, consideration of wireless backhaul in this model is also of high interest. Such an addition has been also coined above (for single-tier case in Chapter 3), whereas here we have multi-tier network with delay being as performance metric. This in fact might result in different expressions and tractability levels, expecting that new conclusions can be done.

- *Mobility*: Incorporation of mobility models into the scenario is yet another addition/extension for this system model, and can be done with the help of existing mobility models in stochastic geometry literature. However, one has to be technically careful with consideration of mobility as downlink performance and evolution of content popularity at the base stations are jointly influenced from movement of users.

- *More curves/tradeoffs*: As similar to our remarks for the single-tier network, more curves and trade-offs for this multi-tier network need to be investigated, especially in different realistic parameter settings. This should help system designers to have a quick decision/comparison for performance assessment of such heterogeneous networks.

In Chapter 5 (Clustered Cellular Networks), the model and analysis can be further detailed by focusing on

- *Tighter approximations*: Our approximations of average delivery rate work pretty well in high target bitrates. However, even though the trends for theoretical and simulation curves of average delivery rate are reasonably identical, the gap between these curves in low target bitrate has to be improved. Due to intra-dependence and inter-dependence between point processes, one might need to do the technical treatment carefully and check out the assumptions/simplifications we have made in the system model.

- *Simpler expressions*: We have so far obtained expressions of average delivery rate in the form of integrals and Laplace transforms. Exact values of these expressions can be evaluated pretty easily with modern software packages, and computations are fairly faster compared to performance evaluations via system level simulations. However, simpler and more elegant expressions with some reasonable assumptions are always desirable, in order to gain quick design insights and use/adapt these expressions for further mathematical models.

- *Other clustering processes*: In our model, Poisson hole process and Matérn cluster process have been considered for modeling of clustered small cells in coverage and capacity aided deployments respectively. An interesting direction in this regard is to check out other point processes (i.e. Ginibre point process) which can capture deployment of clustered cache-enabled small cells. The ultimate goal of looking to these kind of processes is to get design insights for more realistic cache-enabled base station deployments, while keeping analytical tractability.

- *Spatio-temporal dynamics of content popularity*: As small cells are clustered, one natural extension of this work is to assume that content requests are also spatially clustered, since users in some macro cells and small cell hot-spots may have different interests than some other users connected to macro and small cells. Such an extension together with temporal dynamics might lead to interesting observations.

- *SINR and limited backhaul*: The downlink in our scenario was only interference-limited (namely no noise) and the backhaul capacity was scaled with number of users and base stations. One interesting observation in this situation is to introduce SINR metric into the scenario and consider cases where backhaul has ultra high-speed connections. The tractability of expressions for average delivery right might be challenging due to SINR, and also, caching might not be helpful to turn the memory into bandwidth, since the only limiting factor could be the downlink. However, even though this is the case, one can still look the performance of this scenario to see in which regimes caching can be beneficial.

- *MIMO, cache-aware precoding schemes and multicast*: We have considered single antennas at macro cells, mobile user terminals and small cells. A holistic view of the network in stochastic geometry setup can be obtained by considering multiple

antennas at the nodes, enabling multi-cast and designing caching-aware precoding schemes. In this situation, more degree of freedom would be added to the network so that several different caching regimes that allow good performance results can be characterized. In fact, there exist works on this direction, as mentioned in the introduction chapter. However stochastic geometry modeling of such scenarios are still in their infancy, but can be also incorporated into cache-enabled networks after some careful technical treatment.

- *Scaling laws in hierarchical model*: Checking out the behaviour of hierarchical tree model in large parameter settings (i.e., as number of contents in the catalog goes to infinity) can lead to interesting results and help to asses the system performance in such infinite regimes. As mentioned in the introduction, there exist some caching works on this direction, however, a stochastic geometry-based approach is missing.

- *Geographical caching methods/policies*: The hierarchical model we have introduced in this chapter can lead to interesting optimization problems. We have defined the mathematical structures and posed the problem. However, a more detailed investigation supported with numerical analysis is needed. On the other hand, geographical caching methods under clustered spatio-temporal access patterns might be yet another scenario to investigate.

- *Online caching policies*: On top of geographical caching methods mentioned above, online caching policies can also be introduced into the scenario so that more practical insights can be obtained. This requires some simplifications as the interference becomes temporally dependent. As stated before, the works in [166] considers such an online policy.

- *More plots/tradeoffs*: Impact of different system parameters, such as base station density, cluster size, transmit, target can be characterized on top of what we have provided in this chapter.

For the second part, namely Part II, our future directions for content popularity learning and algorithmic aspects of caching can be summarized as follows.

In Chapter 6 (Proactive Caching), more insights can be obtained by looking into

- *Social network metrics*: We have considered eigen-vector centrality to measure the influence of users among the social network. On top of this, other centrality measures can be investigated numerically. Also, Shapley-based metrics might also be introduced for the fairness among the users in the network, as users are caching/sharing contents via D2D communications under limited transmission power constraints.

- *Relationship between D2D and social networks*: The social network we have studied is based on the assumption that links/friends in the social networks are mapped directly on the D2D network. This means that a user who has a friend in the social network has also the same friend in his proximity, so that D2D communications can be enabled. An interesting point to come up with a more detailed model for caching.

- *Collaborative filtering methods*: Regularized-SVD has been choosen for estimation of content popularity matrix. There exist large literature on collaborative filtering methods where one can apply other techniques (user-based, item-based, probabilistic models) to see the gains of caching for such methods.

- *Content dissemination process*: We have used CRP to model the content dissemination process in the social network. Other kind of stochastic process, such as Indian buffet process, can be considered. In this regard, the modeling of content dissemination process in social network is of high interest.

- *Complex network structures*: We have considered caching at base stations and user terminals. An interesting study would be adding caching capabilities not only at the edge of network but in several levels of network, i.e., hierarchical networks, multi-hop networks. A careful estimation of content access statistics is required for better performance results.

In Chapter 7 (Transfer Learning):

- *Analytical characterization of positive/negative transfer*: The analytical characterization of transfer learning in our setup is yet another interesting venue to investigate. The question of how much gain be obtained under a given parameter setting is of interest, especially to reveal positive/negative transfer parametrically. In fact, the recent works in [141, 142] propose new models for learning unknown content popularities via knowledge transfer.

- *Other source domains/real traces*: We have considered social interactions extracted from D2D as a source domain to improve the estimation in the target domain, namely at the base stations. Yet another interesting future work is to assess the performance of TL-based caching using real traces, collected from different sources.

In Chapter 8 (Big Data for Caching), several general directions for this practical setup can be given, such as

- *Detailed characterization of the demand*: An interesting future direction of this work is to conduct a more detailed characterization of the traffic which captures different spatio-temporal content access patterns for caching.

- *Novel machine learning algorithms*: In order to estimate the content access patterns for cache decision, the development of novel machine learning algorithms is yet another interesting direction.

- *Deterministic/randomized caching policies*: Design of new deterministic/randomized cache decision algorithms are required and should not be purely based on content popularity and storing most popular contents, so that higher backhaul offloading can be achieved while satisfying users' requests.

- *Experimental test-bed in realtime*: We have showed our results for content popularity estimation on the big data platform and conduct numerical studies of caching by using information gathered from real traces (i.e., content popularity, time, cell type, etc.). The ultimate goal of this line of work is to develop an online/real-time setup where content popularity estimation and caching at the base stations are simultaneously done in practice.

Bibliography

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019," *White Paper*, 2015. [Online]. Available: http://goo.gl/tZ6QMk

[2] Ericsson, "5G radio access - research and vision," *White Paper, [Online] http://goo.gl/Huf0b6*, 2012.

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

[4] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, 2013.

[5] Intel, "Rethinking the small cell business model," *White Paper, [Online] http://goo.gl/c2r9jX*, 2012.

[6] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[7] V. Etter, M. Kafsi, and E. Kazemi, "Been There, Done That: What Your Mobility Traces Reveal about Your Behavior," in *Mobile Data Challenge by Nokia Workshop, in conjunction with Int. Conf. on Pervasive Computing*, 2012.

[8] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 10–21, 2011.

[9] 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 Selected Areas in Communications*, vol. 30, no. 3, pp. 550–560, 2012.

[10] 3GPP TR 36.839 v11.0.0, "Mobility enhancements in heterogeneous networks (release 11)," *[Online] http://www.3gpp.org/DynaReport/36913.htm*, sept 2014.

[11] M. Simsek, M. Bennis, and I. Guvenc, "Mobility management in hetnets: a learning-based perspective," *EURASIP Journal on Wireless Communications and Networking*, vol. 2015, no. 1, p. 26, 2015.

[12] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. V. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3202–3212, 2013.

[13] M. Simsek, M. Bennis, and I. Guvenc, "Enhanced intercell interference coordination in hetnets: Single vs. multiflow approach," in *IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2013, pp. 725–729.

[14] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, "Dynamic uplink-downlink optimization in tdd-based small cell networks," *arXiv preprint arXiv: 1402.7292*, 2014.

[15] M. Bennis, M. Simsek, A. Czylwik, W. Saad, S. Valentin, and M. Debbah, "When cellular meets WiFi in wireless small cell networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 44–50, June 2013.

[16] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "On/off macrocells and load balancing in heterogeneous cellular networks," *arXiv preprint arXiv: 1305.5585*, 2013.

[17] F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5g," *arXiv preprint arXiv: 1312.0229*, 2013.

[18] E. Baştuğ, M. Bennis, and M. Debbah, "Living on the Edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, August 2014.

[19] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *IEEE International Conference on Computer Communications (INFOCOM'12)*, 2012, pp. 1107–1115.

[20] S. Göbbels, "Disruption tolerant networking by smart caching," *IEEE International Journal of Communication Systems*, pp. 569–595, April 2010.

[21] C. M. R. Inst., "C-RAN: The Road Towards Green RAN," *White Paper, [Online] http://goo.gl/jZt7TR*, 2011.

[22] S. Sezer, S. Scott-Hayward, P.-K. Chouhan, B. Fraser, D. Lake, J. Finnegan, N. Viljoen, M. Miller, and N. Rao, "Are we ready for SDN? implementation challenges for software-defined networks," *Communications Magazine, IEEE*, vol. 51, no. 7, 2013.

[23] L. A. Belady, "A study of replacement algorithms for a virtual-storage computer," *IBM Syst. J.*, vol. 5, no. 2, p. 78–101, 1966.

[24] J. Wang, "A survey of web caching schemes for the internet," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 5, pp. 36–46, October 1999.

[25] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[26] A. Araldo, M. Mangili, F. Martignon, and D. Rossi, "Cost-aware caching: optimizing cache provisioning and object placement in ICN," *arXiv preprint arXiv: 1406.5935*, 2014.

[27] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 26–36, 2012.

[28] M. A. Kader, E. Baştuğ, M. Bennis, E. Zeydan, A. Karatepe, A. S. Er, and M. Debbah, "Leveraging big data analytics for cache-enabled wireless networks," in *IEEE Global Communications Conference (GLOBECOM) Workshop*, San Diego, CA, USA, 2015.

[29] E. Zeydan, E. Baştuğ, M. Bennis, M. A. Kader, A. Karatepe, A. S. Er, and M. Debbah, "Big data caching for networking: Moving from cloud to edge," *IEEE Communications Magazine, 2015*, Submitted.

[30] E. Baştuğ, M. Bennis, E. Zeydan, M. A. Kader, A. Karatepe, A. S. Er, and M. Debbah, "Big data meets telcos: A proactive caching perspective," *Journal of Communications and Networks, Special Issue on Big Data Networking-Challenges and Applications*, vol. 17, no. 6, pp. 549–558, December 2015.

[31] E. Baştuğ, M. Bennis, and M. Debbah, *Proactive Caching in 5G Small Cell Networks*. Wiley, In Press (2015). [Online]. Available: http://goo.gl/vxKNz1

[32] M. Deghel, E. Baştuğ, M. Assaad, and M. Debbah, "On the benefits of edge caching for MIMO interference alignment," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'15)*, Stockholm, Sweden, 2015.

[33] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, "Caching on the edge: a green perspective for 5G networks," in *IEEE International Conference on Communications (ICC'15)*, London, UK, June 2015.

[34] E. Baştuğ, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," in *International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt'15) - D2D Workshop*, Mumbai, India, May 2015.

[35] E. Baştuğ, M. Bennis, and M. Debbah, "Anticipatory caching in small cell networks: A transfer learning approach," in *1st KuVS Workshop on Anticipatory Networks*, Stuttgart, Germany, 09/2014 2014.

[36] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP Journal on Wireless Communications and Networking*, no. 1, p. 41, February 2015.

[37] E. Baştuğ, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *International Symposium on Wireless Communication Systems (ISWCS'14)*, Barcelona, Spain, August 2014.

[38] ——, "Social and spatial proactive caching for mobile data offloading," in *Small Cell and 5G Networks (SmallNets) workshop in conjuction with IEEE International Conference on Communications (ICC'14)*, Sydney, Australia, 2014.

[39] E. Baştuğ, J.-L. Guénégo, and M. Debbah, "Proactive small cell networks," in *20th International Conference on Telecommunications (ICT'13)*, Casablanca, Morocco, May 2013.

[40] E. Baştuğ, J.-L. Guénégo, and M. Debbah", "Cloud storage for small cell networks," in *IEEE 1st International Conference on Cloud Networking (CloudNet'12)*, Paris, France, November 2012.

[41] K. Hamidouche, W. Saad, and M. Debbah, "Many-to-many matching games for proactive social-caching in wireless small cell networks," in *12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2014, pp. 569–574.

[42] K. Hamidouche, E. Baştuğ, M. Bennis, and M. Debbah, "Le caching proactif dans les réseaux cellulaires 5g," *La Revue de l'Electricité et de l'Electronique (REE)*, vol. 2014-4, 2014.

[43] E. Baştuğ, K. Hamidouche, W. Saad, and M. Debbah, "Centrality-based caching for mobile wireless networks," in *1st KuVS Workshop on Anticipatory Networks*, Stuttgart, Germany, 09/2014 2014.

[44] K. Hamidouche, W. Saad, and M. Debbah, "Distributed caching in dense small cell networks," in *IEEE Communication Theory Workshop*, Orange Country, CA, US, May 2015.

[45] A. Abboud, E. Baştuğ, K. Hamidouche, and M. Debbah, "Distributed caching in 5g networks: An alternating direction method of multipliers approach," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC'15)*, Stockholm, Sweden, 2015.

[46] K. Hamidouche, W. Saad, and M. Debbah, *Distributed Caching in C-RAN*. Cambridge University Press, Submitted.

[47] K. Hamidouche, W. Saad, M. Debbah, and H. V. Poor, "Mean-field games for distributed caching in ultra-dense small cell networks," in *American Control Conference*, Submitted 2016.

[48] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, "Content-aware user clustering and caching in wireless small cell networks," in *11th International Symposium on Wireless Communication Systems (ISWCS'14)*, Barcelona, Spain, 2014.

[49] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "In-network caching and content placement in cooperative small cell networks," in *1st International Conference on 5G for Ubiquitous Connectivity (5GU)*, Levi, Finland, November 2014.

[50] ——, "Match to cache: Optimizing user association and backhaul allocation in cache-aware small cell networks," in *IEEE International Conference on Communications (ICC'2015)*, London, UK, June 2015.

[51] A. Kumar and W. Saad, "On the tradeoff between energy harvesting and caching in wireless networks," in *IEEE International Conference on Communications (ICC'2015), Workshop on Green Communications and Networks*, London, UK, June 2015.

[52] S. Tamoor-ul Hassan, M. Bennis, P. H. Nardelli, and M. Latva-Aho, "Modeling and analysis of content caching in wireless small cell networks," *arXiv preprint arXiv: 1507.00182*, 2015.

[53] A. Altieri, P. Piantanida, L. R. Vega, and C. Galarza, "On fundamental trade-offs of device-to-device communications in large wireless networks," *arXiv preprint arXiv: 1405.2295*, 2014.

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

[55] A. Ghorbel, M. Kobayashi, and S. Yang, "Cache-enabled broadcast packet erasure channels with state feedback," *arXiv preprint arXiv: 1509.02074*, 2015.

[56] K. Wan, D. Tuninetti, and P. Piantanida, "On the optimality of uncoded cache placement," *arXiv preprint arXiv: 1511.02256*, 2015.

[57] V. Varma and T. Quek, "Congestion games in caching enabled heterogeneous cellular networks," in *IFIP Networking Conference (IFIP Networking), 2015*, May 2015, pp. 1–6.

[58] S. H. Chae, J. Y. Ryu, T. Q. S. Quek, and W. Choi, "Cooperative transmission via caching helpers," in *IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, USA, 2015.

[59] B. Zhou, Y. Cui, and M. Tao, "Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks," *arXiv preprint arXiv: 1504.04428*, 2015.

[60] H. Zhou, M. Tao, E. Chen, and W. Yu, "Content-centric multicast beamforming in cache-enabled cloud radio access networks," *arXiv preprint arXiv: 1504.05663*, 2015.

[61] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *arXiv preprint arXiv: 1509.06611*, 2015.

[62] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for d2d-assisted wireless caching networks," *arXiv preprint arXiv: 1509.06932*, 2015.

[63] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *arXiv preprint arXiv: 1512.06176*, 2015.

[64] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *arXiv preprint arXiv: 1512.06938*, 2015.

[65] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[66] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *arXiv preprint arXiv: 1308.0178*, 2013.

[67] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. Diggavi, "Hierarchical coded caching," in *IEEE International Symposium on Information Theory (ISIT'14)*, June 2014, pp. 2142–2146.

[68] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *arXiv preprint arXiv: 1311.3646*, 2013.

[69] J. Hachem, N. Karamchandani, and S. Diggavi, "Multi-level coded caching," *arXiv preprint arXiv: 1404.6563*, 2014.

[70] ——, "Content caching and delivery over heterogeneous wireless networks," *arXiv preprint arXiv: 1404.6560*, 2014.

[71] U. Niesen and M. A. Maddah-Ali, "Coded caching for delay-sensitive content," *arXiv preprint arXiv: 1407.4489*, 2014.

[72] Z. Chen, "Fundamental limits of caching: Improved bounds for small buffer users," *arXiv preprint arXiv: 1407.1935*, 2014.

[73] A. Sengupta, R. Tandon, and T. C. Clancy, "Fundamental limits of caching with secure delivery," *arXiv preprint arXiv: 1312.3961*, 2013.

[74] ——, "Improved approximation of storage-rate tradeoff for caching via new outer bounds," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1691–1695.

[75] A. Krishnan, N. S. Prem, V. M. Prabhakaran, and R. Vaze, "Critical database size for effective caching," *arXiv preprint arXiv: 1501.02549*, 2013.

[76] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *arXiv preprint arXiv: 1503.00265*, 2015.

[77] C.-Y. Wang, S. H. Lim, and M. Gastpar, "Information-theoretic caching: Sequential coding for computing," *arXiv preprint arXiv: 1504.00553*, 2015.

[78] S. Sahraei and M. Gastpar, "K users caching two files: An improved achievable rate," *arXiv preprint arXiv: 1512.06682*, 2015.

[79] S. Wang, W. Li, X. Tian, and H. Liu, "Fundamental limits of heterogenous cache," *arXiv preprint arXiv: 1504.01123*, 2015.

[80] W. Huang, S. Wang, L. Ding, F. Yang, and W. Zhang, "The performance analysis of coded cache in wireless fading channel," *arXiv preprint arXiv: 1504.01452*, 2015.

[81] S. D. Jad Hachem, Nikhil Karamchandani, "Effect of number of users in multi-level coded caching," *arXiv preprint arXiv: 1504.05931*, 2015.

[82] R. Timo and M. Wigger, "Joint cache-channel coding over erasure broadcast channels," *arXiv preprint arXiv: 1505.01016*, 2015.

[83] M. A. Maddah-Ali and U. Niesen, "Cache-aided interference channels," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 809–813.

[84] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," *arXiv preprint arXiv: 1512.07856*, 2015.

[85] 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 Communications Magazine*, vol. 51, no. 4, pp. 142–149, 2013.

[86] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *arXiv preprint arXiv: 1305.5216*, 2014.

[87] A. F. M. Mingyue Ji, Giuseppe Caire, "On the average performance of caching and coded multicasting with random demands," in *11th International Symposium on Wireless Communication Systems (ISWCS'14)*, Barcelona, Spain, August 2014.

[88] A. F. Molisch, G. Caire, D. Ott, J. R. Foerster, D. Bethanabhotla, and M. Ji, "Caching eliminates the wireless bottleneck in video-aware wireless networks," *arXiv preprint arXiv: 1405.5864*, 2014.

[89] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *arXiv preprint arXiv: 1405.5336*, 2014.

[90] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *arXiv preprint arXiv: 1502.03124*, 2015.

[91] M. Ji, A. Tulino, J. Llorca, and G. Caire, "Caching-aided coded multicasting with multiple random requests," in *IEEE Information Theory Workshop (ITW)*. IEEE, 2015, pp. 1–5.

[92] S.-W. Jeon, S.-N. Hong, M. Ji, and G. Caire, "On the capacity of multihop device-to-device caching networks," in *IEEE Information Theory Workshop (ITW)*. IEEE, 2015, pp. 1–5.

[93] G. Vettigli, M. Ji, A. M. Tulino, J. Llorca, and P. Festa, "An efficient coded multicasting scheme preserving the multiplicative caching gain," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2015, pp. 251–256.

[94] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite length analysis of caching-aided coded multicasting," in *52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2014, pp. 914–920.

[95] M. Ji, M. F. Wong, A. M. Tulino, J. Llorca, G. Caire, M. Effros, and M. Langberg, "On the fundamental limits of caching in combination networks," in *IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2015, pp. 695–699.

[96] S.-W. Jeon, S.-N. Hong, M. Ji, and G. Caire, "Caching in wireless multihop device-to-device networks," in *IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 6732–6737.

[97] M. Ji, K. Shanmugam, G. Vettigli, J. Llorca, A. M. Tulino, and G. Caire, "An efficient multiple-groupcast coded multicasting scheme for finite fractional caching," in *IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 3801–3806.

[98] B. Azari, O. Simeone, U. Spagnolini, and A. Tulino, "Hypergraph-based analysis of clustered cooperative beamforming with application to edge caching," *arXiv preprint arXiv: 1510.06222*, 2015.

[99] S.-W. Jeon, S.-N. Hong, M. Ji, G. Caire, and A. F. Molisch, "Wireless multihop device-to-device caching networks," *arXiv preprint arXiv: 1511.02574*, 2015.

[100] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," in *IEEE International Symposium on Information Theory Proceedings (ISIT'13)*, July 2013, pp. 1017–1021.

[101] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," *arXiv preprint arXiv: 1402.7314*, 2014.

[102] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Transactions on Communications*, vol. 62, no. 10, pp. 3665–3677, October 2014.

[103] D. Kosmanos, A. Argyriou, and L. Tassiulas, "Optimizing video quality in dense small-cell wireless networks with packet overhearing," in *IEEE Global Communications Conference (GlobeCom'14)*. Austin, TX, USA: IEEE, December 2014.

[104] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Enhancing wireless networks with caching: Asymptotic laws, sustainability & trade-offs," *Computer Networks*, vol. 64, pp. 353–368, 2014.

[105] S. Gitzenis, S. Toumpis, and L. Tassiulas, "Efficient file replication in large wireless networks with dynamic popularity," in *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine'14)*. Rhodes, Greece: IEEE, August 2014.

[106] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Joint caching and base station activation for green heterogeneous cellular networks," in *IEEE International Conference on Communications (ICC'15)*, London, UK, June 2015.

[107] P. Blasco and D. Gündüz, "Learning-based optimization of cache content in a small cell base station," *arXiv preprint arXiv: 1402.3247*, 2014.

[108] ——, "Content-level selective offloading in heterogeneous networks: Multi-armed bandit optimization and regret bounds," *arXiv preprint arXiv: 1407.6154*, 2014.

[109] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, "Learning distributed caching strategies in small cell networks," in *International Symposium on Wireless Communication Systems (ISWCS)*, Barcelona, Spain, August 2014.

[110] M. Gregori, J. Gomez-Vilardebo, J. Matamoros, and D. Gündüz, "Joint transmission and caching policy design for energy minimization in the wireless backhaul link," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1004–1008.

[111] J. Pääkkönen, C. Hollanti, and O. Tirkkonen, "Device-to-device data storage for mobile cellular systems," in *IEEE Globecom Workshops (GC Wrokshops)*, 2013, pp. 671–676.

[112] A. Barreal, C. Hollanti, D. Karpuk, and H.-f. Lu, "Algebraic codes and a new physical layer transmission protocol for wireless distributed storage systems," *arXiv preprint arXiv: 1405.4375*, 2014.

[113] C. Hollanti, D. Karpuk, A. Barreal, and H.-f. F. Lu, "Space-time storage codes for wireless distributed storage systems," *arXiv preprint arXiv: 1404.6645*, 2014.

[114] J. Pääkkönen, C. Hollanti, and O. Tirkkonen, "Device-to-device data storage with regenerating codes," *arXiv preprint arXiv: 1411.1608*, 2014.

[115] J. Pedersen, A. G. i Amat, I. Andriyanova, and F. Brännström, "Repair scheduling in wireless distributed storage with D2D communication," *arXiv preprint arXiv: 1504.06231*, 2015.

[116] A. Liu and V. Lau, "Cache-induced opportunistic MIMO cooperation: A new paradigm for future wireless content access networks," in *IEEE International Symposium on Information Theory (ISIT'14)*, June 2014, pp. 46–50.

[117] ——, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 390–402, January 2014.

[118] ——, "Exploiting base station caching in mimo cellular networks: Opportunistic cooperation for video streaming," *IEEE Transactions on Signal Processing*, vol. 63, no. 1, pp. 57–69, January 2015.

[119] ——, "On the improvement of scaling laws for wireless ad hoc networks with physical layer caching," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 161–165.

[120] ——, "Asymptotic scaling laws of wireless adhoc network with physical layer caching," *arXiv preprint arXiv: 1510.05205*, 2015.

[121] E. Altman, K. Avrachenkov, and J. Goseling, "Coding for caches in the plane," *arXiv preprint arXiv: 1309.0604*, 2013.

[122] M. Mitici, J. Goseling, M. de Graaf, and R. J. Boucherie, "Deployment vs. data retrieval costs for caches in the plane," Enschede, the Netherlands, December 2013. [Online]. Available: http://doc.utwente.nl/88064/

[123] M. Mitici, J. Goseling, M. Graaf, and R. J. Boucherie, "Energy-delay trade-off of wireless data collection in the plane," in *35th WIC Symposium on Information Theory in the Benelux*. Technical University Eindhoven and IEEE, November 2014, pp. 3–10.

[124] J. Gong, S. Zhou, Z. Zhou, and Z. Niu, "Proactive push with energy harvesting based small cells in heterogeneous networks," *arXiv preprint arXiv: 1501.06239*, 2015.

[125] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "GreenDelivery: Proactive content caching and push with energy harvesting based small cells," *arXiv preprint arXiv: 1503.04254*, 2015.

[126] D. Malak and M. Al-Shalash, "Optimal caching for device-to-device content distribution in 5g networks," in *Globecom Workshops (GC Wkshps)*. IEEE, 2014, pp. 863–868.

[127] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Modeling and performance analysis of clustered device-to-device networks," *arXiv preprint arXiv: 1508.02668*, 2015.

[128] P. Ostovari, A. Khreishah, and J. Wu, "Cache content placement using triangular network coding," in *IEEE Wireless Communications and Networking Conference (WCNC)*, 2013, pp. 1375–1380.

[129] V. A. Siris, X. Vasilakos, and G. C. Polyzos, "Efficient proactive caching for supporting seamless mobility," *arXiv preprint arXiv: 1404.4754*, 2014.

[130] V. Siris and D. Dimopoulos, "Multi-source mobile video streaming with proactive caching and d2d communication," in *2015 IEEE 16th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, June 2015, pp. 1–6.

[131] A. Gharaibeh, A. Khreishah, I. Khalil, and J. Wu, "Asymptotically-optimal incentive-based en-route caching scheme," *arXiv preprint arXiv: 1404.4639*, 2014.

[132] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," *arXiv preprint arXiv: 1409.7626*, 2014.

[133] Z. Ming, M. Xu, and D. Wang, "InCan: In-network cache assisted eNodeB caching mechanism in 4G LTE networks," *Computer Networks*, no. 0, pp. –, 2014.

[134] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'14)*. Washington, DC, USA: IEEE, September 2014.

[135] G. Alfano, M. Garetto, and E. Leonardi, "Content-centric wireless networks with limited buffers: when mobility hurts," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 1815–1823.

[136] D. Liu and C. Yang, "Will caching at base station improve energy efficiency of downlink transmission?" in *EEE Global Conference on Signal and Information Processing (GlobalSIP'14)*. Atlanta, Georgia, USA: IEEE, December 2014.

[137] P. Sermpezis, L. Vigneri, and T. Spyropoulos, "Offloading on the edge: Analysis and optimization of local data storage and offloading in hetnets," *arXiv preprint arXiv: 1503.00648*, 2015.

[138] Z. Wang and V. W. Wong, "A novel D2D data offloading scheme for LTE networks," in *IEEE International Conference on Communications (ICC'15)*. London, UK: IEEE, June 2015.

[139] N. K. Purushothama, L. Massoulie, E. Baccelli, A. C. Viana, and D. Towsley, "On the Interaction between Content Caching and Request Assignment in Cellular Cache Networks," INRIA Saclay, Research Report RR-8707, March 2015. [Online]. Available: https://hal.inria.fr/hal-01138204

[140] W. C. Ao and K. Psounis, "Distributed caching and small cell cooperation for fast content delivery," in *ACM MobiHoc*, Hangzhou, China, June 2015.

[141] B. B. Nagaraja and K. G. Nagananda, "Caching with unknown popularity profiles in small cell networks," *arXiv preprint arXiv: 1504.03632*, 2015.

[142] B. Bharath, K. Nagananda, and H. V. Poor, "A learning-based approach to caching in heterogenous small cell networks," *arXiv preprint arXiv: 1508.03517*, 2015.

[143] B. O. Symeon Chatzinotas, Dimitrios Christopoulos, "Cellular-broadcast service convergence through caching for CoMP cloud RANs," *arXiv preprint arXiv: 1504.08274*, 2015.

[144] A. Khreishah and J. Chakareski, "Collaborative caching for multicellcoordinated systems," *CNTCV 2015, In conjunction with IEEE INFOCOM*, 2015.

[145] T. Wang, L. Song, and Z. Han, "Dynamic femtocaching for mobile users," in *IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2015, pp. 861–865.

[146] J. Zhang, X. Lin, C.-C. Wang, and X. Wang, "Coded caching for files with distinct file sizes," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1686–1690.

[147] S. P. Shariatpanahi, H. Shah-Mansouri, and B. Hossein Khalaj, "Caching gain in interference-limited wireless networks," *IET Communications*, vol. 9, no. 10, pp. 1269–1277, 2015.

[148] W. Shin, B.-Y. Min, and D. K. Kim, "Vehicaching: Embracing user request on vehicle route with proactive data transportation," in *IEEE 81st Vehicular Technology Conference (VTC Spring)*. IEEE, 2015, pp. 1–5.

[149] J. Jiang, S. Zhang, B. Li, and B. Li, "Maximized cellular traffic offloading via device-to-device content sharing," *IEEE Journal on Selected Areas in Communications*, vol. PP, no. 99, pp. 1–1, 2015.

[150] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, "On the complexity of optimal routing and content caching in heterogeneous networks," *arXiv preprint arXiv: 1501.00216*, 2014.

[151] M. Gerami, M. Xiao, S. Salimi, and M. Skoglund, "Secure partial repair in wireless caching networks with broadcast channels," *arXiv preprint arXiv: 1507.05533*, 2015.

[152] V. Bioglio, F. Gabry, and I. Land, "Optimizing mds codes for caching at the edge," *arXiv preprint arXiv: 1508.05753*, 2015.

[153] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," *arXiv preprint arXiv: 1509.00558*, 2015.

[154] M. Erol-Kantarci, "Cache-at-relay: energy-efficient content placement for next-generation wireless relays," *International Journal of Network Management*, 2015.

[155] L. Marini, J. Li, and Y. Li, "Distributed caching based on decentralized learning automata," in *IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 3807–3812.

[156] Z. H. Awan and A. Sezgin, "Fundamental limits of caching in d2d networks with secure delivery," in *IEEE International Conference on Communication Workshop (ICCW)*. IEEE, 2015, pp. 464–469.

[157] Y. Ugur, Z. H. Awan, and A. Sezgin, "Cloud radio access networks with coded caching," *arXiv preprint arXiv: 1512.02385*, 2015.

[158] T. A. Johnson and P. Seeling, "Browsing the mobile web: device, small cell, and distributed mobile caches," in *IEEE International Conference on Communication Workshop (ICCW)*. IEEE, 2015, pp. 1025–1029.

[159] A. Ramakrishnan, C. Westphal, and A. Markopoulou, "An efficient delivery scheme for coded caching," in *27th International Teletraffic Congress (ITC 27)*. IEEE, 2015, pp. 46–54.

[160] N. Liu and W. Kang, "The multiple access diamond channel with caching relays," in *IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 2862–2866.

[161] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *arXiv preprint arXiv: 1505.06615*, 2015.

[162] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. PP, no. 99, pp. 1–1, 2015.

[163] C. Yang, Z. Chen, Y. Yao, and B. Xia, "Performance analysis of wireless heterogeneous networks with pushing and caching," in *IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 2190–2195.

[164] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for d2d assisted wireless caching networks," *arXiv preprint arXiv: 1510.07865*, 2015.

[165] X. Zhao, C. Yang, Y. Yao, Z. Chen, and B. Xia, "Cognitive and cache-enabled d2d communications in cellular networks," *arXiv preprint arXiv: 1510.06480*, 2015.

[166] S. A. R. Zaidi, M. Ghogho, and D. C. McLernon, "Information centric modeling for two-tier cache enabled cellular networks," in *IEEE International Conference on Communication Workshop (ICCW)*. IEEE, 2015, pp. 80–86.

[167] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching and file sharing under heterogeneous file preferences," *arXiv preprint arXiv: 1504.05931*, 2015.

[168] B. Chen and C. Yang, "Performance gain of precaching at users in small cell networks," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'15)*. IEEE, 2015, pp. 2190–2195.

[169] Z. Zhou, M. Dong, K. Ota, and Z. Chang, "Energy-efficient context-aware matching for resource allocation in ultra-dense small cells," *IEEE Access*, vol. 3, pp. 1849–1860, 2015.

[170] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the placement delivery array design in centralized coded caching scheme," *arXiv preprint arXiv: 1510.05064*, 2015.

[171] A. Gharaibeh, A. Khreishah, B. Ji, and M. Ayyash, "A provably efficient online collaborative caching algorithm for multicell-coordinated systems," *arXiv preprint arXiv: 1509.02911*, 2015.

[172] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless bc: Interplay of coded-caching and csit feedback," *arXiv preprint arXiv: 1511.03961*, 2015.

[173] J. Li, W. Chen, M. Xiao, F. Shu, and X. Liu, "Efficient video pricing and caching in heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, December 2015.

[174] F. Alotaibi, S. Hosny, J. Tadrous, H. E. Gamal, and A. Eryilmaz, "Towards a marketplace for mobile content: Dynamic pricing and proactive caching," *arXiv preprint arXiv: 1511.07573*, 2015.

[175] L. Xiang, D. W. K. Ng, T. Islam, R. Schober, and V. W. Wong, "Cross-layer optimization of fast video delivery in cache-enabled relaying networks," *arXiv preprint arXiv: 1511.05410*, 2015.

[176] R. Hou, Y. Cheng, L. X. Cai, and H. Zhuang, "Performance evaluation for caching-based content distribution in backhaul-limited small cell networks," in *International Conference on Wireless Communications & Signal Processing (WCSP)*. IEEE, 2015, pp. 1–5.

[177] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, October 2010.

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

[179] E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "Modeling and delay analysis of geographical caching methods in cellular networks," *(to be submitted to a special issue in) IEEE Journal on Selected Areas in Communications*, 2016.

[180] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Modeling and analysis of geographical caching methods in clustered cellular networks," *(to be submitted to) IEEE Transactions on Wireless Communications*, 2016.

[181] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, "Energy consumption aspects of cache-enabled 5g wireless networks," *(to be submitted to) IEEE Transactions on Wireless Communications*, 2016.

[182] F. Dilmi, E. Baştuğ, and M. Debbah, "FlexibleEarth3D : Un kit de visualisation pour les simulations des réseaux 5G," in *En préparation*, 2016.

[183] M. Maso, E. Baştuğ, L. S. Cardoso, M. Debbah, and O. Ozdemir, "Reconfigurable cognitive transceiver for opportunistic networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, 05/2014 2014.

[184] E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "Deployment cost and delay of caching in two-tiered networks," *(to be submitted to) a conference*, 2016.

[185] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Benefits of edge caching in coverage and capacity-aided heteregeneous networks," *(to be submitted to) a conference*, 2016.

[186] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, "Energy consumption aspects of cache-empowered heterogeneous networks: Optimization and analysis," *(to be submitted to) a conference*, 2016.

[187] F. Dilmi, E. Baştuğ, and M. Debbah, "FlexibleEarth3D: A visualization toolkit for 5G networks simulations," in *(to be submitted to) a conference*, 2016.

[188] E. Baştuğ, A. Menafoglio, and T. Okhulkova, "Polynomial chaos expansion for an efficient uncertainty and sensitivity analysis of complex numerical models," in *ESREL 2013*, Amsterdam, Netherlands, 2013.

[189] B. Mawlawi, E. Baştuğ, C. Nerquizian, S. Azarian, and M. Debbah, "Non-invasive green small cell networks," in *46th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, USA, 2012.

[190] C. Cicconetti, G. Wunder, I. Bucaille, N. Cardona, V. Mancuso, M. Dräxler, A. de la Oliva, T. Ratnarajah, L. Dittmann, P. Rost *et al.*, "5G radio network architecture," 2014, [Online] http://goo.gl/quVgsK.

[191] J. Andrews, F. Baccelli, and R. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3122–3134, November 2011.

[192] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[193] J. Hoydis and M. Debbah, "David vs goliath or small cells vs massive mimo," *[Online] http://goo.gl/isfya5*, 2011.

[194] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing.* UK: Cambridge University Press, 2007.

[195] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry and Wireless Networks: Volume 1: THEORY.* Now Publishers Inc, 2009, vol. 1.

[196] H. Inaltekin, M. Chiang, H. V. Poor, and S. B. Wicker, "On unbounded path-loss models: effects of singularity on wireless network performance," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1078–1092, September 2009.

[197] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *IEEE Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'99)*, vol. 1. IEEE, 1999, pp. 126–134.

[198] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*. ACM, 2011, pp. 305–316.

[199] C. Fricker, P. Robert, and J. Roberts, "A versatile and accurate approximation for LRU cache performance," in *Proceedings of the 24th International Teletraffic Congress*, ser. ITC '12. International Teletraffic Congress, 2012, pp. 8:1–8:8.

[200] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies*, ser. CoNEXT '09. New York, NY, USA: ACM, 2009, pp. 1–12.

[201] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1305–1314, September 2002.

[202] F. Baccelli, M. Klein, M. Lebourges, and S. Zuyev, "Stochastic geometry and architecture of communication networks," *Telecommunication Systems*, vol. 7, no. 1-3, pp. 209–227, 1997.

[203] F. Baccelli and S. Zuyev, "Poisson-voronoi spanning trees with applications to the optimization of communication networks," *Operations Research*, vol. 47, no. 4, pp. 619–631, 1999.

[204] S. Lee and K. Huang, "Coverage and economy of cellular networks with many base stations," *IEEE Communications Letters*, vol. 16, no. 7, pp. 1038–1040, July 2012.

[205] N. Deng, W. Zhou, and M. Haenggi, "Heterogeneous cellular network models with dependence," *IEEE Journal on Selected Areas in Communications*, Accepted (2015).

[206] M. Haenggi, *Stochastic geometry for wireless networks*. Cambridge University Press, 2012.

[207] D. Zwillinger, *Table of integrals, series, and products*. Elsevier, 2014.

[208] Netflix, "Netflix prize," *[Online] http://www.netflixprize.com*, 2009.

[209] J. Lee, M. Sun, and G. Lebanon, "A comparative study of collaborative filtering algorithms," *[Online] arXiv: 1205.3193*, 2012.

[210] P. Arkadiusz, "Improving regularized singular value decomposition for collaborative filtering," in *Proceedings of KDD cup and workshop Vol. 2007.*, 2007.

[211] M. Newman, *Networks: an introduction.* Oxford University Press, 2009.

[212] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651 – 666, 2010.

[213] T. L. Griffiths and Z. Ghahramani, "The Indian Buffet Process: An Introduction and Review," *J. Mach. Learn. Res.*, vol. 12, pp. 1185–1224, July 2011.

[214] M. A. I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* National Bureau of Standards Applied Mathematics Series 55. Tenth Printing, 1972.

[215] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[216] C. Lynch, "Big data: How do your data grow?" *Nature*, vol. 455, no. 7209, pp. 28–29, 2008.

[217] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing.* ACM, 2012, pp. 13–16.

[218] T. H. Luan, L. Gao, Z. Li, Y. Xiang, and L. Sun, "Fog computing: Focusing on mobile users at the edge," *arXiv preprint arXiv: 1502.01815*, 2015.

[219] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[220] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[221] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive data download and user demand shaping for data networks," *submitted to IEEE Transactions on Information Theory [Online] arXiv: 1304.5745*, 2014.

[222] Google, "Recommended upload encoding settings (Advanced)," https://goo.gl/KJXfhh, 2015, [Online; accessed 30-August-2015].

[223] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[224] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, no. 8, pp. 30–37, August 2009.

[225] "Apache Hadoop," http://hadoop.apache.org/, 2015, [Online; accessed 02-April-2015].

[226] "Cloudera," http://www.cloudera.com/content/cloudera/en/documentation.html, 2015, [Online; accessed 02-April-2015].

[227] "The Wireshark Network Analyzer 1.12.2," https://www.wireshark.org/docs/man-pages/tshark.html, 2015, [Online; accessed 02-April-2015].

[228] "Apache Hive TM," https://hive.apache.org/, 2015, [Online; accessed 02-April-2015].

[229] Apache, "HttpClient API Tutorial," https://hc.apache.org/httpcomponents-client-ga/tutorial/pdf/httpclient-tutorial.pdf, 2015, [Online; accessed 25-April-2015].

[230] Y. Dong, Q. Ke, Y. Cai, B. Wu, and B. Wang, "Teledata: data mining, social network analysis and statistics analysis system based on cloud computing in telecommunication industry," in *Proceedings of the third international workshop on Cloud data management.* ACM, 2011, pp. 41–48.

[231] H.-D. J. Jeong, W. Hyun, J. Lim, and I. You, "Anomaly teletraffic intrusion detection systems on hadoop-based platforms: A survey of some problems and solutions," in *15th International Conference on Network-Based Information Systems (NBiS).* IEEE, 2012, pp. 766–770.

[232] J. Magnusson and T. Kvernvik, "Subscriber classification within telecom networks utilizing big data technologies and machine learning," in *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications.* ACM, 2012, pp. 77–84.

[233] W. Indyk, T. Kajdanowicz, P. Kazienko, and S. Plamowski, "Mapreduce approach to collective classification for networks," in *Artificial Intelligence and Soft Computing.* Springer, 2012, pp. 656–663.

[234] O. F. Celebi, E. Zeydan, O. F. Kurt, O. Dedeoglu, O. Ileri, B. A. Sungur, A. Akan, and S. Ergut, "On use of big data for enhancing network coverage analysis," in *20th International Conference on Telecommunications (ICT'13)*, Casablanca, Morocco, May 2013.

[235] I. A. Karatepe and E. Zeydan, "Anomaly detection in cellular network data using big data analytics," in *Proceedings of European Wireless 2014.* VDE, 2014, pp. 1–5.

[236] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 1–14.

[237] D. Rossi, G. Rossini *et al.*, "On sizing ccn content stores by exploiting topological information." in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2012, pp. 280–285.

[238] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network–measurements, models, and implications," *Computer Networks*, vol. 53, no. 4, pp. 501–514, 2009.

[239] E. Mucelli Rezende Oliveira, A. Carneiro Viana, K. P. Naveen, and C. Sarraute, "Measurement-driven mobile data traffic modeling in a large metropolitan area," INRIA, Research Report RR-8613, October 2014.

[240] A. Paterek, "Improving regularized singular value decomposition for collaborative filtering," in *Proceedings of KDD cup and workshop*, vol. 2007, 2007, pp. 5–8.

[241] W. C. Cheung, T. Q. Quek, and M. Kountouris, "Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 3, pp. 561–574, April 2012.

[242] F. Baccelli, B. Błaszczyszyn, and P. Mühlethaler, "Stochastic analysis of spatial and opportunistic aloha," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1105–1119, September 2009.

[243] D. Chen, T. Quek, and M. Kountouris, "Backhauling in heterogeneous cellular networks: Modeling and tradeoffs," *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3194–3206, June 2015.

[244] U. Schilcher, C. Bettstetter, and G. Brandner, "Temporal correlation of interference in wireless networks with rayleigh block fading," *IEEE Transactions on Mobile Computing*, vol. 11, no. 12, pp. 2109–2120, December 2012.

[245] R. K. Ganti and M. Haenggi, "Interference and outage in clustered wireless ad hoc networks," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4067–4086, September 2009.

Appendix A

Single-Tier Cellular Networks

A.1 Proof of Theorem 1

In order to prove Theorem 1, we modify some useful results from [191]. Conditioning on the nearest base station at a distance r from the typical user, the outage probability can be written as:

$$p_{\text{out}}(\lambda, T, \alpha, S, \gamma) = \mathbb{E}_r \Big[1 - \mathbb{P}[\ln(1 + \text{SINR}) > T, f_o \in \Delta_{b_o} \mid r] \Big].$$

Since expectation is a linear operator and these two events are independent, the above expression can be decomposed as:

$$p_{\text{out}}(\lambda, T, \alpha, S, \gamma) = 1 - \underbrace{\mathbb{E}_r \Big[\mathbb{P} \left[\ln(1 + \text{SINR}) > T \mid r \right] \Big]}_{(i)} \underbrace{\mathbb{E}_r \Big[\mathbb{P} \left[f_o \in \Delta_{b_o} \mid r \right] \Big]}_{(ii)}. \qquad \text{(A.1)}$$

Proceeding term by term, we first write (i) as:

$$\mathbb{E}_r \left[\mathbb{P} \left[\ln(1 + \text{SINR}) > T \mid r \right] \right]$$

$$= \int_{r>0} \mathbb{P} \left[\ln(1 + \text{SINR}) > T \mid r \right] f_r(r) \mathrm{d}r \qquad \text{(A.2)}$$

$$\stackrel{(a)}{=} \int_{r>0} \mathbb{P} \left[\ln(1 + \text{SINR}) > T \mid r \right] e^{-\pi \lambda r^2} 2\pi \lambda r \mathrm{d}r$$

$$\stackrel{(b)}{=} \int_{r>0} \mathbb{P} \left[\frac{h r^{-\alpha}}{\sigma^2 + I_r} > e^T - 1 \mid r \right] e^{-\pi \lambda r^2} 2\pi \lambda r \mathrm{d}r$$

$$\stackrel{(c)}{=} \int_{r>0} \mathbb{P} \left[h > r^\alpha (e^T - 1)(\sigma^2 + I_r) \mid r \right] e^{-\pi \lambda r^2} 2\pi \lambda r \mathrm{d}r, \qquad \text{(A.3)}$$

where $f_r(r) = e^{-\pi \lambda r^2} 2\pi \lambda r$ is the PDF of r for PPP [191], hence (a) follows from its substitution. The expression in (b) is obtained by plugging the SINR formula and letting it on the left hand side of the inequality, (c) is the result of some algebraic manipulations for keeping fading variable h alone.

149

Conditioning on I_r and using the fact that $h \sim \text{Exponential}(\mu)$, the probability of random variable h exceeding $r^\alpha(e^T - 1)(\sigma^2 + I_r)$ can be written as:

$$\mathbb{P}\left[h > r^\alpha(e^T - 1)(\sigma^2 + I_r) \mid r\right]$$
$$= \mathbb{E}_{I_r}\left[\mathbb{P}\left[h > r^\alpha(e^T - 1)(\sigma^2 + I_r) \mid r, I_r\right]\right]$$
$$= \mathbb{E}_{I_r}\left[\exp\left(-\mu r^\alpha(e^T - 1)(\sigma^2 + I_r)\right) \mid r\right]$$
$$= e^{-\mu r^\alpha(e^T-1)\sigma^2}\mathcal{L}_{I_r}\left(\mu r^\alpha(e^T - 1)\right), \tag{A.4}$$

where $\mathcal{L}(s)$ is the Laplace transform of random variable I_r evaluated at s conditioned on the distance of the nearest base station from the origin. Substituting (A.4) into (A.3) yields the following:

$$\mathbb{E}_r\left[\mathbb{P}\left[\ln(1 + \text{SINR}) > T \mid r\right]\right] = \int_{r>0} e^{-\mu r^\alpha(e^T-1)\sigma^2}\mathcal{L}_{I_r}\left(\mu r^\alpha(e^T - 1)\right) e^{-\pi\lambda r^2}2\pi\lambda r \mathrm{d}r. \tag{A.5}$$

Defining g_i as a random variable of arbitrary but identical distribution for all i, and R_i as the distance from the i-th base station to the tagged receiver, the Laplace transform is written as:

$$\mathcal{L}_{I_r}(s) = \mathbb{E}_{I_r}\left[e^{-sI_r}\right] = \mathbb{E}_{\Phi,\{g_i\}}\left[\exp\left(-s\sum_{i\in\Phi\setminus\{b_o\}} g_i R_i^{-\alpha}\right)\right]$$

$$= \mathbb{E}_{\Phi,\{g_i\}}\left[\prod_{i\in\Phi\setminus\{b_o\}} \exp\left(-sg_i R_i^{-\alpha}\right)\right]$$

$$\overset{(a)}{=} \mathbb{E}_\Phi\left[\prod_{i\in\Phi\setminus\{b_o\}} \mathbb{E}_{\{g_i\}}\left[\exp\left(-sg_i R_i^{-\alpha}\right)\right]\right]$$

$$\overset{(b)}{=} \mathbb{E}_\Phi\left[\prod_{i\in\Phi\setminus\{b_o\}} \mathbb{E}_g\left[\exp\left(-sg R_i^{-\alpha}\right)\right]\right]$$

$$= \exp\left(-2\pi\lambda\int_r^\infty \left(1 - \mathbb{E}_g\left[\exp\left(-sgv^{-\alpha}\right)\right]\right) v\mathrm{d}v\right),$$

where (a) comes from the independence of g_i from the point process Φ, and (b) follows from the i.i.d. assumption of g_i. The last step comes from the probability generating functional (PGFL) of the PPP, which basically says that for some function $f(x)$, $\mathbb{E}\left[\prod_{x\in\Phi} f(x)\right] = \exp\left(-\lambda\int_{\mathbb{R}^2}(1 - f(x))\mathrm{d}x\right)$. Since the nearest interfering base station is at least at a distance r, the integration limits are from r to infinity. Denoting $f(g)$ as the PDF of g, then plugging in $s = \mu r^\alpha(e^T - 1)$ and switching the integration order yields

$$\mathcal{L}_{I_r}\left(\mu r^\alpha(e^T - 1)\right) = \exp\left(-2\pi\lambda\int_0^\infty \left(\int_r^\infty \left(1 - e^{-\mu r^\alpha(e^T-1)v^{-\alpha}g}\right) v\mathrm{d}v\right) f(g)\mathrm{d}g\right).$$

By change of variables $v^{-\alpha} \to y$, the Laplace transform can be rewritten as:

$$\mathcal{L}_{I_r}\left(\mu r^\alpha (e^T - 1)\right) = \exp\Big(\lambda \pi r^2 - \frac{2\pi\lambda \left(\mu(e^T - 1)\right)^{\frac{2}{\alpha}} r^2}{\alpha} \times$$
$$\int_0^\infty g^{\frac{2}{\alpha}}\left[\Gamma\left(-\frac{2}{\alpha}, \mu\left(e^T - 1\right)g\right) - \Gamma\left(-\frac{2}{\alpha}\right)\right] f(g)\mathrm{d}g\Big). \quad \text{(A.6)}$$

Plugging (A.6) into (A.5), using the substitution $r^2 \to v$ and after some algebraic manipulations, the expression becomes

$$\mathbb{E}_r\left[\mathbb{P}\left[\ln(1 + \text{SINR}) > T \mid r\right]\right] = \pi\lambda \int_0^\infty e^{-\pi\lambda v\beta(T,\alpha) - \mu(e^T - 1)\sigma^2 v^{\alpha/2}}\mathrm{d}v, \quad \text{(A.7)}$$

where $\beta(T, \alpha)$ is given as

$$\beta(T, \alpha) = \frac{2\left(\mu(e^T - 1)\right)}{\alpha}\mathbb{E}_g\left[g^{\frac{2}{\alpha}}\left(\Gamma\left(-\frac{2}{\alpha}, \mu\left(e^T - 1\right)g\right) - \Gamma\left(-\frac{2}{\alpha}\right)\right)\right].$$

So far, we have obtained (i) of (A.1). The term (ii) is straightforward to derive. In the system model, as we assume that every small base station caches the same popular files and they have the same storage size, the cache hit probability becomes independent of the distance r. This yields:

$$\mathbb{E}_r\left[\mathbb{P}\left[f_o \in \Delta_{b_o} \mid r\right]\right] = \int_0^{S/L} f_{\text{pop}}(f, \gamma)\mathrm{d}f. \quad \text{(A.8)}$$

Plugging both (A.7) and (A.8) into (A.1) and rearranging the terms, we conclude the proof. ∎

A.2 Proof of Theorem 2

Average achievable delivery rate is $\bar{\tau} = \mathbb{E}\left[\tau\right]$, where the average is taken over the PPP and the fading distribution. It can be shown that

$$\bar{\tau} = \mathbb{E}\left[\tau\right]$$
$$\overset{(a)}{=} \mathbb{E}\left[\mathbb{P}\left[\ln(1 + \text{SINR}) > T\right]\left(T\mathbb{P}\left[f_o \in \Delta_{b_o}\right] + C\left(\lambda\right)\mathbb{P}\left[f_o \notin \Delta_{b_o}\right]\right)\right]$$

$$\overset{(b)}{=} \mathbb{E}\left[\underbrace{\mathbb{P}\left[\ln(1 + \text{SINR}) > T \mid r\right]}_{\tau_1}\right] \times$$
$$\left(\mathbb{E}\left[\underbrace{T\mathbb{P}\left[f_o \in \Delta_{b_o} \mid r\right]}_{\tau_2}\right] + \mathbb{E}\left[\underbrace{C\left(\lambda\right)\mathbb{P}\left[f_o \notin \Delta_{b_o} \mid r\right]}_{\tau_3}\right]\right)$$

$$= \mathbb{E}\left[\tau_1\right]\left(\mathbb{E}\left[\tau_2\right] + \mathbb{E}\left[\tau_3\right]\right), \tag{A.9}$$

where (a) is obtained by plugging the delivery rate as defined in (3.6), and (b) follows from independence of the events and linearity of the expectation operator.

Derivation of $\mathbb{E}[\tau_1]$ can be obtained from the proof of Theorem 1, by following the steps from (A.2) to (A.7). On the other hand, the fact that the cache hit probability is independent of r, $\mathbb{E}_r[\tau_2]$ can be expressed as

$$\mathbb{E}_r[\tau_2] = T \int_0^{S/L} f_{\mathrm{pop}}(f, \gamma) \mathrm{d}f.$$

Using similar arguments, $\mathbb{E}_r[\tau_3]$ is written as:

$$\mathbb{E}_r[\tau_3] = C(\lambda)\left(1 - \int_0^{S/L} f_{\mathrm{pop}}(f, \gamma)\mathrm{d}f\right).$$

Substituting these expressions into (A.9) concludes the proof. ∎

A.3 Proof of Proposition 1

Since Proposition 1 is a special case of Theorem 1, we follow the similar steps. We first rewrite (A.1) as:

$$p_{\mathrm{out}}(\lambda, T, \alpha, S, \gamma) = 1 - \underbrace{\mathbb{E}_r\left[\mathbb{P}\left[\ln(1 + \mathrm{SINR}) > T \mid r\right]\right]}_{(i)} \underbrace{\mathbb{E}_r\left[\mathbb{P}\left[f_o \in \Delta_{b_o} \mid r\right]\right]}_{(ii)}. \tag{A.10}$$

For the proceeding of (i), the proof of Theorem 1 can be followed starting from (A.2) to (A.5). Then, the Laplace transform is written as

$$\begin{aligned}
\mathcal{L}_{I_r}(s) &= \mathbb{E}_\Phi\left[\prod_{i \in \Phi \setminus \{b_o\}} \mathbb{E}_g\left[\exp\left(-sgR_i^{-\alpha}\right)\right]\right] \\
&\overset{(a)}{=} \mathbb{E}_\Phi\left[\prod_{i \in \Phi \setminus \{b_o\}} \frac{\mu}{\mu + sR_i^{-\alpha}}\right] \\
&= \exp\left(-2\pi\lambda \int_r^\infty \left(1 - \frac{\mu}{\mu + sv^{-\alpha}}\right)v\mathrm{d}v\right), \tag{A.11}
\end{aligned}$$

where (a) comes from the new assumption that $g \sim \mathrm{Exponential}(\mu)$. Then, plugging $s = \mu r^\alpha\left(e^T - 1\right)$ yields:

$$\mathcal{L}_{I_r}\left(\mu r^\alpha\left(e^T - 1\right)\right) = \exp\left(-2\pi\lambda \int_r^\infty \frac{e^T - 1}{e^T - 1 + (\frac{v}{r})^\alpha}v\mathrm{d}v\right).$$

Using a change of variables $u = \left(\frac{v}{r(e^T-1)^{\alpha/2}}\right)^2$ results in

$$\mathcal{L}_{I_r}\left(\mu r^\alpha \left(e^T - 1\right)\right) = \exp\left(-\pi r^2 \lambda \rho(T, \alpha)\right), \tag{A.12}$$

where

$$\rho(T, \alpha) = (e^T - 1)^{2/\alpha} \int_{(e^T-1)^{-2/\alpha}}^{\infty} \frac{1}{1 + u^{\alpha/2}} \mathrm{d}u.$$

Substituting (A.12) into (A.5) with $r^2 \rightarrow v$ gives

$$\pi\lambda \int_0^\infty e^{-\pi\lambda v(1+\rho(T,\alpha))-\mu(e^T-1)\sigma^2 v^{\alpha/2}} \mathrm{d}v. \tag{A.13}$$

Since $\alpha = 4$ in our special case, (A.13) simplifies to

$$\pi\lambda \int_0^\infty e^{-\pi\lambda v(1+\rho(T,4))-\mu(e^T-1)\sigma^2 v^2} \mathrm{d}v, \tag{A.14}$$

where

$$\begin{aligned}
\rho(T, 4) &= (e^T - 1)^{2/\alpha} \int_{(e^T-1)^{-2/\alpha}}^{\infty} \frac{1}{1 + u^2} \mathrm{d}u \\
&= (e^T - 1)^{2/\alpha} \left(\frac{\pi}{2} - \arctan\left((e^T - 1)^{-2/\alpha}\right)\right) \\
&= \sqrt{e^T - 1} \left(\frac{\pi}{2} - \arctan\left(\frac{1}{\sqrt{e^T - 1}}\right)\right).
\end{aligned}$$

From this point, (A.14) can be further simplified since it has a form similar to:

$$\int_0^\infty e^{-ax} e^{-bx^2} \mathrm{d}x = \sqrt{\frac{\pi}{b}} \exp\left(\frac{a^2}{4b}\right) Q\left(\frac{a}{\sqrt{2b}}\right),$$

where $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} \mathrm{d}y$ is the standard Gaussian tail probability. Setting $a = \pi\lambda(1 + \rho(T, 4))$ and $b = \mu(e^T - 1)\sigma^2 = (e^T - 1)/\text{SNR}$ gives

$$\frac{\pi^{\frac{3}{2}}\lambda}{\sqrt{\frac{e^T-1}{\text{SNR}}}} \exp\left(\frac{(\lambda\pi(1 + \rho(T, 4)))^2}{4(e^T - 1)/\text{SNR}}\right) Q\left(\frac{\lambda\pi(1 + \rho(T, 4))}{\sqrt{2(e^T - 1)/\text{SNR}}}\right). \tag{A.15}$$

This is the final expression for (i) of (A.10). The term (ii) of (A.10) can be obtained by using similar arguments given for (A.8) in the proof of Theorem 1, meaning that the cache hit probability is independent of distance r. Thus:

$$\begin{aligned}
\mathbb{E}_r\left[\mathbb{P}\left[f_o \in \Delta_{b_o} \mid r\right]\right] &= \int_0^{S/L} f_{\text{pop}}(f, \gamma) \mathrm{d}f \\
&\overset{(a)}{=} \int_1^{1+S/L} (\gamma - 1) f^{-\gamma} \mathrm{d}f \\
&= 1 - \left(\frac{L}{L + S}\right)^{\gamma-1}, \tag{A.16}
\end{aligned}$$

153

where (a) follows from plugging definition of $C(f, \lambda)$ given in Assumption 1 and changing the integration limits accordingly. The last term is the result of the integral. Therefore, we conclude the proof by plugging (A.15) and (A.16) into (A.10). ∎

A.4 Proof of Proposition 2

The proposition is a special case of Theorem 2, thus we have the similar steps. We start by rewriting (A.9) as:

$$\bar{\tau} = \mathbb{E}\left[\underbrace{\mathbb{P}\left[\ln(1 + \mathrm{SINR}) > T \mid r\right]}_{\tau_1}\right] \times$$

$$\left(\mathbb{E}\left[\underbrace{T\mathbb{P}\left[f_o \in \Delta_{b_o} \mid r\right]}_{\tau_2}\right] + \mathbb{E}\left[\underbrace{C\left(\lambda\right)\mathbb{P}\left[f_o \notin \Delta_{b_o} \mid r\right]}_{\tau_3}\right]\right)$$

$$= \mathbb{E}\left[\tau_1\right]\left(\mathbb{E}\left[\tau_2\right] + \mathbb{E}\left[\tau_3\right]\right). \tag{A.17}$$

In this expression, the term $\mathbb{E}\left[\tau_1\right]$ can be obtained from the proof of Proposition 1. More precisely, observe that $\mathbb{E}\left[\tau_1\right]$ is identical to (i) of (A.10). Thus, following the steps from (A.11) to (A.15), we obtain

$$\mathbb{E}\left[\tau_1\right] = \mathbb{E}\left[\mathbb{P}\left[\ln(1 + \mathrm{SINR}) > T \mid r\right]\right]$$

$$= \frac{\pi^{\frac{3}{2}}\lambda}{\sqrt{\frac{e^T - 1}{\mathrm{SNR}}}}\exp\left(\frac{(\lambda\pi(1 + \rho(T, 4)))^2}{4(e^T - 1)/\mathrm{SNR}}\right)Q\left(\frac{\lambda\pi(1 + \rho(T, 4))}{\sqrt{2(e^T - 1)/\mathrm{SNR}}}\right). \tag{A.18}$$

On the other hand, $\mathbb{E}\left[\tau_2\right]$ can be obtained by taking T out of the expectation and plugging (A.16) into the formula, i.e.

$$\mathbb{E}\left[\tau_2\right] = \mathbb{E}\left[T\mathbb{P}\left[f_o \in \Delta_{b_o} \mid r\right]\right]$$

$$= T\left(1 - \left(\frac{L}{L + S}\right)^{\gamma - 1}\right). \tag{A.19}$$

Finally, $\mathbb{E}\left[\tau_3\right]$ is easy to derive as

$$\mathbb{E}\left[\tau_3\right] = \mathbb{E}\left[C\left(\lambda\right)\mathbb{P}\left[f_o \notin \Delta_{b_o} \mid r\right]\right]$$

$$= C\left(\lambda\right)\left(\frac{L}{L + S}\right)^{\gamma - 1}$$

$$= \left(\frac{C_1}{\lambda} + C_2\right)\left(\frac{L}{L + S}\right)^{\gamma - 1}, \tag{A.20}$$

where definition of $C(\lambda)$ follows from Assumption 1. Substituting (A.18), (A.19) and (A.20) into (A.17) concludes the proof. ∎

Appendix B

Multi-Tier Cellular Networks

B.1   Proof of Lemma 5

The proof follows the similar lines as in [191,241] and derived here for sake of completeness. The fact that thermal noise is ignored (hence the transmission is interference-limited), we have

$$\mathbb{P}_{\text{m}} = \int_0^\infty \mathbb{P}\Big(\text{SIR}_{\text{mc}}(x) > \gamma \Big| \|x\|_2 \in \mathrm{d}r\Big) f_{\text{mc}}(r)\mathrm{d}r \tag{B.1}$$

$$\overset{(a)}{=} \int_0^\infty \mathbb{P}\Big(\frac{P_{\text{mc}}h_x\ell(x)}{I_{\text{mm}} + I_{\text{sm}}} > \gamma\Big) f_{\text{mc}}(r)\mathrm{d}r \tag{B.2}$$

$$\overset{(b)}{=} \int_0^\infty \mathbb{E}\Big\{\exp\Big(-\frac{\gamma r^\alpha}{P_{\text{mc}}}(I_{\text{mm}} + I_{\text{sm}})\Big)\Big\} f_{\text{mc}}(r)\mathrm{d}r \tag{B.3}$$

$$\overset{(c)}{=} \int_0^\infty \exp\Big(-\pi r^2 \lambda_{\text{mc}}\rho(\gamma,\alpha)\Big) \times$$

$$\exp\Big(-\pi r^2 (P_{\text{sc}}/P_{\text{mc}})^{2/\alpha}\lambda_{\text{sc}}\gamma^{2/\alpha}A(\alpha)\Big) f_{\text{mc}}(r)\mathrm{d}r \tag{B.4}$$

$$= \exp\Big(-\pi r^2 \big[\lambda_{\text{mc}}\rho(\gamma,\alpha) + (P_{\text{sc}}/P_{\text{mc}})^{2/\alpha}\lambda_{\text{sc}}\gamma^{2/\alpha}A(\alpha)\big]\Big) \tag{B.5}$$

where $\rho(\gamma,\alpha) = \gamma^{2/\alpha}\int_{\gamma^{-2/\alpha}}^\infty \frac{1}{1+u^{\alpha/2}}\mathrm{d}u$ and $A(\alpha) = \frac{2\pi/\alpha}{\sin(2\pi/\alpha)}$. The step (a) follows from Slivnyak Theorem [195]; the step (b) is due to the Laplace Transform of fading power coefficient h_x, which is an Exponential random variable (Rayleigh fading) with $\mathbb{E}[h_x] = 1$. The step (c) comes from the independence of I_{mm} and I_{sm} and the Laplace transform of I_{mm} and I_{sm} [191, 242]. The final expression is obtained by considering the spatial probability distribution function of macro cells, that is $f_{\text{mc}}(r) = 2\lambda_{\text{mc}}\pi r \exp(-\pi\lambda_{\text{mc}}r^2)$, and calculating the integral accordingly. The derived results so far hold for the typical user connecting to the nearest macro cell. In case of connecting to the nearest small cell, one can straightforwardly repeat the above steps to get

$$\mathbb{P}_{\text{s}} = \exp\Big(-\pi r^2 \big[\lambda_{\text{sc}}\rho(\gamma,\alpha) + (P_{\text{mc}}/P_{\text{sc}})^{2/\alpha}\lambda_{\text{mc}}\gamma^{2/\alpha}A(\alpha)\big]\Big). \tag{B.6}$$

Having $\mathbb{P}_{\mathrm{m}}$ from (B.5) and $\mathbb{P}_{\mathrm{s}}$ from (B.6) concludes the proof. ∎

B.2 Proof of Theorem 6

The average delay for a typical user connecting to a macro cell is made of two independent components as defined in (4.14), thus we have

$$\bar{D}_{\mathrm{m}} = \mathbb{E}[D_{\mathrm{m}}] \tag{B.7}$$

$$= \mathbb{E}[D_{\mathrm{dm}}] + \mathbb{E}[D_{\mathrm{bm}}]. \tag{B.8}$$

For derivation of average downlink delay $\mathbb{E}[D_{\mathrm{dm}}]$, we follow the same line of [243]. Observe that the expected delay is at least T_0 with probability 1. The first failure appears with probability $1 - p_{\mathrm{m}}(r)$ and takes T_0 additional time. Given the first failure, the second failure appears with probability $1 - p_{\mathrm{m}}(r)$ and takes T_0 additional time. By proceeding in this way, one can write

$$\mathbb{E}[D_{\mathrm{dm}} \mid r] = T_0 + T_0\big(1 - p_{\mathrm{m}}(r)\big) + T_0\big(1 - p_{\mathrm{m}}(r)\big)^2 +$$
$$\cdots + T_0(1 - p_{\mathrm{m}}(r))^{M-1} \tag{B.9}$$

$$= T_0 \frac{1 - \big(1 - p_{\mathrm{m}}(r)\big)^M}{p_{\mathrm{m}}(r)}. \tag{B.10}$$

Then, the average downlink delay is computed by using law of total expectation such as

$$\mathbb{E}[D_{\mathrm{dm}}] = \mathbb{E}\big[\mathbb{E}[D_{\mathrm{dm}} \mid r]\big] = T_0 \int_0^\infty 2\pi\lambda_{\mathrm{mc}}\exp(-\pi r^2\lambda_{\mathrm{mc}}) \frac{1 - \big(1 - p_{\mathrm{m}}(r)\big)^M}{p_{\mathrm{m}}(r)} \mathrm{d}r \tag{B.11}$$

$$= T_0 \sum_{i=0}^{M-1} (-1)^i \binom{M}{i+1} \times$$
$$\frac{1}{1 + i\big[\rho(\gamma, \alpha) + (P_{\mathrm{sc}}/P_{\mathrm{mc}})^{2/\alpha}(\lambda_{\mathrm{sc}}/\lambda_{\mathrm{mc}})\gamma^{2/\alpha}A(\alpha)\big]}. \tag{B.12}$$

Observe that the temporal correlation of interference is not considered here. Therefore, the derived expression is a lower bound for the downlink delay whereas one can prove this claim by observing that the success probability $p_{\mathrm{m}}(r)$ is a decreasing function and make use of Fortuin-Kasteleyn-Ginibre (FKG) inequality. The readers who are interested to complete characterization of the delay can check out the tools developed in [244]. We now focus on average backhaul delay, that is

$$\mathbb{E}[D_{\mathrm{bm}}] = \bar{\mu}_{\mathrm{bm}} \tag{B.13}$$

$$\overset{(a)}{=} \beta \frac{\lambda_{\mathrm{mc}}}{\lambda_{\mathrm{cs}}} \int_0^\infty 2\lambda_{\mathrm{cs}}\pi r^2 e^{-\pi\lambda_{\mathrm{cs}}r^2} \mathrm{d}r \tag{B.14}$$

$$= \frac{1}{2}\beta\lambda_{\mathrm{mc}}\lambda_{\mathrm{cs}}^{-3/2} \tag{B.15}$$

where step (a) follows from making use of Lemmas 3-4. The parameter β here is a scaling factor that relates the backhaul delay to the downlink delay, meaning that better backhaul infrastructure and methods corresponds to a lower value of β thus less delay.

Writing down the downlink delay (B.12) in the form of B_1 function such as $B_1(T_0, M, \gamma, \alpha, P_{\mathrm{sc}}, P_{\mathrm{mc}}, \lambda_{\mathrm{sc}}, \lambda_{\mathrm{mc}})$ and plugging into (B.8) together with the backhaul delay expression (B.15) give the desired result. ∎

B.3 Proof of Theorem 7

Recalling the definition of delay in (4.15), the average delay for typical small cell user is expressed as

$$\bar{D}_{\mathrm{s}} = \mathbb{E}[D_{\mathrm{s}}] \tag{B.16}$$

$$= \mathbb{E}\big[D_{\mathrm{ds}} + \mathbb{1}_{\{f_{\mathrm{s}} \in \Delta_0\}} D_{\mathrm{ca}} + \big(1 - \mathbb{1}_{\{f_{\mathrm{s}} \in \Delta_0\}}\big) D_{\mathrm{bs}}\big] \tag{B.17}$$

$$= \mathbb{E}\big[D_{\mathrm{ds}}\big] + \mathbb{E}\big[\mathbb{1}_{\{f_{\mathrm{s}} \in \Delta_0\}}\big]\mathbb{E}\big[D_{\mathrm{ca}}\big] + \mathbb{E}\big[D_{\mathrm{bs}}\big] -$$
$$\mathbb{E}\big[\mathbb{1}_{\{f_{\mathrm{s}} \in \Delta_0\}}\big]\mathbb{E}\big[D_{\mathrm{bs}}\big] \tag{B.18}$$

where the final express comes from the independence of the processes and linearity of expectation operator. From now on, we proceed term by term.

Note that the expression for average downlink delay $\mathbb{E}[D_{\mathrm{s}}]$ can be obtained by using the definition of SIR for typical small cell user (see (4.2)) and invoking the steps in proof of Theorem 6 up to (B.12). Thus we have

$$\mathbb{E}[D_{\mathrm{s}}] = T_0 \sum_{i=0}^{M-1} (-1)^i \binom{M}{i+1} \frac{1}{1 + i\big[\rho(\gamma, \alpha) + (P_{\mathrm{mc}}/P_{\mathrm{sc}})^{2/\alpha} \lambda_{\mathrm{mc}} \gamma^{2/\alpha} A(\alpha)\big]} \tag{B.19}$$

and can be written in the form of function B_1 such as $B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}})$. On the other hand, the average caching delay $\mathbb{E}\big[D_{\mathrm{ca}}\big]$ simply follows from its definition, that is

$$\mathbb{E}\big[D_{\mathrm{ca}}\big] = \bar{\mu}_{\mathrm{ca}}. \tag{B.20}$$

Additionally, the backhaul delay for typical macro cell user is given by

$$\mathbb{E}\big[D_{\mathrm{bs}}\big] = \bar{\mu}_{\mathrm{bs}} \tag{B.21}$$

$$\stackrel{(a)}{=} \beta \frac{\lambda_{\mathrm{sc}}}{\lambda_{\mathrm{cr}}} \int_0^\infty 2\lambda_{\mathrm{cs}} \pi r^2 e^{-\pi \lambda_{\mathrm{cs}} r^2} \mathrm{d}r \tag{B.22}$$

$$= \frac{1}{2} \beta \lambda_{\mathrm{sc}} \lambda_{\mathrm{cs}}^{-3/2} \tag{B.23}$$

where the step (a) is due to the Lemmas 3-4. We now focus on the term $\mathbb{E}\big[\mathbb{1}_{\{f_{\mathrm{s}} \in \Delta_0\}}\big]$ which can to be derived for *fixed, distance-dependent* and *load-dependent* content popularities separately.

Fixed content popularity

The fact that small cells have knowledge of the fixed content popularity distribution $f_{\text{pop}}(f, \eta_0)$ and can store the most popular content up to S_{p}, the integration limit for cache hit probability is from 1 to $1 + S_{\text{p}}$. Additionally, the remaining catalogue is stored uniformly at random with storage capacity of S_{u}. Therefore, we have that

$$
\mathbb{E}\left[\mathbb{1}_{\{f_{\text{s}} \in \Delta_0\}}\right]
$$

$$
= \int_1^{1+S_{\text{p}}} f_{\text{pop}}(f, \eta_0)\, \mathrm{d}f + \frac{S_{\text{u}}}{f_0 - S_{\text{p}}} \int_{1+S_{\text{p}}}^{f_0} f_{\text{pop}}(f, \eta_0)\, \mathrm{d}f \tag{B.24}
$$

$$
= \int_1^{1+S_{\text{p}}} (\eta_0 - 1)\, f^{-\eta_0}\, \mathrm{d}f + \frac{S_{\text{u}}}{f_0 - S_{\text{p}}} \int_{1+S_{\text{p}}}^{f_0} (\eta_0 - 1)\, f^{1-\eta_0}\, \mathrm{d}f \tag{B.25}
$$

$$
= \underbrace{1 - \left(1 + S_{\text{p}}\right)^{1-\eta}}_{(i)} + \underbrace{\frac{S_{\text{u}}}{f_0 - S_{\text{p}}} \left(1 - \left(1 + f_0\right)^{1-\eta} + \left(1 + S_{\text{p}}\right)^{1-\eta}\right)}_{(ii)}. \tag{B.26}
$$

The term (i) in the final expression can be written in the form of function B_2 with $B_2(S_{\text{p}}, \eta_0)$ and the term (ii) has form of B_3 with $B_3(S_{\text{u}}, S_{\text{p}}, f_0, \eta_0)$. Plugging (B.19), (B.20), (B.23) and (B.26) into (B.18) gives the desired result in (4.23).

Distance-dependent content popularity

The average distance from a typical user to its connecting base station is given by

$$
\bar{r} = \int_0^\infty r \times 2\lambda_k r \pi \exp(-\pi \lambda_k r^2)\, \mathrm{d}r \tag{B.27}
$$

$$
= \frac{1}{2\sqrt{\lambda}}. \tag{B.28}
$$

Then, using similar arguments as in fixed-content popularity case, we have the following cache hit probability, that is

$$
\mathbb{E}\left[\mathbb{1}_{\{f_{\text{s}} \in \Delta_0\}}\right]
$$

$$
= \int_1^{1+S_{\text{p}}} f_{\text{pop}}(f, \bar{r})\, \mathrm{d}f + \frac{S_{\text{u}}}{f_0 - S_{\text{p}}} \int_{1+S_{\text{p}}}^{f_0} f_{\text{pop}}(f, \bar{r})\, \mathrm{d}f \tag{B.29}
$$

$$
= \int_1^{1+S_{\text{p}}} (\bar{r} - 1)\, f^{-\bar{r}}\, \mathrm{d}f + \frac{S_{\text{u}}}{f_0 - S_{\text{p}}} \int_{1+S_{\text{p}}}^{f_0} (\bar{r} - 1)\, f^{1-\bar{r}}\, \mathrm{d}f \tag{B.30}
$$

$$
= B_2(S_{\text{p}}, \bar{r}) + B_3(S_{\text{u}}, S_{\text{p}}, f_0, \bar{r}). \tag{B.31}
$$

Plugging (B.19), (B.20), (B.23) and (B.31) into (B.18) gives the desired result in (4.22).

Load-dependent content popularity

Cache hit probability under load-dependent content popularity is straightforwardly derived as

$$\mathbb{E}\big[\mathbb{1}_{\{f_s \in \Delta_0\}}\big] = \int_1^{1+S_p} f_{\text{pop}}\left(f, \lambda_{\text{ut}}/\lambda_{\text{sc}}\right) \mathrm{d}f + \frac{S_u}{f_0 - S_p} \int_{1+S_p}^{f_0} f_{\text{pop}}\left(f, \lambda_{\text{ut}}/\lambda_{\text{sc}}\right) \mathrm{d}f \tag{B.32}$$

$$= \int_1^{1+S_p} \left(\frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}} - 1\right) f^{-\frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}}} \mathrm{d}f + \frac{S_u}{f_0 - S_p} \int_{1+S_p}^{f_0} \left(\frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}} - 1\right) f^{-\frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}}} \mathrm{d}f \tag{B.33}$$

$$= B_2(S_p, \frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}}) + B_3(S_u, S_p, f_0, \frac{\lambda_{\text{ut}}}{\lambda_{\text{sc}}}). \tag{B.34}$$

Plugging (B.19), (B.20), (B.23) and (B.34) into (B.18) gives the desired result in (4.23). We therefore conclude the proof. ∎

B.4 Proof of Proposition 3

Recalling Definition 3, suppose that the typical user is connected to the nearest small cell thus $r_{\text{sc}} < \kappa r_{\text{mc}}$ holds. Therefore, conditioning on r_{mc} and averaging over spatial distribution of small cells, then averaging over macro cells yield the desired result, such as

$$p_a = \mathbb{E}\Big[\int_0^{\kappa r_{\text{mc}}} 2\pi\lambda_{\text{sc}} t e^{-\pi\lambda_{\text{sc}} t^2} \mathrm{d}t | r_{\text{mc}}\Big] \tag{B.35}$$

$$= \int_0^\infty (1 - e^{-\pi\lambda_{\text{sc}}\kappa^2 r_{\text{mc}}^2}) 2\pi\lambda_{\text{mc}} r_{\text{mc}} e^{-\pi\lambda_{\text{mc}} r_{\text{mc}}^2} \mathrm{d}r_{\text{mc}} \tag{B.36}$$

$$= \frac{\kappa^2 \lambda_{\text{sc}}}{\lambda_{\text{mc}} + \kappa^2 \lambda_{\text{sc}}}. \tag{B.37}$$

The association probability to a macro cell is then $1 - p_a$. ∎

B.5 Proof of Proposition 4

From Theorem 6, we know that the average delay of typical user connecting to the nearest macro cell is given by

$$B_1(T_0, M, \gamma, \alpha, P_{\text{sc}}, P_{\text{mc}}, \lambda_{\text{sc}}, \lambda_{\text{mc}}) + \frac{1}{2}\lambda_{\text{ut}}\lambda_{\text{mc}}\lambda_{\text{cs}}^{-3/2}. \tag{B.38}$$

When the association policy given in Definition 3 is employed at the network, then, the process of small cell users is approximated with a thinned PPP of density $(1 - p_a)\lambda_{\text{ut}}$ where $(1 - p_a)$ is the association probability given in Proposition 3. Adding TDMA in

their multiple access and using Lemma 4 for average number of nodes in hierarchical model, the total average delay for macro cell users is given by

$$\bar{D}_{\mathrm{m}}^{(\mathrm{tot})} = \frac{(1-p_{\mathrm{a}})^2\lambda_{\mathrm{ut}}^2}{\lambda_{\mathrm{mc}}} B_1(T_0, M, \gamma, \alpha, P_{\mathrm{sc}}, P_{\mathrm{mc}}, \lambda_{\mathrm{sc}}, \lambda_{\mathrm{mc}}) + \frac{1}{2}\beta(1-p_{\mathrm{a}})\lambda_{\mathrm{ut}}\lambda_{\mathrm{mc}}\lambda_{\mathrm{cs}}^{-3/2}. \quad (\mathrm{B.39})$$

On the other hand, from Theorem 7, we have the average delay of typical small cell user. Knowing that small cell users follow a thinned PPP with density $p_{\mathrm{a}}\lambda_{\mathrm{ut}}$, the total average delay for small cell users is obtained by repeating the same steps above, that is

$$\bar{D}_{\mathrm{s}}^{(\mathrm{tot})} = \frac{p_{\mathrm{a}}^2\lambda_{\mathrm{ut}}^2}{\lambda_{\mathrm{sc}}} B_1(T_0, M, \gamma, \alpha, P_{\mathrm{mc}}, P_{\mathrm{sc}}, \lambda_{\mathrm{mc}}, \lambda_{\mathrm{sc}}) +$$
$$p_{\mathrm{a}}\lambda_{\mathrm{ut}}\Big(\bar{\mu}_{\mathrm{ca}} - \frac{1}{2}\beta\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2}\Big)\Big(B_2(S_{\mathrm{p}}, \eta_0) + B_3(S_{\mathrm{u}}, S_{\mathrm{p}}, f_0, \eta_0)\Big) +$$
$$\frac{1}{2}\beta p_{\mathrm{a}}\lambda_{\mathrm{ut}}\lambda_{\mathrm{sc}}\lambda_{\mathrm{cs}}^{-3/2}. \quad (\mathrm{B.40})$$

Combining (B.39) and (B.40) gives the desired result. ∎

B.6 Proof of Proposition 5

Recall that the total network cost is defined in (4.32). Therein, by Proposition (4), we have total average delay. Therefore, it remains to compute the average cost of constructing the backhaul links for macro and small cells. In case of macro cells, this is given by

$$\bar{L}_{\mathrm{mc}} = \lambda_{\mathrm{mc}}\mathbb{E}\big[a_{\mathrm{mc}}r^{b_{\mathrm{mc}}}\big] \quad (\mathrm{B.41})$$

$$= \lambda_{\mathrm{mc}}\int_0^\infty a_{\mathrm{mc}}r^{b_{\mathrm{mc}}}2\pi\lambda_{\mathrm{cs}}re^{-\pi\lambda_{\mathrm{cs}}r^2}\mathrm{d}r \quad (\mathrm{B.42})$$

$$= \lambda_{\mathrm{mc}}a_{\mathrm{mc}}\frac{\Gamma\big(\frac{b_{\mathrm{mc}}}{2}+1\big)}{\big(\pi\lambda_{\mathrm{cs}}\big)^{b_{\mathrm{mc}}/2}} \quad (\mathrm{B.43})$$

where the final expression is obtained by invoking by Lemma 3. In case of small cell users, we similarly have that

$$\bar{L}_{\mathrm{sc}} = \lambda_{\mathrm{sc}}\mathbb{E}\big[a_{\mathrm{sc}}r^{b_{\mathrm{sc}}}\big] \quad (\mathrm{B.44})$$

$$= \lambda_{\mathrm{sc}}\int_0^\infty a_{\mathrm{sc}}r^{b_{\mathrm{sc}}}2\pi\lambda_{\mathrm{cs}}re^{-\pi\lambda_{\mathrm{cs}}r^2}\mathrm{d}r \quad (\mathrm{B.45})$$

$$= \lambda_{\mathrm{sc}}a_{\mathrm{sc}}\frac{\Gamma\big(\frac{b_{\mathrm{sc}}}{2}+1\big)}{\big(\pi\lambda_{\mathrm{cs}}\big)^{b_{\mathrm{sc}}/2}}. \quad (\mathrm{B.46})$$

Substituting (B.41) and (B.44) into (4.32) concludes the proof. ∎

Appendix C

Clustered Cellular Networks

C.1 Proof of Theorem 10

The rate of typical macro cell user is subject to random fluctuations of spatial distribution of cells, fading, interference and capacity of central routers. We start by writing compact form of rate given in (5.10) in Definition (8). Denoting event $\mathcal{A}_1$ as $\log(1 + \mathrm{SIR}_{\mathrm{mu}}) > \tau_{\mathrm{mc}}$, event $\mathcal{A}_2$ as $R'_{\mathrm{mu}} > \tau_{\mathrm{mc}}$, we have

$$\bar{R}_{\mathrm{mu}} = \mathbb{E}\Big[R_{\mathrm{mu}}\Big] \tag{C.1}$$

$$\overset{(a)}{=} \mathbb{E}\Big[\tau_{\mathrm{mc}}\mathbb{P}\big[\mathcal{A}_1\big]\mathbb{P}\big[\mathcal{A}_2\big]\Big] \tag{C.2}$$

$$= \tau_{\mathrm{mc}}\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big] \tag{C.3}$$

where (a) is obtained by plugging compact form of (5.10) into the expression and considering the independence of event $\mathcal{A}_1$ with respect to $\mathcal{A}_2$, and the last step comes from the linear properties of expectation operator. In the above expressions, note that we avoid to specify conditioning for notational clarity. We shall later consider this depending on the context.

In what follows, we first re-express the terms $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$ and $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big]$ in a general manner, then focus on expressions inside these terms and finalize our calculations. Some of results below are obtained by using the proof techniques developed in [205].

We start by re-expressing the term $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$, that is

$$\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big] = \mathbb{E}\Big[\mathbb{P}\big[\log(1 + \mathrm{SIR}_{\mathrm{mu}}) > \tau_{\mathrm{mc}}\big]\Big] \tag{C.4}$$

$$\overset{(a)}{=} \int_0^\infty \mathbb{P}\big[\log(1 + \mathrm{SIR}_{\mathrm{mu}}) > \tau_{\mathrm{mc}}\big|r\big] f_{\mathrm{mc}}(r_{\mathrm{mc}})\mathrm{d}r_{\mathrm{mc}} \tag{C.5}$$

$$\overset{(b)}{=} \int_0^\infty \mathbb{P}\big[\log(1 + \frac{P_{\mathrm{mc}} h_{r_0} r^{-\alpha}}{I_{\mathrm{mm}} + I_{\mathrm{sm}}}) > \tau_{\mathrm{mc}}\big|r_{\mathrm{mc}}\big] f_{\mathrm{mc}}(r_{\mathrm{mc}})\mathrm{d}r_{\mathrm{mc}} \tag{C.6}$$

$$\overset{(c)}{=} \int_0^\infty \mathbb{P}\big[h > \frac{(e^{\tau_{\mathrm{mc}}} - 1)(I_{\mathrm{mm}} + I_{\mathrm{sm}})}{P_{\mathrm{mc}} r^{-\alpha}}\big|r_{\mathrm{mc}}\big] f_{\mathrm{mc}}(r_{\mathrm{mc}})\mathrm{d}r_{\mathrm{mc}} \tag{C.7}$$

$$\overset{(d)}{=} \int_0^\infty \mathbb{E}_{I_{\mathrm{mm}}, I_{\mathrm{sm}}}\Big[\mathbb{P}\big[h > \frac{(e^{\tau_{\mathrm{mc}}} - 1)(I_{\mathrm{mm}} + I_{\mathrm{sm}})}{P_{\mathrm{mc}} r_{\mathrm{mc}}^{-\alpha}}\big|r_{\mathrm{mc}}, I_{\mathrm{mm}}, I_{\mathrm{sm}}\big]\Big] f_{\mathrm{mc}}(r_{\mathrm{mc}})\mathrm{d}r_{\mathrm{mc}}$$

$$\overset{(e)}{=} \int_0^\infty \exp\big(-\frac{(e^{\tau_{\mathrm{mc}}} - 1)(I_{\mathrm{mm}} + I_{\mathrm{sm}})}{P_{\mathrm{mc}} r_{\mathrm{mc}}^{-\alpha}}\big) f_{\mathrm{mc}}(r_{\mathrm{mc}})\mathrm{d}r_{\mathrm{mc}} \tag{C.8}$$

$$= \int_0^\infty e^{-\frac{(e^{\tau_{\mathrm{mc}}} - 1)}{P_{\mathrm{mc}} r_{\mathrm{mc}}^{-\alpha}}} \mathcal{L}_{I_{\mathrm{mm}}}\Big(\frac{e^{\tau_{\mathrm{mc}}} - 1}{P_{\mathrm{mc}} r_{\mathrm{mc}}^{-\alpha}}\Big) \mathcal{L}_{I_{\mathrm{sm}}}\Big(\frac{e^{\tau_{\mathrm{mc}}} - 1}{P_{\mathrm{mc}} r_{\mathrm{mc}}^{-\alpha}}\Big) f_{\mathrm{mc}}(r_{\mathrm{mc}})\mathrm{d}r_{\mathrm{mc}} \tag{C.9}$$

where steps (a) and (b) are obtained by plugging the definition of SIR in (5.6) and considering averaging over the spatial distribution of nodes whose PDF is given by $f_{\mathrm{mc}}(r_{\mathrm{mc}})$. The step (c) is obtained by rearranging the terms which yields the fading coefficient h alone, (d) is due to the averaging over the interference terms I_{mm} and I_{sm}, and the step (e) is due to the fact that h is an Exponentially distributed random variable with mean 1. The final step is obtained by using the definition of Laplace transform with $\mathcal{L}_{I_{\mathrm{mm}}} = \mathbb{E}_{I_{\mathrm{mm}}}[e^{-sI_{\mathrm{mm}}}]$ and $\mathcal{L}_{I_{\mathrm{sm}}} = \mathbb{E}_{I_{\mathrm{sm}}}[e^{-sI_{\mathrm{sm}}}]$, where $s = \frac{e^{\tau_{\mathrm{mc}}} - 1}{P_{\mathrm{mc}} r_{\mathrm{mc}}^{-\alpha}}$. We shall calculate the Laplace transforms for $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$ in the upcoming subsections both for coverage and capacity-aided deployments. Now, let us consider the term $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big]$, that is

$$\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big] = \mathbb{E}\Big[\mathbb{P}\big[R'_{\mathrm{mu}} > \tau_{\mathrm{mc}}\big]\Big] \tag{C.10}$$

$$\overset{(a)}{=} \mathbb{P}\Big[\frac{\gamma C_{\mathrm{cr}}}{\mathbb{E}\big[N_{\mathrm{mc}} N_{\mathrm{mu}}\big]} > \tau_{\mathrm{mc}}\Big] \tag{C.11}$$

$$= \mathbb{P}\Big[C_{\mathrm{cr}} > \frac{\mathbb{E}\big[N_{\mathrm{mc}}\big]\mathbb{E}\big[N_{\mathrm{mu}}\big]\tau_{\mathrm{mc}}}{\gamma}\Big] \tag{C.12}$$

where (a) follows from the definition of rate splitting policy in (5.8) and using the fact that the term inside the outer expectation is constant with respect to the spatial random fluctuations. However, the inner expectations are with respect to the spatial random variables. In the above expression, the last step is obtained by rearranging the terms. For the final expressions, we now focus on coverage and capacity-aided deployments separately.

C.1.1 Coverage-aided deployment

Remember that the typical user located at origin o is connected to the macro cell at y_0, assumed at distance $r_{\mathrm{mc}} = \|y_0 - o\|$. We now proceed by deriving Laplace transforms of I_{mm} and I_{sm}, then focus on the backhaul part and final expressions.

1) Laplace transform of I_{mm}: It can be written as follows

$$\mathcal{L}_{I_{\mathrm{mm}}}(s) \tag{C.13}$$

$$= \mathbb{E}_{\Phi_{\mathrm{mc}}, h}^{!y_0}\Big[\exp\Big(-s\sum_{y\in\Phi_{\mathrm{mc}}} P_{\mathrm{mc}}h_y\ell(y)\Big)\Big] \tag{C.14}$$

$$= \mathbb{E}_{\Phi_{\mathrm{mc}}}^{!y_0}\Big[\prod_{y\in\Phi_{\mathrm{mc}}} \frac{1}{1 + sP_{\mathrm{mc}}\ell(y)}\Big] \tag{C.15}$$

$$\overset{(a)}{=} \exp\Big(-\lambda_{\mathrm{mc}}\int_{\mathbb{R}^2\setminus\mathcal{B}(o,r_{\mathrm{mc}})}\Big(1 - \frac{1}{1 + sP_{\mathrm{mc}}h\ell(y)}\Big)\mathrm{d}y\Big) \tag{C.16}$$

$$\overset{(b)}{=} \exp\Big(\frac{-s\pi\lambda_{\mathrm{mc}}P_{\mathrm{mc}}(2/\alpha)}{1 - 2/\alpha}r_{\mathrm{mc}}^{2-\alpha}F\big(1, 1 - 2/\alpha; 2 - 2/\alpha; -sP_{\mathrm{mc}}r_{\mathrm{mc}}^{-\alpha}\big)\Big) \tag{C.17}$$

where (a) comes from the PGFL of the PPP which states that $\mathbb{E}\big[\prod_{x\in\Phi} f(x)\big] = \exp\big(\int_{\mathbb{R}^2}(1 - f(x))\Lambda(\mathrm{d}x)\big)$ for $f : \mathbb{R}^2 \to [0, 1]$ and the integration region is $\mathbb{R}^2 \setminus \mathcal{B}(o, r_{\mathrm{mc}})$ as the closest interferer as at least at a distance r_{mc}, $F(x, y; z; w)$ is the hypergeometric function [207]. The step (b) is obtained by the help of equation (3.194.5) in [207] and a change to polar coordinates.

2) Laplace transform of I_{sm}: Assume a PPP with λ_{sc} instead of Poisson hole process for clustering process of small cells, and denote the interference of these small cells to the typical user by $\tilde{I}_{\mathrm{sm}}$. Since the small cells with Poisson hole process are at least at R_c distance away from the macro cells, their interference I_{sm} would be stochastically dominated by the interference of this PPP denoted by $\tilde{I}_{\mathrm{sm}}$[1], expect those within distance R_{mc} from the signalling macro cell. Now, let us denote $\mathcal{H}_{\mathrm{mc}}$ as the ball centered at the position of the signalling macro cell with radius R_{mc} and $\mathcal{H}_{\mathrm{mc}}^{\mathcal{C}} = \mathbb{R}^2 \setminus \mathcal{H}_{\mathrm{mc}}$. By using a

---

[1] Assuming two random variables A and B, we suppose that A stochastically dominates B if $\mathbb{P}(A > x) \geq \mathbb{P}(B > x)$ for all a, or in other words, $F_A(x) \leq F_B(x)$ for cumulative distribution functions $F_A(x)$ and $F_B(x)$.

modified path loss function $\tilde{\ell}(x) = \ell(x)\mathbb{1}_{\{x \in \mathcal{H}_{\mathrm{mc}}^{\mathcal{C}}\}}$

$$\mathcal{L}_{I_{\mathrm{sm}}}(s) \approx \mathcal{L}_{\tilde{I}_{\mathrm{sm}}}(s) \tag{C.18}$$

$$= \mathbb{E}_{\Phi_{\mathrm{sc}},g}^{!y_0}\left[\exp\left(-s\sum_{x \in \Phi_{\mathrm{sc}}} P_{\mathrm{sc}}g_x\tilde{\ell}(x)\right)\right] \tag{C.19}$$

$$= \exp\left(-\lambda_{\mathrm{sc}'}\int_{\mathbb{R}^2} 1 - \frac{1}{1 + sP_{\mathrm{sc}}\tilde{\ell}(x)}\mathrm{d}x\right) \tag{C.20}$$

$$= \exp\left(-\lambda_{\mathrm{sc}'}\int_{\mathcal{H}_{\mathrm{mc}}^{\mathcal{C}}} \frac{sP_{\mathrm{sc}}\ell(x)}{1 + sP_{\mathrm{sc}}\ell(x)}\mathrm{d}x\right) \tag{C.21}$$

$$= \exp\left\{-\lambda_{\mathrm{sc}'}\left(\frac{(sP_{\mathrm{sc}})^{2/\alpha}\pi^2(2/\alpha)}{\sin(\pi\frac{2}{\alpha})} - \pi R_{\mathrm{c}}^2 A_{\mathrm{mc}}(s, R_{\mathrm{c}})\right)\right\} \tag{C.22}$$

where

$$A_{\mathrm{mc}}(s, R_{\mathrm{c}}) = \frac{1}{\pi R_{\mathrm{c}}^2}\int_{\mathcal{H}_{\mathrm{mc}}}\left(1 - \frac{1}{1 + sP_{\mathrm{sc}}\ell(x)}\right)\mathrm{d}x \tag{C.23}$$

$$= \frac{1}{\pi R_{\mathrm{c}}^2}\int_0^{2\pi}\int_0^{r_{\mathrm{mc}}\cos\varphi + \sqrt{R_{\mathrm{c}}^2 - r_{\mathrm{mc}}^2\sin^2\varphi}} \frac{r\mathrm{d}r\mathrm{d}\varphi}{1 + s^{-1}P_{\mathrm{sc}}^{-1}r^\alpha} \tag{C.24}$$

We have thus approximated the Laplace transform of I_{sm} by taking the Laplace transform of $\tilde{I}_{\mathrm{sm}}$.

3) Backhaul part: Since macro cells and central rooters are distributed according to two independent PPPs with densities λ_{mc} and λ_{cr} respectively, the average number of macro cells connected to a nearest central router is evaluated by Lemma 9, that is

$$\mathbb{E}[N_{\mathrm{mc}}] = \frac{\lambda_{\mathrm{mc}}}{\lambda_{\mathrm{cr}}}. \tag{C.25}$$

Now, observe that mobile users are distributed according to a PPP with density λ_{cr}. Since these users are connected to macro or clustered small cells based on their nearest distance, we have from Lemma 9 that

$$\mathbb{E}[N_{\mathrm{mu}}] = \lambda_{\mathrm{mc}}\frac{\lambda_{\mathrm{ut}}}{\lambda_{\mathrm{mc}} + \lambda_{\mathrm{sc}}} \tag{C.26}$$

$$= \frac{\lambda_{\mathrm{mc}}\lambda_{\mathrm{ut}}}{\lambda_{\mathrm{mc}} + \lambda_{\mathrm{sc}'}\exp(-\lambda_{\mathrm{mc}}\pi R_{\mathrm{c}}^2)}. \tag{C.27}$$

Observe also that the summation in the denominator of the final expression is due to superposition of two independent PPPs. Even though the spatial distribution of clustered small cells is not a PPP but a Poisson hole process (see Definition 4), the results still hold as this clustered process is based on PPP and is motion-invariant. Also, observe that as $R_{\mathrm{c}} \to 0$, the spatial distribution of small cells becomes a PPP with density $\lambda_{\mathrm{sc}'}$. On the other hand, $R_{\mathrm{c}} \to \infty$ yields $\lambda_{\mathrm{sc}} \to 0$, thus no deployment of small cells. Based on

these observations for $\mathbb{E}[N_{\mathrm{mc}}]$ and $\mathbb{E}[N_{\mathrm{mu}}]$, and using the fact that C_{cr} is Exponentially distributed random variable with mean μ, we have

$$\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big] = \mathbb{P}\Big[C_{\mathrm{cr}} > \frac{\mathbb{E}\big[N_{\mathrm{mc}}\big]\mathbb{E}\big[N_{\mathrm{mu}}\big]\tau_{\mathrm{mc}}}{\gamma}\Big] \tag{C.28}$$

$$= 1 - \mathbb{P}\Big[C_{\mathrm{cr}} \le \frac{\mathbb{E}\big[N_{\mathrm{mc}}\big]\mathbb{E}\big[N_{\mathrm{mu}}\big]\tau_{\mathrm{mc}}}{\gamma}\Big] \tag{C.29}$$

$$= 1 - \exp\Big(-\frac{\tau_{\mathrm{mc}}\lambda_{\mathrm{cr}}\big(\lambda_{\mathrm{mc}} + \lambda_{\mathrm{sc}'}\exp(-\lambda_{\mathrm{mc}}\pi R_{\mathrm{c}}^2)\big)}{\mu\gamma\lambda_{\mathrm{mc}}^2\lambda_{\mathrm{ut}}}\Big). \tag{C.30}$$

4) Final expression: For the evaluation of $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$ in (C.9), we consider that $f_{\mathrm{mc}}(r_{\mathrm{mc}})$ follows a Weibull distribution such that $f_{\mathrm{mc}}(r_{\mathrm{mc}}) = \frac{k}{\nu}\big(\frac{r_{\mathrm{mc}}}{\nu}\big)^{k-1}e^{-(r_{\mathrm{mc}}/\nu)^k}$ where k and ν are shift and shape parameters respectively. Also the integration limit is from 0 to R_{c}. This approximation stems from the fact that Rayleigh distribution (and node distributions) are special cases of Weibull distribution (see [205] for a similar motivation). Plugging expressions of Laplace transforms (namely (C.17) and (C.22)) into (C.9), and considering (C.30) yield the final expression of (C.3). For the notational convenience, we denote $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$ and $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big]$ as $B_1^{(\mathrm{cov})}$ and $B_2^{(\mathrm{cov})}$ respectively. This concludes our approximation for coverage-aided deployment.

C.1.2 Capacity-aided deployment

Following the similar structure as in the previous section, we start by Laplace transforms of I_{mm} and I_{sm}, then focus on the backhaul part and final expressions.

1) Laplace transform of I_{mm}: Since macro cells follow a PPP, we have the similar steps as in previous section. That is to say

$$\mathcal{L}_{I_{\mathrm{mm}}}(s) = \exp\Big(\frac{-s\pi\lambda_{\mathrm{mc}}P_{\mathrm{mc}}(2/\alpha)}{1-2/\alpha}r_{\mathrm{mc}}^{2-\alpha}F\big(1, 1-2/\alpha; 2-2/\alpha; -sP_{\mathrm{mc}}r_{\mathrm{mc}}^{-\alpha}\big)\Big). \tag{C.31}$$

2) Laplace transform of I_{sm}: For a Matérn cluster process, the Laplace transform of the interference is expressed as (see [206], Corollary 4.13)

$$\mathcal{L}_{I_{\mathrm{sm}}}(s) = \exp\Big(-\lambda_{\mathrm{sc}'}\int_{\mathbb{R}^2}\big(1-\exp(-\bar{c}\nu(s,y))\big)\mathrm{d}y\Big), \tag{C.32}$$

where $\nu(s,y) = \int_{\mathbb{R}^2}\frac{f(x)}{1+(sP_{\mathrm{sc}}\ell(x-y))^{-1}}\mathrm{d}x$. Therein, as we have Matérn cluster process, $f(x)$ is the node distribution around the parent point given as

$$f(x) = \begin{cases} \frac{1}{\pi R_{\mathrm{c}}^2}, & \text{if } \|x\| < R_{\mathrm{c}}, \\ 0, & \text{otherwise.} \end{cases} \tag{C.33}$$

3) Backhaul part: Since Φ_{mc} and Φ_{cr} are two independent PPPs, the average number of macro cells connected to the nearest central router can be induced by Lemma 9, that is $\mathbb{E}[N_{\mathrm{mc}}] = \frac{\lambda_{\mathrm{mc}}}{\lambda_{\mathrm{cr}}}$. On the other hand, note that all mobile users modelled by a Cox process with density $\lambda_{\mathrm{ut}} = \lambda_{\mathrm{ut-m}} + \lambda_{\mathrm{ut-s}}$. However, as macro cell users specifically follow a PPP with density $\lambda_{\mathrm{ut-m}}$, the average number of users connected to the nearest macro cell is simply $\mathbb{E}[N_{\mathrm{mu}}] = \frac{\lambda_{\mathrm{ut-m}}}{\lambda_{\mathrm{mc}}}$. Therefore, using similar arguments as in the backhaul part of coverage-aided deployment (see the previous section), we have

$$\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big] = \mathbb{P}\Big[C_{\mathrm{cr}} > \frac{\mathbb{E}\big[N_{\mathrm{mc}}\big]\big[N_{\mathrm{mu}}\big]\tau_{\mathrm{mc}}}{\gamma}\Big] \tag{C.34}$$

$$= 1 - \mathbb{P}\Big[C_{\mathrm{cr}} \leq \frac{\mathbb{E}\big[N_{\mathrm{mc}}\big]\mathbb{E}\big[N_{\mathrm{mu}}\big]\tau_{\mathrm{mc}}}{\gamma}\Big] \tag{C.35}$$

$$= 1 - \exp\Big(-\frac{\tau_{\mathrm{mc}}\lambda_{\mathrm{cr}}}{\mu\gamma\lambda_{\mathrm{ut-m}}}\Big). \tag{C.36}$$

4) Final expression: For the evaluation of $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$ in (C.9), we consider that $f_{\mathrm{mc}}(r_{\mathrm{mc}}) = \frac{k}{\nu}\big(\frac{r_{\mathrm{mc}}}{\nu}\big)^{k-1}e^{-(r_{\mathrm{mc}}/\nu)^k}$ where k and ν are shift and shape parameters respectively. Also the integration limit is from 0 to R_{c}. Plugging expressions of Laplace transforms (namely (C.31) and (C.32)) into (C.9), and considering (C.36) give the final expression of (C.3). For the notational convenience, we denote $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$ and $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big]$ as $B_1^{(\mathrm{cap})}$ and $B_2^{(\mathrm{cap})}$ respectively. This concludes our approximation for capacity-aided deployment.

We therefore conclude the proof of average delivery rate of typical macro cell user both in coverage and capacity-aided deployments. ∎

C.2 Proof of Theorem 11

As similar to the rate of typical macro cell user, the rate of typical small cell user is affected by random fluctuations of spatial distribution of cells, fading, interference and capacity of central routers. Let us first denote event $\mathcal{A}_1$ as $\log(1 + \mathrm{SIR}_{\mathrm{su}}) > \tau_{\mathrm{sc}}$, event $\mathcal{A}_2$ as $R'_{\mathrm{su}} > \tau_{\mathrm{sc}}$ and event $\mathcal{A}_3$ as $f_z \in \Delta_x$. Then, the average rate can be expressed by using the rate given in (5.11) in Definition (8), such as

$$\bar{R}_{\mathrm{su}} = \mathbb{E}\Big[R_{\mathrm{su}}\Big] \tag{C.37}$$

$$\overset{(a)}{=} \mathbb{E}\Big[\tau_{\mathrm{sc}}\mathbb{P}\big[\mathcal{A}_2\big]\Big(\mathbb{P}\big[\mathcal{A}_2\big] + \mathbb{P}\big[\mathcal{A}_3\big] - \mathbb{P}\big[\mathcal{A}_2\big]\mathbb{P}\big[\mathcal{A}_3\big]\Big)\Big] \tag{C.38}$$

$$\overset{(b)}{=} \tau_{\mathrm{sc}}\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big] + \tau_{\mathrm{sc}}\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_3\big]\Big] - \tau_{\mathrm{sc}}\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big]\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_3\big]\Big] \tag{C.39}$$

where (a) comes from the compact rate definition in (5.11), (b) is due to the linear properties of expectation operator and independence of event $\mathcal{A}_1$ with respect to $\mathcal{A}_2$ and $\mathcal{A}_3$

and regrouping the terms. In the above expressions, observe that we skip to mention conditioning for notational clarity. However, we shall later mention in case of need. In this proof, we make use of some proof techniques developed in [205].

In the following, we first refine the expressions for $\mathbb{E}\left[\mathbb{P}[\mathcal{A}_1]\right]$ and $\mathbb{E}\left[\mathbb{P}[\mathcal{A}_2]\right]$, then we focus on crucial parts, then finalize our calculations.

We start by refining the term $\mathbb{E}\left[\mathbb{P}[\mathcal{A}_1]\right]$, that is

$$\mathbb{E}\left[\mathbb{P}[\mathcal{A}_1]\right] = \mathbb{E}\left[\mathbb{P}\left[\log(1 + \text{SIR}_{\text{su}}) > \tau_{\text{sc}}\right]\right] \tag{C.40}$$

$$\stackrel{(a)}{=} \int_0^\infty \mathbb{P}\left[\log(1 + \text{SIR}_{\text{su}}) > \tau_{\text{sc}}\big|r_{\text{sc}}\right] f_{\text{sc}}(r_{\text{sc}}) \mathrm{d}r_{\text{sc}} \tag{C.41}$$

$$\stackrel{(b)}{=} \int_0^\infty \mathbb{P}\left[\log(1 + \frac{P_{\text{sc}} g r^{-\alpha}}{I_{\text{ss}} + I_{\text{ms}}}) > \tau_{\text{sc}}\big|r_{\text{sc}}\right] f_{\text{sc}}(r_{\text{sc}}) \mathrm{d}r_{\text{sc}} \tag{C.42}$$

$$\stackrel{(c)}{=} \int_0^\infty \mathbb{P}\left[g > \frac{(e^{\tau_{\text{sc}}} - 1)(I_{\text{ss}} + I_{\text{ms}})}{P_{\text{sc}} r^{-\alpha}}\big|r_{\text{sc}}\right] f_{\text{sc}}(r_{\text{sc}}) \mathrm{d}r_{\text{sc}} \tag{C.43}$$

$$\stackrel{(d)}{=} \int_0^\infty \mathbb{E}_{I_{\text{ss}}, I_{\text{ms}}}\left[\mathbb{P}\left[g > \frac{(e^{\tau_{\text{sc}}} - 1)(I_{\text{ss}} + I_{\text{ms}})}{P_{\text{sc}} r_{\text{sc}}^{-\alpha}}\big|r_{\text{sc}}, I_{\text{ss}}, I_{\text{ms}}\right]\right] f_{\text{sc}}(r_{\text{sc}}) \mathrm{d}r_{\text{sc}}$$

$$\stackrel{(e)}{=} \int_0^\infty \exp\left(-\frac{(e^{\tau_{\text{sc}}} - 1)(I_{\text{ss}} + I_{\text{ms}})}{P_{\text{sc}} r_{\text{sc}}^{-\alpha}}\right) f_{\text{sc}}(r_{\text{sc}}) \mathrm{d}r_{\text{sc}} \tag{C.44}$$

$$= \int_0^\infty e^{-\frac{(e^{\tau_{\text{sc}}} - 1)}{P_{\text{sc}} r_{\text{sc}}^{-\alpha}}} \mathcal{L}_{I_{\text{ss}}}\left(\frac{e^{\tau_{\text{sc}}} - 1}{P_{\text{sc}} r_{\text{sc}}^{-\alpha}}\right) \mathcal{L}_{I_{\text{ms}}}\left(\frac{e^{\tau_{\text{sc}}} - 1}{P_{\text{sc}} r_{\text{sc}}^{-\alpha}}\right) f_{\text{sc}}(r_{\text{sc}}) \mathrm{d}r_{\text{sc}} \tag{C.45}$$

where steps (a) and (b) follow from the definition of SIR in (5.7) and averaging over the spatial distribution of cells for a given PDF of $f_{\text{sc}}(r_{\text{sc}})$. The step (c) is obtained by rearranging the terms so that the fading coefficient h is left on the left hand side, the step (d) is due to the averaging over the interference terms I_{ss} and I_{ms}, and the step (e) is obtained by considering the fact that g is an Exponentially distributed random variable with mean 1. The final step follows from the definition of Laplace transform with $\mathcal{L}_{I_{\text{ss}}} = \mathbb{E}_{I_{\text{ss}}}[e^{-sI_{\text{ss}}}]$ and $\mathcal{L}_{I_{\text{ms}}} = \mathbb{E}_{I_{\text{ms}}}[e^{-sI_{\text{ms}}}]$, where $s = \frac{e^{\tau_{\text{sc}}} - 1}{P_{\text{sc}} r_{\text{sc}}^{-\alpha}}$. Now, let us refine the term $\mathbb{E}\left[\mathbb{P}[\mathcal{A}_2]\right]$, that is

$$\mathbb{E}\left[\mathbb{P}[\mathcal{A}_2]\right] = \mathbb{E}\left[\mathbb{P}\left[R'_{\text{su}} > \tau_{\text{sc}}\right]\right] \tag{C.46}$$

$$\stackrel{(a)}{=} \mathbb{P}\left[\frac{\gamma C_{\text{cr}}}{\mathbb{E}[N_{\text{sc}}] \mathbb{E}[N_{\text{su}}]} > \tau_{\text{sc}}\right] \tag{C.47}$$

$$= \mathbb{P}\left[C_{\text{cr}} > \frac{\mathbb{E}[N_{\text{sc}}] \mathbb{E}[N_{\text{su}}] \tau_{\text{mc}}}{\gamma}\right] \tag{C.48}$$

where (a) is due to the the definition of rate splitting policy in (5.9) and using the fact that the term inside the outer expectation is constant with respect to the spatial random variations. However, the inner expectations are with respect to the spatial random variables. In the above expression, the last step is obtained by rearranging the terms. Finally,

for the term $\mathbb{E}\left[\mathbb{P}\left[\mathcal{A}_2\right]\right]$ (namely cache hit probability), we have that

$$\mathbb{E}\left[\mathbb{P}\left[f_z \in \Delta_x\right]\right] = \int_0^\infty f_{\mathrm{pop}}\left(f, \gamma\right)\mathrm{d}f \tag{C.49}$$

$$= \int_1^{1+F_{\mathrm{sc}}} f_{\mathrm{pop}}\left(f, \gamma\right)\mathrm{d}f \tag{C.50}$$

$$= 1 - \left(1 + F_{\mathrm{sc}}\right)^{1-\eta} \tag{C.51}$$

After this general refinement, we now turn our attention to the final expressions for coverage and capacity-aided deployments separately.

C.2.1 Coverage-aided deployment

Note that the typical user located at origin o is connected to the nearest small cell x_0, and assumed at distance $r_{\mathrm{sc}} = \|x_0 - o\|$. We start with Laplace transforms of I_{mm} and I_{sm}, then focus on the backhaul part and final expressions.

Laplace transform of I_{ss}: Let us assume that $\tilde{I}_{\mathrm{ss}}$ is the interference due to the points in Φ_{ss} expect the nodes within the distance r_{sc} from the typical small cell user. Therefore, the interference I_{ss} is stochastically dominated by the $\tilde{I}_{\mathrm{ss}}$, and we have

$$\mathcal{L}_{I_{\mathrm{ss}}}(s) \approx \mathcal{L}_{\tilde{I}_{\mathrm{ss}}}(s) \tag{C.52}$$

$$= \exp\left(-\pi\lambda_{\mathrm{sc}'}\int_{r_{\mathrm{sc}}^2}^\infty \left(\frac{1}{1 + s^{-1}P_{\mathrm{sc}}^{-1}x^{\alpha/2}}\right)\mathrm{d}x\right) \tag{C.53}$$

$$= \exp\left(\frac{-s\pi\lambda_{\mathrm{sc}'}P_{\mathrm{sc}}(2/\alpha)}{1 - 2/\alpha}r_{\mathrm{sc}}^{2-\alpha}F\left(1, 1 - 2/\alpha; 2 - 2/\alpha; -sP_{\mathrm{sc}}r_{\mathrm{sc}}^{-\alpha}\right)\right). \tag{C.54}$$

Laplace transform of I_{ms}: Supposing that $\mathcal{H}_{\mathrm{sc}}$ is the ball centered at the position of connected small cell with radius R_{c} and $\mathcal{H}_{\mathrm{sc}}^{\mathcal{C}} = \mathbb{R}^2 \setminus \mathcal{H}_{\mathrm{sc}}$, the Laplace transform can be expressed as

$$\mathcal{L}_{I_{\mathrm{ms}}}(s) = \exp\left\{-\lambda_{\mathrm{mc}}\left(\frac{(sP_{\mathrm{mc}})^{2/\alpha}\pi^2(2/\alpha)}{\sin(\pi\frac{2}{\alpha})} - \pi R_{\mathrm{c}}^2 A_{\mathrm{sc}}(s, R_{\mathrm{c}})\right)\right\}, \tag{C.55}$$

where

$$A_{\mathrm{sc}}(s, R_{\mathrm{c}}) = \frac{1}{\pi R_{\mathrm{c}}^2}\int_0^{2\pi}\int_0^{r_{\mathrm{sc}}\cos\varphi + \sqrt{R_{\mathrm{c}}^2 - r_{\mathrm{sc}}^2\sin^2\varphi}}\frac{r\mathrm{d}r\mathrm{d}\varphi}{1 + s^{-1}P_{\mathrm{mc}}^{-1}r^\alpha}. \tag{C.56}$$

Backhaul part: By Lemma 9 and independence of Φ_{sc} and Φ_{cr}, the average number of small cells connected to a nearest central router is

$$\mathbb{E}[N_{\mathrm{sc}}] = \frac{\lambda_{\mathrm{sc}}}{\lambda_{\mathrm{cr}}} \tag{C.57}$$

$$= \frac{\lambda_{\mathrm{sc}'}\exp(-\lambda_{\mathrm{mc}}\pi R_{\mathrm{c}}^2)}{\lambda_{\mathrm{cr}}}. \tag{C.58}$$

Even though the process of small cells is not a PPP but Poisson hole process, note that the results still hold due to motion-invariant properties of the process. Similarly, the average number of users connected to a small cell is given by

$$\mathbb{E}[N_{\mathrm{mu}}] = \lambda_{\mathrm{sc}} \frac{\lambda_{\mathrm{ut}}}{\lambda_{\mathrm{mc}} + \lambda_{\mathrm{sc}}} \tag{C.59}$$

$$= \frac{\lambda_{\mathrm{sc}'} \exp(-\lambda_{\mathrm{mc}} \pi R_{\mathrm{c}}^2) \lambda_{\mathrm{ut}}}{\lambda_{\mathrm{mc}} + \lambda_{\mathrm{sc}'} \exp(-\lambda_{\mathrm{mc}} \pi R_{\mathrm{c}}^2)}. \tag{C.60}$$

Based on the expressions for $\mathbb{E}[N_{\mathrm{sc}}]$ and $\mathbb{E}[N_{\mathrm{mu}}]$, and using the fact that C_{cr} is Exponentially distributed random variable with mean μ, we have

$$\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big] = \mathbb{P}\Big[C_{\mathrm{cr}} > \frac{\mathbb{E}\big[N_{\mathrm{sc}}\big]\mathbb{E}\big[N_{\mathrm{su}}\big]\tau_{\mathrm{sc}}}{\gamma}\Big] \tag{C.61}$$

$$= 1 - \mathbb{P}\Big[C_{\mathrm{cr}} \leq \frac{\mathbb{E}\big[N_{\mathrm{sc}}\big]\mathbb{E}\big[N_{\mathrm{su}}\big]\tau_{\mathrm{sc}}}{\gamma}\Big] \tag{C.62}$$

$$= 1 - \exp\Big(-\frac{\tau_{\mathrm{sc}}\lambda_{\mathrm{cr}}(\lambda_{\mathrm{mr}} + \lambda_{\mathrm{sc}})}{\mu\gamma\lambda_{\mathrm{sc}}^2\lambda_{\mathrm{ut}}}\Big) \tag{C.63}$$

$$= 1 - \exp\Big(-\frac{\tau_{\mathrm{sc}}\lambda_{\mathrm{cr}}(\lambda_{\mathrm{mr}} + \lambda_{\mathrm{sc}})}{\mu\gamma\lambda_{\mathrm{sc}}^2\lambda_{\mathrm{ut}}}\Big) \tag{C.64}$$

where $\lambda_{\mathrm{sc}} = \lambda_{\mathrm{sc}'} \exp(-\lambda_{\mathrm{mc}} \pi R_{\mathrm{c}}^2)$.

4) Final expression: For the evaluation of $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$ in (C.45), we consider that $f_{\mathrm{sc}}(r_{\mathrm{sc}}) = \frac{k}{\nu}\big(\frac{r_{\mathrm{sc}}}{\nu}\big)^{k-1} e^{-(r_{\mathrm{sc}}/\nu)^k}$ where k and ν are shift and shape parameters respectively [205]. Also the integration limit is from 0 to R_{c}. Plugging expressions of Laplace transforms (namely (C.65) and (C.66)) into (C.45), and considering both (C.70) and (C.51) give the final expression of (C.39). For the notational convenience, we denote $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$, $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big]$ and $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_3\big]\Big]$ as $C_1^{(\mathrm{cov})}$, $C_2^{(\mathrm{cov})}$ and $C_3^{(\mathrm{cov})}$ respectively. This concludes our approximation for coverage-aided deployment.

C.2.2 Capacity-aided deployment

Following the similar organization as in the previous section, we first focus on Laplace transforms of I_{ss} and I_{ms}, then detail the backhaul part. The finall expressions shall follow afterwards.

1) Laplace transform of I_{ss}: Since the typical user is connected to the nearest small cell is (r_{sc} distance far away, there is no small cells in the ball centered at the origin with radius r_{sc}. Using a modified path loss function $\tilde{\ell}(x) = \ell(x) \mathbb{1}_{\{\|x\| > r_{\mathrm{sc}}\}}$ and using results in [245] (see Eq. (34)), the Laplace transform can be expressed as

$$\mathcal{L}_{I_{\mathrm{ss}}}(s) = \exp\Big(-\lambda_{\mathrm{sc}'}\int_{\mathbb{R}^2}\Big(1 - \exp(-\bar{c}\nu(s,x))\Big)\mathrm{d}x\Big)\int_{\mathbb{R}^2}\Big(\exp(-\bar{c}\nu(s,x))\Big)f(x)\mathrm{d}x \tag{C.65}$$

where $\nu(s,x) = \int_{\mathbb{R}^2} \frac{f(y)}{1+(sP_{\text{sc}}\bar{\ell}(y-x))^{-1}} \mathrm{d}y$.

2) Laplace transform of I_{ms}: Since the interference is due to macro cells which is formed by a PPP, we have

$$\mathcal{L}_{I_{\text{ms}}}(s) = \exp\Big(-\lambda_{\text{mc}}\frac{(sP_{\text{mc}})^{2/\alpha}\pi^2(2/\alpha)}{\sin(\pi\frac{2}{\alpha})}\Big). \tag{C.66}$$

3) Backhaul part: Due to independence of Φ_{sc} and Φ_{cr}, the average number of small cells connected to a nearest central router is deduced by Lemma 9, that is to say

$$\mathbb{E}[N_{\text{sc}}] = \frac{\lambda_{\text{sc}'}\bar{c}}{\lambda_{\text{cr}}}. \tag{C.67}$$

Similarly, the average number of users connected to a small cell is given by $\mathbb{E}[M_{\text{su}}] = \frac{\lambda_{\text{ut}-\text{s}}}{\lambda_{\text{sc}'}\bar{c}}$. Under these observations and using similar arguments as in the backhaul part of coverage-aided deployment (see the previous section), we have

$$\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big] = \mathbb{P}\Big[C_{\text{cr}} > \frac{\mathbb{E}\big[N_{\text{sc}}\big]\big[N_{\text{su}}\big]\tau_{\text{sc}}}{\gamma}\Big] \tag{C.68}$$

$$= 1 - \mathbb{P}\Big[C_{\text{cr}} \le \frac{\mathbb{E}\big[N_{\text{sc}}\big]\mathbb{E}\big[N_{\text{su}}\big]\tau_{\text{sc}}}{\gamma}\Big] \tag{C.69}$$

$$= 1 - \exp\Big(-\frac{\tau_{\text{mc}}\lambda_{\text{cr}}}{\mu\gamma\lambda_{\text{ut}-\text{s}}}\Big). \tag{C.70}$$

4) Final expression: For the evaluation of $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$ in (C.45), we consider that $f_{\text{sc}}(r_{\text{sc}}) = \frac{k}{\nu}\big(\frac{r_{\text{sc}}}{\nu}\big)^{k-1}e^{-(r_{\text{sc}}/\nu)^k}$ where k and ν are shift and shape parameters respectively. Also the integration limit is from 0 to R_{c}. Plugging expressions of Laplace transforms (namely (C.65) and (C.66)) into (C.45), and considering both (C.70) and (C.51) give the final expression of (C.39). For the notational convenience, we denote $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_1\big]\Big]$, $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_2\big]\Big]$ and $\mathbb{E}\Big[\mathbb{P}\big[\mathcal{A}_3\big]\Big]$ as $C_1^{(\text{cap})}$, $C_2^{(\text{cap})}$ and $C_3^{(\text{cap})}$ respectively. This concludes our approximation for capacity-aided deployment.

We therefore conclude the proof of average delivery rate of typical small cell user both in coverage and capacity-aided deployments. ∎