The Dissertation Committee for Derya Malak
certifies that this is the approved version of the following dissertation:

Modeling and Analyzing Device-to-Device Content Distribution in Cellular Networks

Committee:

Jeffrey G. Andrews, Supervisor

François Baccelli

Gustavo de Veciana

Alexandros G. Dimakis

Mihai Sîrbu

Modeling and Analyzing Device-to-Device Content Distribution in Cellular Networks

by

Derya Malak

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2017

To Mediha and Ahmet Haşim.

Acknowledgments

Eskiden şaşardık bazı şeylerin yokluğuna
Artık bu yokları var etmeyi usladık
Ağaçları budadık omandan balıkları tuttuk denizden
Hani bazı açılmaz sanılan kapıları omuzladık
Çünkü herkesin elinde bir saat bir sümbülteber

Kayayı Delen İncir (1982) – Turgut Uyar

My mother, Mediha, has always been my teacher. Her guidance has made me the person I am today. If I could become half as kind, patient, sincere and dedicated as her, that would be good enough for me. My father, Ahmet Haşim, has loved me so much and always believed in hard work and sacrifice in achieving. They have been the most devoted and committed people I have ever known. They have made a big sacrifice and abandoned a foreign country leaving behind everything, to raise me and my younger brother Sezer in our home, Turkey. The past four years have been smooth with their affection and help. My brother, Sezer, has always been very mature and humble, and he has taken care of my parents in my absence. Without the unconditional love and support from my family, this journey, which has become my life now, was not possible.

I am indebted to my Ph.D. advisor Prof. Jeffrey G. Andrews. I have been very fortunate to work with him. Thanks to Jeff for being very open minded, for the invaluable support and sympathy. His intuition, vision, high

standards and top-tier contributions to the field, have expanded my horizon. He has been a great role model to me. He has been very encouraging, and I am very happy that he let me work on problems different from the line of work our group has been doing, which would not have been possible without his support. I have also learnt a lot from him in terms of managing my priorities, improving my teaching and writing abilities. Thanks to him, I also have had the chance to have great collaborations both from industry and academia.

I would like to thank my committee Prof. François Baccelli, Prof. Gustavo de Veciana, Prof. Alexandros G. Dimakis and Prof. Mihai Sîrbu for their feedback and invaluable comments on my dissertation.

I would like to take the opportunity to thank to my professors in UT Austin. I have always felt very privileged to be a part of the Wireless Networking and Communications Group (WNCG). I have learnt a lot from Jeff, François, Gustavo, Sanjay, and their incredibly vast knowledge and different technical backgrounds and approaches. They have been very approachable and easy to talk to. I especially enjoyed how unique, sharp and brilliant Jeff has been as an advisor and teacher. Having no rigorous math background, I have been thrilled to take François's classes on stochastic geometry and random graphs. It has been a unique opportunity to take a course on queuing theory from Gustavo, who is a flawless teacher and researcher. I have been very delighted that he also spared his time whenever I have visited his office. Thanks to Sanjay and his course on advanced topics in probability, I developed an interesting model on exchangeability for cache placement. Overall,

all these people with incredible visions and backgrounds, made up my mind to work on problems in the intersection of their fields. It has been a privilege to share the same environment.

I would like to acknowledge the support from Huawei Technologies and UT Graduate School fellowship. I want to take the opportunity to thank my collaborators. Prof. Harpreet S. Dhillon from Virginia Tech, helped and guided me through my first year, and has been very patient with me. My internship experience with my mentor Dr. Mazin Al-Shalash from Huawei Technologies, Plano, TX, during summers 2014 and 2015 has been very prolific. We have started and continued to work on content caching mechanisms in device-to-device networks. Indeed, this laid the foundation of my dissertation. I have already missed the invaluable whiteboard discussions during those internships. I also have had the great opportunity to collaborate with Dr. Howard C. Huang, who is with Bell Labs, Murray Hill, NJ, during summer 2016. Working with someone with experience in both academia and industry has been a pleasant experience. We have worked on random access communication under resource constraints. I acquired practical knowledge and technical skills from each of my mentors.

I would like to thank my M.S. advisor Prof. Özgür B. Akan. I have learnt a lot from him during my M.S. studies. His efforts and support have made it possible for me to attend the Ph.D. program in UT Austin.

I would like to acknowledge the support of the Scientific and Technological Research Council of Turkey (Tubitak).

I thank Melanie, the graduate program coordinator, for being such an amazing, brilliant and humble human being, rendering everything possible and easing my graduate life since my first day at UT Austin. She has something down to a fine art. I will miss her.

It would not have been possible to make my way through this path without any company. I would like to thank my friends in our research group Xingqin, Arthur, Sarabjot, Qiaoyang, Abhishek, Yingzhe, Mandar, Ahmad and Rebal. I am indebted to my colleagues Ankit, Virag and his wife Shaili, Mridula, Rajat, Abishek, Mandar, Sara, Erik, Preeti, Karthikeyan, Soumya, Ethan, Vatsal, Srinadh, Chang-sik, Subhashini, Eirini, Srilakshmi, Jessica. I would like to thank my friend Gonca for her constant support.

Austin has been an authentic home, by all means. I will cherish the time spent alone for the long river walks, the local coffee shops –Caffé Medici and Mozart's–, Whole Foods, and the art museums, and good times spent with my great friends Ankit, Koray, Mandar, Sara, Erik, Rajat, Mridula, Abishek, and Preeti, and I am indebted to Mridula for being warm and uniquely beautiful. I would like to thank my other half Murat, who has been with me throughout the years. Graduate life has its own pearls and pitfalls. My last four years in Austin have thought me a lot and helped me become persistent and patient, and see what I genuinely expect from life. This experience has led me down to a unique path and shaped my future. I look forward to becoming a good teacher, a lifelong learner, seeing what is behind the wall of theory, and adding something from myself.

Modeling and Analyzing Device-to-Device Content Distribution in Cellular Networks

Publication No. _____

Derya Malak, Ph.D.
The University of Texas at Austin, 2017

Supervisor: Jeffrey G. Andrews

Device-to-device (D2D) communication is a promising approach to optimize the utilization of air interface resources in 5G networks, since it allows decentralized proximity-based communication. To obtain caching gains through D2D, mobile nodes must possess content that other mobiles want. Thus, devising intelligent cache placement techniques are essential for D2D. The goal of this dissertation is to provide randomized spatial models for content distribution in cellular networks by capturing the locality of the content, and additionally, to provide dynamic content placement algorithms exploiting the node configurations.

First, a randomized content caching scheme for D2D networks in the cellular context is proposed. Modeling the locations of the devices as a homogeneous Poisson Point Process (PPP), the probability of successful content delivery in the presence of interference and noise is derived. With some

idealized modeling aspects, i.e., given that (i) only a fraction of users to be randomly scheduled at a given time, and (ii) the request distribution does not change over time, it has been shown that the performance of caching can be optimized by smoothing out the request distribution, where the smoothness of the caching distribution is mainly determined by the path loss exponent, and holds under Rayleigh, Ricean and Nakagami fading models.

Second, to take the randomized caching model a step further, a spatially correlated content caching scenario is contemplated. Inspired by the Matérn hard-core point process of type II, which is a first-order pairwise interaction model, D2D nodes caching the same file are never closer to each other than the *exclusion radius*. The exclusion radius plays the role of a substitute for caching probability. The optimal exclusion radii that maximize the *hit probability* can be determined by using the request distribution and cache memory size. Unlike independent content placement, which is oblivious to the geographic locations of the nodes, the new strategy can be effective for proximity-based communication even when the cache size is small.

Third, an auction-aided Matérn carrier sense multiple access (CSMA) policy that considers the joint analysis of scheduling and caching is studied. The auction scheme is distributed. Given a *cache configuration*, i.e., the set of cached files in each user at a given snapshot, each D2D receiver determines the value of its request, by bidding on the set of *potential transmitters* in its communication range. The values of the receiver bids are reported to the potential transmitter, which computes the cumulated sum of these variables taken on

all users in its cell. The potential transmitter then reports the value of the bid sum to other potential transmitters in its contention range. Given the accumulated bids of all potential transmitters, the contention range and the medium access probability, a fraction of the potential transmitters are jointly scheduled, determined by the auction policy, in order to optimize the throughput. Later, a Gibbs sampling-based cache update strategy is proposed to iteratively optimize the hit rate by taking the scheduling scheme into account.

In this dissertation, a variety of distributed algorithms for D2D content caching are proposed. Our results indicate that the geographic locality and the network parameters have a significant role in determining and optimizing the placement strategy. Exploiting the user interactions and spatial diversity, and incentivizing cooperation among D2D nodes are crucial in realizing the full potential of caching. Furthermore, from a network point of view, the scheduling and the caching phases are closely linked to each other. Hence, understanding the interaction between these two phases helps develop novel dynamic caching strategies capturing the temporal and spatial locality of the demand.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Wireless networks are experiencing exploding demand for data services driven by the proliferation of smart devices. Forecasts indicate that cellular networks may need to support a sevenfold increase in capacity from 2016 to 2021. Currently more than half of all wireless data bits are video. By the end of the decade, video is expected to consume 78 percent of wireless bandwidth [1]. Driven by this insatiable demand for wireless capacity, different technologies, such as ultra-high density heterogeneous base station (BS) deployments, and directly communicating data from one device to device (D2D) to another without traversing the network [2], have come to the forefront as candidates for the next (5th) generation of wireless networks.

D2D communication is a promising technique for enabling proximity-based applications involving discovering and communicating with nearby devices. It has the advantage of a limited investment requirement, since the increasing density of users, and increasing capacity of the handheld devices provide high amounts of data being stored locally, and enable the likelihood of finding the desired content locally instead of accessing the BS. D2D also provides increased offloading from the heavily loaded cellular network, which

is justifiable as memory costs continue to plummet, and machine learning approaches are expected to provide an accurate prediction of the demand by facilitating the features extracted from query prefetching history, hence eliminating the need for using large number of resources [3, 4]. In addition to these, content caching is indispensable to D2D because D2D without caching does not exploit how to effectively distribute popular content and is futile.

Caching of popular content at various nodes in the network is a well known technique to optimize the utilization of air-interface resources in cellular networks, and increase content access speed and availability [5]. D2D communications will be an important component of the 5th and 6th generations of wireless networks to meet the growing demand for local wireless services [6]. D2D communication and several use cases are being actively standardized by 3GPP to allow device discovery, decentralized file sharing and public safety applications [7–10]. There are many different mobile applications for content caching and routing that enable smartphones to connect via Bluetooth or through their Wi-Fi interfaces such as Inmobly, Amazon CloudFront, CacheFly Content Distribution Network (CDN) and FireChat [11], [12]. These projects aim to develop technology to create direct connections between cellular phones without the need of a mobile phone operator.

D2D communication intriguing since it allows increased spatial reuse and possibly very high rate communication without increased network infrastructure or new spectrum, but is only viable when the mobile users have content that other nearby users want [13], which allows short-range communica-

tion which is independent of the network infrastructure. Therefore, intelligent caching of popular files is indispensable for D2D to be successful. By caching content directly on the devices, and by exploiting D2D communication, the devices themselves can form an effective CDN.

Rest of the chapter is split into several parts. In Section 1.1, the related work on D2D in the cellular context is discussed. In Section 1.2, an overview of content distribution using D2D caching in wireless networks is given. In Section 1.3, we summarize the network model. In Section 1.4, we discuss different coverage models and their applicability for different network scenarios. In Section 1.6, we describe how to optimize the cache hit probability, and discuss possible approaches to optimize the performance of D2D caching. In Section 1.8, we briefly discuss the key contributions of this dissertation, and in Sect. 1.9, we outline the organization of the dissertation.

1.1 D2D in the Cellular Context

Hybrid networks consisting of both infrastructure-based and ad hoc networks, a more general concept of D2D-enabled cellular networks, have been widely studied in [14–22]. D2D communications in cellular networks have been proposed for relaying purposes to improve the coverage and throughput performance [14–16, 19]. D2D communication has also been studied in the context of Peer-to-Peer (P2P) networking [21, 22], and is a potential efficient component to reduce energy consumption in public safety systems [23]. Energy efficient and user friendly device discovery schemes, resource management for

D2D communication as an overlay or underlay to a cellular network, and D2D mode selection are detailed in [24]. For a detailed survey on D2D communication and an extensive study on integrated cellular and D2D communications, readers are referred to [24, 25].

Unlike general ad hoc networks, D2D can benefit from cellular infrastructure (e.g., network coordinated device discovery, synchronization and enhanced security), and can operate on licensed bands, which makes resource allocation more tractable and reliable [26, 27]. Spectrum sharing for D2D communication in cellular networks is studied in [28]. A framework for providing the optimal resource partitions between D2D and cellular networks, which allows for time-frequency resources to be either shared or orthogonally partitioned between the two networks, are investigated in [26]. Optimal spectrum partition and mode selection in D2D overlaid cellular networks are studied in [29]. A optimization framework for a D2D-enabled downlink cellular network, in which D2D links use a frequency band orthogonal to the cellular users, is developed in [30], in order to determine when the potential D2D users transmit directly, and when they fall back to the cellular mode is proposed.

D2D scheduling and CDNs have been widely studied in [6, 31–33], and caching is utilized to improve the spectral efficiency in D2D wireless network in [13, 34, 35]. Proactive caching has been proposed in [36–41] so that the requests can be tracked, learnt, and predicted ahead of time. Furthermore, with demand shaping and pricing-based models [42–45], the network traffic is smoothed out over time in order to minimize the data delivery costs. Content

dissemination in social networks is explored in [46]. A Transfer learning (TL) approach, which lies in extracting collaborative social behavior information from the source domain to aid in the learning in the target domain, is proposed to learn and transfer the rich contextual information to estimate the large-scale file popularity matrix. Game theory models have also been studied in [47–49] to determine the social optima for data caching models. A profile matching model for proximity-based mobile social networks for user selection has been proposed in [50]. Given the social relations collected by the Evolved Node B (eNB), the traffic offloading process in D2D communication has been optimized in [51].

1.2 An Overview of Content Caching Approaches

Content caching has received significant attention as a means of improving the throughput and latency of networks without requiring additional bandwidth or other technological improvements. Video caching appears particularly profitable and plausible compared to other types of content [41], and is perfectly suited to D2D networks for offloading traffic from congested cellular networks.

Research to date on content caching has been mainly focused on two different perspectives. In one line of work, *given the delivery scheme*, the content placement is optimized by exploiting the statistics of the demands and making popular content available locally, as in [52], [53]. Alternatively, the objective is to optimize the delivery phase *given the cache contents* and for

known demand distribution [54], [55]. In another line of research, researchers have aimed to understand the *fundamental limits of caching gain* with no cooperation. Gain of coded multicasting [56], information theoretic scaling laws for throughput and number of D2D connections, and collaboration distance [57], have been investigated.

Alternatively, as in the current dissertation, there are several studies focusing on *decentralized caching algorithms* that have optimized the caching distribution to maximize the cache hit probability, using deterministic or random caching as in [58], [57] given a BS-user topology. FemtoCaching replaces backhaul capacity with storage capacity at the small cell access points, i.e., helpers, and the optimum way of assigning files to the helpers is analyzed in [59] to minimize the delay. Despite the ongoing research, we still lack a through understanding of the spatial correlations and geographical locality of the demand.

We next detail the D2D network model we utilize in the current dissertation.

1.3 D2D Network Model

We consider a spatial D2D-enabled cellular network setting in which both the D2D user and BS locations are modeled by a Poisson point process (PPP) Φ. Users have limited communication range and finite storage. The D2D users are served by each other if the desired content is cached at a user within its radio range: this is called a *hit*. Otherwise, they are served by

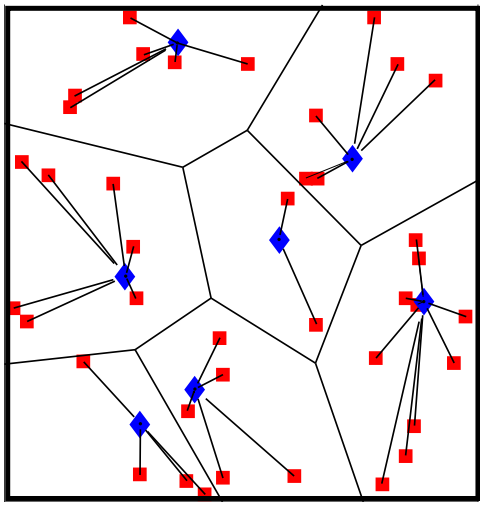


Figure 1.1: Illustration of the nearest BS association in which both the D2D user (square) and BS (diamond) locations are modeled by a Poisson point process (PPP).

the cellular network BS, which is what D2D communication aims to avoid. An example D2D device to BS association model is illustrated in Fig. 1.1, where each D2D is associated with the nearest BS. For clarity, the D2D device associations are not shown.

The associations between the D2D devices depend on their mutual interests and proximity. We next discuss different possible models we can exploit for modeling those interactions.

1.4 Coverage Models

Consider a given realization $\phi = \{x_i\} \subset \mathbb{R}^2$ of the PPP transmitter process Φ. Different coverage models can be used to model the performance of D2D communications. For example, as detailed in [60], three practical

coverage models: (i) the signal-to-interference-plus-noise ratio (SINR) model, (ii) the Boolean model, and (iii) the overlaid network model with orthogonal resources (bandwidth).

The SINR model is well suited to interference-limited networks to describe the coverage quality. The SINR at the reception, $\mathrm{SINR}(x_i)$, when user o at the origin is connected to BS $x_i \in \Phi$ and is defined as

$$\mathrm{SINR}(x_i) = \frac{S_i/l(r_i)}{\sigma^2 + I - S_i/l(r_i)}, \tag{1.1}$$

where S_i is the shadowing experienced between the typical user and the BS at x_i. The parameter $r_i = |x_i|$ is the distance of x_i from o, and $l(r) = r^{-\alpha}$ is the path loss function, with exponent $\alpha > 2$, and the constant σ^2 is the noise power, $I = P\sum_{x_i \in \Phi} S_i/l(r_i)$ is the total received power from the network. The typical user is covered when $\mathrm{SINR}(x_i) > T$, where T is the threshold.

The coverage number $\mathcal{N}(T)$ indicates how many BSs cover the typical user simultaneously and is denoted by the random variable

$$\mathcal{N}(T) = \sum_{x_i \in \Phi} \mathbf{1}[\mathrm{SINR}(x_i) > T]. \tag{1.2}$$

The Boolean model (BM) is tractable for the noise-limited regime [60], where the interference is small compared to the noise [61, Ch. 3]. Specifically, given a transmit power P, if we only consider path loss, no fading and no interference, the received signal at the boundary should be larger than a threshold to guarantee coverage, i.e., $Pr^{-\alpha} \geq T$, yielding $r \leq \mathrm{R_{D2D}} = (P/T)^{\alpha}$. Hence, D2D users can only communicate within a finite range, which we call the D2D

radius, denoted by $\mathrm{R}_{\mathsf{D2D}}$, and the coverage area of the BM is determined by a *fixed communication radius*. A file request is fulfilled by the D2D users within $\mathrm{R}_{\mathsf{D2D}}$ if one has the file; else the D2D user is served by a BS.

In overlaid networks, the coverage distribution is the convolution of the coverage probability distributions of the individual networks given that they are independent. Interested readers can refer to [60] for further details.

1.5 Caching Distribution

Given storage size N, same for all nodes, let Y_{m_i} be the indicator random variable that takes the value 1 if file m is available in the cache located at $x_i \in \Phi$ and 0 otherwise. Thus, the storage constraint is given as

$$\sum_{m=1}^{M} Y_{m_i} \leq N, \quad x_i \in \Phi, \tag{1.3}$$

i.e., Y_{m_i}'s are inherently dependent. Optimal content placement is a binary problem where the cache placement satisfies (1.3). However, the optimization of the cache hit problem given this constraint is combinatorial and is NP-hard.

The caching probability of file m in cache i is given by $p_c(m, x_i) = \mathbb{P}(Y_{m_i} = 1)$. For tractability reasons, we take the expectation of this relation and obtain our relaxed cache placement constraint

$$\sum_{m=1}^{M} p_c(m, x_i) \leq N, \quad x_i \in \Phi. \tag{1.4}$$

Later, we show in Chapters 2 and 3 that there are feasible solutions to the relaxed problem filling up all the cache slots.

1.6 Cache Hit Probability

A key objective is to maximize the *cache hit probability*, which is the probability that a given D2D node can find a desired file at another node's cache within its communication range. Intuitively, given a finite amount of storage at each node, popular content should be seeded into the network in a way that maximizes the hit probability that a given D2D device can find a desired file – selected at random according to a request distribution – within its radio range. We explore this problem quantitatively in this dissertation by considering different spatial content models and deriving, optimizing and comparing the hit probabilities for each of them.

The cache hit probability is expressed as follows:

$$\mathrm{P}_{\mathsf{Hit}} = 1 - \sum_{m=1}^{M} p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}(T) = k)\, \mathrm{P}_{\mathrm{Miss}}(m, k), \tag{1.5}$$

where $\mathbb{P}(\mathcal{N} = k)$ is the coverage distribution, i.e., the probability that k transmitters (caches) cover the typical receiver. The parameter $p_r(m)$ models the request or demand distribution, and $\mathrm{P}_{\mathrm{Miss}}(m, k)$ is the probability that k caches cover a receiver, and none has file m, i.e., the probability of cache miss. Cache misses occur due to limited communication range and finite storage constraint.

We next briefly discuss different components of caching, and how to develop independent or distributed placement techniques, in order to optimize the cache hit probability by maximizing (1.5) subject to the probabilistic placement constraint given in (1.4).

1.6.1 Demand Traffic Models

Assume that there are M total files in the network, where all files have the same size, and each user has the same cache size $N < M$. Depending on its cache state, each user makes requests for new files based on a general popularity distribution over the set of the files. The popularity of such requests is modeled by the Zipf distribution, which has probability mass function (pmf)

$$p_r(i) = \frac{1}{i^{\gamma_r}} \Big/ \sum_{m=1}^{M} \frac{1}{m^{\gamma_r}}, \quad i = 1, \ldots, M, \tag{1.6}$$

where γ_r is the Zipf exponent that determines the skewness of the distribution. The distribution is shown in Fig. 1.2. The demand profile is Independent Reference Model (IRM), i.e., the standard synthetic traffic model in which the request distribution does not change over time. If the objective is to maximize the average cache hit probability of the PPP model, it is sufficient to consider a snapshot of the network, in which the D2D user realization is given and requests are independent and identically distributed (i.i.d.) over the space.

Extensions to also incorporate the temporal correlation of real traffic traces can be done by exploiting models like the Shot-Noise Model (SNM) [62]. This overcomes the limitations of the IRM by explicitly accounting for the temporal locality in requests for contents. However, in that case, the problem under study will have an additional dimension to optimize over, and to do so, online learning algorithms should be developed to both learn the demand and optimize the spatial placement.

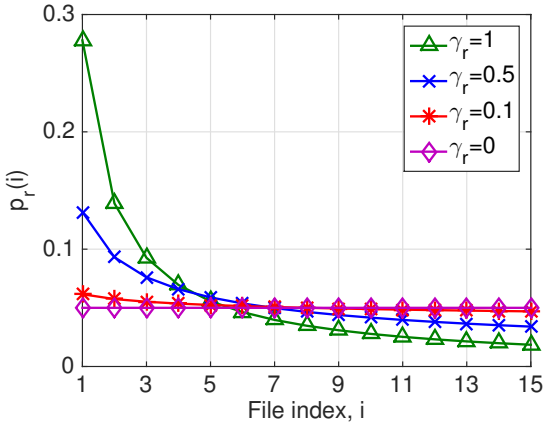


Figure 1.2: Zipf(γ_r) popularity distribution with respect to file index i for different values of γ_r.

In reality, the spatial-temporal distribution of demand should be captured accurately to pave the way for effectively placing content in devices. Video caching and pre-fetching appears particularly profitable and plausible versus other types of content [41], and is perfectly suited to D2D networks for offloading traffic from congested cellular networks. Popularity of videos has strong spatial-temporal correlation. Video has several interesting characteristics. For example, large files, consumed at a near constant rate over a fairly long time. A few key portals, Netflix, Youtube, Hulu, Facebook, etc, serve most of it. Some videos are often shared locally or via social network. Deployment characteristics and densities of BSs in urban and rural regions can be very different. Furthermore, users might be clustered in hotspot zones, such as coffee shops, restaurants, airports, stadium and campus. Therefore, to optimize the performance of wireless caching, the spatial and temporal variation of demand profile and the impact of BS or cache locations should be

accurately modeled.

1.6.2 Spatial Cache Placement

Cache placement can be implemented in an independent manner or in a correlated way. Independent caching is a probabilistic placement model, in which the caches do not cooperate, and the files are independently placed in the cache memories of different nodes according to the same distribution [60], [63], and [52]. Special cases of this model include caching most popular content and geographic content placement (GCP) in [60]. However, it is not usually optimal to cache files independently. In network scenarios, better approaches can be implemented by developing cooperative, i.e., spatially correlated, cache placement strategies rather than independently placing the files, which can improve the cache hit rate. However, it is not trivial to design a *joint placement distribution over the geographic domain*. One of the contributions of this dissertation is that we have devised a spatially correlated probabilistic placement policy, in which the D2D caches are loaded in a distributed manner via additional marks attached to them without accounting for any cost, in a timescale that is much shorter than the time over which the device locations are predicted.

An example D2D enabled cellular network scenario that considers the possible interactions between the D2D users and the BS, where coverage is modeled by the Boolean model, is illustrated in Fig. 1.3, in which each D2D device has a cache size of $N = 2$. Different content types are denoted by

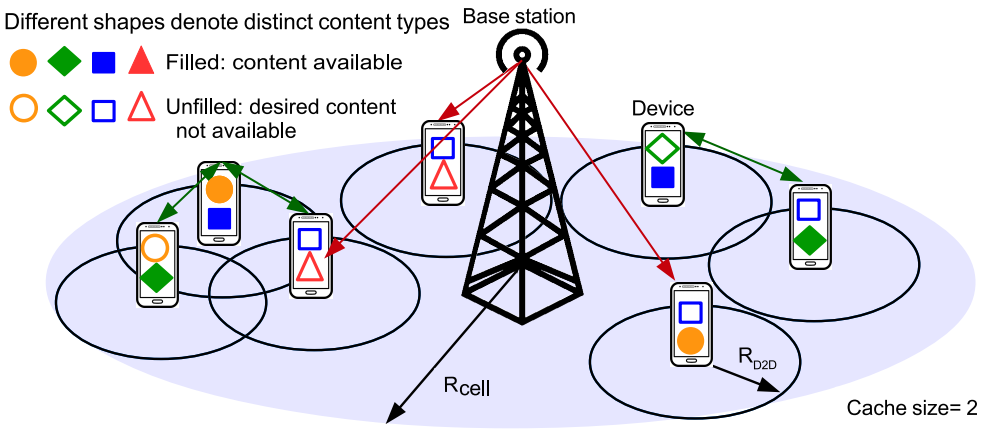


Figure 1.3: An illustration of spatial content placement for a D2D enabled cellular network model.

distinct shapes. If the content is available in a cache, then the corresponding shape is filled, and vice versa. The parameters R_{D2D} and R_{cell} denote the ranges for D2D and cellular communications, respectively. If the devices are within range, they can obtain the desired contents from each other. Otherwise, the BS serves the requests.

1.7 Spatial-Temporal Dynamics

The local demand profile of receivers change over time. To develop an efficient caching algorithm, it is required to estimate the popularity profile, then optimize the caching strategy, i.e., the admission and extinction policies, in order to maximize the cache hit probability and balance the load at the same time.

From a temporal perspective, prefetching and proactive caching [37] have been shown to provide significant caching gains. From a geographic perspective, distributed solutions are required for scalability and to improve different utility metrics. For example, a node (cache) can decide what to store and when to update its configuration based on the side information, i.e., content availabilities, of its nearest neighbors. File insertion (or similarly eviction) rate should be determined according to (i) the popularity profile of the file, (ii) the cache configurations of the neighbors, and (iii) the amount of storage. There are well-known models to model the short-range interactions. Dynamical models, such as Gibbs point processes (GPPs) [64, Ch. 5.5], Ising-Glauber models [65], and mean-field approximation as an effective field that represents a substitute for the local interactions between cache states [66] can be utilized. Exploiting those, spatial-temporal models that reach an equilibrium state within a specified time can be designed.

1.7.1 Modeling and Algorithmic Challenges

To capture the impact of (i) *the temporal variations* and correlation of content requests, models like the SNM should be exploited, in order to account for the temporal locality in requests for contents and correctly predict the per file popularities and the overall request distribution, and (ii) *the geographic locality* of content can be tailored to provide file selection diversity to users in order not to under or over-cache a file in a given area, and captured to estimate the future spatial request distribution, and build a local empirical

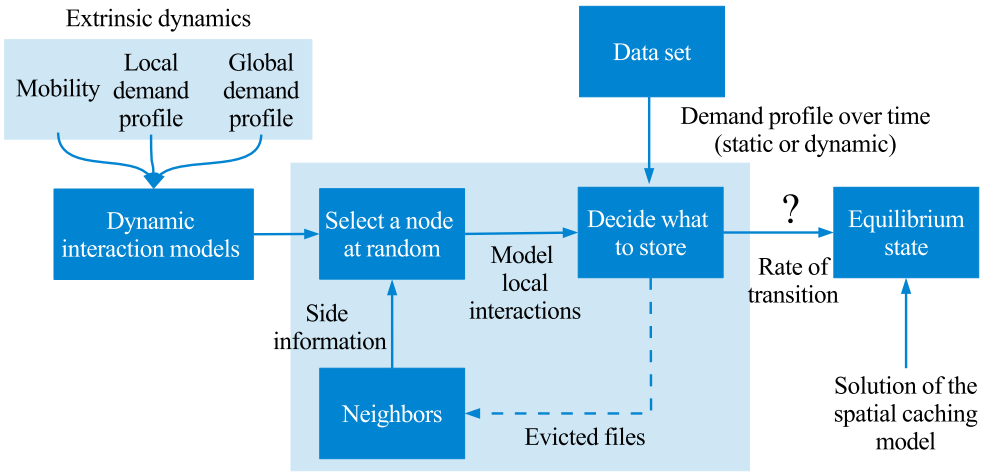


Figure 1.4: Dynamic caching algorithm.

request distribution based on the local demand behavior, and determine what files to cache where.

In Fig. 1.4, we give an outline for a spatial-temporal algorithm that captures the interactions between users, compares with the extrinsic demand dynamics, in order to devise a caching algorithm that can reach to an equilibrium state (solution) within a desired duration. In Chapter 4 of this dissertation, adapted from this outline, we propose a dynamic caching model capturing neighboring interactions in order to maximize the cache hit probability.

1.7.2 Testing Theory with Data Set

We will use proprietary data on movie requests and ratings over time. Although our current data has no geographic information, there are empirical

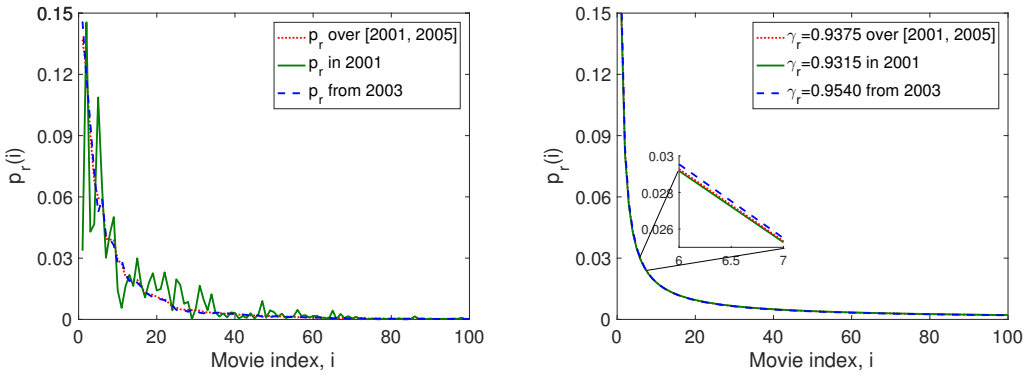


Figure 1.5: Proprietary data on movie requests (left), and the Zipf distribution approximation on movie requests (right).

models to predict the relationship between the traffic density and the spatial distribution of base stations [67]. Zipf distribution is a good approximation for modeling the static (IRM) demand, but no longer valid when demand distribution changes over time. Demand distribution can have a high variation over time, which can be seen from Fig. 1.5 (left). Some files are requested at a lower rate but their popularities do not fade away over time. On the other hand, some files have instantaneous popularity and their popularities fade away quickly. Therefore, it is important to estimate the variation of popularity over time. A minimum mean square error (MMSE) estimator for approximating the proprietary data in Fig. 1.5 with the Zipf distribution is given is Fig. 1.5 (right). As can be seen, the Zipf distribution does not give a good approximation for the variation of demand distribution over short time intervals. In Fig. 1.6, the linear regression between the data and popularity distribution (both in logarithmic scale) is illustrated to demonstrate the

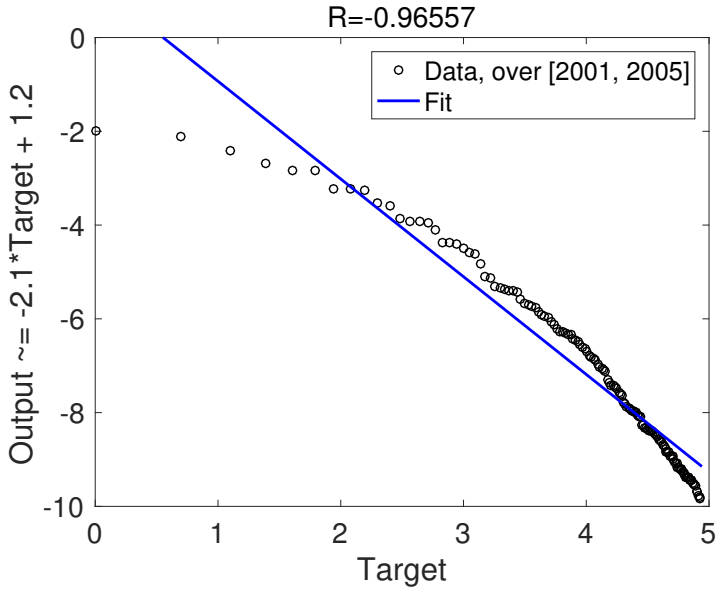


Figure 1.6: Linear regression between the logarithm of file indices and the corresponding popularities in logarithmic scale.

accuracy of Zipf approximation for long time intervals.

1.8 Contributions

This dissertation focuses on the analysis and design of content aggregation and caching approaches for the D2D networks in the cellular context. Specifically, we study optimal content caching strategies to maximize the density of successful receptions as a function of the coverage distribution in D2D networks, propose to investigate a caching model for D2D by incorporating the spatial, or geographic, and spatial-temporal characteristics and network dynamics, and analyze an energy efficient multi-hop data aggregation model for MTC uplink, and propose a delay-sensitive RA scheme. The following are

the contributions of my dissertation.

Optimizing Density of Successful Receptions with D2D Caching: In this dissertation, we use results from stochastic geometry to derive the probability of successful content delivery in the presence of interference and noise. We employ a general transmission strategy where multiple files are cached at the users and different files can be transmitted simultaneously throughout the network. We then formulate an optimization problem, and find the caching distribution that maximizes the density of successful receptions (DSR) under a simple transmission strategy where a single file is transmitted at a time throughout the network. We model file requests by a Zipf distribution with exponent γ_r, which results in an optimal caching distribution that is also a Zipf distribution with exponent γ_c, which is related to γ_r through a simple expression involving the path loss exponent. We also develop strategies to optimize content caching for the more general case with multiple files, and bound the DSR for that scenario.

Spatially Correlated Caching for D2D Communications: We study optimal geographic content placement for device-to-device (D2D) networks in which each file's popularity follows the Zipf distribution. The locations of the D2D users (caches) are modeled by a Poisson point process (PPP) and have limited communication range and finite storage. We initially propose a spatially exchangeable content placement technique to prioritize the caches for content placement. We demonstrate that exchangeable placement actually performs worse than the baseline independent content placement. Later,

inspired by the Matérn hard-core (type II) point process that captures pairwise interactions between nodes, we devise a novel spatially correlated caching strategy called *hard-core placement* (HCP) such that the D2D nodes caching the same file are never closer to each other than the *exclusion radius*. The exclusion radius plays the role of a substitute for caching probability. We derive and optimize the exclusion radii to maximize the *hit probability*, which is the probability that a given D2D node can find a desired file at another node's cache within its communication range. Contrasting it with independent content placement, which is used in most prior work, our analysis shows that our HCP strategy often yields a significantly higher cache hit probability. We further demonstrate that the HCP strategy is effective for small cache sizes and a small communication radius, which are likely conditions for D2D.

A Distributed Auction Policy for User Association in D2D Caching Networks: Unlike the randomized content caching models, which do not capture the network dynamics and spatial characteristics of D2D networks, next, in Chapter 4, we contemplate a more sophisticated caching model to achieve desirable hit rates by jointly determining how to cache the files and schedule the transmissions in D2D networks. We propose a distributed bidding-aided Matérn carrier sense multiple access (CSMA) policy for device-to-device (D2D) content distribution. The network is composed of D2D receivers and potential D2D transmitters, i.e., transmitters are turned on or off by the scheduling algorithm. Each D2D receiver determines the value of its request, by bidding on the set of potential transmitters in its communi-

cation range. Given a medium access probability, a fraction of the potential transmitters are jointly scheduled, i.e., turned on, determined by the auction policy. The bidding-aided scheduling algorithm exploits (i) the local demand distribution, (ii) spatial distribution of D2D node locations, and (iii) the cache configurations of the potential transmitters. We contrast the performance of the bidding-aided CSMA policy with other well-known CSMA schemes that do not take into account (i)-(iii), demonstrate that our algorithm achieves a higher spectral efficiency in terms of the number of bits transmitted per unit time per unit bandwidth per user. The gain becomes even more visible under randomized configurations and requests rather than more skewed placement configurations and deterministic demand distributions. Incorporating the Gibbs sampling method for cache updates into the scheduling policy, we later aim to iteratively maximize the cache hit rate.

1.9 Organization

The contributions of the dissertation are covered in Chapters 2 through 4. Chapter 2 proposes a new probabilistic content caching model that maximizes the density of successful receptions in D2D networking. Chapter 3 discusses a spatially correlated caching model for D2D. Chapter 4 focuses a spatial-temporal content caching model for D2D communications that jointly considers the optimization of user associations and content placement by incorporating the cache dynamics and geographic characteristics of D2D users. The dissertation is concluded in Chapter 5 and the proposed research is outlined.

Chapter 2

Optimizing Content Caching to Maximize the Density of Successful Receptions in D2D Networking

Wireless networks are experiencing a well-known ever-rising demand for enhanced high rate data services, in particular wireless video, which is forecast to consume over three-fourths of wireless bandwidth by 2021 [1]. Non-real-time video in particular is expected to comprise half of this amount [68], and comprises large files that can be cached in the network. Meanwhile, preliminary D2D techniques have been standardized by 3GPP to allow decentralized file sharing and public safety applications [10]. D2D is intriguing since it allows increased spatial reuse and possibly very high rate communication without increased network infrastructure or new spectrum, but is only viable when the mobile users have content that other nearby users want. Thus, it is clear that smart content caching is essential for D2D[1].

Caching popular content is a well known technique to reduce resource usage, and increase content access speed and availability [5]. Infrastructure-

[1]This chapter has been published in [69], [70], [52]. I am the primary author of these works. Coauthor Dr. Mazin Al-Shalash has provided many valuable discussions and insights to this work, and Dr. Jeffrey G. Andrews is my supervisor.

based caching can reduce delay and when done at the network edge, also reduce the impact on the backhaul network, which in many cases is the bottleneck in wireless networks [59]. However, this type of caching does not reduce the demand on spectral resources. To gain spectral reuse and increase the area spectral efficiency, the content must be cached on wireless devices themselves, which allows short-range communication which is independent of the network infrastructure. D2D communication can enable proximity-based applications involving discovering and communicating with nearby devices [28]. Synchronized distributed network architectures for D2D communications are designed, e.g., FlashLinQ [6] and ITLinQ [34], and caching is shown to provide increased spectral reuse in D2D-enabled networks [13]. Although order optimal solutions for optimal content placement is known under certain channel conditions [71–73], it is not known how to best cache content in a D2D network. Intuitively, popular content should be seeded into the users' limited storage resources in a way that maximizes the probability that a given D2D device can find a desired file within its radio range. Exploring this problem quantitively is the goal of this chapter.

2.1 Related Work

Different aspects of D2D content distribution are studied. Scalability in ad hoc networks is considered [74], where decentralized algorithms for message forwarding are proposed by considering a Zipf product form model for message preferences. Throughput scaling laws with caching have been widely studied

[56,57,75]. Optimal collaboration distance, Zipf distribution for content reuse, best achievable scaling for the expected number of active D2D interference-free collaboration pairs for different Zipf exponents is studied [76]. With a heuristic choice (Zipf) of caching distribution for Zipf distributed requests, the optimal collaboration distance [58] and the Zipf exponent to maximize number of D2D links are determined [75]. However, in general, the caching pmf is not necessarily same as the request pmf. This brings us to the one of the main objectives in this chapter, which is to find the best caching pmf that achieves the best density of successful receptions (DSR) in D2D networks.

Under the classical protocol model of ad hoc networks [77], for a grid network model, with fixed cache size M, as the number of users n and the number of files m become large with $nM \gg m$, the order optimal[2] caching distribution is studied and the per-node throughput is shown to behave as $\Theta(M/m)$ [71,78]. The network diameter is shown to scale as $\sqrt{n}$ for a multi-hop scenario [72]. It is shown that local multi-hop yields per-node throughput scaling as $\Theta(\sqrt{M/m})$ [73].

Spatial caching for a client requesting a large file that is stored at the caches with limited storage, is studied [79]. Using Poisson point process (PPP) to model the user locations, optimal geographic content placement and outage in wireless networks are studied [60]. The probability that the typical user finds the content in one of its nearby base stations (BS)s is optimized using

[2]The order optimality in [71,78] is in the sense of a throughput-outage tradeoff due to simple model used.

the distribution of the number of BSs simultaneously covering a user [80]. Performance of randomized caching in D2D networks from a DSR maximization perspective has not been studied, which we study in this chapter.

Although the work conducted in [75,76] focused on the optimal caching distribution to maximize the average number of connections, the system model was overly simplistic. They assumed a cellular network where each BS serves the users in a square cell. The cell is divided into small clusters. D2D communications are allowed within each cluster. To avoid intra-cluster interference, only one transmitter-receiver pair per cluster is allowed, and it does not introduce interference for other clusters. In this chapter, we aim to overcome these serious limitations using a more realistic D2D network model that captures the simultaneous transmissions where there is no restriction in the number of D2D pairs.

2.2 Contributions

This chapter develops optimal content caching strategies that aim to maximize the average density of successful receptions so as to address the demands of D2D receivers. The contributions are as follows.

Physical channel modeling using PPP. We introduce the network model in Sect. 2.3, in which the locations of the D2D users are modeled as a homogeneous PPP. Different from the grid-based model in [71, 78], we consider the actual physical channel model. PPP modeling makes our analysis tractable because unlike the cluster-based model in [58], where only a

pair of users are allowed to communicate in a square region, we require no constraint on the link distance and allow a random number of simultaneous transmissions. All analysis is for a typical mobile node which is permissible in a homogeneous PPP by Slivnyak's theorem [64]. The interference due to simultaneously active transmitters, noise and the small-scale Rayleigh fading are incorporated into the analysis. Any transmission is successful as long as the Signal-to-Interference-plus-Noise Ratio (SINR) is above a threshold.

Density of successful receptions (DSR**).** We propose a new file caching strategy exploiting stochastic geometry and the results of [81], and we introduce the concept of the density of successful receptions (DSR). Although in this chapter, we do not investigate the throughput-outage tradeoff as in [71,78], the DSR is closely related to the outage probability, obtained through the scaling of the coverage, i.e., the complement of the outage probability, with the number of receivers per unit area.

Maximizing the DSR **for the sequential multi-file model.** We study a randomized transmission model for D2D users with storage size 1 in Sect. 2.3. We propose techniques for randomized content caching based on the possible ways of prioritizing different files. In Sect. 2.4, we start with a baseline model with single file to determine the optimal fractions of transmitters γ_1 and receivers γ_2 in the D2D network model with PPP distributed user locations that maximizes the DSR. In Sect. 2.5, we consider the more general sequential multi-file transmission scenario, where we investigate the maximum DSR in terms of the optimal fractions of γ_1 and γ_2 derived in Sect. 2.4, to determine

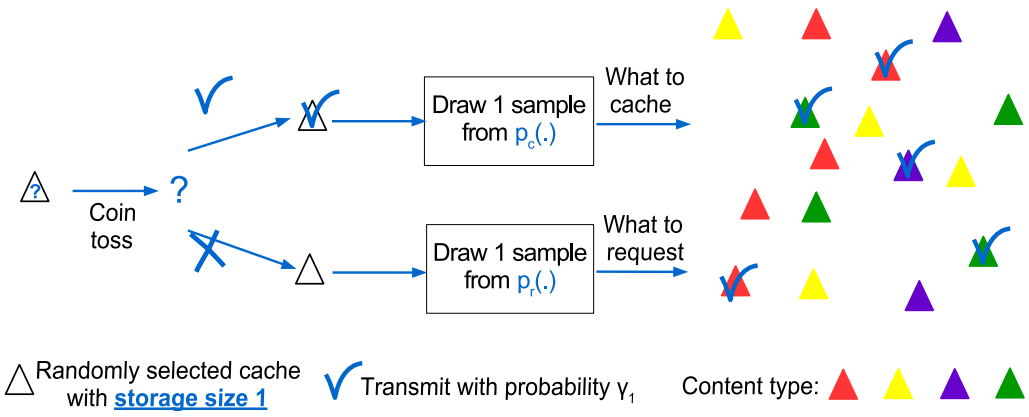


Figure 2.1: A randomized caching model, in which the placement distribution is independent and identical over the spatial domain.

the DSR, and optimize the caching pmf based on the randomized model.

Small-scale fading DSR results. We formulate an optimization problem in Sect. 2.5.1 to find the best caching distribution that maximizes the DSR under a simple transmission strategy where single file is transmitted at a time throughout the network, assuming user demands are modeled by a Zipf distribution with exponent γ_r. This scheme yields a certain fraction of users to be active at a time based on the distribution of the requests. In Sect. 2.5.2, we optimize the DSR of users for the multi-file setup, where the small-scale fading is Rayleigh distributed. We consider several special cases corresponding to 1) small but non-zero noise, 2) arbitrary noise and 3) an approximation for arbitrary noise allowing the path loss exponent $\alpha = 4$. For case 1), we show that the optimal caching strategy also has a Zipf distribution but with exponent $\gamma_c = \frac{\gamma_r}{\alpha/2+1}$ where $\alpha > 2$. For case 2), we show that the same result

holds based on an approximation of the SINR coverage justified numerically in Sect. 2.5.2. This relation implies that γ_c is smaller than γ_r, i.e., the caching distribution should be more uniform compared to the request distribution, yet more popular files should be cached at a higher number of D2D users. For case 3), we obtain a distribution similar to Benford's law (detailed in Sect. 2.5.2) that optimizes the caching pmf. We also extend our results to the "general request distributions", and show that cases 1) and 2) are also valid for Ricean and Nakagami fading distributions in Sect. 2.5.2.

In general, the optimal DSR and the optimal caching distribution might not be tractable. Therefore, assuming the request and caching probabilities are known a priori, we weight the caching pmf to provide iterative techniques to optimize the DSR under different settings. We propose caching strategies that consider maximizing the DSR of the least desired file and of all files as detailed in Sect. 2.6.2.

Maximizing the DSR **for the simultaneous multi-file model.** In Sect. 2.7, we extend our study to the simultaneous transmissions of different files and define popularity-based and global strategies. The popularity-based strategy is in favor of the transmission of popular files and discards unpopular files. On the other hand, the global strategy schedules all the files simultaneously, which leads to lower coverage than the sequential model does. Optimization of the DSR in these cases is very intricate compared to the case of sequential modeling. Therefore, we numerically compare the proposed caching models in Sect. 2.7, and observe that the optimal solutions become skewed

towards the most popular content in the network. Thus, we infer that under different models, the optimal caching distribution may not be a Zipf distribution as also found in [71–73].

Insights. Our results show that the optimal caching strategy exhibits less locality of the reference (abbreviated as locality) compared to the input stream of requests, i.e., the demand distribution[3]. We also analyze the special case of $\alpha = 4$ using a tight approximation for standard Gaussian Q-function. Using this approach we show that the optimal caching distribution can be approximated by Benford's law, which is a special bounded case of Zipf's law [84]. In Sect. 2.8, we validate that both Zipf distribution and Benford's law have very similar distributional characteristics, further validating the generality of the results. For the multiple file case, we extend our results by finding lower and upper bounds for the DSR in Sect. 2.6. Simulations show that the bounds are very accurate approximations for particular γ_r values.

2.3 System Model

We consider a mobile network model in which D2D users are spatially distributed as a homogeneous PPP Φ of density λ, where a randomly selected user can transmit or receive information. In the multiple file scenario, the

[3]The performance of demand-driven caching depends on the locality exhibited by the stream of requests. The more skewed the popularity pmf, (i) the stronger the locality and the smaller the miss rate of the cache[82], and (ii) good cache replacement strategies are expected to produce an output stream of requests exhibiting less locality than the input stream of requests [83]. In [82], authors showed that (i) and (ii) hold for caches operating under random on-demand replacement algorithms.

randomized caching model we propose is shown in Fig. 2.1. The model can be summarized as follows. At any time slot, only a fraction of the D2D users scheduled. Any user transmits with probability γ_1 and receives with probability $\gamma_2 = 1 - \gamma_1$ independently of other users. Each user has a cache with storage size 1. If it is selected as a receiver at a time slot, it draws a sample from the request distribution $p_r(\cdot)$, which is assumed to be Zipf distributed. If it is selected as transmitter at a time slot, it draws a sample from the caching distribution $p_c(\cdot)$. The selection of request distribution and the optimization of caching distribution will be detailed in Sect. 2.5. At any time slot, each receiver is scheduled based on closest transmitter association.

A system model for the D2D content distribution network with multiple files is illustrated in Fig. 2.2. For illustration purposes, different types are separated on the plot. However, transmissions of different files can occur simultaneously. For multiple file case, different from the single file case, where the D2D content distribution network is like a downlink cellular network since nearest transmitter has the content, a farther transmitter is often the one with the file required by the receiver.

General models for the multi-cell SINR using stochastic geometry were developed in [81], where the downlink coverage probability was derived as:

$$\mathrm{p_{cov}}(\mathrm{T}, \lambda, \alpha) \triangleq \mathbb{P}[\mathsf{SINR} > \mathrm{T}] = \pi\lambda \int_0^\infty e^{-\pi\lambda r \beta(\mathrm{T},\alpha) - \mu\,\mathrm{T}\,\sigma^2 r^{\alpha/2}}\,\mathrm{d}r, \qquad (2.1)$$

where $\beta(\mathrm{T}, \alpha) = \frac{2(\mu\mathrm{T})^{\frac{2}{\alpha}}}{\alpha}\mathbb{E}\big[g^{\frac{2}{\alpha}}(\Gamma(-2/\alpha, \mu\mathrm{T}g) - \Gamma(-2/\alpha))\big]$. The expectation is with respect to the interference power distribution g, the transmit power is

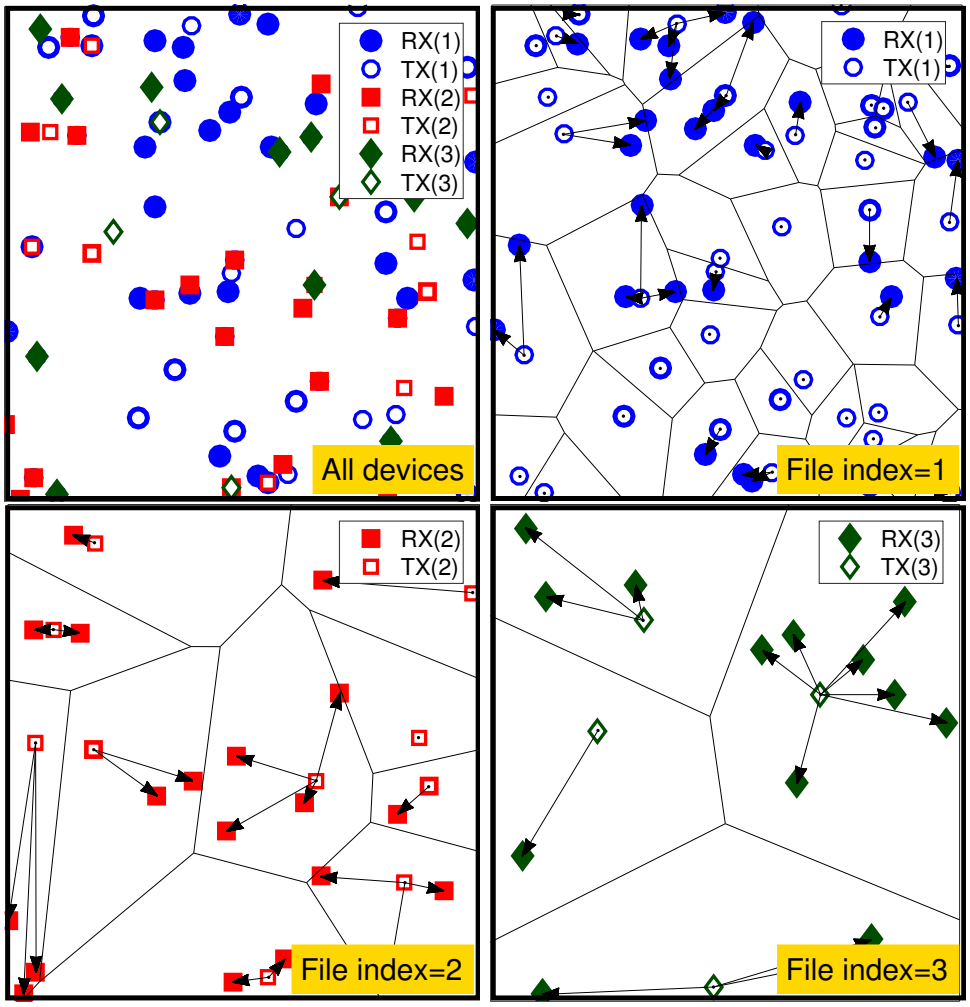


Figure 2.2: System model for D2D users with multiple files. Each receiver is associated to its closest transmitter that contains the requested file, where TX(k) and RX(k) denote the set of transmitters and receivers corresponding to file k.

$1/\mu$, and Signal-to-Noise Ratio (SNR) is defined at a distance of $r = 1$ and is $\mathsf{SNR} = 1/(\mu\sigma^2)$. A summary of the symbol definitions and important network parameters are given in Table 2.1.

Definition 1. Density of successful receptions (DSR). *The performance*

of a randomly chosen receiver is determined by its SINR *coverage. For the homogeneous* PPP Φ *with density* λ*, let* γ_1 *fraction of all users be the transmitter process* Φ_t*, and* γ_2 *fraction of users be the receiver process* Φ_r*, where* $0 < \gamma_1, \gamma_2 < 1$*. The coverage probability of a randomly chosen receiver is* $\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1, \alpha)$*, which is the same for all receivers, and the total average number of receivers is proportional to the density* $\lambda\gamma_2$*. Hence, the* DSR*, which denotes the mean number of successful receptions per unit area, equals*

$$\begin{aligned} \mathsf{DSR} &= \lambda\gamma_2\, \mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1, \alpha) \\ &= \lambda\gamma_2 \Big(\pi\lambda\gamma_1 \int_0^\infty e^{-\pi\lambda\gamma_1 r\beta(\mathrm{T},\alpha) - \mu\,\mathrm{T}\,\sigma^2 r^{\alpha/2}}\, \mathrm{d}r \Big), \end{aligned} \tag{2.2}$$

where $\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1, \alpha)$ *is obtained by combining (2.1) with the thinning property of the* PPP*, i.e.,* Φ_t*, which is obtained through the thinning of* Φ*, is a homogeneous* PPP *with density* $\lambda\gamma_1$ *[61, Ch. 1].*

We consider the generalized file caching problem in PPP networks where every user randomly requests or caches some files based on the availabilities. Our goal is to maximize the DSR in (2.2) for single file and multiple files. We discuss the details of our optimization problem in Sects. 2.4 and 2.5.

2.4 DSR For a Single File

We first assume that there is a single file in the network. The single file case is the baseline model for the more general multi-file model presented in Sect. 2.5. Sampled uniformly at random from the PPP Φ, a fraction γ_1 of the users form the process Φ_t of the users possessing the file, and a fraction

γ_2 of the users form the process Φ_r of the users who want the same file. The receivers communicate with the nearest transmitter while all other transmitters act as interferers, and each transmitter can serve multiple receivers. A receiver is in coverage when its SINR from its nearest transmitter is larger than some threshold T. Given the total density of receivers is given by $\lambda\gamma_2$, and each receiver is successfully covered with probability $\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1, \alpha)$, the DSR, i.e., DSR, is given by their product. In the single file scenario, since there is only 1 file being transmitted in the network, there is no caching pmf. Our objective in this section is to determine the optimal fractions of transmitters γ_1 and receivers γ_2 in the PPP network that maximizes the DSR. In Sect. 2.5, we consider the multiple file transmission scenario, where we use the optimal fractions of transmitters and receivers γ_1 and γ_2, respectively, derived in this section, to determine the DSR, and optimize the caching pmf based on the randomized model outlined in Sect. 2.3. We formulate the following optimization problem to determine γ_1 and γ_2:

$$\begin{aligned} \overset{*}{\mathsf{DSR}} = \max_{\gamma_1>0,\,\gamma_2>0} \quad & \lambda\gamma_2\, \mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1, \alpha) \\ \text{s.t.} \quad & \gamma_1 + \gamma_2 = a, \quad 0 < a \leq 1, \end{aligned} \tag{2.3}$$

where $\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1, \alpha)$ is the coverage probability of a typical user, and $a \leq 1$ is the total fraction of transmitting and receiving users in a PPP network Φ with density λ.

Lemma 1. *The fraction of transmitters should be less than that of receivers, i.e., the solution of (5.12) satisfies the following relation:* $\gamma_1 < a/2 < \gamma_2 < a \leq 1$.

Symbol	Definition
T; $\alpha > 2$	SINR threshold; Path loss exponent
γ_1; γ_2	Fraction of transmitting users; fraction of receiving users
Φ	Homogeneous PPP of all D2D users
Φ_t; Φ_r	PPP transmitter process; PPP receiver process
λ; λ_t	Intensity of Φ; intensity of Φ_t
μ^{-1}; σ^2	The constant transmit power; Noise variance
$g \sim \exp(\mu)$	Interference power distribution
γ_r; γ_c	Zipf request parameter; Zipf caching parameter
M; 1	Size of the file catalog; storage size of any user
$p_r(\cdot)$; $p_c(\cdot)$	Popularity pmf; caching pmf
$\mathrm{p_{cov}}(\mathrm{T}, \lambda, \alpha)$	Coverage probability for the sequential transmission model
$\mathcal{P}_{\mathrm{cov}}(\mathrm{T}, \lambda, \alpha)$	Coverage probability for the general transmission model
$\beta(\mathrm{T}, \alpha)$	A function of interference in the exponent of $\mathrm{p_{cov}}$
$F_B(\cdot)$	The pmf of the Benford's distribution
DSR	Density of successful receptions
$\mathsf{DSR_S}; \mathsf{DSR_P}; \mathsf{DSR_G}$	Sequential; popularity-based; global model DSR
Q-function	The tail probability of the standard normal distribution
$\Theta(\cdot)$; $o(\cdot)$	Big O notation; Little-o notation

Table 2.1: Notation for Chapter 2.

Proof. See Appendix A in [52]. □

Lemma 2. *The maximum* DSR *for arbitrary noise and* $\alpha = 4$ *is given by*

$$\overset{*}{\mathsf{DSR}} = \frac{\lambda(a-\gamma_1)}{\left(\frac{1}{\gamma_1}\left[\frac{1}{\gamma_1} - \frac{1}{a-\gamma_1}\right]\frac{2\mu T\sigma^2}{(\pi\lambda)^2\beta(\mathrm{T},4)} + \beta(\mathrm{T},4)\right)}.$$

Proof. See Appendix B in [52]. □

Corollary 1. Low SNR case, $\alpha = 4$**.** *As* $\sigma^2 \to \infty$*, the coverage can be approximated as* $\mathrm{p_{cov}}(\mathrm{T}, \lambda, \alpha) = \mathbb{P}[\mathsf{SINR} > \mathrm{T}] \approx \mathbb{P}[\mathsf{SNR} > \mathrm{T}] = \pi\lambda \int_0^\infty e^{-\pi\lambda r - \mu\,\mathrm{T}\,\sigma^2 r^{\alpha/2}}\,\mathrm{d}r$.

Hence, the maximum DSR *is given as*

$$\overset{*}{\mathsf{DSR}} = \lambda(a-\gamma_1)\Big/\Big(\frac{1}{\gamma_1}\left[\frac{1}{\gamma_1}-\frac{1}{a-\gamma_1}\right]\frac{2\mu T\sigma^2}{(\pi\lambda)^2}+1\Big), \tag{2.4}$$

where optimal γ_1 *satisfies* $\frac{a-3a\gamma_1+3\gamma_1^2}{\gamma_1^3(a-\gamma_1)} = \frac{(\pi\lambda)^2}{4\mu T\sigma^2}$.

Corollary 2. No noise (degenerative) case. *For no noise,* $\mathrm{p_{cov}}(\mathrm{T},\lambda,\alpha) = \beta(\mathrm{T},\alpha)^{-1}$. *Maximum* DSR *for single file for* $0 < a \leq 1$, *Rayleigh fading, no noise, and* $\alpha > 2$ *is* $\mathsf{DSR}^* = \max_{\gamma_1>0} \lambda(a-\gamma_1)\frac{1}{\beta(\mathrm{T},\alpha)} = \frac{\lambda(a-\gamma_1^*)}{\beta(\mathrm{T},\alpha)}$, *obtained for the optimal value of* γ_1, *i.e.,* $\gamma_1^* = \varepsilon > 0$ *so that there is one transmitter*[4].

Next, we consider the low noise approximation of the success probability that is more easily computable than the constant noise power expression and more accurate than the no noise approximation for $\sigma^2 = 0$. Using the expansion $\exp(-x) = 1 - x + o(x)$ for $\sigma^2 \neq 0$ as $x \to 0$, the term $\mathrm{p_{cov}}(\mathrm{T},\lambda,\alpha)$ for small but non-zero noise case can be calculated after an integration by parts of (2.1) as follows

$$\mathrm{p_{cov}}(\mathrm{T},\lambda,\alpha) = \frac{1}{\beta(\mathrm{T},\alpha)} - \frac{\mu\,\mathrm{T}\,\sigma^2\,(\lambda\pi)^{-\frac{\alpha}{2}}}{\beta(\mathrm{T},\alpha)^{\frac{\alpha}{2}+1}}\Gamma\left(1+\frac{\alpha}{2}\right)+o\left(\sigma^2\right).$$

Lemma 3. *The maximum* DSR *for a single file for* $a = 1$, *Rayleigh fading, small noise is equal to*

$$\overset{*}{\mathsf{DSR}} = \frac{\lambda\alpha}{\beta(\mathrm{T},\alpha)}\left[\frac{1}{\alpha}-\frac{(\gamma_1^*-1)}{\alpha+\gamma_1^*(2-\alpha)}o(\sigma^2)\right].$$

[4]In the no noise case the single file result is trivial. In multiple file case, there will be interference due to the simultaneous transmissions of multiple files, which will be discussed in Sect. 2.5.

Proof. See Appendix C in [52]. □

For $\alpha = 4$, there is a closed form expression for $\beta(\mathrm{T}, 4)$ as follows: $\beta(\mathrm{T}, 4) = 1 + \sqrt{\mathrm{T}} \arctan(\sqrt{\mathrm{T}})$, which we use for the derivation of Lemma 4.

Lemma 4. *The maximum* DSR *for small but non-zero noise and* $\alpha = 4$ *is*

$$\overset{*}{\mathsf{DSR}} = \frac{2\lambda(a - \gamma_1)}{(1 + \sqrt{\mathrm{T}} \arctan(\sqrt{\mathrm{T}}))} \left[1 - \frac{\mu \, \mathrm{T} \, \sigma^2 a}{\mu \, \mathrm{T} \, \sigma^2 (2a - \gamma_1) + o(\sigma^2)} \right] + o(\sigma^2). \quad (2.5)$$

Proof. See Appendix D in [52]. □

Discussion. In Fig. 2.3 (a), we illustrate the relation between DSR^* and SNR for $\mathrm{T} = \mathsf{SNR}/2$, λ=0.1. To simplify the notation, we assume that $\gamma_1 + \gamma_2 = 1$ and let $\gamma = \gamma_1$ and $\gamma_1^* = \gamma_{\text{opt}}$. As SNR increases for $\mathrm{T} = \mathsf{SNR}/2$, the DSR decreases and γ_{opt} decreases. Note that the solid lines denote the simulation results for the PPP model. In Fig. 2.3 (b), the variation of DSR^* with respect to T for $\mathsf{SNR} = 10$, λ=0.1 is shown. The coverage $\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1, \alpha)$ is monotonically decreasing in T and a concave increasing function of γ_1. For increasing T, the value of DSR becomes very small, and to maximize the DSR, a higher fraction of the users should be transmitters (i.e., higher γ_1) to compensate the outage. For low T, to maximize the DSR, the fraction of the receivers γ_2 should be higher. Therefore, as T decreases, the DSR increases and becomes right-skewed, but γ_{opt} decreases only slightly, which is negligible[5].

[5]This follows from the separability assumption of $\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1, \alpha)$ in $\lambda\gamma_1$ and T, thus insensitivity of the DSR maximization problem to the value of T, which is further detailed in Assumption 1 of Sect. 2.5.2, and verified in Appendix F in [52]

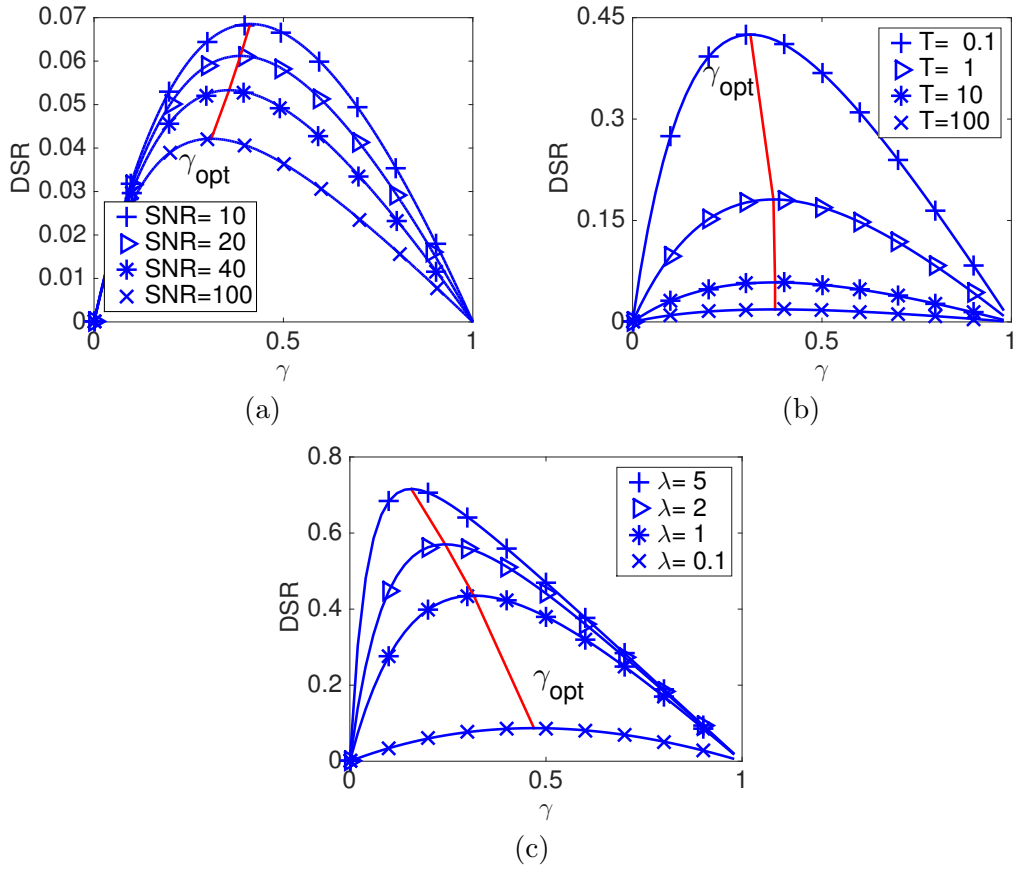


Figure 2.3: DSR for single file versus γ with respect to SNR, T and λ. (a) DSR, $\mathrm{T} = \mathsf{SNR}/2$, λ=0.1, where the dashed curves correspond to the respective Monte Carlo simulations, (b) DSR, $\mathsf{SNR} = 20$, λ=0.1, and (c) DSR, $\mathsf{SNR} = .1$, $\mathrm{T} = .05$.

Thus, we conclude that γ_{opt} is largely invariant to T and mainly determined by SNR. In Fig. 2.3 (c), we show the variation of DSR^* with λ. The DSR increases with λ. On the other hand, γ_{opt} decreases as the density of users increases and transmissions from increased number of users cause high interference.

Although the single file case is trivial in the sense that it boils down to the optimization of the fractions of the transmitters and receivers that maximizes the DSR, it is the baseline model for the multiple file case where the main objective is to determine the optimal caching distribution over the set of files. We discuss the multiple file setup next.

2.5 Density of Successful Receptions of the Sequential Serving Model with Multiple Files

We determine the optimal caching distribution for the transmitters to maximize the DSR for the sequential serving-based strategy, in which one type of file is transmitted at a time. Later, in Sect. 2.7, we study the general case, where the transmissions of different files can take place simultaneously.

File Popularity Distribution. To model the file popularity in a general PPP network, we use Zipf distribution for p_r, which is commonly used in the literature [76]. Then, the popularity of file i is given by $p_r(i) = \frac{1}{i^{\gamma_r}} \Big/ \sum_{j=1}^{M} \frac{1}{j^{\gamma_r}}$, for $i = 1, \ldots, M$, where γ_r is the Zipf exponent and there are M files in total. The demand distribution $p_r \sim$Zipf(γ_r) is the same for all receivers of the PPP model.

2.5.1 Sequential Serving-based Model

In this model, only the set of transmitters having a specific file transmits simultaneously. Hence, this is the special case where only one file is transmitted at a time network-wide. This is illustrated in Fig. 2.1 in Sect. 2.3. If a user is selected as a receiver at a time slot, it draws a sample from the request distribution $p_r(\cdot)$, which is known. If any user is randomly selected as the transmitter at a time slot with probability γ_1, it draws a sample from the caching distribution $p_c(\cdot)$, which is not known yet. At any time slot, each receiver is scheduled based on closest transmitter association. According to this model, since file i is available at each transmitter with $p_c(i)$, using the thinning property of the PPP [61, Ch. 1], the probability of coverage for file i is

$$\mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha) = \pi \lambda_t p_c(i) \int_0^\infty e^{-\pi \lambda_t p_c(i) r \beta(\mathrm{T},\alpha) - \mu\, \mathrm{T}\, \sigma^2 r^{\alpha/2}}\, \mathrm{d}r, \tag{2.6}$$

where $\lambda_t = \lambda \gamma_1$ is the total density of the transmitting users.

Given that the requests are modeled by the Zipf distribution, our objective is to maximize the DSR of users for the sequential serving-based model, denoted by $\mathsf{DSR_S}$ for a PPP model with density λ:

$$\begin{aligned} \max_{p_c} \quad & \mathsf{DSR_S} \\ \text{s.t.} \quad & \sum_{i=1}^{M} p_c(i) = 1 \\ & p_r(i) = \frac{1}{i^{\gamma_r}} \Big/ \sum_{j=1}^{M} \frac{1}{j^{\gamma_r}}, \quad i = 1, \ldots, M, \end{aligned} \tag{2.7}$$

where $\mathsf{DSR_S} = \lambda\gamma_2 \sum_{i=1}^{M} p_r(i)\, \mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1 p_c(i), \alpha)$, the first constraint is the total probability law for the caching distribution, and the second constraint is the demand distribution modeled as Zipf with exponent γ_r, and $\gamma_2 = 1 - \gamma_1$, and M is the number of files.

Note that $\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1 p_c(i), \alpha)$ in (2.7) is obtained for a sequential transmission or scheduling model and it is same as the formulation given in (2.1) which follows from Theorem 1 of [81]. This model can be generalized to different scheduling schemes. For example, in Sect. 2.7, we introduce a more general model where multiple files are simultaneously transmitted, and obtain a coverage expression $\mathcal{P}_{\mathrm{cov}}(\mathrm{T}, \cdot, \alpha)$ that is different from $\mathrm{p_{cov}}(\mathrm{T}, \cdot, \alpha)$ in (2.7), which is detailed in Theorem 2 of Sect. 2.7.

Similar to the optimal fractions of the transmitter and receiver processes calculated in Sect. 2.4 for the single file case, optimal values of γ_1 and $\gamma_2 = 1 - \gamma_1$ for multi-file case can be found by taking the derivative of (2.7) with respect to γ_1, which yields the following expression:

$$\sum_{i=1}^{M} \lambda p_r(i) p_c(i) \Big\{ \int_0^{\infty} \Big[\frac{1}{\gamma_1} - \frac{1}{1-\gamma_1} - \pi\lambda p_c(i)\beta(\mathrm{T}, \alpha) r \Big] e^{-\pi\lambda\gamma_1 p_c(i) r \beta(\mathrm{T},\alpha) - \mu \mathrm{T}\sigma^2 r^{\frac{\alpha}{2}}}\, \mathrm{d}r \Big\} = 0, \quad (2.8)$$

where optimal value of γ_1 and the pmf $p_c(\cdot)$ are coupled. Therefore, we first solve (2.7) by optimizing the pmf $p_c(\cdot)$ and then, determine the γ_1 value that satisfies (2.8).

We now investigate different special network scenarios where significant

simplification is possible.

2.5.2 Rayleigh Fading DSR Results

We optimize the DSR of users for the multi-file setup, where interference fading power follows an exponential distribution with $g \sim \exp(\mu)$. We consider several special cases corresponding to 1) small but non-zero noise, 2) arbitrary noise and 3) an approximation for arbitrary noise allowing the path loss exponent $\alpha = 4$. We find the optimal caching distribution corresponding to each scenario.

Lemma 5. Small but non-zero noise, $\alpha > 2$**.** *The optimal caching distribution is* $p_c(i) = \frac{1}{i^{\gamma_c}} \Big/ \sum_{j=1}^{M} \frac{1}{j^{\gamma_c}}$, $i = 1, \ldots, M$, *which is also Zipf distributed, where* $\gamma_c = \frac{\gamma_r}{\alpha/2+1}$ *is the Zipf exponent for the caching pmf.*

Proof. See Appendix E in [52]. □

Assuming $\alpha > 2$, the caching pmf exponent satisfies $\gamma_c < \frac{\gamma_r}{2}$, which implies that the optimal caching pmf that maximizes the DSR has a more uniform distribution exhibiting less locality of reference compared to the request distribution that is more skewed towards the most popular files.

Assumption 1. Separability of coverage distribution. *For Rayleigh, Ricean and Nakagami small-scale fading distributions, the function* $\beta(\mathrm{T}, \alpha)^{\alpha/2}$ *can be approximated as a linear function of* T *as shown in Fig. 2.4. This*

relation[6] *greatly simplifies the analysis of the optimization problem in (2.7).*

Lemma 6. Arbitrary Noise, $\alpha > 2$**.** *For arbitrary noise, from Assumption 1, the optimal caching distribution* $p_c(\cdot)$ *can be approximated as a Zipf distribution given by*

$$p_c(i) \approx \frac{1}{i^{\gamma_c}} \Big/ \sum_{j=1}^{M} \frac{1}{j^{\gamma_c}}, \quad i = 1, \ldots, M, \tag{2.9}$$

where $\gamma_c = \frac{\gamma_r}{\alpha/2+1} < \frac{\gamma_r}{2}$ *is the Zipf exponent for the caching pmf assuming* $\alpha > 2$*.*

Proof. See Appendix A.1. □

Interestingly, this result is the same as Rayleigh fading with small but non-zero noise model developed in Sect. 2.5.2, which follows from the monotonic transformation [85] caused by increasing the noise power σ^2 in (2.6). According to the pmf given in (2.9), the optimal caching strategy exhibits less locality of reference than the input stream of requests. Therefore, it is a good caching strategy, which will be further verified in Sect. 2.8. Lemma 6 suggests that files with higher popularity should be cached less frequently than the demand for this file, and unpopular files should be cached more frequently than the demand for the file. However, high popularity files should be still

[6] Although the expression $\beta(\mathrm{T}, \alpha)^{\alpha/2}/\,\mathrm{T}$ is not analytically tractable, we can approximate $\beta(\mathrm{T}, \alpha)^{\alpha/2}$ as a linear function of T because the lower incomplete Gamma function has light-tailed characteristics. Since the channel power distribution -which is exponential due to Rayleigh fading- is also light tailed, we can expect to observe such a linear approximation in our numerical results.

cached at more locations compared to the low popularity files. The path loss evens out the file popularities and the caching distribution should be more uniform compared to the request distribution. The sequential transmission model shows that for a Zipf request distribution with exponent γ_r, which is skewed towards the most popular files, the optimal caching pmf should be also Zipf distributed with the relation $\gamma_c < \frac{\gamma_r}{2}$ for $\alpha > 2$, implying that the caching pmf is more uniform than the request pmf.

The next result generalizes Lemma 6 to any request distribution $p_r(\cdot)$ rather than the Zipf distribution, and is derived from Appendix F in [52] using the separability of coverage from Assumption 1.

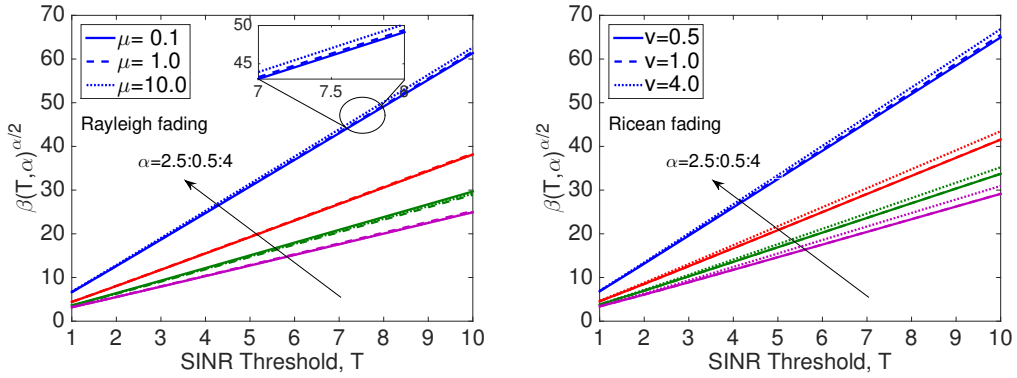
Theorem 1. *For arbitrary noise, if the small-scale fading is Rayleigh, Nakagami or Ricean distributed, from Assumption 1, for a general request pmf, $p_r(\cdot)$, the optimal caching pmf is approximated as*

$$p_c(i) \approx \frac{p_r(i)^{\frac{1}{(\alpha/2+1)}}}{\sum_{j=1}^{M} p_r(j)^{\frac{1}{(\alpha/2+1)}}},\ i = 1, \ldots, M. \tag{2.10}$$

From (2.10), it is required to flatten the request pmf to optimize the caching performance. Examples include the case of uniform demands, where the optimal caching distribution should be also uniform, and Geometric(p) request distribution, for which the caching distribution satisfies Geometric(q), where $q = 1 - (1-p)^{\frac{1}{(\alpha/2+1)}}$. In the case of Zipf demands, we can derive the same result as in Lemma 6. These example distributions are summarized in Table 2.2.

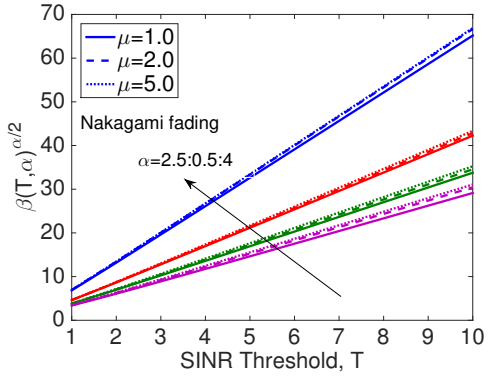
Popularity Distribution	**Caching Distribution**
Uniform	Uniform
Geometric(p)	Geometric(q), $\quad q = 1 - (1-p)^{\frac{1}{(\alpha/2+1)}}$
Zipf(γ_r)	Zipf(γ_c), $\quad \gamma_c = \frac{\gamma_r}{\alpha/2+1}$

Table 2.2: Relation between the example popularity distributions and their corresponding optimal caching distributions for Chapter 2.



(a) Rayleigh fading, param. μ.

(b) Ricean fading, param. v.



(c) Nakagami fading, shape param. μ.

Figure 2.4: The linear relation between $\beta(\mathrm{T},\alpha)^{\alpha/2}$ and T.

Lemma 7. An Approximation for Arbitrary Noise with $\alpha = 4$. *For a total number of files M and arbitrary noise with $\alpha = 4$, the optimal caching pmf is*

$$p_c(i) = a_i + b \log\left(\frac{i+1}{i}\right),\ i = 1, \ldots, M, \tag{2.11}$$

where $b = \frac{\sqrt{\mu \, \mathrm{T} \, \sigma^2 \gamma_r}}{\pi \lambda_t \beta(\mathrm{T},4)}$, $a_i = \frac{1}{M} + \frac{b}{M} \sum_{j=1}^{M} \log\left(\frac{j}{i+1}\right)$, and the pmf is valid only if $b \leq [M \log(M) - \log(M!)]^{-1}$.

Proof. See Appendix G in [52]. □

The distribution $p_c(\cdot)$ in (2.11) of Lemma 7 is a variety of Benford's law [84], which is a special bounded case of Zipf's law. Benford's law refers to the frequency distribution of digits in many real-life sources of data and is characterized by the pmf $F_B(i) = \log_{10}\left(\frac{i+1}{i}\right),\ i \in \{1, \ldots, 9\}$. In distributed caching problems, the number of files, M, is generally much greater than 9. Therefore, we generalize the law as $F_B(i) = \log_{M+1}\left(\frac{i+1}{i}\right)$, $i \in \{1, \ldots, M\}$. The result in (2.11) has a very similar form as the Benford law with shift parameter a_i for file i and a scaling parameter b, as determined in Lemma 7.

2.6 Bounds on the DSR and Different Caching Strategies

The analysis of the DSR becomes intractable for the multiple file case when the caching pdf does not have a simple form. Therefore, we derive a lower and upper bound to characterize the DSR for the sequential serving model and provide two different caching strategies to maximize $\mathsf{DSR_S}$.

2.6.1 Bounds on $\mathsf{DSR_S}$

We provide a lower and upper bound for $\mathsf{DSR_S}$, the DSR of the sequential serving-based transmission model with multiple files. We discussed the optimal file caching problem for multiple file scenarios in [69]. Here, we compare our solution to the several bounds and other caching strategies.

2.6.1.1 Upper Bound (UB)

Using the concavity of $\mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha)$ in $p_c(i)$, a UB is found as

$$\begin{aligned}\sum_{i=1}^{M} p_r(i)\, \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha) &\overset{(a)}{<} \mathrm{p_{cov}}\left(\mathrm{T}, \lambda_t \sum_{i=1}^{M} p_r(i) p_c(i), \alpha\right) \\ &\overset{(b)}{\leq} \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_r(1), \alpha), \end{aligned} \tag{2.12}$$

where (a) follows from Jensen's inequality, and (b) follows from the assumption $p_r(1) > p_r(i)$ for $1 < i \leq M$ that yields $\sum_{i=1}^{M} p_r(i)p_c(i) < p_r(1)\sum_{i=1}^{M} p_c(i) = p_r(1)$, where $p_r(1) = \left(\sum_{j=1}^{M} j^{-\gamma_r}\right)^{-1}$.

2.6.1.2 Lower Bound (LB)

Using the fact that given $p_r(\cdot)$ is Zipf distributed, the optimal $p_c(\cdot)$ also has Zipf distribution as proven in Lemma 6 as a solution of the $\mathsf{DSR_S}$ maximization problem in (2.7). As a result, any distribution that is not skewed towards the most popular files will yield a suboptimal $\mathsf{DSR_S}$. Hence a uniform caching distribution performs worse than the Zipf law, and a LB is found as

$$\sum_{i=1}^{M} p_r(i)\, \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha) \;>\; \sum_{i=1}^{M} p_r(i)\, \mathrm{p_{cov}}\left(\mathrm{T}, \frac{\lambda_t}{M}, \alpha\right)$$

$$= \mathrm{p}_{\mathrm{cov}}\left(\mathrm{T}, \frac{\lambda_t}{M}, \alpha\right). \tag{2.13}$$

2.6.2 Caching Strategies with Multiple Files

We propose two optimization formulations to maximize $\mathsf{DSR_S}$ in the presence of multiple files, where the request and caching probabilities are known a priori because in general the optimal $\mathsf{DSR_S}$ and the optimal caching distribution is not tractable. The first strategy, where we maximize the DSR for the least popular file, favors the least desired file, i.e., the file with the lowest popularity, to prevent from fading away in the network. Therefore, we introduce the variables $0 \leq \rho_i \leq 1$ for files $i \in \{1, \cdots, M\}$ to weight the caching pmf $p_c(\cdot)$. The second strategy aims to maximize the DSR of all files by optimizing the fraction ρ_i's of the users for each file type. We assume the caching distribution is given. Then, we provide iterative techniques to solve the problems presented in this section.

2.6.2.1 Maximum DSR of the Least Desired File

Our motivation behind maximizing the DSR of the least desired file is to prevent the files with low popularity from fading away in the network.

Lemma 8. *The caching probability of each file is weighted by $\rho_i < 1$ so that the total fraction of transmissions for all files, denoted by ξ satisfies $\xi = \sum_{i=1}^{M} \rho_i p_c(i) \leq 1$. Given $\eta = \max\limits_{i,\ \rho_i=1} p_r(i)p_c(i) = p_r(j)p_c(j)$ for some j, the optimal solution is given by $\rho_i = 1_{\{i \geq j\}} + \frac{\eta}{p_r(i)p_c(i)} 1_{\{1 \leq i < j\}}$.*

Proof. See Appendix H in [52]. □

2.6.2.2 Maximum DSR of All Files

We maximize the DSR for all files without any prioritization.

Lemma 9. *The optimal solution to maximize the* DSR *for all files is given by* $\rho_i = 1$ *for all* i.

Proof. See Appendix I in [52]. □

As well as maximizing the DSR for the sequential model, one might wish to select a file with a particular request probability, and use D2D to distribute this file and all files with higher probability or simultaneously cache all files using D2D as detailed in Sect. 2.7. In the next section, we describe the simultaneous transmission of multiple files, and derive expressions for SINR distribution and DSR.

2.7 Simultaneous Transmissions of Different Files

We consider the multiple file case, where a typical receiver requires a specific set of files, and the set of its transmitter candidates are the ones that contain any of the requested files. Each receiver gets the file from the closest transmitter candidate. The rest of the active transmitters that do not have the files requested are the interferers. We provide a detailed analysis for the SINR coverage next.

Assume that each receiver has a state, determined by the set of files it requests. For a receiver in state j, the set of requested files is $f_r(j)$. Let the tagged receiver be $y \in \Phi_r$ and in state j, and $\Phi_t(j)$ be the set of transmitters that a receiver in state j can get data from. Hence, the set of transmitter candidates for user in state j is the superposition given by $\Phi_t(j) = \sum_{i \in f_r(j)} \Phi_{t,i}$, where $\Phi_{t,i}$ is the set of transmitters containing file i. Let λ_j be the density of $\Phi_t(j)$, where $\lambda_j = \lambda_t p_j = \lambda \gamma_1 p_j$. The rest of the transmitters, i.e., $\sum_{i \notin f_r(j)} \Phi_{t,i}$, is an independent process with density $\lambda_t - \lambda_j = \lambda_t(1 - p_j) = \lambda\gamma_1(1 - p_j)$.

The sum $p_j = \sum_{i \in f_r(j)} p_c(i)$ gives the probability that the user has at least one of the files requested by any receiver in state j. Hence, the density of the transmitter candidates λ_j for a receiver in state j equals the product of $\lambda\gamma_1$ and $\sum_{i \in f_r(j)} p_c(i)$, i.e., $\lambda_j = \lambda_t p_j = \lambda\gamma_1 \sum_{i \in f_r(j)} p_c(i)$. Using the nearest neighbor distribution of the typical receiver in state j, the distance to its nearest transmitter is distributed as $\mathsf{Rayleigh}(\sigma_j) \sim \frac{r}{\sigma_j^2} \exp\left(-\frac{r^2}{2\sigma_j^2}\right)$, for $\sigma_j = \frac{1}{\sqrt{2\pi\lambda_j}}$ and $r \geq 0$.

We assume that all users experience Rayleigh fading with mean 1, and constant transmit power of $1/\mu$. Assuming user y is at o, in state j and is a receiver, and x is the tagged transmitter denoted by b_o, and the distance between them is r, then the SINR at user y is $\mathsf{SINR}_j = \frac{hr^{-\alpha}}{\sigma^2 + I_{r(j)}}$, where h is the channel gain parameter between x and y, σ^2 is the white Gaussian noise, and $I_{r(j)}$ is the total interference at node y in state j, and given by the following expression: $I_{r(j)} = \sum_{z \in \Phi_t \setminus b_o} g_z r_z^{-\alpha} = \sum_{z \in \Phi_t(j) \setminus b_o} g_z r_z^{-\alpha} + \sum_{z \in \Phi_t \setminus \Phi_t(j)} g_z r_z^{-\alpha}$, where g_z is the channel gain from the interferer z and the receiver y, r_z is the

interferer z to receiver distance, on RHS, the first term is the interference due to the set of transmitters that has the files requested by the receiver, and the second term is the interference due to the rest of the transmitters that do not have any of the desired files by the receiver. The total interference depends on the transmission scheme. Compared to the nearest user association [81], it is hard to characterize the interference in dynamic caching models with different association techniques.

Theorem 2. *The probability of coverage of a typical user conditioned on being at state j is given by*[7]

$$\mathcal{P}_{\text{cov}}(\mathrm{T}, \lambda_j, \alpha) = \pi\lambda_j \int_0^\infty e^{-\pi\lambda_j v(1-\rho_2(\mathrm{T},\alpha))} \times e^{-\pi\lambda_t v(\rho_1(\mathrm{T},\alpha)+\rho_2(\mathrm{T},\alpha))} e^{-\mathrm{T}\sigma^2 v^{\alpha/2}} \,\mathrm{d}v, \quad (2.14)$$

where

$$\rho_1(\mathrm{T}, \alpha) = \mathrm{T}^{2/\alpha} \int_{\mathrm{T}^{-2/\alpha}}^\infty \frac{1}{1+u^{\alpha/2}} \,\mathrm{d}u,$$
$$\rho_2(\mathrm{T}, \alpha) = \mathrm{T}^{2/\alpha} \int_0^{\mathrm{T}^{-2/\alpha}} \frac{1}{1+u^{\alpha/2}} \,\mathrm{d}u.$$

Proof. See Appendix J in [52]. □

We now consider the special case of the path loss exponent $\alpha = 4$, which is more tractable.

[7] The definition of $\mathcal{P}_{\text{cov}}(\mathrm{T}, \lambda_j, \alpha)$ here is different from the definition of the classical downlink coverage probability $\mathrm{p}_{\text{cov}}(\mathrm{T}, \lambda, \alpha)$ given in (2.1) due to the possibility of simultaneous transmissions of different file types.

Corollary 3. *The probability of coverage of a typical user conditioned on being at state j for the special case of $\alpha = 4$ and $\mu = 1$ is given by*

$$\mathcal{P}_{\text{cov}}(\mathrm{T}, \lambda_j, 4) = \pi \lambda_t p_j \sqrt{\frac{\pi}{\mathrm{T}\,\sigma^2}} e^{\frac{H(\mathrm{T}, \lambda_t, p_j)^2}{2}} \mathrm{Q}\left(H(\mathrm{T}, \lambda_t, p_j)\right), \tag{2.15}$$

where we let

$$H(\mathrm{T}, \lambda_t, p_j) = \Big(\frac{p_j}{\sqrt{\mathrm{T}}} - p_j \tan^{-1}\left(\frac{1}{\sqrt{\mathrm{T}}}\right) + \frac{\pi}{2}\Big) \frac{\pi \lambda_t}{\sqrt{2\sigma^2}}.$$

Proof. See Appendix K in [52]. □

Since the term $\sqrt{\mathrm{T}} \tan^{-1}\left(\frac{1}{\sqrt{\mathrm{T}}}\right)$ is increasing in T and converges to 1 in the limit as T goes to infinity, $H(\mathrm{T}, \lambda_t, \cdot)$ is increasing in p_j, and positive. Furthermore, $\mathcal{P}_{\text{cov}}(\mathrm{T}, \lambda_j, \alpha)$ is monotonically increasing in p_j. This observation is essential in the characterization of the DSR under different user criteria.

We consider two different strategies for the simultaneous transmission of multiple files, namely popularity-based and global models, which differ mainly in the set of files cached at the transmitters.

2.7.1 **Popularity-based** DSR

In this approach, a set of files corresponding to the most popular ones in the network is cached simultaneously at all transmitters. We define $\mathsf{DSR_P}$, which stands for the DSR of the popularity-based approach, and is calculated over the set of most popular files as

$$\mathsf{DSR_P} = \lambda \gamma_2 \sum_{k \in \mathcal{K}} p_r(k)\, \mathcal{P}_{\text{cov}}(\mathrm{T}, \xi_l, \alpha), \tag{2.16}$$

where $\mathcal{K}$ is the set of the K most popular files, and $\xi_l = \lambda\gamma_1 \sum_{i\in\mathcal{L}} p_c(i)$, where $\mathcal{L}$ is a set corresponding to the most popular K files cached at the transmitters among the set of available files in the caches.

Consider the special case of (2.16), where only the most popular file in the network is cached at all the transmitters if available, i.e., $|\mathcal{K}| = 1$, which modifies (2.16) as

$$\begin{aligned}\mathsf{DSR_P} &= \lambda\gamma_2 p_r(k)\, \mathcal{P}_{\text{cov}}(\mathrm{T}, \lambda\gamma_1 p_c(k), \alpha) \\ &\overset{(a)}{=} \lambda\gamma_2 p_r(k)\, \mathrm{p}_{\text{cov}}(\mathrm{T}, \lambda\gamma_1 p_c(k), \alpha),\end{aligned}$$

where (a) follows from the fact that for $|\mathcal{K}| = 1$, the coverage probability becomes same as the sequential serving-based model in Sect. 2.5, and the most popular file index k can be found from the demand distribution and is given by $k = \underset{i\in\{1,\ldots,M\}}{\arg\max}\ p_r(i)$, and hence the corresponding density of the transmitters is $\lambda\gamma_1 p_c(k)$, where $p_r(k) \geq p_r(l)$ for all $l = 1, \ldots, M$.

2.7.2 **Global** DSR

Global DSR is defined as the average performance of all users in the network, which is determined by the spatial characteristics of file distributions and the coverage of a typical user. The DSR function in our model is state dependent since the coverage probability of a user is determined according to the files requested by the user. The expected global DSR is given as follows:

$$\mathsf{DSR_G} = \lambda\gamma_2 \sum_{i=1}^{M} p_r(i)\, \mathcal{P}_{\text{cov}}(\mathrm{T}, \gamma_1\lambda p_c(i), \alpha). \tag{2.17}$$

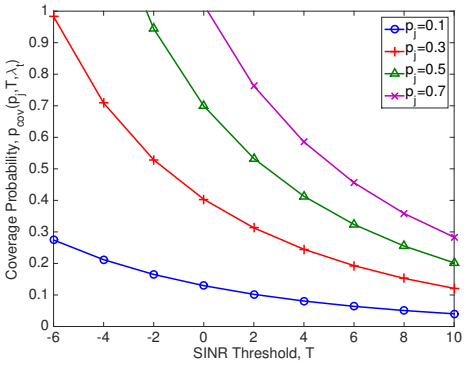


Figure 2.5: Analytical model for the SINR coverage probability for different transmitter densities, $\lambda = 1$ and $\gamma_1 = 0.4$.

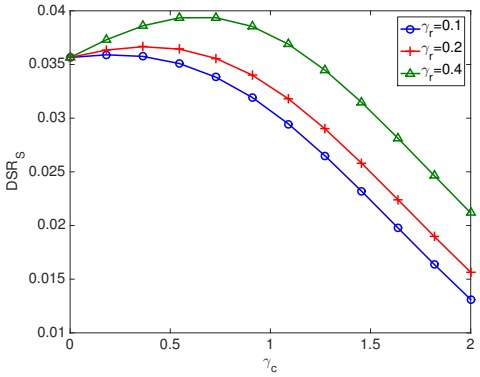


Figure 2.6: Average DSR for the sequential model, $\mathsf{DSR_S}$ versus γ_c for Zipf request and Zipf caching distributions.

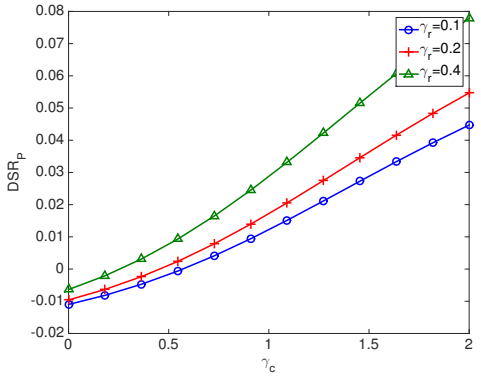


Figure 2.7: Average DSR for the popularity-based model, $\mathsf{DSR_P}$ versus γ_c.

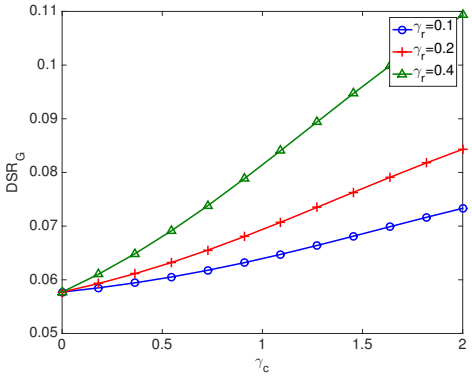


Figure 2.8: Average DSR for the global model, $\mathsf{DSR_G}$ versus γ_c.

A Discussion on the Various Transmission Models. Popularity-based transmission and global model in this section do not depend on the cache states. Instead, they both depend on the global file popularity distributions, and have similar characteristics as given in (2.16) and (2.17). It is intuitive to observe that the optimal caching distributions in both models follow similar

trends as the request distribution. Sequential serving-based model in Sect. 2.5.1 boils down to the scenario characterized in [81] where only a subset of transmitters and their candidate receivers are active simultaneously. Hence, this model mitigates interference and provides higher coverage than the other models. However, since the DSR is a weighted function of the file transmit pmf $p_c(\cdot)$, the DSR of the model is reduced.

Now, we present some numerical results on the general transmission models discussed and present results related to the popularity-based DSR, global DSR and sequential DSR.

State dependent coverage probability. We illustrate the SINR coverage probability for varying p_j for a fixed fraction of transmitters ($\gamma_1 = 0.4$) in Fig. 2.5. The coverage probability is state dependent[8] and for the receiver in state j, the density of transmitters is given by $\lambda_j = \lambda p_j$ where $p_j = \gamma_1 \sum_{i \in f_r(j)} p_c(i)$. If the requested files are available in the set of transmitters, then the receiver has higher coverage. Therefore, for higher fraction of transmitters γ_1, the coverage probability is higher.

Caching performance of the proposed transmission models. The optimal caching strategies that maximize the caching problems of Sect. 2.7 given in (2.16) and (2.17) are not necessarily Zipf distributed. However, without the Zipf distribution assumption, the optimization formulations become intractable since $\mathrm{p_{cov}}(\mathrm{T}, \lambda_j, \alpha)$ in (2.14) is nonlinear in the density of the

[8]The receiver's state refers to the collection of files it requests.

users. Therefore, for simulation purposes, we find the optimal Zipf caching exponents that maximize the proposed functions.

DSR **comparison.** We investigate the variation of the sequential model $\mathsf{DSR_S}$ with respect to the caching parameter γ_c. From Fig. 2.6, we observe that γ_c increases with the request distribution parameter γ_r, assuming both distributions are Zipf. In Figs. 2.7 and 2.8, we illustrate the variation of the popularity-based model $\mathsf{DSR_P}$ and the global model $\mathsf{DSR_G}$ with γ_c. In both figures, it is clearly seen that as the requests become more skewed (higher γ_r), the DSR increases. It also increases with γ_c, which implies that the optimal caching distribution should also be skewed towards the highly popular files.

2.8 Numerical Results and Discussion

We evaluate the optimal caching distributions that maximize the DSR. The simulation results are based on Sects. 2.5 and 2.6. We consider a general PPP network model with Rayleigh fading distribution with $\mu = 1$ and $\alpha = 4$ for small and general noise solutions. The requests are modeled by Zipf(γ_r).

Benford versus Zipf distributions. In Figs. 2.9 and 2.10, we illustrate the trend of optimal Zipf caching distribution and the Benford law developed in Sect. 2.5 for different numbers of total files. As seen from Fig. 2.9, these two distributions have similar characteristics. However, as γ_r increases, the range of M for which Benford caching distribution in (2.11) and Zipf laws are comparable becomes narrower. For $\gamma_r > 0.3$, it is not practical to approximate the Benford law with a Zipf distribution. In fact, as described

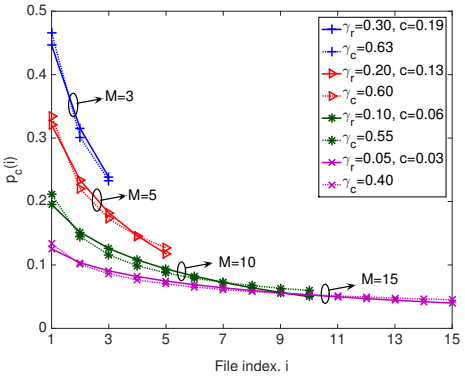


Figure 2.9: For a Zipf(γ_r) popularity distribution, Benford law and approximate Zipf(γ_c) caching pmf for various M.

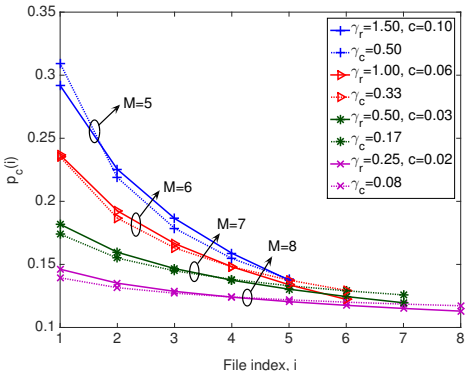


Figure 2.10: For a Zipf(γ_r) popularity distribution, Benford law and optimal Zipf(γ_c) pmf ($\mathsf{SNR} = 30$ dB).

in Sect. 2.5, as the noise level decreases, i.e., $b = \sqrt{\mu \, \mathrm{T} \, \sigma^2}\gamma_r/(\pi\lambda_t\beta(\mathrm{T},4))$ drops, the optimal caching strategy converges to Zipf distribution. As seen in Fig. 2.10, for small noise, i.e., for high SNR, these laws behave similarly for relatively high γ_r values compared to the general noise case.

We now compare the DSR of the sequential serving model for various γ_r based on the optimal solutions that are also Zipf distributed, as derived in Sect. 2.5, and the lower and upper bounds obtained in Sect. 2.6. The numerical solutions are obtained by calculating the DSR of various (random) caching pmfs and picking the best one that achieves the highest DSR.

Zipf caching with $\gamma_c = \frac{\gamma_r}{(\alpha/2+1)}$ is a good approximation to maximize the DSR**.** In Fig. 2.11, we compare the performances of different caching strategies for a Zipf request distribution with parameter $\gamma_r = 0.5$ and $\mathsf{SNR} = 1$. The Zipf caching distribution with parameter $\gamma_r/3$ is very close to the optimal solution evaluated numerically that is also very close to the sim-

ple lower bound derived in (2.13). Furthermore, Benford distribution has very similar characteristics as the optimal caching distribution solution. There is a huge gap between the UB and the no noise in terms of the DSR, and the DSR for the no noise case is the highest among all for all SNR or T values.

LB and UB get closer together as the SNR increases. In Fig. 2.12, we compare the performance of the caching distributions for a Zipf request pmf with parameter $\gamma_r = 0.5$ and $\mathsf{SNR} = 10$. At high SNR, the UB and LB are closer. Still, the numerical solution and the Zipf caching pmf with parameter $\gamma_r/3$ give similar densities of successful communication, which is very close to the lower bound because for that choice of γ_r, the request distribution converges to a uniform distribution. Benford caching distribution does not perform as well as the Zipf caching distribution, and is even worse than the LB. In Fig. 2.13, where $\gamma_r = 2$ and $\mathsf{SNR} = 1$, the Zipf caching pmf with parameter $\gamma_r/3$ does not have the same performance as the optimal solution evaluated numerically. Benford distribution has also similar performance as the Zipf caching pmf. In Fig. 2.14, we also show that Zipf caching pmf and Benford distributions have similar performance as the numerical solution for $\gamma_r = 2$ and $\mathsf{SNR} = 10$.

Transmit Diversity. In the sequential serving model, where only one file is transmitted at a time network-wide, as discussed in Sect. 2.5, using a transmitter diversity scheme will improve the DSR. For the second scenario presented in Sect. 2.7, in which different files are transmitted simultaneously, a similar diversity scheme can be applied instead of treating the other trans-

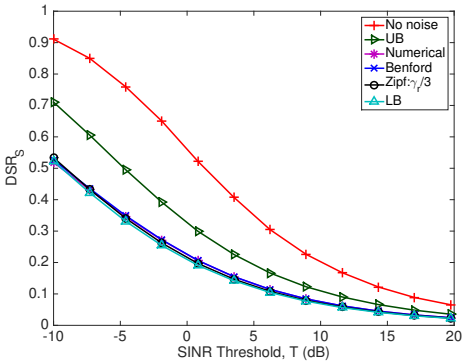


Figure 2.11: Bounds and approximations to the optimal $\mathsf{DSR_S}$ for $M = 10$, $\mathsf{SNR} = 1, \lambda = 1$, Zipf request pmf with $\gamma_r = 0.5$.

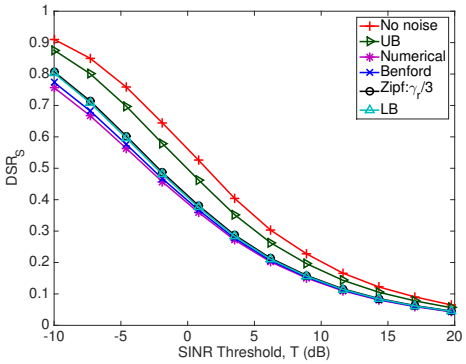


Figure 2.12: Bounds and approximations to the optimal $\mathsf{DSR_S}$ for $M = 10$, $\mathsf{SNR} = 10, \lambda = 1$, Zipf request pmf with $\gamma_r = 0.5$.

mitters as interferers. Diversity combining techniques include the maximal-ratio combining (e.g., of the k closest transmitters [80]), where the received signals are weighted with respect to their SINR and then summed, equal-gain combining, where all the received signals are summed coherently, i.e., the shot-noise model [61, Ch. 2], and the selection combining, which is based on the strongest $\mathsf{D2D}$ user association, in which the received signal power (e.g., from the k strongest users [80]) is considered.

Although diversity can decrease the outage probability, how to achieve this in practice is a critical issue. Diversity would seem to require synchronization of all transmitting devices at the physical layer unless higher layer coding is used, which might not be very practical for content distribution. Assuming full synchronization provides an upper bound on what could be achieved, due to space constraints, we leave such analysis to future work.

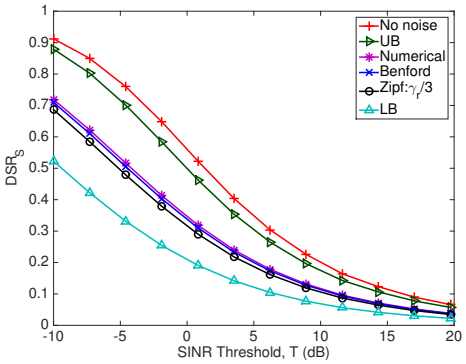


Figure 2.13: Bounds and approximations to the optimal $\mathsf{DSR_S}$ for $M = 10$, $\mathsf{SNR} = 1, \lambda = 1$, Zipf request pmf with $\gamma_r = 2$.

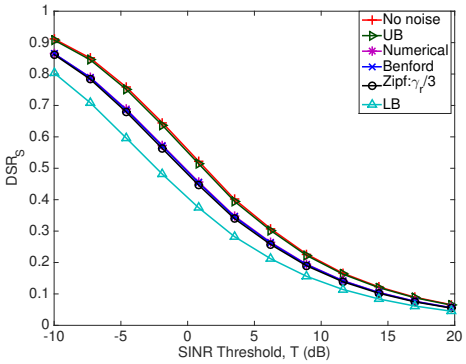


Figure 2.14: Bounds and approximations to the optimal $\mathsf{DSR_S}$ for $M = 10$, $\mathsf{SNR} = 10, \lambda = 1$, Zipf request pmf with $\gamma_r = 2$.

2.9 Summary

Content distribution using direct D2D communications is a promising approach for optimizing the utilization of air-interface resources in 5G network. This work is the first attempt to derive closed form expressions for the optimal content caching distribution and the optimal caching strategies providing maximum DSR in terms of the optimal fractions of transmitters and receivers in a D2D network by using a homogeneous PPP model with realistic noise, interference and Rayleigh fading. We derive the SINR coverage for different transmission strategies in D2D networks with some idealized modeling aspects, i.e., simultaneous scheduling of the users containing the same type of files and Zipf distributed content caching assumption for the general multi-file transmissions. Our results for the sequential transmission model show that the optimal caching pmf can also be modeled using the Zipf law and its exponent

γ_c is related to γ_r through a simple expression involving the path loss exponent: $\gamma_c = \frac{\gamma_r}{(\alpha/2+1)}$. The optimal content placement for more general demand profiles under Rayleigh, Ricean and Nakagami fading distributions suggests to flatten the request distribution to optimize the caching performance.

The limitations of the model can be overcome by investigating the optimal caching distributions that maximize the DSR for the more general transmission settings incorporating the transmit diversity, and developing intelligent scheduling techniques, which are left as future work. The dynamic settings capturing the changes in the file popularities over time and the interference caused by simultaneous transmissions should also be considered. Future issues include the minimization of backhaul transmissions and BS overhead to optimize resource utilization through D2D collaboration. Future work could also include the design of distributed caching strategies to maximize the hit probability for users by using an SINR coverage model or a distance-based coverage process given the limited range of D2D.

Chapter 3

Spatially Correlated Content Caching for Device-to-Device Communications

D2D communication is a promising technique for enabling proximity-based applications and increased offloading from the heavily loaded cellular network, and is being actively standardized by 3GPP [10]. The efficacy of D2D caching networks relies on users possessing content that a nearby user wants. Therefore, intelligent caching of popular files is critical for D2D to be successful[1]. Caching has been shown to provide increased spectral reuse and throughput gain in D2D-enabled networks [13], and the optimal way to cache content is studied from different perspectives, e.g. using probabilistic placement [60], maximizing cache-aided spatial throughput [88], but several aspects of optimal caching exploiting spatial correlations for network settings have not been explored. Intuitively, given a finite amount of storage at each node, popular content should be seeded into the network in a way that maximizes the *hit probability* that a given D2D device can find a desired file – selected at random according to a request distribution – within its radio range. We explore this

[1]This chapter has been published in [86], [87]. I am the primary author of these works. Coauthor Dr. Mazin Al-Shalash has provided many valuable discussions and insights to this work, and Dr. Jeffrey G. Andrews is my supervisor.

problem quantitatively in this chapter by considering different spatial content models and deriving, optimizing and comparing the hit probabilities for each of them.

Content caching has received significant attention as a means of improving the throughput and latency of networks without requiring additional bandwidth or other technological improvements. Video caching appears particularly profitable and plausible compared to other types of content [1], [41], and is perfectly suited to D2D networks for offloading traffic from congested cellular networks.

3.1 Related Work and Motivation

Research to date on content caching has been mainly focused on two different perspectives. On one hand, researchers have attempted to understand the fundamental limits of caching gain. The gain offered by local caching and broadcasting is characterized in the landmark paper [56]. Although this work does not deal with D2D communications and the caches cannot cooperate, it provides the first attempt to characterize the gain offered by local caching. Scaling of the number of active D2D links and optimal collaboration distance with D2D caching are studied in [57], [76]. Combining random independent caching with short-range D2D communications can significantly improve the throughput [71]. Capacity scaling laws in wireless ad hoc networks are investigated in [77], featuring short link distances, and cooperative schemes for order optimal throughput scaling is proposed in [89]. Capacity scaling laws

for single [71], [56] and multi-hop caching networks [73] are also investigated. Physical layer caching is studied in [90] to mitigate the interference, and in [91] to achieve linear capacity scaling. Finite-length analysis of random caching schemes that achieve multiplicative caching gain is presented in [92], [93].

Alternatively, as in the current chapter, there are several studies focusing on decentralized caching algorithms that have optimized the caching distribution to maximize the cache hit probability, using deterministic or random caching as in [58], [57] given a base station (BS)-user topology. FemtoCaching replaces backhaul capacity with storage capacity at the small cell access points, i.e., helpers, and the optimum way of assigning files to the helpers is analyzed in [59] to minimize the delay. There are also geographic placement models focusing on finding the cache locally such as [60], in which the cache hit probability is maximized for SINR, Boolean and overlaid network coverage models, and [52], in which the density of successful receptions is maximized using probabilistic placement. Although most of these strategies suggest that the caching distribution should be skewed towards the most popular content and exploit the diversity of content, and it is not usually optimal to cache just the most popular files, as pointed out in [57], [76]. Further, as this chapter will show, unlike the probabilistic policies, where the files are independently placed in the cache memories of different nodes according to the same distribution [60], [63], and [52]; it is not usually optimal to cache files independently. For larger transmission range and higher network density, we will quantify and see that the hit-maximizing caching strategy can be increasingly skewed away from

independently caching the popular files.

Recent studies also address problems at the intersection of the hit probability and the spatial throughput. The spatial throughput in D2D networks is optimized by suitably adjusting the proportion of active devices in [94]. Exploiting stochastic geometry, a Poisson cluster model is proposed in [95] and the area spectral efficiency is maximized assuming that the desired content is available inside the same cluster as the typical device. Some of the existing work focuses on mitigating excessive interference to maximize the throughput or capacity, as in [91], [13], [90]. Employing probabilistic caching, cache-aided throughput, which measures the density of successfully served requests by local device caches, is investigated in [88]. The optimal caching probabilities obtained by cache-aided throughput optimization provide throughput gain, particularly in dense user environments compared with the cache-hit-optimal case.

Challenges for the adoption of caching for wireless access networks also include making timely estimates of varying content popularity [96]. Cache update algorithms exploiting the temporal locality of the content have been well studied [97]. Inspired from the Least Recently Used (LRU) replacement principle, a multi-coverage caching policy at the edge-nodes is proposed in [98], where caches are updated in a way that provides content diversity to users who are covered by more than one node. Although [98] combines the temporal and spatial aspects of caching and approaches the performance of centralized policies, it is restricted to the LRU principle.

3.2 Contributions and A High Level Summary

We consider a spatial D2D network setting in which the D2D user locations are modeled by a Poisson point process (PPP), and users have limited communication range and finite storage. The D2D users are served by each other if the desired content is cached at a user within its radio range: this is called a *hit*. Otherwise, they are served by the cellular network base station, which is what D2D communication aims to avoid.

We concentrate exclusively on the content placement phase in the above setting in order to maximize the cache hit probability via exploiting the spatial diversity. We do not focus on the transmission phase that incorporates the path loss, fading or interference. The coverage process of the proposed scheme is represented by a Boolean model (BM). The BM is tractable for the noise-limited regime [60], where the interference is small compared to the noise. The coverage area of the BM is determined by a *fixed communication radius*, as will be detailed in Sect. 3.3.

Spatial caching, pairwise interactions and Matérn hard-core-inspired placement. We introduce a spatial content distribution model for a D2D network, and describe the cache hit probability maximization problem in Sect. 3.3. Our aim is to extend the independent content placement strategy, also known as geographic content placement (GCP) [60], where there is no spatial correlation in placement, which we discuss in Sect. 3.4. We propose a spatially exchangeable content placement technique to prioritize the caches for content placement, which is detailed in Sect. 3.5. Exchangeable placement ac-

tually performs worse than the baseline independent content placement. Next, exploiting the Matérn hard-core (MHC) models, we propose novel spatially correlated cache placement strategies that enable spatial diversity to maximize the D2D cache hit probability. In Sect. 3.6, we detail the MHC placement and analyze two different MHC placement strategies: (i) HCP-A that can provide a significantly higher cache hit probability than the GCP scheme in the small cache size regime and (ii) HCP-B that has a higher hit probability than GCP for short ranges.

The key differences from the independent placement model. The device locations follow the PPP distribution, which provides a random deployment instead of a fixed pattern, and hence it is possible to have cache clusters and isolated caches [81], and the content placement distribution is optimized accordingly. Unlike the independent placement model, where the cache placement distribution is independent and identically distributed (i.i.d.) over the spatial domain, the MHC model captures the pairwise interactions between the D2D nodes and yields a negatively correlated placement. The caches storing a particular file are never closer to each other than some given distance, called the exclusion radius, meaning that neighboring users are not likely to cache redundant content. Hence, the radius of exclusion plays the role of a substitute for caching probability.

Comparisons and design insights. Sect. 3.7 provides a simulation study to compare the performance between the different content placement strategies. Independent content placement does not exploit D2D interactions

at the network level, and our results show that geographic placement should exploit locality of content, which is possible through negatively correlated placement. For short range communication and small cache sizes, HCP is preferred, and when the network intensity is fixed, the cache hit rate gain of the HCP model over the GCP and caching most popular content schemes can reach up to 37% and 50%, respectively when the communication range is improved, as demonstrated in Sect. 3.7.

3.3 System Model and Problem Formulation

The locations of the D2D users are modeled by a PPP Φ with density λ_{t} as in [28]. We assume that there are M total files in the network, where all files have the same size, and each user has the same cache size $N < M$. Depending on its cache state, each user makes requests for new files based on a general popularity distribution over the set of the files. The popularity of such requests is modeled by the Zipf distribution, which has probability mass function (pmf) $p_r(n) = \frac{1}{n^{\gamma_r}} / \sum_{m=1}^{M} \frac{1}{m^{\gamma_r}}$, for $n = 1, \ldots, M$, where γ_r is the Zipf exponent that determines the skewness of the distribution. The demand profile is Independent Reference Model (IRM), i.e., the standard synthetic traffic model in which the request distribution does not change over time. Our objective is to maximize the average cache hit probability performance of the proposed caching model. Therefore, it is sufficient to consider a snapshot of the

network[2], in which the D2D user realization is given and requests are i.i.d. over the space. We devise a spatially correlated probabilistic placement policy, in which the D2D caches are loaded in a distributed manner via additional marks attached to them without accounting for any cost, in a timescale that is much shorter than the time over which the locations are predicted, as will be detailed in Sect. 3.6.

Consider a given realization $\phi = \{x_i\} \subset \mathbb{R}^2$ of the PPP Φ. The coverage process of the proposed model can be represented by a Boolean model (BM) [61, Ch. 3]. Specifically, given a transmit power P, if we only consider path loss (with exponent α), no fading and no interference, the received signal at the boundary should be larger than a threshold to guarantee coverage, i.e., $Pr^{-\alpha} \geq T$, yielding $r \leq \mathrm{R}_{\mathsf{D2D}} = (P/T)^{\alpha}$. Hence, D2D users can only communicate within a finite range, which we call the D2D radius, denoted by $\mathrm{R}_{\mathsf{D2D}}$. A file request is fulfilled by the D2D users within $\mathrm{R}_{\mathsf{D2D}}$ if one has the file; else the D2D user is served by a BS.

The BM is driven by the independently marked PPP on $\mathbb{R}^2$ $\tilde{\Phi} = \sum_i \delta_{(x_i, B_i(\mathrm{R}_{\mathsf{D2D}}))}$, whose points x_i's denote the germs, and on disc-shaped grains $B_i(\mathrm{R}_{\mathsf{D2D}})$ – a closed ball of fixed radius $\mathrm{R}_{\mathsf{D2D}}$ centered at x_i – that model the

[2]Extension of the model to also incorporate the temporal correlation of real traffic traces can be done by exploiting models like the Shot-Noise Model (SNM). This overcomes the limitations of the IRM by explicitly accounting for the temporal locality in requests for contents [62]. However, in that case, the problem under study will have an additional dimension to optimize over, and to do so, online learning algorithms should be developed to both learn the demand and optimize the spatial placement. The study of the temporal dynamics of the request distribution and the content transmission phase is left as future work.

coverage regions of germs. The BM is a tractable model for the noise-limited regime [60]. The coverage process of the D2D transmitters driven by the BM is given by the union $V_{\mathrm{BM}} = \bigcup_i (x_i + B_0(\mathrm{R}_{\mathsf{D2D}}))$ [61, Ch. 3]. For the interference-limited regime, there is no notion of communication radius, and the analysis of the coverage becomes more involved. SINR coverage models as in [60] can be exploited to determine the distribution of the coverage number, i.e., the number of D2D users covering the typical receiver. However, this is beyond the scope of the current chapter.

To characterize the successful transmission probability, one needs to know the number of users that a typical node can connect to, i.e., the coverage number. Exploiting the properties of the PPP, the distribution of the number of transmitters covering the typical receiver is given by $\mathcal{N}_P \sim \mathsf{Poisson}(\lambda_{\mathrm{t}}\, \pi \mathrm{R}_{\mathsf{D2D}}^2)$. Therefore,

$$\mathbb{P}(\mathcal{N}_P = k) = e^{-\lambda_{\mathrm{t}}\, \pi \mathrm{R}_{\mathsf{D2D}}^2} \frac{(\lambda_{\mathrm{t}}\, \pi \mathrm{R}_{\mathsf{D2D}}^2)^k}{k!}, \quad k \geq 0. \tag{3.1}$$

3.3.1 Cache Hit Probability

Assume that the cache placement at the D2D users is done in a dependent manner. Given $\mathcal{N}_P = k$ transmitters cover the typical receiver, let Y_{m_i} be the indicator random variable that takes the value 1 if file m is available in the cache located at $x_i \in \phi$ and 0 otherwise. Thus, the caching probability of file m in cache i is given by $\mathrm{p}_{\mathrm{c,X}}(m, x_i) = \mathbb{P}(Y_{m_i} = 1)$. Optimal content placement is a binary problem where the cache placement constraint $\sum_{m=1}^{M} Y_{m_i} \leq N$ is satisfied for all $x_i \in \phi$, i.e., Y_{m_i}'s are inherently dependent. However, the

original problem is combinatorial and is NP-hard. For tractability reasons, we take the expectation of this relation and obtain our relaxed cache placement constraint: $\sum_{m=1}^{M} \mathrm{p_{c,X}}(m, x_i) \leq N$. Later, we show there are feasible solutions to the relaxed problem filling up all the cache slots.

The maximum average total cache hit probability, i.e., the probability that the typical user finds the content in one of the D2D users it is covered by, for a content placement strategy X can be evaluated by solving the following optimization formulation:

$$\begin{aligned} \max_{\mathrm{p_{c,X}}} \quad & \mathrm{P_{Hit,X}} \\ \text{s.t.} \quad & \sum_{m=1}^{M} \mathrm{p_{c,X}}(m, x_i) \leq N, \quad x_i \in \Phi, \end{aligned} \tag{3.2}$$

where the hit probability is given by the following expression:

$$\mathrm{P_{Hit,X}} = 1 - \sum_{m=1}^{M} p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_X = k)\, \mathrm{P_{Miss,X}}(m, k), \tag{3.3}$$

where $\mathbb{P}(\mathcal{N}_X = k)$ is the probability that k transmitters (caches) cover the typical receiver, and $\mathrm{P_{Miss,X}}(m, k)$ is the probability that k caches cover a receiver, and none has file m.

We propose different strategies to serve the D2D requests that maximize the cache hit probability. Assuming a transmitter receives one request at a time and multiple transmitters can potentially serve a request, the selection of an active transmitter depends on the caching strategy. A summary of the symbol definitions and important network parameters are given in Table 4.1.

3.3.2 Repulsive Content Placement Design

Optimizing the marginal distribution for content caching by decoupling the caches of D2D users in a spatial network scenario is not sufficient to optimize the joint performance of the caching. The performance can be improved by developing spatially correlated content placement strategies that exploit the spatial distribution of the D2D nodes, as we propose in this chapter.

Negatively correlated spatial placement corresponds to a distance-dependent thinning of the transmitter process so that neighboring users are less likely to have matching contents. This kind of approach is promising from an average cache hit rate optimization perspective. Therefore, we mainly focus on negatively dependent or repulsive content placement strategies.

We next define negative dependence for a collection of random variables.

Definition 2. *Random variables $Y_1, \ldots, Y_k$, $k \geq 2$, are said to be negatively dependent, if for any numbers $y_1, \ldots, y_k \in \mathbb{R}$, we have that [99]*

$$\mathbb{P}\Big(\bigcap_{i=1}^{k} Y_i \leq y_i\Big) \leq \prod_{i=1}^{k} \mathbb{P}(Y_i \leq y_i),$$
$$\mathbb{P}\Big(\bigcap_{i=1}^{k} Y_i > y_i\Big) \leq \prod_{i=1}^{k} \mathbb{P}(Y_i > y_i).$$

Next, in Prop. 1, we state the benefit of negatively correlated placement, which is the basis of future spatially correlated policies including our proposed policy in the current chapter.

Proposition 1. *Negatively dependent content placement provides a higher average cache hit probability than the independent placement strategies.*

Proof. See Appendix A in [87]. □

In the remainder of this chapter, we first discuss the independent content placement model in Sect. 3.4, which is a special case of the geographic content placement (GCP) problem using the Boolean model first proposed in [60].

We then ask the following question: Given the coverage number k and file m, how large cache hit rates can we achieve, i.e., how small can $\mathrm{P}_{\mathrm{Miss,N}}(m,k) \leq \mathbb{P}(Y_m = 0)^k$ get for a spatial content placement setting, or what is the best negatively dependent content placement strategy? To answer that, we consider a negatively dependent content placement strategy inspired from the Matérn hard-core processes MHC (type II), which we call as the hard-core content placement (HCP). We detail the HCP model in Sect. 3.6.

3.4 Independent Content Placement Design

Independent cache placement design is the baseline model where the files are cached at the D2D users identically and independently of each other. Let $\mathrm{p}_{\mathrm{c,I}}(m) = p_c(m,x_i) = \mathbb{P}(Y_m = 1)$ be the caching probability of file m in any cache, which is the same at all points $x_i \in \phi$.

The maximum average total cache hit probability, i.e., the probability that the typical user finds the content in one of the D2D users it is covered by,

Symbol	Definition
General System Model Parameters	
Baseline PPP with transmitter density λ_{t}	Φ
A realization of the PPP	$\phi = \{x_i\} \subset \mathbb{R}^2$
D2D communication radius	$\mathrm{R}_{\mathsf{D2D}}$
closed ball centered at x_i with radius $\mathrm{R}_{\mathsf{D2D}}$	$B_i(\mathrm{R}_{\mathsf{D2D}})$
The coverage process of the D2D transmitters driven by the BM	$V_{\mathrm{BM}} = \bigcup_i (x_i + B_0(\mathrm{R}_{\mathsf{D2D}}))$
File request distribution; Zipf request exponent	$p_r(\cdot) \sim \mathrm{Zipf}(\gamma_{\mathrm{r}})$; γ_r
Caching probability of file m in cache i	$\mathrm{p}_{\mathrm{c,X}}(m, x_i)$
Density of receivers; density of D2D users	λ_{r}; λ_{t}
Number of D2D users covering a receiver under strategy X	$\mathcal{N}_X$
Hit probability for placement strategy X	$\mathrm{P}_{\mathsf{Hit,X}}$
Miss probability of file m given k users cover the typical receiver for placement strategy X	$\mathrm{P}_{\mathrm{Miss,X}}(m,k)$
Total number of files; cache size	$M; N < M$
Independent Content Placement Design	
The caching distribution for	
independent placement	$\mathrm{p}_{\mathrm{c,I}}(m)$
geographic content placement (GCP) strategy in [60]	$\mathrm{p}_{\mathrm{c,G}}(m)$
caching most popular content (MPC)	$\mathrm{p}_{\mathrm{c,MPC}}(m) = 1_{m \leq N}$
Hard-Core Content Placement (HCP) Design	
HCP-A model constructed from the underlying PPP Φ	Φ_M
Exclusion radius of file m for the HCP-A model	r_m
The density of the HCP-A model for file m	$\lambda_{\mathsf{HCP\text{-}A}}(m)$
The number of neighboring transmitters in $B_0(r_m)$	$C_m \sim \mathsf{Poisson}(\bar{C}_m)$
	$\bar{C}_m = \lambda_{\mathrm{t}}\, \pi r_m^2$
The number of transmitters containing file m in $B_0(\mathrm{R}_{\mathsf{D2D}})$	$\tilde{C}_m$
$2k$ dimensional bounded region $[0, D]^{2k}$	$\mathcal{D}^k = [0, D]^{2k}$
The cache miss region given there exists k nodes	$\mathcal{V}^k = [0, D]^{2k} \backslash [0, \mathrm{R}_{\mathsf{D2D}}]^{2k}$
Second-order product density for file m	$\rho_m^{(2)}(r)$

Table 3.1: Notation for Chapter 3.

can be evaluated by solving

$$\begin{aligned} \max_{\mathrm{p_{c,I}}} \quad & \mathrm{P_{Hit,I}} \\ \text{s.t.} \quad & \sum_{m=1}^{M} \mathrm{p_{c,I}}(m) \leq N, \end{aligned} \tag{3.4}$$

and $\mathrm{P_{Miss,I}}(m,k) = (1-\mathrm{p_{c,I}}(m))^k$, which is related to $\mathrm{P_{Hit,I}}$ through the $\mathrm{P_{Hit,X}}$ expression in (3.3).

First, we consider the following trivial case of independent placement, which is clearly suboptimal.

Proposition 2. Caching most popular content MPC. *The baseline solution is to store the most popular files only. Letting* $Y_m = 1_{m\leq N}$*, i.e.,* $\mathrm{p_{c,MPC}}(m) = 1_{m\leq N}$*, the miss probability is* $\mathrm{P_{Miss,MPC}}(m,k) = 1_{N<m\leq M}$ *for all* m *when* $k \geq 1$*, and* $\mathrm{P_{Miss,MPC}}(m,k) = 1$ *when* $k = 0$*. Hence, the average cache hit probability for the* MPC *scheme is* $\mathrm{P_{Hit,MPC}} = \mathbb{P}(\mathcal{N}_X \geq 1)\sum_{m=1}^{N} p_r(m)$.

The independent cache design problem in this chapter is a special case of the geographic content placement (GCP) problem using the Boolean model as proposed in [60]. The optimal solution of the GCP problem [60] is characterized by Theorem 3.

Theorem 3. Geographic Content Placement (GCP**) [60, Theorem 1].** *The optimal caching distribution for the independent placement strategy is*

given as follows

$$\mathrm{p}^{*}_{\mathrm{c,G}}(m) = \begin{cases} 1, & \mu^* < p_r(m)\mathbb{P}(\mathcal{N}_P = 1) \\ \frac{1}{\lambda_{\mathrm{t}}\,\pi \mathrm{R}^2_{\mathsf{D2D}}} \log\left(\frac{p_r(m)\,\lambda_{\mathrm{t}}\,\pi \mathrm{R}^2_{\mathsf{D2D}}}{\mu^*}\right), & p_r(m)\mathbb{P}(\mathcal{N}_P = 1) \leq \mu^* \leq p_r(m)\mathbb{E}[\mathcal{N}_P], \\ 0, & \mu^* > p_r(m)\mathbb{E}[\mathcal{N}_P] \end{cases} \tag{3.5}$$

where $\mathbb{P}(\mathcal{N}_P = 1) = e^{-\,\lambda_{\mathrm{t}}\,\pi \mathrm{R}^2_{\mathsf{D2D}}}(\lambda_{\mathrm{t}}\,\pi \mathrm{R}^2_{\mathsf{D2D}})$, $\mathbb{E}[\mathcal{N}_P] = \lambda_{\mathrm{t}}\,\pi\,\mathrm{R}^2_{\mathsf{D2D}}$. *The placement probabilities satisfy*

$$p_r(j) \sum_{m=1}^{M} \mathbb{P}(\mathcal{N}_P = m) m (1 - \mathrm{p}^{*}_{\mathrm{c,G}}(j))^{m-1} = \mu^*, \quad j \in \{1, \ldots, M\}. \tag{3.6}$$

The optimal variable μ^* *satisfies the equality* $\sum_{m=1}^{M} \mathrm{p}^{*}_{\mathrm{c,G}}(m) = N$.

Thus, the optimal value of the average cache hit probability for the GCP *model is given by*

$$\mathrm{P}_{\mathsf{Hit,G}} = \sum_{m=1}^{M} p_r(m)[1 - \exp\left(-\,\lambda_{\mathrm{t}}\,\mathrm{p}^{*}_{\mathrm{c,G}}(m)\pi \mathrm{R}^2_{\mathsf{D2D}}\right)]. \tag{3.7}$$

Proof. It follows from the use of the Lagrangian relaxation method [60, Theorem 1]. The solution is found numerically using the bisection method. □

Throughout the chapter we use the terms independent cache placement and GCP interchangeably.

A Linear Approximation to Independent Cache Design. Given that each cache can store $N < M$ files[3], our objective is to determine the number of files L that should be stored in the cache with probability 1, and

[3]Swapping the contents within a cache does not change cache's state.

the maximum number of distinct files K that can be stored as a function of the design parameters, e.g., $\mathrm{R}_{\mathsf{D2D}}$, λ_{t} and N. We uniquely determine (L, K) that approximate the optimal content placement pmf in (3.5).

Proposition 3. A linear approximation to GCP. *The following linear content placement model approximates (3.5):*

$$\mathrm{p}_{\mathrm{c,G}}^{\mathrm{Lin}}(m) = \min\left\{1, \left(1 - \frac{m-L}{K-L}\right)^{+}\right\}, \tag{3.8}$$

where $y^{+} = \max\{y, 0\}$ *and the pair* (L, K) *can be determined using (B.1) and (B.3).*

Proof. See Appendix B.1. □

We next demonstrate that this linear model is a good approximation. We compare the optimal solution $\mathrm{p}_{\mathrm{c,G}}^{*}(m)$ (3.5) and our linear approximation (3.8) $\mathrm{p}_{\mathrm{c,G}}^{\mathrm{Lin}^{*}}(m)$ in Fig. 3.1, and observe that our linear solution is indeed a good approximation of the optimal solution. Keeping γ_r constant, by increasing $\mathrm{R}_{\mathsf{D2D}}$, we expect to see a more diverse set of requests from the user, L to decrease and K to increase. The converse is also true. When we keep $\mathrm{R}_{\mathsf{D2D}}$ fixed, and increase γ_r, since the requests become more skewed towards the most popular files, the optimal strategy for the user is to store the most popular files in its cache. Keeping $\mathrm{R}_{\mathsf{D2D}}$ and γ_r fixed, and increasing λ has a similar effect as increasing $\mathrm{R}_{\mathsf{D2D}}$, as illustrated. From these plots, although it is clear that independent placement favors the most popular contents, it is not always optimal to cache the most popular contents everywhere.

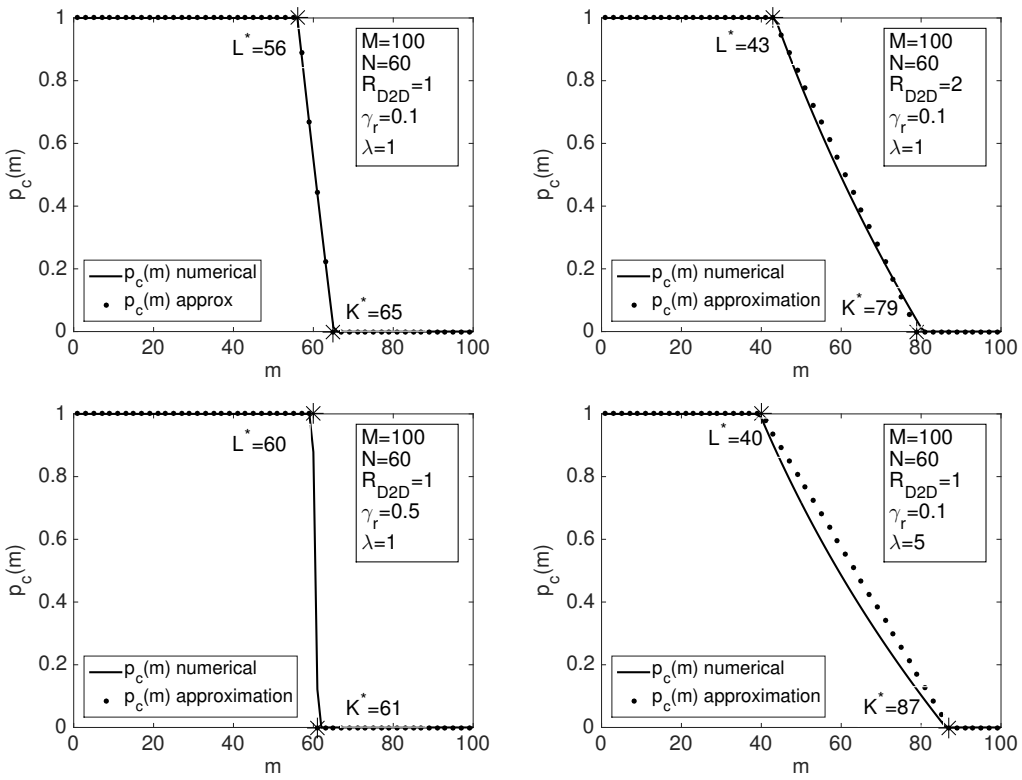


Figure 3.1: Optimal cache placement (independently at each user) with more focused content popularity.

Next, in Sect. 3.5, we consider a simple spatially dependent content placement strategy that is inspired from exchangeability.

3.5 Spatially Exchangeable Content Placement Design

From a user's perspective, the exact location of cached content is not important as long as it is available within $\mathrm{R_{D2D}}$. This is illustrated in Fig. 3.2 by an example with two equivalent models. In both models, the number of caches having any content type is the same. However, the locations where

the content is cached are different. More generally, from the typical user's perspective, any finite permutation of any content type among the caches within $\mathrm{R}_{\mathsf{D2D}}$ of the user is equivalent.

Consider a spatially exchangeable cache model defined as follows. For an ordered set of n transmitters covering a typical receiver with desired content m, the binary sequence $Y_{m_1}, \ldots, Y_{m_n}$ denotes the availability of the content in the respective caches: Y_{m_i} takes the value 1 if file m is available in cache i and 0 otherwise. The sequence $\{Y_{m_i}\}$ is *exchangeable* in the spatial domain.

Definition 3. An exchangeable sequence $Y_1, Y_2, \ldots, Y_n$ of random variables *is such that for any finite permutation r of the indices $1, 2, \ldots, n$, the joint probability distribution of the permuted sequence $Y_{r(1)}, Y_{r(2)}, \ldots, Y_{r(n)}$ is the same as the joint distribution of the original sequence [100].*

A theoretical description of exchangeability is given now.

Theorem 4. de Finetti's theorem. *A binary sequence $Y_1, \ldots, Y_n, \ldots$ is exchangeable if and only if there exists a distribution function F on $[0,1]$ such that for all n $p(y_1, \ldots, y_n) = \int_0^1 \theta^{t_n}(1-\theta)^{n-t_n}\,\mathrm{d}F(\theta)$, where $p(y_1, \ldots, y_n) = \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n)$ is the joint pmf and $t_n = \sum_{i=1}^{n} y_i$. It further holds that F is the distribution function of the limiting frequency, i.e., if $X = \lim_{n\to\infty} \sum_i Y_i/n$ a.s., then $\mathbb{P}(X \leq x) = F(x)$ and by conditioning with $X = \theta$, we obtain*

$$\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n | X = \theta) = \theta^{t_n}(1-\theta)^{n-t_n}. \tag{3.9}$$

Future samples behave like earlier samples, meaning formally that any order (of a finite number of samples) is equally likely. This formalizes the

notion of the future being predictable on the basis of past experience. To give more intuition on exchangeability, we next give an example.

Example 1. Sampling Without Replacement [101]. *Fix the number transmitters n covering a receiver with desired content m, and consider any permutation $Y_{r(m_1)}, \ldots, Y_{r(m_n)}$. Conditionally place the content to cache: $\mathbb{P}(Y_{r(m_k)} = 0 | Y_{r(m_1)} = 0, \ldots Y_{r(m_{k-1})} = 0) = \frac{k}{k+1}$ for $1 \leq k \leq n$. Hence, the miss probability for file m given k caches cover a receiver is $P_{\text{Miss,E}}(m,k) = \frac{1}{2} \times \frac{2}{3} \times \ldots \times \frac{k}{k+1} = \frac{1}{k+1}$. In this example, the limiting random variables are uniformly distributed on $[0,1]$, i.e., $X_m \sim F_m = U[0,1]$ for $m \in \{1, \ldots, M\}$. Hence, (3.11) gives the same result for $P_{\text{Miss,E}}(m,k)$.*

The formulation to maximize the cache hit for an exchangeable placement strategy becomes

$$\begin{aligned} \max_{f_{X_m}} \ & \mathrm{P}_{\mathsf{Hit,E}} \\ \text{s.t.} \quad & \sum_{m=1}^{M} \mathbb{E}[X_m] \leq N, \\ & \int_0^1 \mathrm{d}F_{X_m}(\theta) = 1, \quad m \in \{1, \ldots, M\}. \end{aligned} \tag{3.10}$$

The constraints are such that the distribution functions F_{X_m} for $m \in \{1, \ldots, M\}$ are on $[0,1]$, and $\mathbb{E}[X_m] = \int_0^1 \theta f_{X_m}(\theta)\, \mathrm{d}\theta$ is the probability a cache contains file m and each cache contains N files in total on average.

From Theorem 4, the average cache miss probability $P_{\text{Miss,E}}(m,k)$ is given by

$$P_{\text{Miss,E}}(m,k) = \int_0^1 (1-\theta)^k f_{X_m}(\theta)\, \mathrm{d}\theta = \mathbb{E}[(1-X_m)^k], \tag{3.11}$$

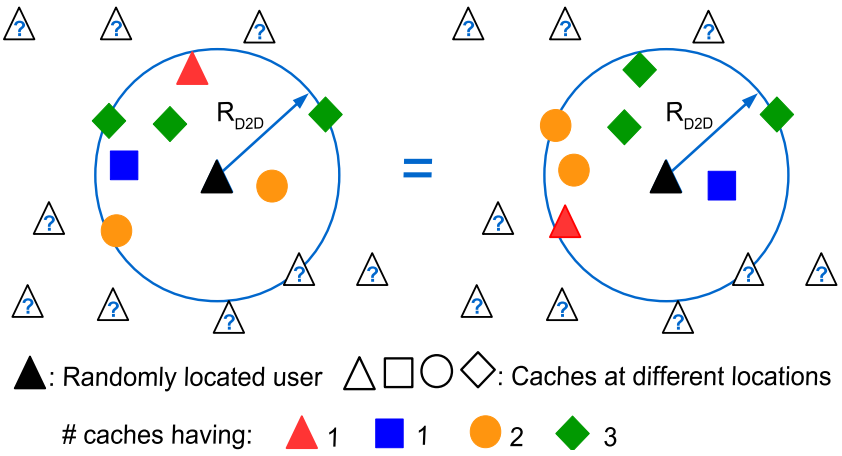


Figure 3.2: Exchangeable cache placement with two equivalent models, where the same set and multiplicity of files are permuted among the caches within $\mathrm{R}_{\mathsf{D2D}}$ of the randomly located user.

which is related to $\mathrm{P}_{\mathsf{Hit,E}}$ through (3.3). Hence, $\mathrm{P}_{\mathsf{Hit,E}}$ in (3.10) is equal to

$$\begin{aligned}\mathrm{P}_{\mathsf{Hit,E}} &= \sum_{m=1}^{M} p_r(m) \int_0^1 \left[1 - \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_t = k)(1-\theta)^k\right] f_{X_m}(\theta)\, \mathrm{d}\theta \\ &= 1 - \sum_{m=1}^{M} p_r(m) \mathbb{E}[e^{-\lambda_{\mathrm{t}} \pi \mathrm{R}_{\mathsf{D2D}}^2 X_m}]. \end{aligned} \tag{3.12}$$

Proposition 4. *Any exchangeable content placement strategy is worse than independent placement in terms of the average cache hit probability.*

Proof. Using the convexity of exponential function, we rewrite (3.12) as

$$\begin{aligned}\mathrm{P}_{\mathsf{Hit,E}} &= 1 - \sum_{m=1}^{M} p_r(m) \mathbb{E}\left[e^{-\lambda_{\mathrm{t}} \pi \mathrm{R}_{\mathsf{D2D}}^2 X_m}\right] \\ &\leq 1 - \sum_{m=1}^{M} p_r(m) e^{-\lambda_{\mathrm{t}} \pi \mathrm{R}_{\mathsf{D2D}}^2 \mathbb{E}[X_m]}. \end{aligned} \tag{3.13}$$

Hence, the hit probability of the exchangeable placement model is lower than the hit probability of the independent placement, for which $p_c(m) = \mathbb{E}[X_m]$ is the placement probability. □

The next result generalizes Proposition 4 to any kind of coverage distribution $\mathbb{P}(\mathcal{N}_t = \cdot)$.

Lemma 10. *Given any coverage distribution, which include the Boolean model and the Signal-to-Interference-and-Noise-Ratio (*SINR*) model or any other coverage model, the exchangeable placement strategy always performs worse than the independent placement strategy.*

Proof. Let X_m's be the limiting random variables for the exchangeable model. From (3.11), $P_{\text{Miss,E}}(m,k) = \mathbb{P}(\bigcap_{m=1}^{k}\{Y_m = 0\}) = \mathbb{E}[(1-X_m)^k]$, and from exchangeability, the distribution function of X_m, i.e., F_{X_m} is on $[0,1]$. From the convexity of $(1-X_m)^k$ for $k \in \mathbb{Z}_{\geq 0}$, $P_{\text{Miss,E}}(m,k) \geq (1-\mathbb{E}[X_m])]^k$. The hit probability for the exchangeable model is bounded by

$$\mathrm{P}_{\mathsf{Hit,E}} \leq 1 - \sum_{m=1}^{M} p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_t = k)(1-\mathbb{E}[X_m])^k,$$

where $\mathbb{E}[X_m] = \mathbb{P}(Y_m = 1)$ denotes the caching probability of file m for the independent placement model, which gives a higher average hit probability than the exchangeable strategy. □

We showed that spatially exchangeable placement yields a positively correlated spatial distribution of content, and is suboptimal in terms of the

cache hit probability compared to independent placement. However, as the coverage number [80] –the number of transmitters simultaneously covering a user– increases, the performance of exchangeable placement approaches the performance of independent placement.

A wide class of random processes exhibit exchangeability, which include combinatorial stochastic processes, Markov chains, coalescent processes, Poisson-Dirichlet processes, Erdős-Rényi graphs, the Chinese restaurant process, and a large collection of statistical mechanical systems on complete graphs. Interested reader can refer to [100] and [102] for further examples.

3.6 Hard-Core Content Placement Design

We next consider the hard-core regime, which provides useful insights for the development of spatial content placement for the regime relevant to D2D communications. Matérn's hard-core (MHC) model is a spatial point process whose points are never closer to each other than some given distance. We provide two different spatially correlated content placement models both inspired from the Matérn hard-core (MHC) (type II): (i) HCP-A which is an optimized placement model to maximize the average total cache hit probability in (3.2), and (ii) HCP-B which has the same marginal content placement probability as the GCP model in [60], and is sufficient for achieving a higher cache hit probability than the GCP model.

3.6.1 Hard-Core Placement Model I (HCP-A)

We propose a content placement approach to pick a subset of transmitters based on some exclusion by exploiting the spatial properties of MHC (type II) model, which we call HCP-A. This type of MHC model is constructed from the underlying PPP Φ modeling the locations of the D2D user caches by removing certain nodes of Φ depending on the positions of the neighboring nodes and additional marks attached to those nodes [61, Ch. 2.1]. Each transmitter of the BM V_{BM} is assigned a uniformly (i.i.d.) distributed mark $U[0,1]$. A node $x \in \Phi$ is selected if it has the lowest mark among all the points in $B_x(R)$, given exclusion radius R. A realization of the MHC point process Φ_M is illustrated in Fig. 3.3.

The HCP-A placement model is motivated from the MHC model and implemented as follows. For each file type, there is a distinct exclusion radius (r_m for file m) instead of having a fixed exclusion radius R, and the exclusion radii are determined by the underlying file popularity distribution. Given a realization ϕ of the underlying PPP modeling the locations of the transmitters with intensity λ_{t}, we sort the file indices in order of decreasing popularity. For given file index m and radius r_m, we implement the steps (a)-(d) described in Fig. 3.3 to determine the set of selected transmitters to place file m. For the same realization ϕ, we implement this procedure for all files. Once a cache is selected N times, then it is full, and no more file can be placed even if it is selected. The objective is to determine the file radii to optimize the placement.

Definition 4. Configuration probability. *The probability density function*

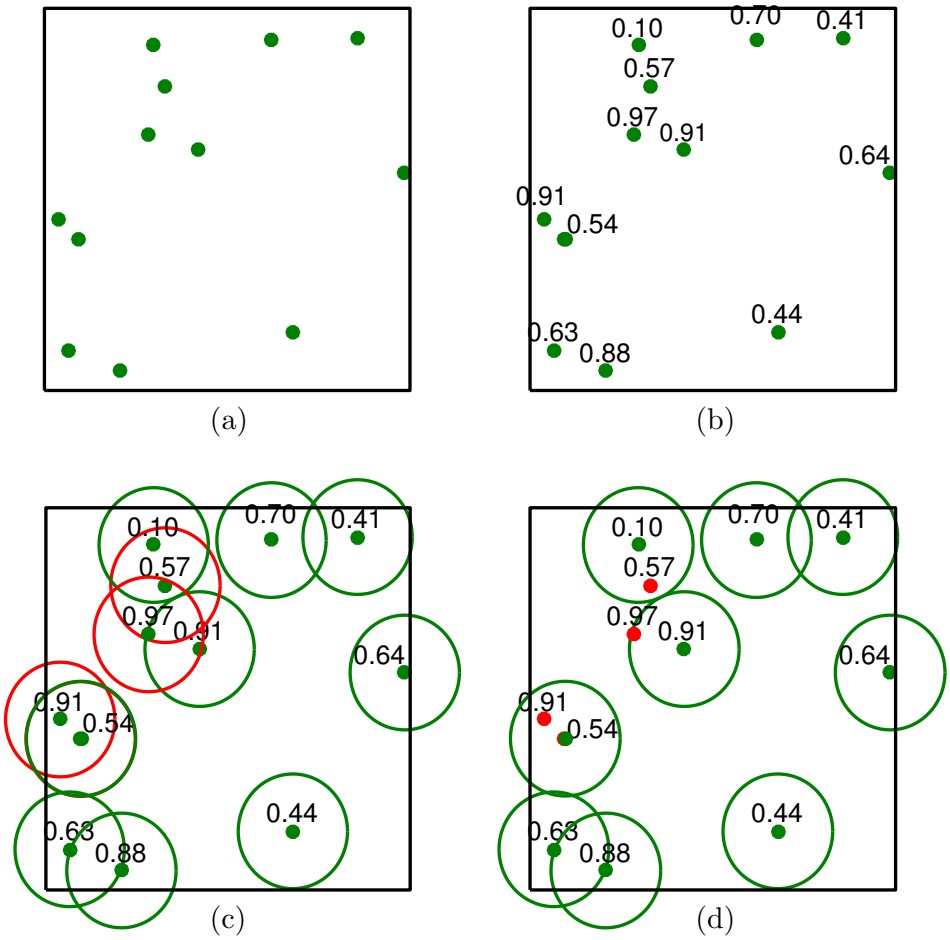


Figure 3.3: MHC point process realization for a given exclusion radius R: (a) Begin with a realization of PPP, ϕ. (b) Associate a uniformly distributed mark $U[0,1]$ to each point of ϕ independently. (c) A node $x \in \phi$ is selected if it has the lowest mark inside $B_x(R)$. (d) Set of selected points for a given realization of the PPP.

(pdf) of the MHC *point process* Φ_M *with exactly* k *points in a bounded region* $\mathcal{D} = [0, D]^2 \in \mathbb{R}^2$ *that denotes the set retained caches that contain file* m *is*

given by $f : \mathbb{R}^{2k} \to [0, \infty)$ *[64, Ch. 5.5] so that*

$$f_m(\varphi) = \begin{cases} a_m, \ \textit{if } s_\varphi(r_m) = 0, \\ 0, \ \textit{otherwise.} \end{cases} \tag{3.14}$$

which is also known as the configuration probability, i.e., the probability that the hard-core model Φ_M *takes the realization* φ*. In the above,* $\varphi = \{x_1, \ldots, x_k\} \subset \mathcal{D}$ *denotes the set of* k *points,* a_m *is a normalizing constant and* $s_\varphi(r)$ *is the number of inter-point distances in* φ *that are equal or less than* r*. This yields a uniform distribution*[4] *of a subset of* k *points with inter-point distances at least* r_m *in* $\mathcal{D}$*.*

We optimize the exclusion radii to maximize the total hit probability. The exclusion radius of a particular file r_m depends on the file popularity in the network, transmitter density and the cache size and satisfies $r_m < \mathrm{R}_{\mathsf{D2D}}$. Otherwise, once r_m exceeds $\mathrm{R}_{\mathsf{D2D}}$, as holes would start to open up in the coverage for that content, the hit probability for file m would suffer. We consider the following cases: (i) if the file is extremely popular, then many transmitters should simultaneously cache the file, yielding a small exclusion radius, and (ii) if the file is not very popular, then fewer (or zero) transmitters would be sufficient for caching the file, yielding a larger exclusion radius. Therefore, intuitively, we expect the exclusion radius to decrease with increasing file popularity. Our analysis also supports this conclusion that the exclusion radius is

[4]The pdf of the retained process (3.14) is a scaled version of the pdf of the PPP Φ in which there is no point within the exclusion range of the typical cache. This yields a uniform distribution of k points in $\mathcal{D}$, i.e., $f(\varphi) = a$, where a is a normalizing constant.

inversely related to the file popularity, i.e., the most popular files are stored in a high number of caches with higher marginal probabilities unlike the files with low popularity that are stored with lower marginals, and with larger exclusion radius.

By the Slivnyak Theorem, the Palm distribution of the PPP Φ seen from its typical point (cache) located at 0 corresponds to the law of $\Phi \cup \{0\}$ under the original distribution [61, Ch. 1.4]. Since the typical node (which is at the origin) of Φ has C_m neighbors distributed as $C_m \sim \mathsf{Poisson}(\bar{C}_m)$ with $\bar{C}_m = \lambda_{\mathrm{t}}\, \pi r_m^2$, given the exclusion radius r_m for file m of the HCP-A model, and the file may be placed at most at only one cache within this circular region. Hence, the probability of a typical D2D transmitter to get the minimum mark in its neighborhood to qualify to cache file m, equivalently, the caching probability of file m at a typical transmitter is

$$\mathrm{p}_{\mathrm{c},\mathsf{HCP\text{-}A}}(m) = \mathbb{E}\Big[\frac{1}{1+C_m}\Big] = \frac{1-\exp(-\bar{C}_m)}{\bar{C}_m}. \tag{3.15}$$

From (3.15), we can easily observe that there is a one-to-one relationship between r_m and $\mathrm{p}_{\mathrm{c},\mathsf{HCP\text{-}A}}(m)$. The inverse relationship between r_m and $\mathrm{p}_{\mathrm{c},\mathsf{HCP\text{-}A}}(m)$ can be seen by taking the following limits:

$$\lim_{r_m \to 0} \mathrm{p}_{\mathrm{c},\mathsf{HCP\text{-}A}}(m) = 1, \quad \lim_{r_m \to \infty} \mathrm{p}_{\mathrm{c},\mathsf{HCP\text{-}A}}(m) = 0, \tag{3.16}$$

which implies that the popular files have small r_m, hence are cached more frequently, and unpopular files have larger exclusion radii, and are stored at fewer locations.

We denote the density of the HCP-A model for file m by

$$\lambda_{\text{HCP-A}}(m) = \frac{[1 - \exp(-\bar{C}_m)]}{\pi r_m^2} = \mathrm{p}_{\mathrm{c,HCP\text{-}A}}(m)\, \lambda_{\mathrm{t}}\,. \tag{3.17}$$

Let $\tilde{C}_m$ be the number of transmitters containing file m within a circular region of radius $\mathrm{R}_{\mathsf{D2D}}$. At most one transmitter is allowed to contain a file within the exclusion radius. Therefore, when $r_m \geq \mathrm{R}_{\mathsf{D2D}}$, we have $\tilde{C}_m \in \{0, 1\}$, and when $r_m < \mathrm{R}_{\mathsf{D2D}}$, we have $\tilde{C}_m \in \{0, 1, 2, \cdots\}$.

Proposition 5. *The* MHC *placement is a negatively dependent placement technique.*

Proof. See Appendix B in [87]. □

As the file popularity increases, the exclusion radius gets smaller. Hence, the average number of transmitters within the exclusion region, i.e., $\bar{C}_m^*$, decreases, and the chance of having at least one transmitter caching that file within $\mathrm{R}_{\mathsf{D2D}}$ increases, i.e., $\mathbb{P}(\tilde{C}_m \geq 1) > \mathbb{P}(\tilde{C}_n \geq 1)$ for $m < n$. This yields a higher $\mathrm{p}_{\mathrm{c,HCP\text{-}A}}(\cdot)$ for more popular files from (3.15). If the demand distribution is uniform over the network, then each file has the same caching probability, i.e., $\mathrm{p}_{\mathrm{c,HCP\text{-}A}}(m)$ is the same for all m, yielding the same r_m for all m, which is intuitive. When the demand distribution is skewed towards the more popular files, then $\lambda_{\text{HCP-A}}(m)$ scales with the request popularity and r_m is inversely proportional to $p_r(m)$, i.e., less popular files will end up being stored in fewer locations, and popular files will be guaranteed to be available over a larger geographic area, which is intuitive.

In the HCP-A model, using the pdf in (3.14) that denotes the configuration of the retained transmitters, the miss probability of file m given k users cover a typical receiver is

$$\mathrm{P}_{\mathrm{Miss,MA}}(m,k) = \int \cdots \int_{\mathcal{V}^k} f_m(x_1,\ldots,x_k)\,\mathrm{d}x_1 \ldots \mathrm{d}x_k, \tag{3.18}$$

where the region $\mathcal{V}^k$ characterizes the cache miss region given there exists k D2D nodes, i.e., it is the $2k$ dimensional region denoted by $\mathcal{V}^k = [0,D]^{2k} \backslash [0,\mathrm{R}_{\mathsf{D2D}}]^{2k}$.

The maximum hit probability for the HCP-A model is given by the solution of

$$\begin{aligned} \max_{\mathrm{p_{c,HCP\text{-}A}}(m)} \quad & \mathrm{P}_{\mathsf{Hit,HCP\text{-}A}} \\ \text{s.t.} \quad & \sum_{m=1}^{M} \mathrm{p_{c,HCP\text{-}A}}(m) \leq N, \end{aligned} \tag{3.19}$$

and $\mathrm{P}_{\mathrm{Miss,MA}}(m,k)$ is given in (3.18), which is related to $\mathrm{P}_{\mathsf{Hit,HCP\text{-}A}}$ through the $\mathrm{P}_{\mathsf{Hit,X}}$ expression given in (3.3) of the original optimization formulation in (3.2).

Proposition 6. *The average cache hit probability for the* HCP-A *model is*

$$\mathrm{P}_{\mathsf{Hit,HCP\text{-}A}} = \sum_{m=1}^{M} p_r(m) \mathbb{P}(\tilde{C}_m > 0 | r_m), \tag{3.20}$$

where the term $\mathbb{P}(\tilde{C}_m > 0|r_m)$ *is essential in determining the cache hit probability and given as*

$$\mathbb{P}(\tilde{C}_m > 0|r_m) \begin{cases} \geq 1 - \exp(-\lambda_{\mathsf{HCP\text{-}A}}(m)\pi \mathrm{R}^2_{\mathsf{D2D}}), & r_m < \mathrm{R}_{\mathsf{D2D}}, \\ = \lambda_{\mathsf{HCP\text{-}A}}(m)\pi\, \mathrm{R}^2_{\mathsf{D2D}}, & r_m \geq \mathrm{R}_{\mathsf{D2D}}\,. \end{cases} \tag{3.21}$$

Proof. See Appendix B.2. □

The optimal solution of the HCP-A model in (3.19) is characterized by Theorem 5.

Theorem 5. Hard-Core Content Placement (HCP). *The optimal caching distribution for the* HCP *model is given as follows*

$$\mathrm{p}^{*}_{\mathrm{c},\mathsf{HCP\text{-}A}}(m) = \begin{cases} {\lambda_{\mathrm{t}}}^{-1}\, W(cp_r(m)), & m \leq \mathrm{m_c}, \\ {\lambda_{\mathrm{t}}}^{-1}\, cp_r(m), & m > \mathrm{m_c}, \end{cases} \tag{3.22}$$

where W *is the Lambert function, and* $\mathrm{m_c} = \underset{m\in\{1,\cdots,M\}}{\arg\max}\, \{r_m | r_m < \mathrm{R_{D2D}}\}$*, and the relation*

$$\sum_{m=1}^{\mathrm{m_c}} W(cp_r(m)) - cp_r(m) = N\,\lambda_{\mathrm{t}} - c \tag{3.23}$$

can be used to determined the value of c*. Hence, we determine* $\lambda^{*}_{\mathsf{HCP\text{-}A}}(m)$ *and the optimal value of the exclusion radius, i.e.,* r^{*}_m*, from (3.23) as a function of the request pmf* $p_r(m)$*, cache size* N *and the transmitter density* λ_{t}*.*

Proof. See Appendix D in [87]. □

Consider a ball centered at origin and of radius D, i.e., $B_0(D)$, with $D \gg \max_m\{r_m\}$, let the number of users in $B_0(D)$ be Poisson with $\mathbb{P}(\mathcal{N}_P(D) = k) = e^{-\bar{C}_D}\frac{(\bar{C}_D)^k}{k!}$, where $\bar{C}_D = \lambda_{\mathrm{t}}\,\pi D^2$ is the average number of transmitters within $B_0(D)$. Due to the limited storage capacity of the caches, the mean total number of files that can be cached in $B_0(D)$ is upper bounded by $N\bar{C}_D$. To determine the average number of users containing a desired file type in region $B_0(D)$, we use the second-order product density of the MHC process Φ_M, which is defined next.

Definition 5. Second-order product density [64, Ch. 5.4]. *For a stationary point process Φ_M, the second-order product density is the joint probability that there are two points of Φ_M at locations x and y in the infinitesimal volumes dx and dy, and given by*

$$\rho_m^{(2)}(r) = \begin{cases} \lambda^2_{\mathsf{HCP\text{-}A}}(m), & r \geq 2r_m \\ \dfrac{2V_{r_m}(r)[1-e^{-\lambda_{\mathrm{t}}\pi r_m^2}] - 2\pi r_m^2[1-e^{-\lambda_{\mathrm{t}}V_{r_m}(r)}]}{\pi r_m^2 V_{r_m}(r)[V_{r_m}(r)-\pi r_m^2]}, & r_m < r < 2r_m \\ 0, & r \leq r_m \end{cases} \quad (3.24)$$

where $\lambda_t^{-2}\rho_m^{(2)}(r)$ is the two-point Palm probability that two points of Φ separated by distance r are both retained to store file m [64, Ch. 5.4], and $V_{r_m}(r) = 2\pi r_m^2 - 2r_m^2\cos^{-1}\left(\frac{r}{2r_m}\right) + r\sqrt{r_m^2 - \frac{r^2}{4}}$ is the area of the union of two circles with radius r_m and separated by distance r. Pairwise correlations between the points separated by $r > r_m$ are modeled using the second-order product density –$\rho_m^{(2)}(r)$ for file m– of the MHC *process.*

Using the Campbell's theorem [61, Ch. 1.4], we deduce that the average number of transmitters of the stationary point process Φ_M –conditioned on there being a point at the origin but not counting it– contained in the ball $B_0(\mathrm{R}_{\mathsf{D2D}})$ is given by

$$\mathbb{E}^{!\circ}\left[\sum_{x\in\Phi_M} 1(x \in B_0(\mathrm{R}_{\mathsf{D2D}}))\right] = {\lambda_{\mathrm{t}}}^{-1}\int_{B_0(\mathrm{R}_{\mathsf{D2D}})} \rho_m^{(2)}(x)\,dx. \quad (3.25)$$

An upper bound on the probability that a user requesting file m is covered is given by the following expression:

$$\mathbb{P}(\tilde{C}_m \geq 1 | r_m < \mathrm{R}_{\mathsf{D2D}}) \overset{(a)}{\leq} \mathbb{E}[\tilde{C}_m | r_m < \mathrm{R}_{\mathsf{D2D}}]$$

$$\stackrel{(b)}{=} 1 - \exp(-\lambda^*_{\mathsf{HCP\text{-}A}}(m)\pi \mathrm{R}^2_{\mathsf{D2D}}) + {\lambda_{\mathrm{t}}}^{-1} \int_{B_0(\mathrm{R}_{\mathsf{D2D}})} \rho^{(2)}_m(x)\mathrm{d}x, \tag{3.26}$$

where (a) follows from using Markov inequality, and (b) from using (3.25), to deduce the average number of caches that stores file m in $B_0(\mathrm{R}_{\mathsf{D2D}})$.

Proposition 7. *The maximum cache hit probability for the* HCP-A *model is approximated by the following lower and upper bounds:*

$$\begin{aligned}
\mathrm{P}^{\mathsf{LB}}_{\mathsf{Hit},\mathsf{HCP\text{-}A}} &= \sum_{m=1}^{\mathrm{m_c}} p_r(m)[1 - e^{-\lambda^*_{\mathsf{HCP\text{-}A}}(m)\pi \mathrm{R}^2_{\mathsf{D2D}}}] \\
&\quad + \sum_{m=\mathrm{m_c}+1}^{M} p_r(m)\, \lambda^*_{\mathsf{HCP\text{-}A}}(m)\pi\, \mathrm{R}^2_{\mathsf{D2D}}, \\
\mathrm{P}^{\mathsf{UB}}_{\mathsf{Hit},\mathsf{HCP\text{-}A}} &= \mathrm{P}^{\mathsf{LB}}_{\mathsf{Hit},\mathsf{HCP\text{-}A}} + \sum_{m=1}^{\mathrm{m_c}} p_r(m){\lambda_{\mathrm{t}}}^{-1} \int_{r^*_m}^{\mathrm{R}_{\mathsf{D2D}}} \rho^{(2)}_m(x)\mathrm{d}x,
\end{aligned} \tag{3.27}$$

where $\bar{C}^*_m = \lambda_{\mathrm{t}}\, \pi (r^*_m)^2$ *with* r^*_m *denoting the optimal value of the radius* r_m*, and* $\lambda^*_{\mathsf{HCP\text{-}A}}(m)$ *follows from (3.17).*

Proof. See Appendix E in [87]. □

To compare the performance of the GCP and the HCP models in terms of their average cache hit probabilities, we next consider an example.

Example 2. Cache hit rate comparison for GCP and HCP. *Consider a simple caching scenario with* $M = 2$ *files and a cache size of* $N = 1$*, and the request distribution satisfies* $p_r(1) = 2/3$ *and* $p_r(2) = 1/3$*. Let* $\lambda_{\mathrm{t}}\, \pi = 1$ *and assume* $\mathrm{R}_{\mathsf{D2D}}$ *is given.*

- *In the* GCP *model, from Theorem 3, given the product* $\lambda_{\mathrm{t}}\,\pi\,\mathrm{R}_{\mathsf{D2D}}^2$*, the values of* $\mathbb{P}(\mathcal{N}_P = 1)$*,* $\mathbb{E}[\mathcal{N}_P]$ *can be computed. Checking the conditions in (3.5), the optimal value of* μ*, and* $\mathrm{p}_{\mathrm{c,G}}^*(1)$ *and* $\mathrm{p}_{\mathrm{c,G}}^*(2)$ *can be determined. Thus, from (3.7), the optimal cache hit probability for the* GCP *model becomes* $\mathrm{P}_{\mathsf{Hit,G}}^* = \sum_{m=1}^{2} p_r(m)[1 - \exp(-\,\mathrm{p}_{\mathrm{c,G}}^*(m)\,\lambda_{\mathrm{t}}\,\pi\,\mathrm{R}_{\mathsf{D2D}}^2)]$.

- *In the* HCP *model, from (3.17), we have* $\lambda_{\mathsf{HCP\text{-}A}}(m) = \frac{[1-\exp(-\bar{C}_m)]}{\pi r_m^2} = \mathrm{p}_{\mathrm{c},\mathsf{HCP\text{-}A}}(m)\,\lambda_{\mathrm{t}}$ *for* $m = 1, 2$*. Using the cache constraint,* $\sum_{m=1}^{2} \lambda_{\mathsf{HCP\text{-}A}}(m) = \lambda_{\mathrm{t}}$*. Thus, from (3.20), the cache hit probability for the* GCP *model becomes* $\mathrm{P}_{\mathsf{Hit,HCP\text{-}A}} = 2/3\mathbb{P}(\tilde{C}_1 > 0|r_1) + 1/3\mathbb{P}(\tilde{C}_2 > 0|r_2)$*, where from (3.21), we compute* $\mathbb{P}(\tilde{C}_m > 0|r_m)$ *using the lower bound in Prop. 7.*

  The optimal values $\mathrm{P}_{\mathsf{Hit,G}}^*$*,* $\mathrm{P}_{\mathsf{Hit,HCP\text{-}A}}^{\mathsf{LB}*}$ *for different* $\mathrm{R}_{\mathsf{D2D}}$ *are tabulated in Table 3.2, where the results for the* HCP *model are obtained by optimizing* $\mathrm{P}_{\mathsf{Hit,HCP\text{-}A}}^{\mathsf{LB}}$ *in (3.27) of Proposition 7. For* $\mathrm{R}_{\mathsf{D2D}}$ *high, as the lower bound of the* HCP *model is very close to* $\mathrm{P}_{\mathsf{Hit,G}}^*$*, both models perform similarly. However, for small* $\mathrm{R}_{\mathsf{D2D}}$*, the* HCP *model outperforms (with a cache hit rate gain up to* 25% *using the lower bound) because it can exploit the spatial diversity.*

Ideally, when a cache placement strategy is applied, the files need to be placed at a cache in a way that all the cache slots are occupied. In the GCP model in [60], authors propose a probabilistic placement policy to fill the caches. However, in the case of HCP-A placement, due to the random assignment of the marks in each cache independently for distinct files, it is not guaranteed that all the caches are full in the HCP-A approach, which causes

R_{D2D}	μ^*	$p^*_{c,G}(1, 2)$	$P^*_{Hit,G}$	r_1^*, r_2^*	$\lambda^*_{HCP\text{-}A}(1, 2)$	$P^{LB*}_{Hit,HCP\text{-}A}$
$\sqrt{0.5}$	0.1836	1, 0	**0.2623**	0.7071, 1.7117	0.2813, 0.0370	**0.3140**
$\sqrt{0.75}$	0.2430	0.9621, 0.0379	**0.352**	0.866, 1.4283	0.2428, 0.0756	**0.4407**
1	.28592	0.8466, 0.1534	**0.4282**	1, 1.257	0.201, 0.1174	**0.5438**
$\sqrt{2}$	0.3468	0.6733, 0.3267	**0.6532**	0.8718, 1.4178	0.2411, 0.0772	**0.6818**
$\sqrt{3}$	0.3156	0.6155, 0.3845	**0.7896**	1.0149, 1.2410	0.1961, 0.1222	**0.7896**
$\sqrt{10}$	0.0318	0.5347, 0.4653	**0.9936**	1.0909, 1.1576	0.1704, 0.1479	**0.9936**
10	$9.0926e^{-21}$	0.5035, 0.4965	**1**	1.1225, 1.1225	0.1592, 0.1592	**1**

Table 3.2: Numerical results in Chapter 3 for Example 2, with $M = 2$, $N = 1$ and $p_r(1) = 2/3$ $p_r(2) = 1/3$.

underutilization of the caches as detailed next.

Proposition 8. Cache underutilization. *The* HCP *placement model causes underutilization of the caches, i.e., on average, the fraction of the* D2D *nodes of* Φ *that contain* N *distinct files is always less than* 1. *This can be formally stated as follows:*

$$\frac{1}{N\mathbb{E}[\mathcal{N}_P]} \sum_{m=1}^{M} \mathbb{E}[\tilde{C}_m] \leq 1, \tag{3.28}$$

where $\mathbb{E}[\mathcal{N}_P] = \lambda_t \, \pi \, R^2_{D2D}$.

Proof. See Appendix F in [87]. □

The storage size N and the exclusion radius r_m have an inverse relationship. As N drops, because it is not possible to cache the files at all the transmitters, the exclusion radius should increase to bring more spatial diversity into the model. From the storage constraint in (3.19), as N drops, r_m

increases ($r_m \to \infty$ as $N \to 0$). Hence, a typical receiver won't be able to find its requested files within its range. When N increases sufficiently, r_m can be made smaller so that more files can be cached at the same transmitter ($r_m \to 0$ as $N \to \infty$). Hence, the typical receiver will most likely have the requested files within its range.

Proposition 9. A sufficient condition for the HCP-A placement model. *The* HCP-A *performs better than the independent placement model (*GCP*) [60] in terms of hit probability if the following condition is satisfied:*

$$\lambda_{\text{HCP-A}}(m) \geq \begin{cases} \lambda_{\text{t}}\, \mathrm{p}^{*}_{\text{c,G}}(m), & r_m < \mathrm{R}_{\text{D2D}}, \\ \dfrac{1-\exp(-\lambda_{\text{t}}\, \mathrm{p}^{*}_{\text{c,G}}(m)\pi \mathrm{R}^2_{\text{D2D}})}{\pi\, \mathrm{R}^2_{\text{D2D}}}, & r_m \geq \mathrm{R}_{\text{D2D}}, \end{cases} \tag{3.29}$$

where $\mathrm{p}^{*}_{\text{c,G}}(m)$ *is the optimal caching distribution for the* GCP*.*

Proof. See Appendix B.3. □

In the regime where r_m is chosen to satisfy the inequality in (3.29), for all m, the HCP-A placement model performs better than independent placement, and the volume fraction occupied by the transmitters caching file m, i.e., the proportion of space covered by the union $\bigcup_{x_i \in \Phi_M} (x_i + B_0(\mathrm{R}_{\text{D2D}}))$ pertaining to file m, is lower bounded by $\frac{\lambda_{\text{HCP-A}}(m)}{\lambda_{\text{t}}} \geq \frac{1-e^{-\lambda_{\text{t}}\, \mathrm{p}^{*}_{\text{c,G}}(m)\pi\, \mathrm{R}^2_{\text{D2D}}}}{\lambda_{\text{t}}}$. When the selection of $\lambda_{\text{HCP-A}}(m)$ does not satisfy (3.29), the volume fraction pertaining to the caches storing file m is upper bounded by $\frac{\lambda_{\text{HCP-A}}(m)}{\lambda_{\text{t}}} < \mathrm{p}^{*}_{\text{c,G}}(m)$.

From (3.29), the density parameter $\lambda_{\text{HCP-A}}(m)$ decreases with R_{D2D}, hence, the exclusion radius r_m increases with R_{D2D}, which is intuitive because

as the number of transmitters within the communication range increases a smaller fraction of them should cache the desired content. The exclusion radius decreases with popularity, i.e., r_m decreases as $p_r(m)$ increases. It also decreases with λ_t and the cache size N.

We consider two regimes of caching controlled by the cache size N, which determines the optimal cache placement solutions for the independent and HCP-A placement models. The spatial diversity of the content is captured by the optimal placement distribution for given N. As N increases, content diversity per cache increases and less spatial diversity is required. Therefore, when N is sufficiently large, independent placement is better than HCP-A placement. For the HCP-A placement model, the exclusion radii decrease with the file popularity. However, for small N, a higher exclusion radii are required for all files, which will increase the spatial diversity. Therefore, in the regime where N is small, for sufficiently large R_{D2D}, HCP-A placement performs better than independent placement (GCP).

We next detail another MHC-based model called HCP-B and provide sufficient conditions for achieving a higher cache hit probability than the GCP model of [60].

3.6.2 Hard-Core Placement Model II (HCP-B)

In this section, we propose a new MHC-inspired placement model called HCP-B. We seek a spatially correlated content caching model that improves the performance of the independent placement model of Sect. 3.4 based on

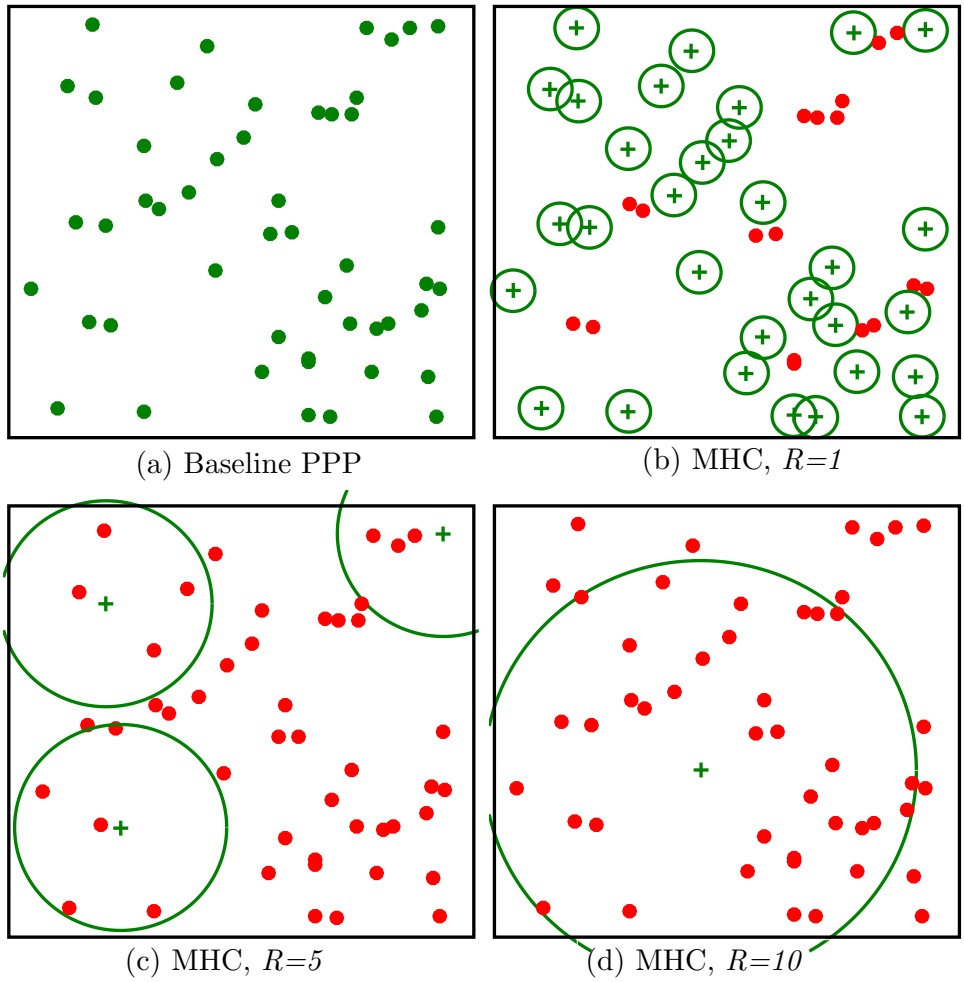
the GCP problem in [60] using the same marginal caching probabilities, i.e., on average the fraction of the users containing a file is equal to its optimal placement probability of the GCP model.

Different from the HCP-A model in Sect. 3.6.1, where we maximize the average cache hit probability given the finite cache storage constraint, in this section we optimize the exclusion radii using the caching distribution in (3.5) of the GCP model in Theorem 3, and provide sufficient conditions so that the HCP-B model is at least as good as the GCP scheme of [60].

The proposed content placement model is slightly different from the MHC point process transmission model with fixed radius. Instead, for each file type, there exists a different exclusion radius. For each file type, a circular exclusion region is created around each active transmitter to prevent all the transmitters located in a circular region from caching a particular content simultaneously. The exclusion radii are determined by the file popularity, which is detailed next.

The critical exclusion radius should be inversely proportional to the popularity of the requests, which is mainly determined by the skewness parameter γ_r. As γ_r increases, the distribution becomes more skewed and higher variability is observed in the exclusion radii of different files.

In Fig. 3.4, we illustrate the trend of the MHC process for different exclusion radii R. Each node is associated a uniformly distributed mark $U[0,1]$ independently. Node $x_i \in \phi$ is selected if it has the lowest mark in $B_i(R)$. As



(a) Baseline PPP (b) MHC, *R=1*

(c) MHC, *R=5* (d) MHC, *R=10*

Figure 3.4: MHC versus the exclusion radii R. (a) Begin with a realization of PPP, ϕ. Set of selected points (denoted by plus sign) for a given realization of the PPP for an exclusion radius of (b) $R = 1$, (c) $R = 5$ and (d) $R = 10$.

the exclusion radius R increases, the intensity of the retained nodes, i.e., λ_{MHC} of HCP-B process, decreases.

Proposition 10. *The exclusion radius for content m for the* HCP-B *model is given as*

$$r_m^* = \sqrt{\frac{1}{\lambda_t \pi} W\Big(- \frac{\exp(-1/\,\mathrm{p}_{\mathrm{c,G}}^*(m))}{\mathrm{p}_{\mathrm{c,G}}^*(m)}\Big) + \frac{1}{\lambda_t \pi\, \mathrm{p}_{\mathrm{c,G}}^*(m)}}, \quad n \in \mathbb{Z}, \tag{3.30}$$

where $\mathrm{p}_{\mathrm{c,G}}^*(\cdot)$ is the optimal caching distribution for GCP *and W is the Lambert function.*

Proof. See Appendix H in [87]. □

From Prop. 10, given the same marginal caching distributions for the GCP and the HCP-A models, the relation (3.30) guarantees the HCP-A model to outperform the independent content placement model in terms of the average cache hit rate performance.

Using the second order properties of the hard-core models, the variance of the HCP model is approximated by [64, Ch. 4.5]

$$\mathrm{Var}_{\mathsf{HCP\text{-}A}} \simeq \lambda_{\mathsf{HCP\text{-}A}} + 2\pi \int_0^\infty \big(\rho^{(2)}(\mathrm{r}) - \lambda_{\mathsf{HCP\text{-}A}}^2\big)\, \mathrm{rdr}.$$

Hence, using (3.24) the variance of the MHC model for file m can be approximated as

$$\begin{aligned}\mathrm{Var}_{\mathsf{HCP\text{-}A}}(\mathrm{m}) \simeq\ & \lambda_{\mathsf{HCP\text{-}A}}(m) - 4\,\lambda_{\mathsf{HCP\text{-}A}}(m)[1 - \exp(-\,\lambda_{\mathrm{t}}\,\pi r_m^2)] \\ & + 2\pi \int_{r_m}^{2r_m} \rho_m^{(2)}(r) r \mathrm{d}r. \end{aligned} \tag{3.31}$$

Note that r_m decreases, and $\lambda_{\mathsf{HCP\text{-}A}}(m)$ and $\rho_m^{(2)}(r)$ increase with popularity. Therefore, we can observe that there is a higher variability for popular files, which means that popular files are placed more randomly than unpopular files, and for unpopular files the placement distribution becomes more regular. This implies that randomized caching is in fact good for popular files, and more deterministic placement techniques are required for unpopular files.

3.7 Numerical Comparison of Different Content Placement Models

We showed that the HCP techniques detailed in Sect. 3.6 yield negatively correlated placement, and can provide a higher cache hit than independent placement (GCP). In this section, we verify our analytical expressions and provide a performance comparison between the GCP of [60], summarized in Sect. 3.4, and the HCP of Sect. 3.6 by contrasting the average cache hit rates, as discussed in Sect. 3.3. For tractability, in our simulations we assume $M = 2$ and $N = 1$. The D2D nodes form realizations of a PPP Φ over the region $[-10, 10]^2$ with an intensity λ_t per unit area. We assume there is a typical receiver at the origin which samples a request from the distribution satisfying $p_r(1) = 2/3$ and $p_r(2) = 1/3$. To compute the average cache hit probability performance of different models, we run 10^5 iterations, where at each iteration, we consider a realization ϕ of PPP Φ.

Cache hit rate with respect to λ_t. We illustrate the cache hit probability trends of the MPC policy, the GCP model in [60], and the HCP-A

and HCP-B placement models together with the bounds for the HCP-A model with respect to the intensity λ_t for $\mathrm{R_{D2D}} = 10$ in Fig. 3.5. It has already been numerically demonstrated in Fig. 3 of [60] that the hit probability of GCP outperforms MPC policy, especially for low SINR thresholds, corresponding to large $\mathrm{R_{D2D}}$ values. Therefore, we use GCP as benchmark for the comparison. The lower and upper bounds for the hit probability of the HCP-A placement in (3.27) of Prop. 7 is also shown. Compared to the GCP model in [60], the HCP-A and HCP-B placement models provide higher cache hit probabilities, which we demonstrate next. From Fig. 3.5, we observe that the average cache hit probability for all cases improves with λ_t, GCP improves with increasing λ_t, and the performance gap between the HCP models and the GCP is higher at high λ_t. The respective cache hit gains of the HCP-B and HCP-A models over GCP can be up to 30% and 37%, and the gain of HCP-A over MPC is 50% for this particular example.

Cache hit rate with respect to $\mathrm{R_{D2D}}$. The numerical comparison for the GCP and the HCP-A models for varying $\mathrm{R_{D2D}}$ and fixed λ_t in Example 2 is tabulated in Table 3.2. Now, we illustrate the dependence of the average cache hit probability of different cache placement models on the communication radius $\mathrm{R_{D2D}}$ in Fig. 3.6. The lower and upper bounds for the hit probability of the HCP-A placement in (3.27) of Prop. 7 is also shown. For high $\mathrm{R_{D2D}}$, both models perform similarly. However, when $\mathrm{R_{D2D}}$ is small, HCP performs better because it exploits the spatial diversity of the D2D caches. For small $\mathrm{R_{D2D}}$, feasible for the D2D regime, MHC-inspired approaches are a

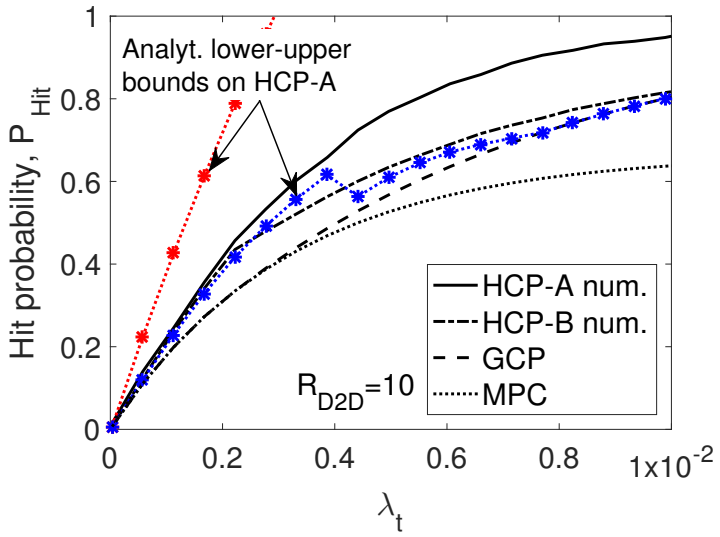


Figure 3.5: Maximum cache hit probabilities of the MPC, GCP and HCP model for varying D2D node intensity λ_t.

better alternative[5].

Cache utilization ratio. As discussed in Proposition 8, the HCP placement model causes underutilization of the caches. We numerically investigate the cache utilization ratio for the HCP-A sufficient condition given in Prop. 9, which is shown in Fig. 3.7. As $\mathrm{R}_{\mathsf{D2D}}$ increases, the utilization drops because there will be more D2D caches around the typical receiver and hence, the required number of cache slots decreases. For small λ_t, the values taken by $\lambda_{\mathsf{HCP\text{-}A}}(m)$ are small that yields a low utilization ratio when $\mathrm{R}_{\mathsf{D2D}}$ is large,

[5]One disadvantage of the HCP-B model is that the excluded files' cache space is not reused, which can be resolved by jointly assigning marks. Therefore, we need to vectorize the marks to jointly determine the set of cached files and to avoid the problems caused by cache underutilization or overuse. The calculation of the cache underutilization or the overuse probability is left as future work.

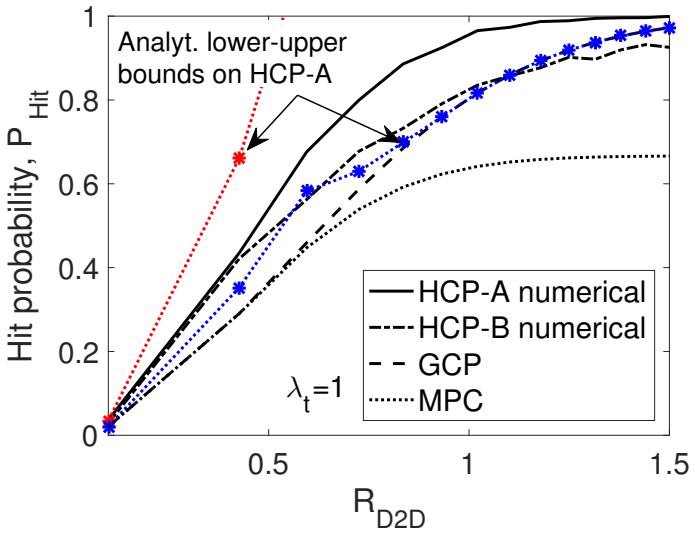


Figure 3.6: Maximum cache hit probabilities of the MPC, GCP and HCP models for varying communication radius.

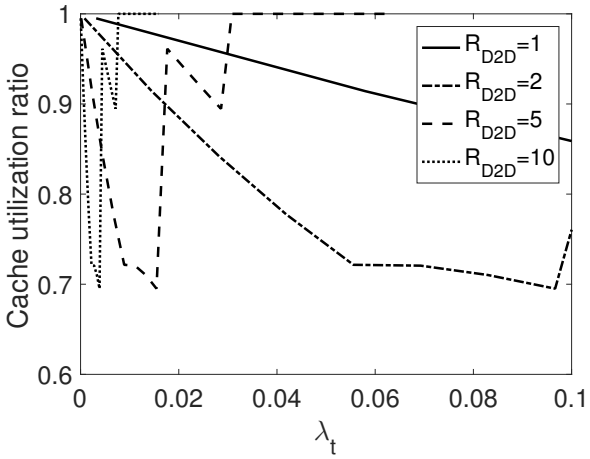


Figure 3.7: The cache underutilization (follows from Prop. 9 of Chapter 3).

which follows from (3.29). However, the utilization can be improved by jointly determining the values of $\lambda_{\text{HCP-A}}(m)$ and R_{D2D}.

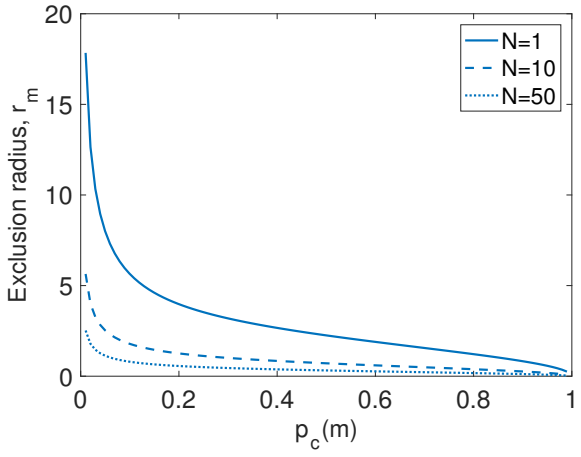


Figure 3.8: Characterization of the exclusion radii of HCP-B for $N = [1, 10, 50]$ and $\mathrm{R}_{\mathsf{D2D}} = 1$ as a function of $p_c(m)$.

Cache size. The performance of the independent and the HCP models is mainly determined by the cache size. Hence, the analysis boils down to finding the critical cache size that determines which model outperforms the other in terms of the hit probability under or above the critical size. In Fig. 3.8, we show the trend of the optimal exclusion radius r_m of the HCP-B model with respect to the caching pmf $\mathrm{p}_{\mathrm{c,X}}(m)$. As we expect from (3.29), the exclusion radius r_m decays with the popularity and the cache size N. Note that the HCP model compensates the small cache size at the cost of communication radius.

Refinement to soft-core models. The thinning leading to the MHC process can be refined such that higher intensities $\lambda_{\mathsf{HCP\text{-}A}}$ are possible [64, Ch. 5.4], at the price of more complicated algorithms [103] and [104]. For refinement of the hard-core models, models based on Gibbs point processes (GPPs)

with repulsive potentials can be developed to generate soft-core[6] placement models [61, Ch. 18]. The study of soft-core models inspired from GPPs, and the maximum caching gain due to the spreading of content in geographic settings is left as future work.

3.8 Summary

We proposed spatially correlated content caching models to maximize the hit probability by incorporating strategies to enable spatial diversity, e.g., spatially exchangeable cache model, and hard-core placement strategies that capture the pairwise interactions to enable spatial diversity.

Our findings on spatial content caching suggest that the following design insights should enable more efficient caching models for D2D*-enabled wireless networks:*

Repulsive cache placement. Negatively correlated content placement rather than independent placement is required to maximize the cache hit probability. Due to the isotropy of the PPP process, we contemplate a rotation invariant caching model. To satisfy negative spatial correlation, geographical separation of the content within the neighborhood of a typical receiver is required. Thus, in caching protocol design, it is important to incorporate an exclusion region around each cache, such that nodes in this region are not allowed to cache simultaneously. We show that high cache hit rates in a PPP

[6] In the case of a soft-core point process, thinning is stronger the closer point pairs of the initial PPP are, but any pair distance still has non-vanishing probability.

network can be achieved through a MHC-inspired placement model.

Towards soft-core placement models. We analyzed the HCP model, where the exclusions are determined by the hard-core radii. Future studies include more general solutions inspired from the GPP or Ising models capturing the pairwise interactions using soft-core potentials. The shape and scale of the potential should be determined accordingly. The pairwise potential function is promising because it can characterize the spatial and temporal dynamics of the file popularities at different geographic locations adaptively. Hence, the soft-core placement incorporating pairwise correlations can be exploited to improve the cache hit rate. This can can pave the way for the development of spatial cache placement and eviction policies to decide what content to discard, when to discard the content and where (to which neighbor) to relay the content, and provide practical design insights into how to adapt to geographical and temporal changes without compromising the accuracy.

Possible extensions also include hierarchical models for content delivery [97], multi-hop routing to improve the hit probability, distributed scheduling and content caching with bursty arrivals and delay constraints, and smoothing the cellular traffic by minimizing the peak-to-average traffic ratio with D2D transmissions.

Chapter 4

Resource Allocation for Content Caching in D2D-Enabled Cellular Networks

Content caching is the key enabling design technique for offloading from the cellular infrastructure to decentralized device-to-device (D2D) communication. Caching aims to maximize the probability that the desired content can be found in a nearby device, i.e., the local hit rate. Due to potentially high density of devices, novel ways of scheduling concurrent D2D transmissions are required in order to avoid interference and optimize the caching performance[1].

Power control is an effective approach to handle interference. Different power control algorithms to either optimize resource utilization for D2D have been proposed in [107], or to maximize the coverage probability of the cellular link as detailed in [108]. Interference analysis in carrier sense multiple access (CSMA) wireless networks is implemented in [109]. A synchronous P2P signaling and a concomitant scheduling protocol is designed in [6] that enables efficient channel aware spatial resource allocation and achieves significant gains over a CSMA system.

[1]This chapter will be published in [105], [106]. I am the primary author of these works. Coauthor Dr. Mazin Al-Shalash has provided many valuable discussions and insights to this work, and Dr. Jeffrey G. Andrews is my supervisor.

Distributed solutions have been proposed for scalability and to improve different utility metrics. For example, a Gibbs sampling approach for scheduling to minimize the total interference and the delay is proposed in [110], and to learn how to optimize the placement to maximize the cache hit rate of cellular networks is analyzed in [111]. Femtocaching using small cell access points, i.e., helpers, to minimize total delay is studied in [59].

Content placement and delivery should be jointly designed to maximize the offloading gain of D2D caching [112]. Fair traffic association is required to balance the total load among the nodes. When the traffic demand and the location of caches are regular enough, the strategy of selecting the nearest cache can actually be close to optimal, as demonstrated in [113]. If the locations are not regular, load balancing can result in a maximum load of order $\Theta(\log \log n)$, where n is the number of servers and requests, as shown in [114]. This is an exponential improvement in a maximum load compared to the scheme which assigns each request to the nearest available replica. Our distributed solution is motivated from load balancing in the context of caching, which also captures the local demand popularity and cache configurations, unlike prior work.

We consider a spatial caching network in which the D2D receivers and the potential transmitters are uniformly distributed. We assume the content placement configuration of the potential transmitters as given. For this system model, we propose a totally distributed scheduling policy for the potential transmitter process by capturing the local demand profile of the receivers, the spatial distribution and the availabilities of the transmitters, with the objective

of maximizing the spectral efficiency.

Our model is an auction-based dynamic scheduling policy in which each receiver bids on the set of potential transmitters in its communication range. A fraction of the transmitters are jointly scheduled based on an on-off scheduling strategy given a medium access probability (MAP). The scheduling is not done uniformly at random, rather it depends on the cache configurations. The proposed solution captures (i) the cache configurations, (ii) the signal-to-interference-and-noise-ratio (SINR) coverage probability conditioned on the potential transmitter process, and (iii) the file popularity via the distribution of the local requests. We demonstrate the performance of our model for a given configuration in terms of the average rate per user under independent reference model (IRM) traffic, then test its robustness under different popularity profiles.

4.1 System Model

We envision a D2D caching network model in which the locations of the receiver process Φ_r and the potential transmitter process Φ are assumed to form a realization of two independent homogeneous two-dimensional spatial Poisson point process (PPPs) with densities λ_r and λ_t, respectively.

We assume that the catalog size of the network is M and $\mathcal{M} = \{1, \ldots, M\}$ denotes the set of files. Each transmitter has a cache of finite size $N < M$. Each receiver makes a file request based on a general popularity distribution over the set of the files. The document requests are modeled according to the Independent Reference Model (IRM), and the popularity distribution is

modeled by the pmf $p_r(n)$, $n \in \mathcal{M}$.

We have the following additional assumptions in the model.

- Consider a snapshot of the set of D2D nodes at a tagged time slot where a subset of the potential transmitters Φ simultaneously access the channel given a MAP p_A.
- The cache configuration is given, i.e., at a given snapshot the set of cached files is revealed to the users.
- Each receiver makes a request for one file randomly sampled from p_r, and can associate with any transmitter within its communication range.
- A transmission is successful only if the received SINR is above the threshold T, given that the potential transmitter is on and it caches the desired file.

High level summary. Each receiver is allowed to communicate with any potential transmitter in its communication range and needs to choose a link. Receiver u is associated with potential transmitter x, estimates the link SINR, and bids on x if the desired content is available in x's cache. The values of the receiver bids are reported to potential transmitter x, and x computes the cumulated sum of these variables taken on all users in its cell. The potential transmitter x then reports the value of the bid sum to other potential transmitters in its contention range. Given the accumulated bids of all po-

tential transmitters, the exclusion (or contention) range[2] and the MAP, the algorithm determines the set of active transmitters.

Let $\tilde{\Phi} = \{(x, m_x, \mathbf{P}_x)\}$ be an independently marked PPP with intensity λ_t, where i) $\Phi = \{x\}$ denotes the locations of potential transmitters, ii) $\{m_x\}$ are the marks of $\tilde{\Phi}$, and iii) $\mathbf{P}_x = (P_x^y : y)$ denotes the virtual power emitted by node x to node y provided it is authorized by the MAC mechanism. The random variables $\mathbf{P}_x$ are i.i.d., exponential with mean μ^{-1}.

Definition 6. Neighborhoods. *The neighborhood system on* Φ *is the family* $N = \{\mathcal{N}(x)\}_{x\in\Phi}$ *of subsets of* Φ *such that for all* $x \in \Phi$*, we have* $x \notin \mathcal{N}(x)$*, and* $z \in \mathcal{N}(x) \implies x \in \mathcal{N}(z)$*. The subset* $\mathcal{N}(x)$ *is called the neighborhood of node* x*.*

For $x \in \Phi$, let the neighbors of node x be

$$\mathcal{N}(x) = \{(y, m_y, \mathbf{P}_y) \in \tilde{\Phi} : P_y^x / l(|x - y|) \geq P_0, y \neq x\}, \tag{4.1}$$

i.e., the nodes in its contention domain. If we only consider path loss and no fading, the received signal at the boundary should be larger than the threshold, equivalent to $D = (\mu P_0)^{-1/\alpha}$ for a fixed transmit power of μ^{-1}. Thus, $P_y^x / l(|x - y|) \geq P_0$ will be equivalent to $y \in B_x(D)$, where $B_x(D)$ is a ball centered at x with contention radius D.

[2]If a transmitter has other transmitters in its contention domain, its channel capacity will be a fraction of the medium capacity due to sharing of resources.

Symbol	Definition
PPP distributed D2D receivers; potential transmitters	Φ_r; Φ
Medium access probability; set of active transmitters	p_A; Φ_t
Density of receivers; density of potential transmitters	λ_r; λ_t
Intensity of the set of active transmitters	$\lambda = p_A\,\lambda_\mathrm{t}$
Signal-to-Noise-Ratio (SNR) at the receiver	σ^{-2}
Signal-to-Interference-and-Noise-Ratio (SINR) threshold	T
Ball centered at node x with radius R	$B_x(R)$
D2D radius; exclusion radius for the Matérn CSMA	$\mathrm{R}_{\mathsf{D2D}}$; D
File request distribution	$p_r \sim \mathrm{Zipf}(\gamma_\mathrm{r})$
Total number of files; cache size; set of all files	M; N; $\mathcal{M}$
File requested by $u \in \Phi_r$; cache config. of $x \in \Phi$	c_u; $\mathcal{C}_x$
Path loss exponent; power law path loss function	α; $l(r) = r^{-\alpha}$
Accumulated bid of transmitter x	$\mathcal{B}_\phi(x)$
A realization of the point process Φ with K nodes	ϕ
Voronoi cell of x with respect to the point measure ϕ_t	$\mathcal{V}_x(\phi_\mathrm{t})$
$M \times K$ binary matrix denoting the cache states	b
j^{th} column of b	$b_{:,j}, x_j \in \phi$
Indicator of availability of file m in cache $x_j \in \phi$	$b_{m,j} = 1$
Set of all feasible cache states	B
On-off powers of potential transmitters	$P_j = 1_{x_j \in \phi_\mathrm{t}}$
A configuration with set of devices ϕ and a cache state matrix b	$\boldsymbol{z}(\{P_j, \mathbf{b}_j\})$
Cache hit rate averaged over the set of requests given a configuration $\boldsymbol{z}$	$\mathsf{R}_{\mathsf{Hit}}(\boldsymbol{z})$

Table 4.1: Notation for Chapter 4.

The medium access indicators $\{e_x\}_x$ are additional dependent marks of the points of Φ as follows:

$$e_x = \mathbb{1}\left(\forall_{y \in \mathcal{N}(x)} m_x < m_y\right). \tag{4.2}$$

The set of transmitters retained by CSMA as a non-independent thin-

ning of the PPP Φ, and denoted by

$$\Phi_t = \{x \in \Phi | e_x = 1\}. \tag{4.3}$$

The probability of medium access of a typical node equals $p_A = \mathbb{E}^0[e_x]$, where $\mathbb{E}^0$ is the expectation with respect to Φ's Palm probability $\mathbb{P}^0$; i.e., $\mathbb{P}^0(\Phi(\{0\}) \geq 1) = 1$ [61, Ch. 4].

Next, by incorporating the SINR coverage characteristics in a realistic D2D network setting with contention prevention provided by the MHC-II model, we envisage a bidding-aided scheduling policy in Sect. 4.2.

4.2 Bidding-Aided Policy for User Associations

Using the potential transmitter model just described, the potential received SINR of a receiver located at u covered by $x \in \Phi$ is expressed as

$$\mathsf{SINR}_{x,u} = \frac{P_{xu} l(|x-u|)}{\sigma^2 + \sum_{z \in \Phi \setminus \{x\}} P_{zu} l(|z-u|)}, \tag{4.4}$$

where $r = |x-u|$ is the distance between the potential transmitter located at $x \in \Phi$ and the receiver u, and for a fixed path loss exponent α, $l(r) = r^{-\alpha}$ under OPL3 [61, Ch. 2.3], and r and $r_z = |z-u|, z \in \Phi$ denote the distance between the potential transmitter and the receiver, and the interferers and the receiver, respectively, and σ^2 is the noise power at the receiver side. Similarly, $\{P_{zu}\}_{z \in \Phi}$ are random variables that denote the on-off powers of potential transmitters, i.e.,

$$P_{zu} = \begin{cases} 1, & z \in \Phi_t \\ 0, & z \in \Phi \setminus \Phi_t \end{cases},$$

where Φ_t is a repulsive point process that models the retained process of transmitters. The procedure to decide the set of retained and silent transmitters will be detailed in this section.

We develop a bidding-based user association algorithm such that receivers are associated in a way to maximize the "local cache hit probability". We introduce an on-off distributed scheduling method with coordination between the neighboring transmitters for the D2D caching framework[3]. For a fixed probability of medium access[4], the bidding algorithm determines which links to activate by capturing the matchings between the availability of the caches and the local demand.

Each receiver $u \in \Phi_r$ bids on the potential transmitters $x \in \Phi$ in its range $\mathrm{R}_{\mathsf{D2D}}$ based on their virtual SINR coverage probability characteristics. Each $x \in \Phi$ accumulates bids from all receivers. Because the local demand and the coverage characteristics will be similar, the transmitters located at similar geographic locations collect similar bids. Upon the assignment of the bids of all the potential transmitters, $x \in \Phi$ is scheduled if it has the highest bid inside a circular exclusion region $B_x(D)$. Hence, the process of retained transmitters Φ_t will be obtained as a dependent thinning of Φ, in contrast with the Matérn hard-core (MHC) model where the potential transmitters are assigned i.i.d. marks. We next discuss the technical details of the bidding approach.

[3]On-off scheduling requires the CSI knowledge about the direct link between the transmitter and its corresponding receiver [108]. We only consider long term CSI (ignore fading).

[4]Only a certain fraction of transmitters is to be activated to control interference and provide the D2D users with high spectral efficiency.

4.2.1 Accumulated Bid of a Potential Transmitter

For given realizations ϕ of Φ, and ϕ_r of Φ_r, the total bid collected at a potential transmitter $x \in \phi$ is determined using the following expression:

$$\mathcal{B}_\phi(x) = \sum\nolimits_{u \in \mathcal{U}_x} p_r^x(c_u) \mathbb{P}(\mathsf{SINR}_{x,u} > T),\ x \in \phi, \tag{4.5}$$

where for the general coverage model with noise and interference, we denote by

$$\mathcal{U}_x = \{u \in \phi_r \cap B_x(\mathrm{R}_{\mathsf{D2D}}) | x \in \phi, c_u \in \mathcal{C}_x\} \tag{4.6}$$

is the set of receivers that bid on the potential transmitter x.

Note that (4.5) is a weighted sum of the virtual SINR coverage distributions of the set of receivers inside the coverage region with radius $\mathrm{R}_{\mathsf{D2D}}$ of the potential transmitter x. The parameter c_u (sampled i.i.d. from p_r) denotes the index of the file requested by receiver u, and $\mathcal{C}_x$ denotes the set of files available in the cache of transmitter $x \in \Phi$, i.e., the cache configuration of x. The local request distribution observed at $x \in \phi$, i.e., the request distribution conditioned on the cache configuration of $x \in \phi$, is given as

$$p_r^x(m) = |\mathcal{U}_x(m)|/|\mathcal{U}_x|, \quad x \in \phi,\ m \in \mathcal{C}_x, \tag{4.7}$$

where

$$\begin{aligned} |\mathcal{U}_x(m)| &= \sum\nolimits_{u \in \phi_r \cap B_x(\mathrm{R}_{\mathsf{D2D}})} \mathbb{1}(c_u = m) \mathbb{1}(m \in \mathcal{C}_x), \quad x \in \phi \\ |\mathcal{U}_x| &= \sum\nolimits_{u \in \phi_r \cap B_x(\mathrm{R}_{\mathsf{D2D}}) | x \in \phi} \mathbb{1}(c_u \in \mathcal{C}_x) \end{aligned} \tag{4.8}$$

are the number of receivers in the coverage of x that request file $m \in \mathcal{C}_x$, and the cardinality of the set of users associated to $x \in \phi$, respectively.

The bidding formulation in (4.5) captures the

- cache availability through the conditioning on the set $\mathcal{U}_x$,
- SINR coverage conditioned on the potential transmitter process ϕ, and
- file popularity through the local request distribution p_r^x as defined in (4.7).

Using this bidding formulation, we analyze the bidding algorithm in Sect. 4.2.2 to determine the set of retained transmitters ϕ_t. We illustrate the bidding algorithm in Fig. 4.1.

Consider the network setup in Fig. 4.1-(a) with the set of potential transmitters and receivers. In Fig. 4.1-(b), we show the interactions between the potential transmitter centered at origin, where the solid (dashed) circle shows the communication (exclusion) range. A receiver can bid on the potential transmitter only if it is in the communication range. Fig. 4.1-(c) shows the system-level interactions that might overlap depending on the potential transmitter locations. Fig.4.1-(d) shows the set of retained transmitters selected based on the bidding algorithm.

Similar to a hard-core process, ϕ_t has an exclusion radius of D possibly different from the communication radius $\mathrm{R}_{\mathsf{D2D}}$ that will be determined in Sect. 4.2.3.

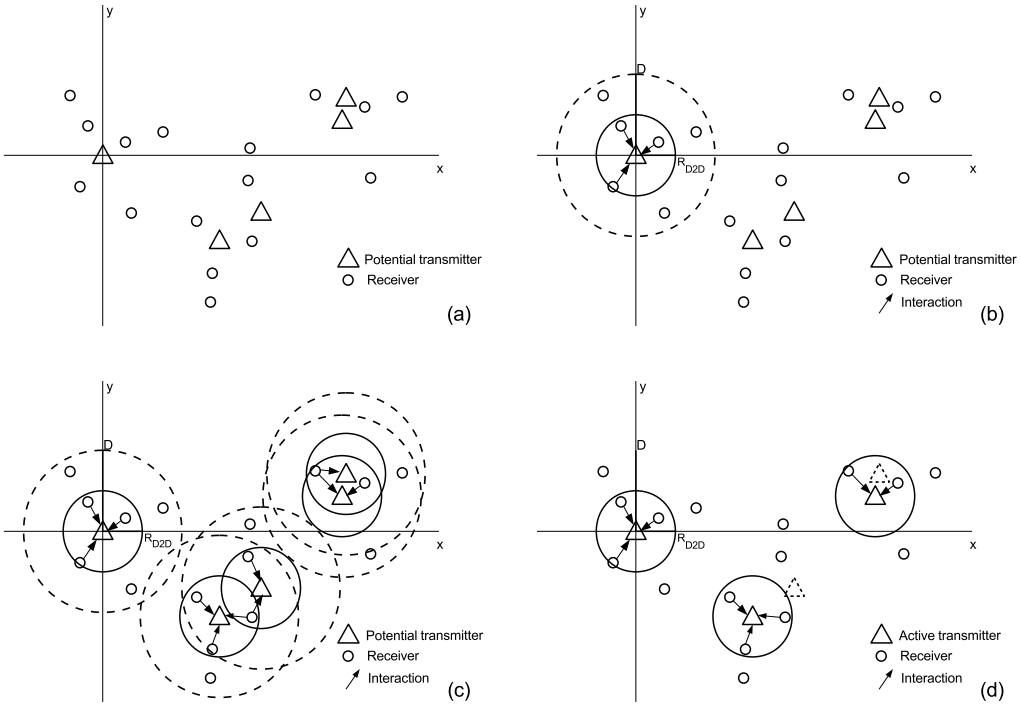


Figure 4.1: A visualization of the bidding algorithm on the receiver and the potential transmitter processes.

The cardinality of receivers that bid on $x \in \phi$, i.e., $|\mathcal{U}_x|$, is distributed as Poisson$(\lambda_r^x \pi \, \mathrm{R}_{\mathsf{D2D}}^2)$, where the intensity of receivers that bid on transmitter x is given by $\lambda_r^x = \lambda_r \sum_{m \in \mathcal{C}_x} p_r(m)$. Hence, the average number of receivers associated to $x \in \phi_t$ is given by $\mathbb{E}[|\mathcal{U}_x|] = \lambda_r^x \pi \, \mathrm{R}_{\mathsf{D2D}}^2$, and the distribution of $|\mathcal{U}_x|$ satisfies

$$\mathbb{P}(|\mathcal{U}_x| = n) = \exp\left(-\lambda_r^x \pi \, \mathrm{R}_{\mathsf{D2D}}^2\right) \frac{(\lambda_r^x \pi \, \mathrm{R}_{\mathsf{D2D}}^2)^n}{n!}. \tag{4.9}$$

The rest of this section is mainly devoted to the special case of the homogenous PPP approximation for the bidding algorithm and its distributional

characteristics. Note that this does not imply that the thinning of Φ is done independently. We will detail how the results will differ for various bidding models in Sect. 4.3.

4.2.2 Analysis of Bidding with Homogeneous PPP Transmitters

The process of transmitters Φ_t arranged according to some homogeneous PPP of intensity $\lambda = p_A\,\lambda_{\mathrm{t}}$ in the Euclidean plane. For the general SINR regime, the probability of coverage of a typical randomly located receiver in the general cellular network model, where the transmitters are arranged according to some homogeneous PPP is evaluated in [81]. The coverage probability of a user u (assuming that the user is associated to the nearest transmitter) is given as

$$\mathbb{P}[\mathsf{SINR} > T] = e^{-\mu T\sigma^2/l(r)}\mathcal{L}_{I_r}(\mu T/l(r)), \tag{4.10}$$

where r denotes the distance from the receiver to the serving transmitter, and $\mathcal{L}_{I_r}(s)$ is the Laplace transform of the interference and is given by

$$\mathcal{L}_{I_r}(s) = \exp\Big(-\pi\lambda\int_{r^2}^{\infty}\frac{1}{1+\mu s^{-1}t^{\alpha/2}}\mathrm{d}t\Big).$$

Hence, we can compute $\mathcal{L}_{I_r}(\mu T/l(r))$ as

$$\mathcal{L}_{I_r}(\mu T/l(r)) \overset{(a)}{=} \exp\Big(-\pi\lambda\rho(T,\alpha)r^2\Big), \tag{4.11}$$

where (a) follows from employing a change of variables $z = t/(T^{2/\alpha}r^2)$, where $\rho(T,\alpha) = T^{2/\alpha}\int_{T^{-2/\alpha}}^{\infty}\frac{1}{1+z^{\alpha/2}}\mathrm{d}z$,

Cumulated bid (4.5) of potential transmitter $x \in \phi$ can be rewritten using the SINR distribution given in (4.10) as

$$\begin{aligned}\mathcal{B}_\phi(x) &= \sum_{u\in\mathcal{U}_x} p_r^x(c_u) e^{-\mu T\sigma^2/l(r_{xu})} \mathcal{L}_{I_{r_{xu}}}(\mu T/l(r_{xu})),\\ &= \sum_{u\in\mathcal{U}_x} p_r^x(c_u) \exp\left(-\mu T\sigma^2/l(r_{xu}) - \pi\lambda\rho(T,\alpha) r_{xu}^2\right)\\ &\overset{(a)}{=} \sum_{u\in\mathcal{U}_x} p_r^x(c_u) \left(1 - \mu T\sigma^2/l(r_{xu}) - \pi\lambda\rho(T,\alpha) r_{xu}^2\right), \end{aligned} \tag{4.12}$$

where $r_{xu} = |x-u|$, and (a) is required for analytical tractability. The total bid expression in (4.12) is a random variable as a function of the local request distribution $p_r^x(c_u)$ of $u \in \mathcal{U}_x$. Conditioning on the value of $|\mathcal{U}_x|$, $u \in \mathcal{U}_x$ are independently and uniformly distributed in the ball $B_x(\mathrm{R}_{\mathsf{D2D}})$.

The spatial distribution of the bids can be calculated using a similar approach to the one proposed in [115]. In Theorem 6, we give the moment-generating function (MGF) of $\mathcal{B}_\phi(x)$, which fully characterizes the bid distribution.

Theorem 6. *The MGF of the cumulated bid of transmitter x, i.e., $\mathcal{B}_\phi(x)$ expression in (4.12), is given as*

$$M_{\mathcal{B}_\phi(x)}(t) = \exp\left(\lambda_r^x \pi\, \mathrm{R}_{\mathsf{D2D}}^2 (a(t) - 1)\right), \tag{4.13}$$

where $a(t)$ is given by

$$a(t) = \frac{1}{\mathrm{R}_{\mathsf{D2D}}^2} \int_0^{\mathrm{R}_{\mathsf{D2D}}^2} \exp\left(t p_r^x(c_u)\left(1 - \gamma/l(v^{1/2}) - \beta v\right)\right) \mathrm{d}v, \tag{4.14}$$

where the parameters are given as $\gamma = \mu T\sigma^2$ and $\beta = \pi\lambda\rho(T,\alpha)$.

Proof. See Appendix C.1. □

Two special cases of Theorem 6, i.e., the noise-limited regime, $I \to 0$, and the interference-limited regime, $\sigma^2 \to 0$, can be obtained by evaluating the integral in (4.14) and incorporating the different SINR regimes in (4.12), which are given next.

Corollary 4. *The bid distribution for the noise-limited regime is characterized by (4.13), where $a(t)$ is given as*

$$a(t) = \frac{e^t p_r^x(c_u)}{\mathrm{R}_{\mathsf{D2D}}^2} \frac{2/\alpha}{(t p_r^x(c_u)\gamma)^{\frac{2}{\alpha}}} \left[\Gamma\left(\frac{2}{\alpha}\right) - \Gamma\left(\frac{2}{\alpha}, \frac{t p_r^x(c_u)\gamma}{l(\mathrm{R}_{\mathsf{D2D}})}\right) \right], \; I \to 0. \quad (4.15)$$

Corollary 5. *The bid distribution for the interference-limited regime is characterized by (4.13), where $a(t)$ is given as*

$$a(t) = \frac{e^t p_r^x(c_u)}{\mathrm{R}_{\mathsf{D2D}}^2} \frac{1}{t p_r^x(c_u)\beta} [1 - e^{-t p_r^x(c_u)\beta \, \mathrm{R}_{\mathsf{D2D}}^2}], \; \sigma^2 \to 0. \quad (4.16)$$

The bid-based approach can be generalized using more general processes. Some other examples include non-homogeneous PPP approximation for MHC models [116], or a modified MHC model [117], or a more general non-homogeneous PPP approximation for the Matérn CSMA [61, Ch. 18.5]. In this section, we only discussed the bidding algorithm under the PPP approximation. In Sect. 4.3, we also discuss the non-homogeneous models for the bidding algorithm.

4.2.3 Communication Range versus Exclusion Range

Given a contention-based model, the interference measured at the typical point depends on the range of the contention domain. Hence, the range at which the communication is successful, i.e., $\mathsf{SINR} \geq T$, is determined by the exclusion radius. Using the SINR expression in (4.4), we rewrite the SINR for noise-limited and interference-limited regimes as follows:

$$\mathsf{SINR} = \begin{cases} hl(r)/\sigma^2, \ I \to 0 \\ hl(r)/\bar{I}, \ \sigma^2 \to 0, \end{cases} \tag{4.17}$$

where h is the exponential channel gain with parameter μ. The communication range is defined by $\mathrm{R}_{\mathsf{D2D}}$ such that $r \leq \mathrm{R}_{\mathsf{D2D}} \implies \mathsf{SINR} \geq T$.

Using (4.17), and neglecting the small scale Rayleigh fading variability, it is easy to note that in the noise-limited regime, there is a one-to-one mapping between T and $\mathrm{R}_{\mathsf{D2D}}$. Unlike the noise-limited regime, $\mathrm{R}_{\mathsf{D2D}}$ for the interference-limited regime is variable. To ease the analysis in the interference-limited regime, we approximate the interference I by its mean $\bar{I}$. Hence, one can derive the communication range

$$\mathrm{R}_{\mathsf{D2D}} = \begin{cases} (\mu T \sigma^2)^{-1/\alpha}, \ I \to 0, \\ (\mu T \bar{I})^{-1/\alpha}, \ \sigma^2 \to 0, \end{cases}$$

respectively for the noise- and interference-limited regimes.

We benefit from a very useful approximation to characterize $\bar{I}$, which is first suggested in [116]. The excess interference ratio (EIR) as defined in [116] is the mean interference measured at the typical point of a stationary hard-core point process of intensity λ with minimum distance D relative to

the mean interference in a Poisson process of intensity $\lambda(r) = \lambda \mathbf{1}_{[D,\infty)}(r)$. Their analysis shows that the excess interference ratio for Matérn processes of type II (MHC-II) never exceeds 1 dB. Thus, using a modified path loss law $\tilde{l}(r) = l(r)\mathbf{1}_{r>D}$, the mean interference is approximated as

$$\bar{I} \approx \lambda \int_{\mathbb{R}^2} \tilde{l}(|y|)\mathrm{d}y = 2\pi\lambda \int_D^\infty r^{-\alpha+1}\,\mathrm{d}r = \frac{2\pi\lambda}{\alpha-2} D^{2-\alpha},$$

using which $\mathrm{R}_{\mathsf{D2D}}$ can be approximated as a function of the exclusion radius D as the interference varies.

In addition to the homogeneous PPP model, there are different methods of estimating the SINR for the thinned transmitter process. For example, exploiting the non-homogeneous PPP model, the intensity of the transmitters becomes $\Lambda(x) = \lambda_{\mathrm{t}}\, k(x)$ [61, Ch. 18.5], where $k(x)$ is the two-point Palm probability that two points of Φ separated by distance r are both retained [64, Ch. 5.4]. Another approach is to utilize the modified Matérn hard-core model proposed in [117]. Technical discussions of these models will be given next.

4.3 Generalized Bidding Models

In this section, we consider more general bidding algorithms using different spatial distributions to model the locations of active transmitters.

For the analytical approximations, we exploit the special case of the Matérn hard-core type-II (MHC-II) process, where $\{m_x\}$ are i.i.d. marks over

$x \in \Phi$, uniformly distributed on $[0, 1]$. The first-order and second-order moment characteristics of the MHC-II process are given as follows.

Definition 7. *The intensity of active transmitters of the* MHC-*II model* Φ_{M} *equals*

$$\lambda_{\mathsf{MHC}} = p_A \lambda_t = \frac{1 - \exp(-\bar{\mathcal{N}})}{\pi D^2}, \tag{4.18}$$

where p_A *is the probability of medium access and* $\bar{\mathcal{N}} = \lambda_{\mathrm{t}}\, \pi D^2$ *is the expected number of neighbors of the typical node.*

The second-order product density of the MHC-*II process* Φ_{M} *is given by [64, Ch. 5.4], [116] as*

$$\begin{aligned} \rho^{(2)}(r) &= {\lambda_{\mathrm{t}}}^2 k(r) \\ &= \begin{cases} \lambda_{\mathsf{MHC}}^2, & r \geq 2D \\ \frac{2V_D(r)[1-\exp(-\lambda_{\mathrm{t}}\, \pi D^2)] - 2\pi D^2[1-\exp(-\lambda_{\mathrm{t}}\, V_D(r))]}{\pi D^2 V_D(r)[V_D(r) - \pi D^2]}, & D < r < 2D \\ 0, & r < D \end{cases}, \end{aligned} \tag{4.19}$$

where $k(r)$ *is the two-point Palm probability that two points of* Φ *separated by distance* r *are both retained [64, Ch. 5.4], and* $V_D(r) = 2\pi D^2 - 2D^2 \cos^{-1}\left(\frac{r}{2D}\right) + r\sqrt{D^2 - \frac{r^2}{4}}$ *denotes the area of the union of two circles having radius* D *and separated by distance* r*.*

4.3.1 Non-Homogeneous PPP Approximation for MHC

Using the first- and second-order statistics of the MHC model given in (4.18) and (4.19), respectively, we can approximate the MHC with a non-

homogeneous PPP model. In that case, the intensity of the transmitters becomes $\Lambda(x) = \lambda_{\mathrm{t}}\, k(x)$ [118, Ch. 18.5], where $k(x)$ is given in (4.19). Hence, the Laplace transform of the interference for the non-homogeneous PPP can be given as

$$\mathcal{L}_{I_r}^{\mathsf{Nppp}}(s) = \exp\Big(- \lambda_{\mathrm{t}} \int_0^{\infty} \int_0^{2\pi} \frac{\tau k(\tau)}{1 + T^{-1} l(r)/l(v)} \mathrm{d}\theta \mathrm{d}\tau \Big), \tag{4.20}$$

where $v = \sqrt{D^2 + r^2 - 2Dr\cos(\theta)}$ as can be seen from Fig. 4.2.

Non-Homogeneous PPP-based Bidding Algorithm. Using the first- and second-order statistics of the MHC model to approximate it by a non-homogeneous PPP and using its Laplace transform given in (4.20), and letting $s_{xu} = \mu T / l(r_{xu})$, we can derive the accumulated bid $\mathcal{B}_\phi$ given in (4.5) of potential transmitter $x \in \phi$ as follows:

$$\begin{aligned} \mathcal{B}_\phi^{\mathsf{Nppp}} &= \sum_{u \in \mathcal{U}_x} p_r^x(c_u) e^{-\mu T \sigma^2 / l(r_{xu})}\, \mathcal{L}_{I_{r_{xu}}}^{\mathsf{Nppp}}(\mu T / l(r_{xu})) \\ &\geq \sum_{u \in \mathcal{U}_x} p_r^x(c_u) \left(1 - \frac{\mu T \sigma^2}{l(r_{xu})} - \lambda_{\mathrm{t}} \int_0^{\infty} \int_0^{2\pi} \frac{\tau k(\tau)}{1 + T^{-1} l(r_{xu})/l(v_{xu})} \mathrm{d}\theta \mathrm{d}\tau \right), \end{aligned} \tag{4.21}$$

where $r_{xu} = |x - u|$ is the distance from the receiver u to the serving transmitter x.

4.3.2 A Modified MHC Model

A modified MHC point process for modeling transmitters is proposed in [117], where transmitters are never closer than some given distance D. The modified MHC can be considered as an approximated grid model, where each

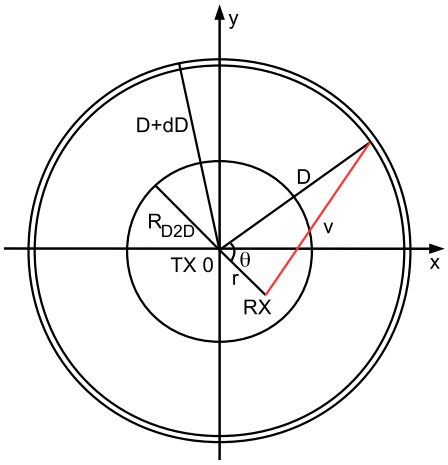


Figure 4.2: Illustration of the coverage area of TX located at 0. The receiver is located at a distance r from the TX. The shortest distance between receiving users and interfering TXs is denoted by v.

transmitter has a coverage radius $\mathrm{R_{D2D}}$ and other interfering transmitters are randomly deployed outside the coverage area. The users are located according to a stationary point process which is independent of Φ, and each user is associated with its closest transmitter.

The proposed model in [117] approximates the point process of the transmitters to three joint processes, illustrated in Fig. 4.2.

- Each user is in the coverage area of a transmitter, denoted by TX 0 (i.e., there exists one transmitter whose distance from the user is less than $\mathrm{R_{D2D}}$);
- Outside the circle whose center is TX 0 and radius is D, denoted by $B_0(D)$, the other transmitters are deployed as a homogeneous PPP Φ_{r_1} with intensity λ_t;

- In the ring area whose center is TX 0 with inner radius D and outer radius $D + \mathrm{d}D$, where $dD \to 0$, the transmitters are deployed as a PPP Φ_{r_2} with intensity $\frac{\lambda_{\mathrm{t}} \times (\pi D^2 - \pi \, \mathrm{R}_{\mathsf{D2D}}^2)}{2\pi D \mathrm{d}D}$. To make the density of transmitters uniform, we have $\pi \, \mathrm{R}_{\mathsf{D2D}}^2 \, \lambda_{\mathrm{t}} = 1$ and thus $\mathrm{R}_{\mathsf{D2D}} = \sqrt{\frac{1}{\pi \, \lambda_{\mathrm{t}}}}$. Therefore, the intensity in the ring area should be $\frac{D^2 / \, \mathrm{R}_{\mathsf{D2D}}^2 - 1}{2\pi D \mathrm{d}D}$.

Given the distance r between the serving transmitter and the user, the Laplace transforms of $I_{r,1}$ and $I_{r,2}$, which denote the interference from regions outside $B_0(D)$ and the ring area, respectively, are given as

$$\mathcal{L}_{I_{r,1}}^{\mathsf{Mhc}}(s) = \exp\Big(\int_0^{2\pi} \int_v^{\infty} \frac{-\lambda_{\mathrm{t}} \, \tau}{1 + \mu s^{-1} / l(\tau)} \mathrm{d}\tau \mathrm{d}\theta \Big), \tag{4.22}$$

$$\mathcal{L}_{I_{r,2}}^{\mathsf{Mhc}}(s) = \exp\Big(\frac{1}{2\pi} \int_0^{2\pi} \frac{-(D^2 / \, \mathrm{R}_{\mathsf{D2D}}^2 - 1)}{1 + \mu s^{-1} / l(v)} \mathrm{d}\theta \Big), \tag{4.23}$$

where $v = \sqrt{D^2 + r^2 - 2Dr\cos(\theta)}$. We denote their product by $\mathcal{L}_{\mathrm{I_r}}^{\mathsf{Mhc}}(s) = \mathcal{L}_{I_{r,1}}^{\mathsf{Mhc}}(s) \mathcal{L}_{I_{r,2}}^{\mathsf{Mhc}}(s)$.

Thus, the SINR coverage probability for a randomly selected user is given as

$$\mathbb{P}[\mathsf{SINR} > T] = e^{-\mu T \sigma^2 r^{\alpha}} \, \mathcal{L}_{\mathrm{I_r}}^{\mathsf{Mhc}}(s), \tag{4.24}$$

where $s = \mu T r^{\alpha}$. To calculate the coverage probability averaged over the distribution of the distance between the transmitter and the user, (4.24) needs to be multiplied by $f_R^{\mathsf{Mhc}}(r)$, which denotes the conditional distribution of the distance from receiver to the serving transmitter given that $r < \mathrm{R}_{\mathsf{D2D}}$, i.e., given there is at least one transmitter in the coverage of the receiver, and then

integrated. Hence, the distance distribution is given as

$$f_R^{\mathsf{Mhc}}(r) = \frac{2\pi\,\lambda_{\mathrm{t}}\, r e^{-\,\lambda_{\mathrm{t}}\,\pi r^2}}{1 - e^{-\,\lambda_{\mathrm{t}}\,\pi\,\mathrm{R}_{\mathsf{D2D}}^2}}, \quad 0 \le r \le \mathrm{R}_{\mathsf{D2D}}\,. \tag{4.25}$$

A Modified Hard-Core Model-based Bidding Algorithm. The modified hard-core model proposed in [117] captures the repulsion between the transmitters at the cost of some additional computational effort. Hence, we can develop a better bidding algorithm exploiting the modified hard-core approach.

Using a similar approach as in (4.12), and letting $s_{xu} = \mu T / l(r_{xu})$, we can derive the accumulated bid $\mathcal{B}_\phi$ given in (4.5) of potential transmitter $x \in \phi$ as follows:

$$\begin{aligned}
\mathcal{B}_\phi^{\mathsf{Mhc}} &= \sum_{u \in \mathcal{U}_x} p_r^x(c_u) e^{-\mu T \sigma^2 / l(r_{xu})}\, \mathcal{L}_{\mathrm{I_{r_u}}}^{\mathsf{Mhc}}(s_{xu}) \\
&\ge \sum_{u \in \mathcal{U}_x} p_r^x(c_u) \left(1 - \frac{\mu T \sigma^2}{l(r_{xu})} - \frac{\lambda_{\mathrm{t}}}{2} \int_0^{2\pi} v_{xu}^2 \rho\Big(T l\Big(\frac{v_{xu}}{r_{xu}}\Big), \alpha\Big) \mathrm{d}\theta \right. \\
&\quad \left. - \frac{1}{2\pi} \int_0^{2\pi} \frac{(\frac{D}{\mathrm{R}_{\mathsf{D2D}}})^2 - 1}{1 + T^{-1} l\Big(\frac{r_{xu}}{v_{xu}}\Big)} \mathrm{d}\theta \right),
\end{aligned} \tag{4.26}$$

where $r_{xu} = |x - u|$ is the distance from the receiver u to the serving transmitter x, and $v_{xu} = \sqrt{D^2 + r_{xu}^2 - 2 D r_{xu} \cos(\theta)}$.

4.3.3 Non-homogeneous PPP approximation for the Matérn CSMA

Conditional on the event $0 \in \Phi_{\mathrm{M}}$, and using the non-homogeneous PPP approximation with intensity $\lambda_{\mathrm{t}}\, h$ for the law of $\Phi_{\mathrm{M}} \setminus 0$ given in [118, Ch. 18.5],

i.e., letting $x = 0$ so that $|z| = r$ in (4.4), the probability that a transmitter covers its receiver is given as

$$\mathbb{P}[\mathsf{SINR} > T] \approx \exp\Big(- s\sigma^2 - \lambda_{\mathrm{t}} \int_{\mathbb{R}^+} \int_0^{2\pi} \frac{\tau h(\tau, P_0)}{1 + \mu/(l(\sqrt{\tau^2 + r^2 - 2r\tau\cos(\theta)})s)} \mathrm{d}\theta \mathrm{d}\tau \Big), \tag{4.27}$$

where $s = \mu T / l(r)$, and the function h is defined in [118, Cor. 18.4.3] as

$$h(r, P_0) = \frac{\frac{2}{c(r,P_0) - \bar{\mathcal{N}}} \left(p_A - \frac{1 - e^{-c(r,P_0)}}{c(r,P_0)} \right) (1 - e^{-P_0 \mu / l(r)})}{p_A - e^{-P_0 \mu / l(r)} \left(\frac{1 - e^{-\bar{\mathcal{N}}}}{\bar{\mathcal{N}}^2} - \frac{\exp(-\bar{\mathcal{N}})}{\bar{\mathcal{N}}} \right)}, \tag{4.28}$$

where $\bar{\mathcal{N}} = \lambda_{\mathrm{t}} \pi D^2$, the detection threshold P_0 satisfies $D = (\mu P_0)^{-1/\alpha}$, the medium access probability is given by $p_A = \frac{1 - e^{-\bar{\mathcal{N}}}}{\bar{\mathcal{N}}}$, and the function $c(r, P_0)$ is given as follows:

$$c(r, P_0) = 2\bar{\mathcal{N}} - \lambda_{\mathrm{t}} \int_{\mathbb{R}^+} \int_0^{2\pi} \exp\Big(-P_0 \mu \Big(l^{-1}(\tau) + l^{-1}\big(\sqrt{\tau^2 + r^2 - 2r\tau\cos(\theta)}\big) \Big) \Big) \tau \mathrm{d}\theta \mathrm{d}\tau.$$

In the above, (4.28) denotes given there are two links in the network and one link is active the conditional probability that these links are both active.

The MHC-II model assigns each transmitter a uniformly distributed and i.i.d., and can capture the repulsion among the transmitters, but it has limitations due to (i) no use of the SINR coverage to assign marks and (ii) selection of a user randomly in this area. Therefore, it fails to capture the attraction between the transmitter and receiver pairs in a content caching scenario. This motivates us to use approaches similar to Gibbs fields.

Exploiting the coverage model for the presented in Sect. 4.3.3, the $\mathcal{B}_\phi$ for the non-homogeneous PPP approximation model can be derived in a similar manner as the other models.

In Sects. 4.2 and 4.3, we have considered general bidding algorithms and provided a distributed auction scheme. Next, in Sect. 4.4, we discuss how to model the process of retained transmitters through on-off scheduling exploiting the auction-based policy.

4.4 Process of Retained Transmitters

Let $\{m_x\}$ be random variables (marks) over $x \in \Phi$ that are i.i.d. and uniformly distributed on $[0, 1]$. Consider the following scheduling policies:

Random selection. In this model, each transmitter is randomly activated with probability p_A, where there is no exclusion region around the transmitters. This case is equivalent to assigning marks $\{m_x\}$ to $x \in \tilde{\Phi}$. Thus, the medium access indicator of node x is

$$e_x^R = \mathbb{1}\left(m_x < p_A\right). \tag{4.29}$$

Matérn CSMA. In the case of MHC thinning, the potential transmitters $x \in \tilde{\Phi}$ are assigned marks $\{m_x\}$, and a transmitter is retained if it has the "lowest mark" or "highest mark" within the exclusion region. Hence, we have

$$e_x^M = \mathbb{1}\left(\forall_{y \in \mathcal{N}(x)} m_x < m_y\right), \tag{4.30}$$

where the parameter D is determined using the first-order characteristics $p_A\lambda_t$.

Bidding-aided Matérn CSMA. Consider the following bidding-aided Matérn CSMA thinning model, where instead of assigning i.i.d. and uniformly distributed marks $\{m_x\}$ on $[0,1]$ to each of $x \in \tilde{\Phi}$, we compare the cumulated bid values $\{\mathcal{B}_\phi(x)\}_x$ and retain the transmitters that have the highest bid value within the exclusion region. Hence, we have

$$e_x^B = \mathbb{1}\left(\forall_{y\in\mathcal{N}(x)}\, \mathcal{B}_\phi(x) > \mathcal{B}_\phi(y)\right), \tag{4.31}$$

where $\mathcal{B}_\phi$ can be determined using either of the models in (4.11) for hom. PPP, (4.20) for non-hom PPP, (4.22), (4.23) for modified MHC model, (4.27) for Matérn CSMA, and the exclusion range parameter D is determined using (4.18) given a MAP p_A.

Bid ordering. In this scheme, given a realization ϕ of Φ with cardinality $|\phi| = N$, bids are sorted in descending order. The sorted bid vector is given as $\mathcal{B}_{\phi,\mathrm{S}} = \mathsf{sort}_{x\in\Phi}(\mathcal{B}_\phi(x))$. For a given probability of medium access p_A, node x is retained if its bid rank is at most $\lfloor p_A N \rfloor$. The medium access indicator is given as

$$e_x^O = \mathbb{1}\left(\mathcal{B}_\phi(x) \geq \mathcal{B}_{\phi,\mathrm{S}}(\lfloor p_A N \rfloor)\right). \tag{4.32}$$

Spectral Efficiency. Spectral efficiency gives the number of bits transmitted per unit time per unit bandwidth. For tractability, we assume that each

transmitter allocates equal time-frequency resources to its users, i.e., each user gets rate proportional to the spectral efficiency of its downlink channel from the serving transmitter. For total effective bandwidth W Hz, the average downlink rate in bits/sec of a typical user is

$$\mathbb{E}[R|N>0] = \mathbb{E}\left[\frac{W}{\tilde{N}}\log_2(1+\mathsf{SINR})\mathbb{1}_{\mathsf{SINR}\geq T}\right], \tag{4.33}$$

where N is the number of users served by the tagged transmitter. The distribution of N (for the PPP BS setting) is characterized in [119]. Given there is at least one user associated to the tagged transmitter, which occurs with probability $\mathbb{P}(N>0) = 1-\exp(-\Lambda_r)$, where $\Lambda_r = \lambda_r \pi \, \mathrm{R}_{\mathsf{D2D}}^2$ is the average number of receivers in the communication range of the transmitter, the conditional probability of having $N=k$ receivers is given as

$$\mathbb{P}(\tilde{N}=k) = \frac{\Lambda_r^k \exp(-\Lambda_r)}{k!(1-\exp(-\Lambda_r))}.$$

We can derive the average spectral efficiency as

$$\mathbb{E}[R|N>0] = \mathbb{P}[\mathsf{SINR}>T]\int_{r\geq 0}\mathbb{P}\Big(\mathsf{SINR}>2^{\frac{r\tilde{N}}{W}}-1\Big)\mathrm{d}r.$$

We obtain a simple upper bound under the following assumptions: (i) each receiver is associated to the nearest transmitter, (ii) the nearest transmitter is active and within the communication range $\mathrm{R}_{\mathsf{D2D}}$, and (iii) there is only one interferer $z\in\Phi\backslash\{x\}$ at a distance D_z from the typical receiver such that $\|z-x\|\geq D$.

$$\mathsf{SINR} = \frac{h/l(r)}{\sigma^2 + g_z/l(D_z)}, \tag{4.34}$$

where the distribution of r, i.e., $f_R^{Mod}(r)$, is given by (4.25). An upper bound for the Laplace transform of the interference is hence given as

$$\mathcal{L}_{I_r}(s) \leq \mathbb{E}[e^{-sg\rho D_z^{-\alpha}}] = \frac{\mu}{\mu + s\rho D_z^{-\alpha}} = \frac{1}{1 + T(r/D_z)^{\alpha}}. \tag{4.35}$$

Hence, an upper bound on the spectral efficiency can be derived as

$$R_{\mathrm{UB}} = \frac{\lambda_{\mathrm{t}}\, p_A}{\lambda_{\mathrm{r}}} \sum_{m=1}^{M} p_r(m) \mathbb{E}\left[W \log_2(1+\mathsf{SINR}) \mathbb{1}_{\mathsf{SINR} \geq T} | b_{m,j} = 1, \forall x_j \in \phi_{\mathrm{t}}\right], \tag{4.36}$$

where the distribution of SINR is derived assuming nearest transmitter association and the nearest active transmitter has the desired file.

This section has mainly focused on how to model the process of retained transmitters and on the calculation of the spectral efficiency. Later, in Sect. 4.6, we will provide a performance comparison between the bidding-aided CSMA policy and the other popular algorithms summarized above (Sect. 4.4) in terms of their spectral efficiencies as defined in (4.33). Next, we detail an online cache update scheme for the process of retained transmitters.

4.5 Online Cache Update Model using Gibbs Sampler

In this section, we propose a cache update scheme depending on the configuration, determined by the cache state, i.e., whether or not the desired content is available in the cache, and the medium access indicator, i.e., whether or not the device is transmitting.

The Gibbs sampling approach has been proposed to optimize different objectives like channel selection and user association as in [110], and hit probability as in [111], where the authors only focus on the caching problem given

the set of active nodes. Different from [111], we combine the on-off scheduling problem with the cache placement problem. For a given on-off scheduling realization, we propose an online cache update rule, which in turn determines the on-off scheduling exploiting the bidding algorithm in Sect. 4.2.

4.5.1 Cache Hit Rate Maximization given On-Off Scheduling

Consider the *finite set* $\phi = \{x_j\}$, which is a realization of the point process Φ with K nodes (sites). We consider the random field on the finite set ϕ called the *phase space*, and denoted as $\zeta = \{0,1\} \times \mathbf{b}$, where $\{0,1\}$ is the medium access indicator and $\mathbf{b}$, a binary vector of size M such that $\sum_{m=1}^{M} b_m = N$, denotes the cache state. A random field on ϕ with phases in ζ is a collection $Z = \{Z(x)\}_{x\in\phi}$ of random variables with values in ζ [120, Ch. 7.1]. It can be regarded as a random variable taking its values in the *configuration space* ζ^{ϕ}. A configuration $\boldsymbol{z} \in \zeta^{\phi}$ is of the form $\boldsymbol{z} = (\boldsymbol{z}(x), x \in \phi)$, where $\boldsymbol{z}(x) \in \zeta$ for all $x \in \phi$. For a given configuration $\boldsymbol{z}$ and a given subset $A \subset \phi$,

$$\boldsymbol{z}(A) = (\boldsymbol{z}(x), x \in A) \tag{4.37}$$

denotes the restriction of $\boldsymbol{z}$ to A. If $\phi \backslash A$ denotes the complement of A in ϕ, then $\boldsymbol{z} = (\boldsymbol{z}(A), \boldsymbol{z}(\phi \backslash A))$. For fixed $x \in \phi$, $\boldsymbol{z} = (\boldsymbol{z}(x), \boldsymbol{z}(\phi \backslash \{x\}))$.

Given a medium access probability of p_A, the set $\phi_{\mathrm{t}} \subset \phi$ denotes the set of active devices for the current realization, i.e., $P_j = 1$ if and only if $x_j \in \phi_{\mathrm{t}}$ and vice versa, i.e., $\phi_{\mathrm{t}} = \{x_j \in \phi : P_j = 1\}$. Denote the cache states by b, which is a $M \times K$ binary matrix, i.e., $b_{m,j} = 1$ if file m is available in

cache $j \in \{1, \ldots, K\}$. The cache constraint is $\sum_{m=1}^{M} b_{m,j} \leq N$ for a given cache constraint. Denote the set of all feasible cache states by B. Therefore, a configuration $\boldsymbol{z} \in \zeta^{\phi}$ is of the form $\boldsymbol{z} = (\boldsymbol{z}(\{P_j, \mathbf{b}_j\}), j = 1, \ldots, K)$, where $\boldsymbol{z}(\{P_j, \mathbf{b}_j\}) \in \{0,1\} \times b_{:,j}$ for all $x_j \in \phi$, where $b_{:,j}$ is the j^{th} column of the cache state matrix $b \in B$. Therefore, the state space cardinality is $|\zeta^{\phi}| = \binom{|\phi|}{|\phi_{\mathrm{t}}|}\binom{M}{N}^{|\phi|}$.

Given a configuration $\boldsymbol{z}$ with set of active devices ϕ_{t} and a cache state matrix $b \in B$, the cache hit rate averaged over the requests is given as

$$\mathsf{R}_{\mathsf{Hit}}(\boldsymbol{z}) = \sum_{x_j \in \phi} \sum_{u \in \phi_r \cap \mathcal{V}_{x_j}(\phi_{\mathrm{t}}(\boldsymbol{z}))} \mathbb{1}(\mathsf{SINR}_{x_j,u}(\boldsymbol{z}) \geq T)\mathbb{1}(b_{c_u,j} = 1), \quad \boldsymbol{z} \in \zeta^{\phi}, \tag{4.38}$$

where ϕ_{t} is a subset such that $\phi_{\mathrm{t}} \in \{\Phi_{\mathrm{t}}\} = \{\Phi_{\mathrm{t}} : \{X_i \in \Phi : \mathbb{E}^0[e_0] = p_A\}\}$ with a slight abuse of notation, where e_x and Φ_{t} are given in Sect. 4.1, respectively in (4.2) and in (4.3). Thus, $\{\Phi_{\mathrm{t}}\}$ denotes the set of configurations of transmitters that satisfy $\mathbb{E}^0[e_0] = p_A$ for any possible marking configuration. The term $\mathcal{V}_{x_j}(\phi_{\mathrm{t}}(\boldsymbol{z}))$ denotes the Voronoi cell of x_j with respect to $\phi_{\mathrm{t}}(\boldsymbol{z})$ under configuration $\boldsymbol{z}$, and is given by $\mathcal{V}_{x_j}(\phi_{\mathrm{t}}(\boldsymbol{z})) = \{y \in \mathbb{R}^2 : |y - x_j| < \inf_{x_i \in \phi_{\mathrm{t}}\boldsymbol{z}, x_i \neq x_j} |y - x_i|\}$. The term $\mathsf{SINR}_{x_j,u}(\boldsymbol{z})$ denotes the SINR of $u \in \phi_r$ assuming that the user is associated to the nearest active transmitter of the process $\phi_{\mathrm{t}}(\boldsymbol{z})$. Hence, if $u \in \mathcal{V}_{x_j}(\phi_{\mathrm{t}}(\boldsymbol{z}))$, then $\mathsf{SINR}_{x_j,u}(\boldsymbol{z}) = 0$ if and only if $P_j = 0$.

We seek to design a randomized iterative cache update rule to find an optimal scheme that achieves

$$\max_{\boldsymbol{z} \in \zeta^{\phi}} \mathsf{R}_{\mathsf{Hit}}(\boldsymbol{z}), \tag{4.39}$$

where $\mathsf{R}_{\mathsf{Hit}}(\boldsymbol{z})$ is given in (4.38). This is a combinatorial optimization problem and difficult to solve for large networks [110]. We next detail how to solve (4.39) using the Gibbs sampler.

4.5.2 The Gibbs Sampler

A Gibbs field Z is a Markov random field with respect to the neighborhood system N because for all sites (nodes) $x \in \phi$ the random variables $Z(x)$ and $Z(\phi \backslash \tilde{\mathcal{N}}_x)$, where $\tilde{\mathcal{N}}_x = \mathcal{N}_x \cup \{x\}$, are independent given $Z(\mathcal{N}_x)$ [120, Ch. 7.2, Theorem 2.1].

A Gibbs potential on ζ^ϕ relative to the neighborhood system N is a collection $\{V_C\}_{C \subset \phi}$ of functions $V_C : \zeta^\phi \to \mathbb{R} \cup \{+\infty\}$ such that (i) $V_C = 0$ if C is not a clique[5], and (ii) for all $\boldsymbol{z}, \boldsymbol{z}' \in \zeta^\phi$ and all $C \subset \phi$, $(\boldsymbol{z}(C) = \boldsymbol{z}'(C)) \implies (V_C(\boldsymbol{z}) = V_C(\boldsymbol{z}'))$. The function V_C depends only on the phases at the sites inside the subset C.

The energy function $\mathcal{E} : \zeta^\phi \to \mathbb{R} \cup \{+\infty\}$, associates a real number $\mathcal{E}(\boldsymbol{z})$ to each configuration. When $\mathcal{E}$ derives from the potential V, it can be written as

$$\mathcal{E}(\boldsymbol{z}) = \sum\nolimits_C V_C(\boldsymbol{z}). \tag{4.40}$$

The local energy at node x of configuration $\boldsymbol{z}$ is given by $\mathcal{E}_x(\boldsymbol{z}) = \mathcal{E}_x(\boldsymbol{z}(x), \boldsymbol{z}(\phi \backslash \{x\})) = \sum_{C \ni x} V_C(\boldsymbol{z})$, where the notation $\sum_{C \ni x}$ means that the sum extends over the sets C that contain the node x.

[5] A subset $C \subset S$ with more than one element is called a clique of the graph (S, N) if and only if any two distinct sites of C are mutual neighbors [120, Ch. 7.1].

The probability distribution

$$\pi_\beta(\boldsymbol{z}) = \frac{1}{Z_\beta} e^{-\beta \mathcal{E}(\boldsymbol{z})}, \quad \boldsymbol{z} \in \zeta^\phi \tag{4.41}$$

is called a Gibbs distribution, where β is the inverse temperature parameter, $\mathcal{E}(\boldsymbol{z})$ is the energy of configuration $\boldsymbol{z}$, and Z_β is the normalizing constant, called the partition function. If the energy function is given as in (4.40), then it is possible to find one of the states that minimizes the energy function by using a Gibbs sampler. Note that $\pi_\beta(\boldsymbol{z})$ in (4.41) (i) favors the configurations with small energy, and (ii) arises as the stationary probability distribution of a Markov random field [120, Ch. 7.6]. If one can identify an irreducible aperiodic MC $\{Z_t\}_{t\geq 0}$ with state space ζ^ϕ and stationary distribution (4.41), then for any initial distribution, the total variation distance[6] satisfies $\lim_{t\to\infty} d_{\mathrm{TV}}(\mathbb{P}(Z_t = \cdot), \pi) = 0$, i.e., its distribution at a large time n will be close to π, and one will therefore have simulated π.

The Gibbs sampler is a procedure where each node updates its own state according to the conditional distribution, called the local specification, which will be given in (4.43). The local specification only depends on the state of the neighbors of node x_j. Hence, the Gibbs sampler is a distributed procedure. The local specification also takes care of the "bidding algorithm" discussed in Sect. 4.2 via the on-off transmit powers. In practice, the updated nodes are not chosen at random, but instead in a well determined order

[6]The total variation distance between two probability distributions μ and ν on Ω is defined by $d_{\mathrm{TV}}(\mu,\nu) = ||\mu - \nu||_{\mathrm{TV}} = \max_{A\subset\Omega} |\mu(A) - \pi(A)|$ [121, Ch. 4.1].

$s(x_1), s(x_2), \ldots, s(x_K)$, where $\{s(x_j)\}_{1 \le j \le K}$ is an enumeration of all the nodes of ϕ, called the scanning policy. The nodes are visited in this order periodically [120, Ch. 7.6].

Defining the energy function as $\mathcal{E}(\boldsymbol{z}) = -\,\mathsf{R}_{\mathsf{Hit}}(\boldsymbol{z})$ for given configuration $\boldsymbol{z} \in \zeta^{\phi}$, we obtain

$$\pi_{\beta}(\boldsymbol{z}) = \mathbb{P}\left(Z(\phi) = \boldsymbol{z}\right) = \frac{e^{\beta\, \mathsf{R}_{\mathsf{Hit}}(\boldsymbol{z})}}{\sum\limits_{\boldsymbol{z}' \in \zeta^{\phi}} e^{\beta\, \mathsf{R}_{\mathsf{Hit}}(\boldsymbol{z}')}}, \quad \boldsymbol{z} \in \zeta^{\phi}. \tag{4.42}$$

Using the Gibbs sampler procedure, we can demonstrate that the transitions to states of smaller local energy, i.e., higher cache hit rate, are favored compared to states of higher energy, i.e., lower cache hit rate. Hence, we can find an optimal state that achieves (4.39). The performance of the Gibbs sampler can be improved by "annealing", i.e., a slow increase of β. When β increases to ∞ with time $t > 0$ like $\log(1+t)$, we get convergence to a collection of states of minimal global energy [110].

The local specification of the Gibbs distribution at node $x_j \in \phi$ is the function $\pi_{\beta}^{j} : \zeta^{\phi} \to [0,1]$ defined by [120, Theorem 2.1]

$$\begin{aligned}
\pi_{\beta}^{j}(\boldsymbol{z}) &= \mathbb{P}\left(Z(x_j) = \boldsymbol{z}(x_j) | Z(\phi \backslash \{x_j\}) = \boldsymbol{z}(\phi \backslash \{x_j\})\right) \\
&\overset{(a)}{=} \frac{\pi_{\beta}(\boldsymbol{z})}{\sum_{z' \in \zeta} \pi_{\beta}(z', \boldsymbol{z}(\phi \backslash \{x_j\}))} \\
&\overset{(b)}{=} \frac{e^{-\sum_{C \ni x_j} V_C(\boldsymbol{z})}}{\sum_{z' \in \zeta} e^{-\sum_{C \ni x_j} V_C(z', \boldsymbol{z}(\phi \backslash \{x_j\}))}} \\
&\overset{(c)}{=} \frac{e^{\beta \sum_{n \in \mathcal{N}(x_j)} \mathsf{R}_{\mathsf{Hit}\, n}(\boldsymbol{z})}}{\sum_{z' \in \zeta} e^{\beta \sum_{n \in \mathcal{N}(x_j)} \mathsf{R}_{\mathsf{Hit}\, n}(z', \boldsymbol{z}(\phi \backslash \{x_j\}))}}, \ x_j \in \phi, \ \boldsymbol{z} \in \zeta^{\phi},
\end{aligned} \tag{4.43}$$

where (a) follows from the definition of conditional probability, (b) from (4.41) and that if C is a clique and x is not in C, then $V_C(z', \boldsymbol{z}(\phi \backslash \{x\})) = V_C(\boldsymbol{z})$ and is independent of $z' \in \zeta$, and (c) from using (4.41). This denotes the conditional distribution of the network configuration conditioned on the restriction[7] of configuration $\boldsymbol{z}$ to all devices except $x_j \in \phi$, under the joint distribution $\pi_\beta(\boldsymbol{z})$. Furthermore, the cache hit rate provided by device n is given as

$$\mathsf{R}_{\mathsf{Hit}n}(\boldsymbol{z}) = \sum_{u \in \phi_r \cap \mathcal{V}_{x_n}(\phi_{\mathrm{t}}(\boldsymbol{z}))} \mathbb{1}(\mathsf{SINR}_{x_n,u}(\boldsymbol{z}) \geq T) \mathbb{1}(b_{c_u,n} = 1), \quad x_n \in \phi, \quad \boldsymbol{z} \in \zeta^\phi, \tag{4.44}$$

and $\mathsf{R}_{\mathsf{Hit}n}(z', \boldsymbol{z}(\phi \backslash \{x_j\}))$ denotes the cache hit rate provided by device n under configuration $z' \in \zeta$ conditioned on the restriction of the configuration $\boldsymbol{z}$ to all devices except $x_j \in \phi$:

$$\begin{aligned} &\mathsf{R}_{\mathsf{Hit}n}(z', \boldsymbol{z}(\phi \backslash \{x_j\})) \\ &= \sum_{u \in \phi_r \cap \mathcal{V}_{x_n}(\phi_{\mathrm{t}}(\boldsymbol{z}))} \mathbb{1}(\mathsf{SINR}_{x_n,u}(\boldsymbol{z}) \geq T) \mathbb{1}(b_{c_u,n} = 1), \quad x_n \in \phi, \quad z' \in \zeta. \end{aligned} \tag{4.45}$$

A finite state irreducible aperiodic MC has a unique stationary distribution π_β on a finite state space ζ^ϕ, and regardless of the initial state, as $t \to \infty$, the distribution of the chain converges to π_β. Let P and $\mathcal{P}$ denote the transition matrix and the collection of all probability distributions on ζ^ϕ of an ergodic MC, respectively. Next, we investigate how large t should get so that the distribution of the chain is close to π_β.

[7] This can be obtained by deleting the j^{th} column of $b \in B$ and deleting P_j.

Definition 8. Mixing time [121, Ch. 4.5]. *The mixing time* $t_{\text{mix}}(\varepsilon)$ *is defined as the smallest time such that for any starting state* Z_0 *with distribution* μ*, the distribution of the state* Z_t *at time* t *is within total variation distance* ε *of* π*:*

$$t_{\text{mix}}(\varepsilon) \triangleq t_{\text{mix}}(1/4) = \min\{t : d_{\text{TV}}(\mu P^t, \pi_\beta) \leq 1/4\}. \tag{4.46}$$

Proposition 11. *The mixing time can be upper bounded as*

$$t_{\text{mix}} \leq \frac{0.5 \log(4\chi^2(\mu, \pi))}{\log((1 - e^{-L\beta\Delta})^{-1})}, \tag{4.47}$$

where the terms $\chi^2(\mu, \pi) = \sum_{i \in \zeta^\phi} \frac{(\mu(i) - \pi(i))^2}{\pi(i)}$*, where* μ *being the distribution for the starting state, and the parameters* $\delta_x = \sup\{\mathsf{R}_{\mathsf{Hit}}(\boldsymbol{z}) - \mathsf{R}_{\mathsf{Hit}}(\boldsymbol{z}')\,;\, \boldsymbol{z}(\phi \backslash \{x\}) = \boldsymbol{z}'(\phi \backslash \{x\})\}$*, and* $\Delta = \sup_{x \in \phi} \delta_x$ *follow from [120, Ch. 7.6], and* L *is the period such that the nodes of* ϕ *are visited in an order periodically.*

Proof. The final result can be obtained using similar techniques as in [120, Ch. 7.6]. □

4.5.3 Cache Admission and Extinction Policy

The cache admission (or content insertion) and the cache extinction (or content ejection) policies are implemented exploiting the Gibbs sampling approach outlined in Sect. 4.5.2, determined by the local specification given in (4.43). The cache admission policy is based on the local demand that cannot be served by a D2D transmitter node. Once a local demand is not served by the set of transmitters that cover it, a file is inserted, i.e., acquired from the

BS[8], based on the conditional distribution rule in (4.43). To create space for the inserted file, a file is evicted from the cache. This can be done by selecting at random or discarding the least recently used (LRU) items first.

Consider a given configuration $z \in \zeta^{\phi}$ such that $z = (z(\{P_i, \mathbf{b}_i\}), x_i \in \phi)$, with the realization of active D2D transmitters denoted by $\phi_{\mathrm{t}} = \{x_i \in \phi : P_i = 1\}$ and $b \in B$.

Assume that the scanning policy picks the potential transmitter node $x_j \in \phi$. Assume that there exists $x_i \in \phi_{\mathrm{t}}$ such that $x_i \in \mathcal{N}(x_j)$ when $|\mathcal{N}(x_j)| \neq 0$. The cache update rule for node x_j is determined by (4.43). Upon the request of file $m \in \{1, \ldots, M\}$, depending on the current state of x_j, one of the following events occur:

1. x_j is scheduled, i.e., $x_j \in \phi_{\mathrm{t}}$:

   (i) When the desired file is available, $b_{m,j} = 1$, it is transmitted. No cache update is required.

   (ii) Desired file is not available in $x_j \cup \mathcal{N}(x_j)$. In this case, since x_j transmits and the desired file is not available, a cache update is required. The update rule is determined by $\pi^j_{\beta}(z)$.

2. x_j is not scheduled, i.e., $x_j \notin \phi_{\mathrm{t}}$:

[8]Similarly, if the demand is not served, the file can also be acquired from the neighboring D2D transmitters. This is left as future work.

(i) At most either of $b_{m,i}$ or $b_{m,j}$ is one. In this case, either the desired file is available but not transmitted, or the desired file is available in a neighbor who transmits, or desired file is not available in $x_j \cup \mathcal{N}(x_j)$. No update is required.

(ii) When $b_{m,i} = b_{m,j} = 1$, a cache update in x_j is required to prevent the redundancy of caching the same file in $\mathcal{N}(x_j)$. The update rule is given by the local specification $\pi^j_\beta(z)$.

In the cases of 1)(ii) and 2)(ii), a cache update in x_j is required. For the case when $|\mathcal{N}(x_j)| = 0$, the update rule for x_j is oblivious to the other nodes. The scenarios 1)(i), 1)(ii), 2)(i) will still be valid.

The main focus of this section was to provide a Gibbs sampler-based update scheme for caches in order to iteratively maximize the cache hit rate given a scheduling configuration. Next, in Sect. 4.6, by incorporating the different models in Sect. 4.4 for the set of retained transmitters, we provide an evaluation in terms of the spectral efficiency in the units of bits/sec/Hz/User and the evolution of the cache hit rate under different bidding algorithms as proposed in Sects. 4.2 and 4.3.

4.6 Performance Evaluation

We consider a realization ϕ of PPP Φ over the region $S = [-5, 5]^2$ with an intensity of $\lambda_t = 3$ per unit area. The catalog size is $M = 100$ files and each potential transmitter $x \in \phi$ can store up to $N = 10$ files. We consider

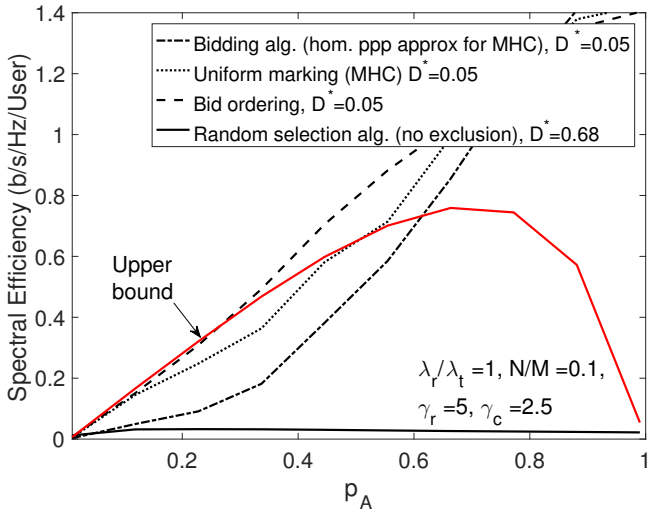


Figure 4.3: Spectral efficiency comparison of the bidding-aided CSMA model with other scheduling policies: skewed cache configurations and requests.

an IRM traffic scenario, where the popularity of requests is modeled by the Zipf distribution, which has pmf $p_r(n) = \frac{1}{n^{\gamma_r}} \Big/ \sum_{m=1}^{M} \frac{1}{m^{\gamma_r}}$, for $n \in \mathcal{M}$, where $\gamma_r \in (0, 1)$ the Zipf exponent that determines the skewness of the distribution. File requests are generated over S according to a time and space homogeneous PPP with intensity $\lambda_r = 3$ requests per unit time per unit area, and file requests are uniform and independent over the space, and any new request can be for $m \in \mathcal{M}$ with probability $p_r(m)$. The rest of the network parameters are chosen as follows. Path loss exponent is $\alpha = 4$, SINR threshold is $T = 0.01$, $\sigma^{-2} = .1$, and the fading parameter is $\mu = 1$.

Next, we consider the homogeneous PPP model of Sect. 4.2, and the non-homogeneous PPP, modified hard-core model, and a non-homogeneous PPP approximation for the Matérn CSMA, as detailed in Sect. 4.3. Then, we

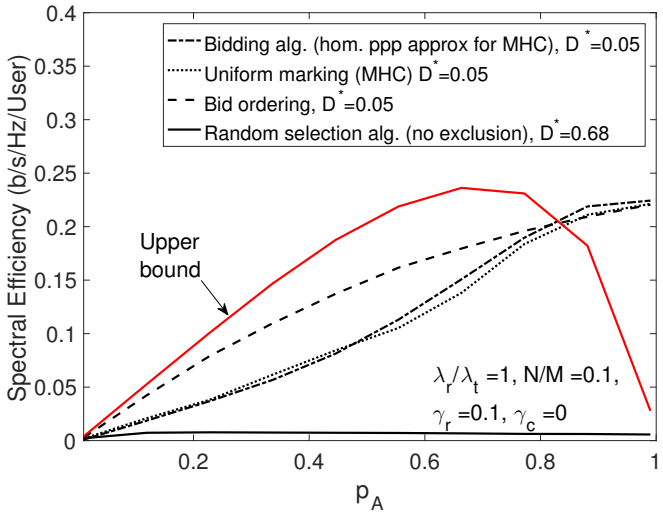


Figure 4.4: Spectral efficiency comparison of the bidding-aided CSMA model with other scheduling policies: randomized cache configurations and requests.

illustrate the performance of different scheduling algorithms as a function of the MAP p_A. In Fig. 4.3, we have a skewed placement configuration $p_c \sim$ Zipf(2.5) and $p_r \sim$ Zipf(5). In Fig. 4.4, we have $p_c \sim$ Zipf(0) and $p_r \sim$ Zipf(0.1). We also compare against the analytical upper bound in (4.36) for the low contention regime of CSMA. The bidding algorithm provides higher throughputs than random selection and uniform marking. For skewed placement, the spectral efficiency performance is very close to the upper bound for the low contention regime.

For cache placement, the medium access probability is fixed to be $p_A = 0.45$. The catalog size is $M = 3$ and the cache size is $N = 1$. We compare the performance of the LRU, in which the least recently used item is discarded, and the online cache update model using Gibbs sampler as detailed in Sect.

4.5.

In order to get convergence to a collection of states with minimal global energy, i.e., maximal cache hit rate, we use the annealed Gibbs sampler[9], where the inverse temperature parameter β slowly increases over time following the relation $\beta = \beta_0 \log(1 + t)$, where $\beta_0 = 10$.

We compare the performance of the LRU scheme and the Gibbs sampler in terms of their cache hit probabilities in Fig. 4.5. Starting with a totally randomized initial configuration of the caches over the set of files in the catalog, and a Zipf distributed request distribution with $\gamma_r = 0.1$ with density $\lambda_r = 0.3$, caches are updated over time, where the nodes are visited in an order periodically. At each iteration of both algorithms, if the selected cache is scheduled, it is updated only if it does not contain the desired file from any of the receivers in its communication range. We observe that both algorithms can behave similarly under random scheduling of the potential transmitters. However, when the transmitters are scheduled according to the bidding algorithm detailed in Sects. 4.2 and 4.3, the Gibbs sampler, unlike the LRU model, captures the local interactions among the nodes to optimize the cache hit rate, hence provides a better average hit rate.

[9]The plain sampler can also be developed, in which β is fixed, and the state updates are not randomized but always chosen to minimize the local energy, i.e., it is a greedy algorithm. The plain sampler minimizes the local energy observed for each transition, only converges to a random state distributed according to the Gibbs distribution, and can get blocked in a local minimum (of the energy). Its speed of convergence is geometric. The plain sampler hence trades the long-term efficiency for the speed of convergence [110].

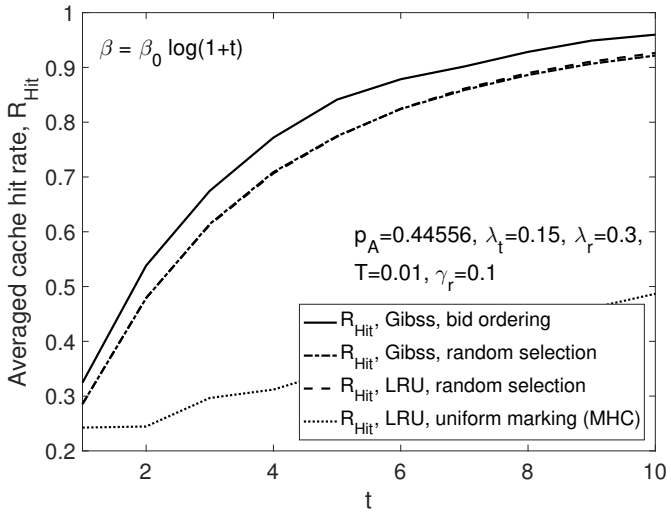


Figure 4.5: Comparison of the Gibbs sampling based caching strategy and LRU cache placement strategy, for a PPP distributed potential transmitter process Φ over the region $S = [-5, 5]^2$ with $\lambda_t = 0.15$, given a MAP $p_A = 0.45$, receiver process Φ_r with density $\lambda_r = 0.3$, catalog size $M = 3$, and cache size $N = 1$.

4.7 Summary

We developed a bidding-aided distributed scheduling policy for D2D users by capturing the local demand profile, the spatial distribution and the configurations of the transmitters, with the objective of maximizing the spectral efficiency in the units of bits/sec/Hz/User. We demonstrated and contrasted the performance of our bidding-aided algorithm with other well-known CSMA policies. The key takeaways include that rather than solely balancing the traffic according to the locations of caches, exploiting the cache configurations and local demand distribution, higher throughput gains can be achieved, and our approach provides new insights into designing dynamic bidding-aided caching algorithms. Possible directions include the extension of the schedul-

ing algorithm to develop dynamic caching algorithms that capture the network configuration in order to achieve higher throughput scaling gains with caching.

Chapter 5

Conclusion

5.1 Summary

In this dissertation, we focused on the modeling and the analysis of device-to-device (D2D) content aggregation and distribution (caching) in the context of cellular networks. Intuitively, the optimal placement of content into the caches should not be spatially independent, since if the file is already cached nearby, it is less useful to cache the file again. We proposed randomized D2D content distribution schemes that capture the actual physical channel model in Chapter 2, which is different from the grid-based model in [71], [78]. We incorporated the interference due to simultaneously active transmitters, noise and the small-scale Rayleigh fading into the analysis such that any transmission is successful as long as SINR is above a threshold. Contrasting with the probabilistic policies, where the files are independently placed in the cache memories of different nodes according to the same distribution [60], [52], and [63], or other approaches that do not consider network-level interactions, our approach in Chapter 3 and Chapter 4 i) captures the spatial, or geographic, correlation of the nodes, to bring spatial diversity in order to increase the hit probability, ii) is distributed and scalable, hence, will pave the way for the design of D2D content distribution systems, and iii) captures the spatial-

temporal interactions among devices via bidding. The contributions of this dissertation are summarized next.

In Chapter 2, we developed a spatially independent and randomized D2D content caching model. We derived the probability of successful content delivery in the presence of interference and noise [69], [70], [52], in which the locations of the D2D caches are modeled by a PPP. We computed the caching distribution that maximizes the density of successful receptions (DSR) under a simple transmission strategy where a single file is transmitted at a time throughout the network. For Zipf distributed request profile, the optimal caching distribution is also modeled using the Zipf law and the caching exponent linearly scales with the request exponent, and inversely proportional to the path loss exponent, which leads to the smoothing effect. Similarly, for more general demand profiles under Rayleigh, Ricean and Nakagami small-scale fading distributions, it is required to flatten the request distribution to optimize the caching performance.

In Chapter 3, we studied optimal geographic content placement for D2D networks in which the locations of the D2D caches are modeled by a PPP and have limited communication range. Inspired by the Matérn hard-core (type II) point process, we devised a novel spatially correlated strategy called hard-core placement (HCP) such that the D2D nodes caching the same file are never closer to each other than the exclusion radius. The exclusion radius plays the role of a substitute for caching probability. We optimized the exclusion radii to maximize the cache hit probability. Contrasting it with the independent

content placement, our HCP strategy often yields a significantly higher cache hit probability. We demonstrated that the HCP strategy is effective for small cache sizes and a small communication radius, which are likely conditions for D2D.

In Chapter 4, we proposed a distributed bidding-aided Matérn carrier sense multiple access (CSMA) policy for a D2D content distribution network with D2D receivers and "potential" D2D transmitters, i.e., transmitters are turned on or off by the scheduling algorithm. Each D2D receiver determines the value of its request, by bidding on the set of potential transmitters in its communication range. Given a medium access probability, a fraction of the potential transmitters are turned on, determined jointly by the auction policy and the power control scheme. We contrasted the performance of the bidding-aided CSMA policy with other well-known CSMA schemes, demonstrated that our algorithm achieves a higher spectral efficiency in terms of the number of bits transmitted per unit time per unit bandwidth per user. The gain becomes even more visible under randomized configurations and requests rather than more skewed placement configurations and deterministic demand distributions. Later, we considered a Gibbs sampling approach for cache updates in order to iteratively maximize the cache hit rate. The update scheme depends on the cache configuration, i.e., whether or not the desired content is available in the cache, and the on-off scheduling algorithm.

5.2 Future Research Directions

A future research objective is to determine how to best allocate the resources under varying traffic, i.e., to balance the busy-hour and average-hour wireless network traffic, through smart content caching techniques, using tools from stochastic processes, wireless communications and networks, stochastic geometry and optimization. Our second objective is to characterize the effect of different transmit diversity and receiver combining techniques on content caching, and study the gain of the cache hit rate through diversity. We next discuss the proposed directions.

5.2.1 Duality of Scheduling and Caching and Model Validation

Others aim to develop a novel spatial-temporal content caching model for D2D communications under renewal traffic to capture how the spatial diversity of the content can be incorporated to improve the caching performance.

To the best of our knowledge, the current research efforts lack a thorough understanding of the connections between content caching and scheduling in the cellular context. On one hand, with content caching, the content should be spread over the network in order to *maximize the cache hit rate*, and on the other hand, with scheduling, the objective is to bring the content in proximity to the user in order to *maximize the throughput*. Therefore, content caching and scheduling problems are in fact closely associated with each other. We will jointly consider these problems, and extend the caching model detailed in Chapter 4.

For a given scheduling algorithm, e.g., the game-theoretic auction model proposed in Chapter 4, and given a realization of cache configuration, the goal is to decide how to update the placement of the content by incorporating the SINR coverage characteristics of the network into caching. As future work, our goal is to model the pairwise interactions among the nodes using the Gibbs point processes (GPPs) that can model more general interactions rather than the HCP model proposed in Chapter 3, which only considered the first-order interactions. We are interested in the regimes for which the Gibbs model provides a higher cache hit probability than other popular models [60], [87].

Content caching exploiting pairwise interactions. GPPs are mathematical models of particle interactions in statistical mechanics, and are characterized by a potential function, modeling the interactions –e.g., attraction or repulsion– among nodes. They are good models for patterns with some degree of regularity, i.e., more regular than MHC processes, or for moderate clustering, but can be deficient in cases of strong clustering [64, Ch. 5.5]. Special cases are the Ising model [122], Markov point processes, spatial birth-and-death processes, cluster processes such as the Neyman-Scott processes, and repulsive processes such as the Strauss model and hard-core processes [64, Ch. 5.5].

The GPP brings a strategy for the placement of nodes, and the content placement is done at the existing nodes generated by the GPP. Consider a GPP Φ_G of distribution P on $[\mathbb{N}, \mathcal{N}_G]$ with exactly k points in a bounded

region $\mathcal{D} = [0, D]^2 \in \mathbb{R}^2$ [64, Ch. 5.5]. Assume that the distribution of the point process is given by a probability density function $f : \mathbb{R}^{2k} \to [0, \infty)$ so that

$$P(\Phi_G \in Y) = \int_{\{x_1,\cdots,x_k\}\in Y} \cdots \int f(x_1, \ldots, x_k)\, \mathrm{d}x_1 \ldots \mathrm{d}x_k, \quad Y \in \mathcal{N}_{G,k}(\mathcal{D}), \tag{5.1}$$

where $\mathcal{N}_{G,k}(\mathcal{D})$ denotes the trace of $\mathcal{N}_G$ on the set of all point processes with k points in $\mathcal{D}$. Because point processes are an unordered set of points, $f(x_1, \ldots, x_k)$ does not depend on the order of the arguments, and is given by

$$f(x_1, \ldots, x_k) = \frac{\exp\left(-E(x_1, \ldots, x_k)\right)}{Z}, \tag{5.2}$$

where the function $E : \mathbb{R}^{2k} \to \mathbb{R} \cup \{\infty\}$ is called the energy function, which does not depend on the order of the arguments, and Z is a normalization constant, which is called the configurational partition function. These terms come from statistical mechanics [64, Ch. 5.5].

Pair potential function. The energy function E is frequently chosen as

$$E(x_1, \ldots, x_k) = \beta \sum_{1\le i<j\le k} \sum \theta(\|x_i - x_j\|), \tag{5.3}$$

where $\theta : [0, \infty) \to (-\infty, \infty]$ is the pair potential, and $\beta = T^{-1}$ is called the inverse temperature [64, Ch. 5.5].

The pair potential characterizes the GPP of density f constructed as above. A typical example is shown in Fig. 5.1. Potential $\theta(r)$ shown in the figure is infinite for $r \le R$, i.e., the inter-node distance can never be less than

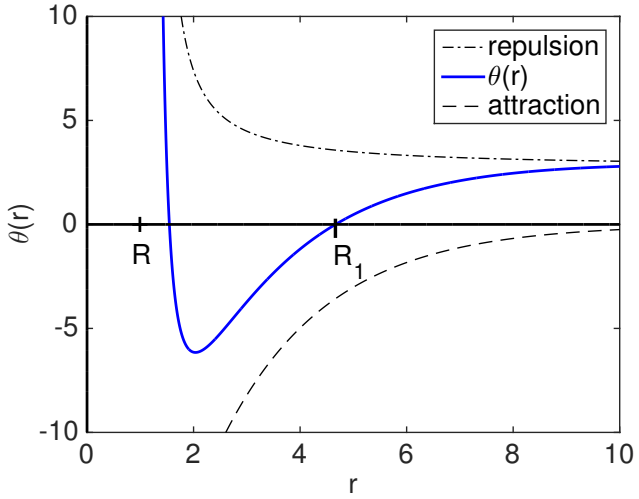


Figure 5.1: A typical pair potential, the result of superposition of attractive and repulsive forces.

R. Therefore, the point process is in fact a hard-core model. For $r > R$, $\theta(r) = \exp\left(\frac{r}{r-R}\right) - 100\exp\left(-\frac{r}{2} - R\right)$. Since $\theta(r)$ is large when r is slightly larger than R, such inter-point distances exist with a low probability. Inter-node distances for which $\theta(r)$ takes its minimum, i.e., the inter-point distances close to R_1, should occur relatively frequently.

GPP-inspired placement design. A caching network modeled by a Gibbs distribution might not require centralized coordination since it captures the pairwise interactions among the nodes in a distributed manner. Although it is hard to characterize GPPs in their most generic form to optimize the performance of caching, in this section we formulate the general hit probability maximization problem for the GPPs.

The cache hit probability for the GPP-inspired placement is given by

$$\mathrm{P}_{\mathsf{Hit,G}} = 1 - \sum_{m=1}^{M} p_r(m) \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_G(T) = k) P_{\mathrm{Miss,G}}(m,k), \tag{5.4}$$

where $\mathbb{P}(\mathcal{N}_G = k)$ is the coverage distribution, i.e., the probability that k transmitters (caches) cover the typical receiver. The parameter $p_r(m)$ models the request or demand distribution, and $P_{\mathrm{Miss,G}}(m,k)$ is the probability that k caches cover a receiver, and none has file m, i.e., the probability of cache miss, and it is given as

$$P_{\mathrm{Miss,G}}(m,k) = \int \cdots \int_{\mathcal{V}^k} f_m(x_1,\ldots,x_k)\,\mathrm{d}x_1 \ldots \mathrm{d}x_k. \tag{5.5}$$

Given k transmitters, the probability that content m is cached at a transmitter is equal to $f_m(x_1,\ldots,x_k)$ for a realization with k transmitters. The expression $\mathbb{E}_{\mathcal{N}_G}[f_m(x_1,\ldots,x_{\mathcal{N}_G})]$ denotes the average of the distribution over all realizations of the GPP. Therefore, the average placement probability of file m in a cache is given by $\mathbb{E}_{\mathcal{N}_G}\left[f_m(x_1,\ldots,x_{\mathcal{N}_G})\right] \leq 1$. Since there are at most N files to be stored in each cache, the cache constraint $\sum_{m=1}^{M} \mathbb{E}_{\mathcal{N}_G}\left[f_m(x_1,\ldots,x_{\mathcal{N}_G})\right] \leq N$ follows. The region $\mathcal{V}^k$ characterizes the cache miss region given there exists k nodes, i.e., it is the $2k$ dimensional region $[0,D]^{2k} \backslash [0,R_{\mathsf{D2D}}]^{2k}$.

The modeling and algorithmic challenges in designing optimal caching strategies include capturing the impact of i) the temporal locality of content to estimate how the popularity changes over time and infer the request distribution, which might be non-stationary, and ii) the geographic locality of content to provide diversity to users who have potential to get the desired content from

more than one user. Using tools from stochastic geometry, queueing theory, and optimization, our objective is to design efficient caching algorithms incorporating the spatial and temporal dynamics in cellular networks, and develop practical use cases for content caching and incentives for its realization.

We will use proprietary data on movie requests and ratings over time (see Fig. 1.5 in Chapter 1) for testing recent theoretical results and algorithms in this dissertation, and also developing and comparing some algorithms that use real data on request distributions over time. Predicting the popularity profile of users through machine learning algorithms, we will run adaptive caching algorithms, similar to the Gibbs sampling approach as proposed in Chapter 4, to demonstrate gains from what we have been doing theory on, and investigate practical use cases for the initial theoretical results we obtained.

5.2.2 Content Caching using Diversity Combining Techniques

A possible direction is to extend the randomized caching model in Chapter 2 by incorporating diversity combining techniques to improve the quality of received signal. The cache hit rate can be improved by using different transmit diversity and receiver combining techniques. We first propose to analyze the effect of (i) the transmit diversity for equal-gain combining and selection combining, and (ii) the receiver diversity using chase combining.

Shot noise (Equal-gain combining). Using the mobile network model in Chapter 2, in which D2D users are spatially distributed as a homogeneous PPP

Φ of density λ, the distribution function of the shot noise from its Laplace transform can be derived via the Plancherel-Parseval Theorem [61, Ch. 2]. Defining $I = \sum_{i\in\Phi} g_i \left\|x_i\right\|^{-\alpha}$, where x_i and g_i are the distance and the channel power gain of the D2D transmitter i, the Laplace functional of the shot noise $\mathcal{L}_I(z)$ equals

$$\begin{aligned}
\mathcal{L}_I(z) &= \mathbb{E}\left[\exp\left(-z\sum_{i\in\Phi} g_i {R_i}^{-\alpha}\right)\right] \\
&= \mathbb{E}\left[\prod_{i\in\Phi} e^{-zg_i {R_i}^{-\alpha}}\right] \\
&\overset{(a)}{=} \mathbb{E}_\Phi\left[\prod_{i\in\Phi} \mathbb{E}_g\left[e^{-zg{R_i}^{-\alpha}}\right]\right] \\
&\overset{(b)}{=} \exp\left(-2\pi\lambda\gamma_1 p_c(i)\int_0^\infty \left(1-\mathbb{E}_g\left[\exp(-zgv^{-\alpha})\right]\right) v\mathrm{d}v\right),
\end{aligned}$$

where (a) follows from the iid distribution of g_i and its further independence from the point process Φ, and (b) follows from the probability generating functional (PGFL) [64] of the PPP.

With the assumption of Rayleigh fading, i.e., $g \sim \exp(\mu)$, we can rewrite $\mathbb{E}_g[\exp(-zgv^{-\alpha})]$ as

$$\mathbb{E}_g[\exp(-zgv^{-\alpha})] = \int_0^\infty e^{-zgv^{-\alpha}}\mu e^{-\mu g}\mathrm{d}g = \frac{1}{\mu^{-1}zv^{-\alpha}+1}.$$

Thus,

$$\begin{aligned}
\mathcal{L}_I(z) &= \exp\left(-2\pi\lambda\gamma_1 p_c(j)\int_0^\infty \tfrac{1}{1+\mu z^{-1}v^\alpha} v\mathrm{d}v\right) \\
&= \exp\left(-\pi\lambda\gamma_1 p_c(j)\int_0^\infty \tfrac{1}{1+\mu z^{-1}y^{\alpha/2}}\mathrm{d}y\right).
\end{aligned}$$

Given the Laplace transform pairs $f(t) \overset{LT}{\longleftrightarrow} F(z)$, using the property $\frac{\mathrm{d}f(t)}{\mathrm{d}t} \overset{LT}{\longleftrightarrow} zF(z) - f(0^-)$, the Laplace transform associated with the complementary cumulative distribution function (CCDF) of the shot noise is $\bar{\mathcal{L}}_I(z) = 1/z - \mathcal{L}_I(z)/z$. Hence, the CCDF of the shot noise is given by [123] as

$$\mathbb{P}(I > t) = \frac{2e^{at}}{\pi} \int_0^\infty \mathrm{Re}(\bar{\mathcal{L}}_I(a + iu)) \cos(ut)\, \mathrm{d}u,$$

where $t = \mathrm{T}\,\sigma^2$ and $\mathrm{Re}(z)$ is the real part of z and $z = a$ is any vertical line contour, i.e., is real valued, and it should be selected such that $\bar{\mathcal{L}}_I(z) = \bar{\mathcal{L}}_I(a+iu)$ has no singularities on or to the right of it. Letting $a = 0$, we derive $\mathbb{P}(I > t)$ as follows:

$$\mathbb{P}(I > t) = \frac{2}{\pi} \int_0^\infty \mathrm{Re}\left(\exp\big(\log\big(\frac{i}{u}\big) \right.$$
$$\left. -\pi\lambda\gamma_1 p_c(j) \int_0^\infty \frac{1}{1 + \mu(iu)^{-1} y^{\alpha/2}} \mathrm{d}y \big) \right) \cos(ut)\, \mathrm{d}u,$$

where i is the imaginary unit. Although this expression can be numerically solved to determine the optimal caching distribution, and through some approximations, we are able to show that the optimal caching distribution can be approximated by a Zipf distribution, the analysis is not very tractable.

We next consider another approach, which is practical to implement and tractable.

Strongest signal (Selection combining). Consider the association to the strongest user, in which, the coverage $\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1 p_c(j), \alpha)$ is bounded by

$$\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1 p_c(j), \alpha) = \mathbb{P}\left(\max_{x \in \Phi} \mathsf{SINR}(x) > \mathrm{T} \right)$$

$$
\begin{aligned}
&= \mathbb{E}\left[1\left(\bigcup_{x\in\Phi} \mathsf{SINR}(x) > \mathrm{T}\right)\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\sum_{x\in\Phi} 1(\mathsf{SINR}(x) > \mathrm{T})\right] \\
&= \mathbb{E}\left[\sum_{x\in\Phi} 1(h_x \left\|x\right\|^{-\alpha} > \mathrm{T}(I_x + \sigma^2))\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{x\in\Phi} \mathbb{P}(h_x > \mathrm{T}(I_x + \sigma^2) \left\|x\right\|^{\alpha})\right] \\
&\overset{(c)}{=} \lambda\gamma_1 p_c(j) \int_{x\in\mathbb{R}^2} \mathbb{E}_{I_x}\left[e^{-\mu \mathrm{T}(I_x+\sigma^2)\|x\|^{\alpha}}\right] \mathrm{d}x \\
&\overset{(d)}{=} 2\pi\lambda\gamma_1 p_c(j) \int_0^{\infty} \mathcal{L}_{I_r}(\mu \mathrm{T} r^{\alpha}) e^{-\mu \mathrm{T}\sigma^2 r^{\alpha}} r \mathrm{d}r,
\end{aligned}
$$

where (a) follows from the union bound, $I_x = \sum_{y\in\Phi\setminus x} g_y \left\|y\right\|^{-\alpha}$ is the interference received by the typical user when it is connected to the user located at x, and (b) from that since the channel power of the direct link is independent of everything else, we can take the expectation h_x inside, and (c) from the Rayleigh fading assumption with $h_x \sim \exp(\mu)$ and Campbell-Mecke Theorem [64], and (d) from the definition of the Laplace transform and converting the integral from Cartesian into polar coordinates. This upper bound is shown to be tight in [124]. Hence, it is reasonable to approximate the coverage using this upper bound. For this model, the Laplace transform of interference equals

$$
\begin{aligned}
\mathcal{L}_I(s) &= \mathbb{E}_{I_x}\left[e^{-sI}\right] = \mathbb{E}_I\left[e^{-s\sum_{y\in\Phi} g_y \|y\|^{-\alpha}}\right] \\
&= \exp\left(-\pi\lambda\gamma_1 p_c(j) \int_0^{\infty} \frac{1}{1 + s^{-1}\mu v^{\alpha/2}} \mathrm{d}v\right), \qquad (5.6)
\end{aligned}
$$

which follows from the fact that channel powers are independent of the users locations, employing the PGFL of PPP [64], and a change of variables.

Thus, we can upper bound the coverage probability as follows:

$$\begin{aligned}
\mathrm{p_{cov}}(\mathrm{T}, \lambda\gamma_1 p_c(j), \alpha) &\leq 2\pi\lambda\gamma_1 p_c(j) \int_0^\infty \mathcal{L}_{I_r}(\mu \mathrm{T} r^\alpha) e^{-\mu \mathrm{T}\sigma^2 r^\alpha} r \mathrm{d}r \\
&\overset{(a)}{=} \pi\lambda\gamma_1 p_c(j) \int_0^\infty e^{-\pi\lambda\gamma_1 p_c(j) \int_0^\infty \frac{1}{1+T^{-1}(\frac{v}{r})^{\alpha/2}} \mathrm{d}v - \mu T \sigma^2 r^{\alpha/2}} \mathrm{d}r \\
&\overset{(b)}{=} \frac{\pi\lambda\gamma_1 p_c(j)}{h(\alpha)\mathrm{T}^{\frac{2}{\alpha}}} \int_0^\infty e^{-\pi\lambda\gamma_1 p_c(j) v - \mu\left[\frac{1}{h(\alpha)}\right]\sigma^2 v^{\alpha/2}} \mathrm{d}v, \qquad (5.7)
\end{aligned}$$

where (a) follows from (5.6) and a change of variables, and (b) follows from a simple change of variables, $\int \frac{1}{1+x^a}\mathrm{d}x = x\,{}_2F_1\left(1, \frac{1}{a}; 1+\frac{1}{a}; -x^a\right)$, and letting $h(\alpha) = \lim_{x\to\infty} x\,{}_2F_1\left(1, \frac{2}{\alpha}; 1+\frac{2}{\alpha}; -x^{\alpha/2}\right)$, where ${}_2F_1$ is the Gauss hypergeometric function. Note that in the original formulation where there is no diversity, by employing a change of variables $v = r\beta(\mathrm{T}, \alpha)$, we can rewrite (2.1) in Definition 2.1 as in (A.6) in Appendix A.1. From Assumption 1, the ratio $\frac{\mathrm{T}}{\beta(\mathrm{T},\alpha)^{\alpha/2}}$ is fixed for a given α, i.e., is a function of α only. We note that in the case of the above proposed model, the coverage probability expression in (5.7) is in the same form as the original model in (A.6). Hence, for diversity with selection combining, we expect to get a similar optimal solution as in the original model without capturing the diversity.

A different diversity combining technique we contemplate is a retransmission based strategy, in which the D2D receiver uses maximum-ratio combining to combine the received bits with the same bits from previous transmissions. This technique is known as Chase combining.

Receiver diversity (Chase combining). Given a caching application with a delay constraint T, we propose a simple retransmission-based strategy.

Consider a caching model in which randomly arriving devices transmit their payload to a receiver, where the device locations are assumed to form realization of a homogeneous two-dimensional spatial PPP. Given a fixed latency constraint per device, we propose a slotted ALOHA (SA) scheme with multiple frequency bins, in which the slot duration and the number of bins are adjustable, i.e., a frame of duration T, determined by the delay constraint, is segmented into M slots with equal length. The bandwidth W is also evenly partitioned into B subbands. In addition, our SA scheme has memory. The payload can be transmitted in multiple retransmission attempts by selecting one bin at random at each attempt, where the resulted SNR of each attempt is combined at the receiver. To prevent decoding failure, i.e., outage, the payload needs to be transmitted in multiple attempts by selecting one bin at random at each attempt, where the resulted SNR of each transmission attempt is combined at the receiver at the origin.

We assume the content placement distribution of the transmitters is i.i.d, i.e., the files are independently placed in the cache memories of different nodes according to the same distribution. We denote this distribution by p_c and the request distribution by p_r. At each time slot, the receiver is associated with the nearest transmitter that has the desired content.

The aggregate process of transmitters from the original transmissions and due to the failed transmission attempts are assumed to occur at the beginning of each slot. In the example of Fig. 5.2, a system model for $M = 4$ retransmissions and $B = 3$ bins is shown. A device arriving during sub-slot

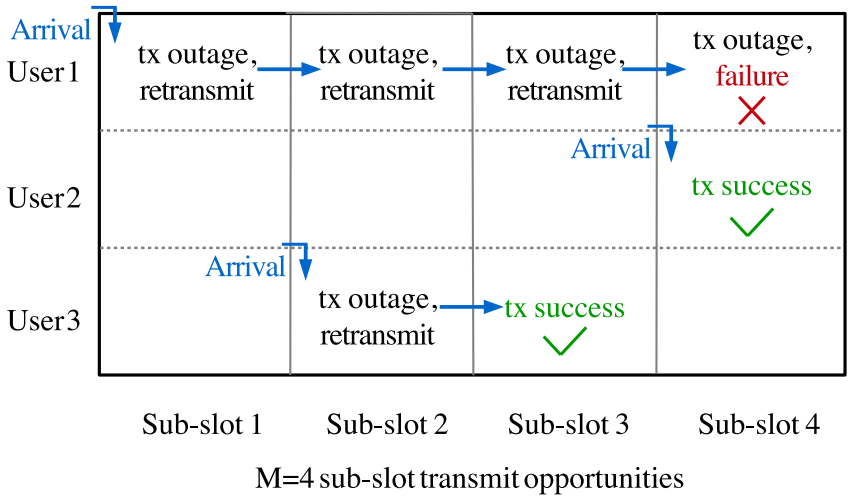


Figure 5.2: An illustration of the retransmission process. The packet success and failure events are highlighted for $M = 4$.

1 has until sub-slot 4 to transmit. Another device arriving during sub-slot 2 has until sub-slot 5 to transmit. This emphasizes the fact that devices can arrive during each sub-slot. On the same plot, an illustration of the proposed retransmission process with packet success and failure events are also given.

An outage occurs when the user fails to receive the desired content by a deadline, corresponding to M consecutive attempts, i.e., a decoding error occurs if SINR across multiple transmissions is below the threshold T. A device fails on m^{th} attempt if the SINR in that attempt, i.e., $\mathsf{SINR}(\mathcal{K}_{\mathsf{m}})$, is below the threshold T. Given a target SINR outage rate δ per device, outage occurs if more than a certain number of users share the same resources. Given M,

denote the set of retransmission indices by $\mathcal{M} = \{1, \ldots, M\}$, and the average aggregate device arrival rate per slot by λ_M, which is the sum of the rates of the original arrivals per slot, i.e., λ, and the arrivals occurring as a result of failed transmissions up to a maximum of M consecutive attempts.

We define $\mathbb{P}_{\text{out}}$ to be the probability of outage on the m^{th} attempt for a given SINR threshold T, which is given as

$$\mathrm{P}_{\mathsf{Fail}}(m) = \mathbb{P}[\mathsf{SINR}_1 < \Gamma, \mathsf{SINR}_2 < \Gamma, \ldots, \mathsf{SINR}_m < \Gamma]. \tag{5.8}$$

Chase combining is used to aggregate the received signals across multiple transmissions, resulting in maximal ratio combining of the desired signal. For tractability, we assume there are no errors or delay in the feedback, so there is immediate retransmission on the next sub-slot after a failure. Therefore, our scheme gives an upper bound on the best achievable performance given a target outage rate.

Our objective is to characterize the performance in terms of the maximum hit rate that can be achieved for a fixed maximum delay for a given number of resource symbols.

Definition 9. Chase combiner. *If $M > 1$, a device is allowed to retransmit if the preceding one fails, for a total of M transmissions. In general, if the received signal vector during transmission $i = 1, \ldots, M$ is $\mathbf{r}_i = a_i\mathbf{s} + \mathbf{n}_i$, where $\mathbf{s} \in \mathbb{C}^n$ is the desired signal, a_i is the complex amplitude, and $\mathbf{n}_i$ is an n-dimensional complex Gaussian noise vector with zero mean and covariance*

$\sigma_i^2\mathbf{I}_n$, *then the SNR at the output of the chase combiner*[1] *is [125]*

$$\left(\sum_{i=1}^{M}|a_i|^2\right)^2 \Big/ \sum_{i=1}^{M}|a_i|^2\sigma_i^2. \tag{5.9}$$

We assume that the encoded blocklength n symbols is sufficiently large that we can exploit the Shannon limit to characterize the performance. The Shannon capacity as a function of the SINR is expressed as $C(\mathsf{SINR}) = \log_2(1+\mathsf{SINR})$. For different number of retransmission attempts, we will investigate the performance for both no fading and Rayleigh fading cases.

Given maximum number of retransmission attempts M, let $u_i(m) \sim \mathsf{Poisson}\left(\frac{\lambda_M}{M}\right)$ be the number of arrivals[2] at the specified sub-slot where $i \leq t$ is the time slot index, m is the retransmission attempt number, and λ_M is the aggregate arrival rate per slot with up to M total transmissions, and is given by

$$\lambda_M = \lambda\left[1 + \sum_{m=1}^{M-1}\mathrm{P}_{\mathsf{Fail}}(m)\right]. \tag{5.10}$$

For ease of notation, the number of arrivals on the m^{th} attempt is denoted by k_m, and the set of arrivals up to including m^{th} attempt is denoted by $\mathcal{K}_m \coloneqq \{k_1, \ldots, k_m\}$.

[1]It follows from maximum-ratio combining of the signal powers at the receiver as a result of M transmission attempts given that SNR per user is ρ for all transmission attempts $i = 1, \ldots, M$ and users, and the noise power is computed by treating interference as noise. The details of the proof are omitted. Interested reader can refer to [125].

[2]For tractability, we inherently have the Poisson distribution assumption for the composite arrival process. From [126] and [127], this assumption is justifiable when the number of retransmissions is not too large.

For the case of small scale Rayleigh fading with parameter $\mu = 1$, let k_m arrivals choose a given resource bin, with SNR ρ per user[3], during transmission $m \in \mathcal{M}$. Using (5.9) and incorporating the channel power distributions, and from Prop. 9, the chase combiner output SINR from (5.9) as a result of $m \in \mathcal{M}$ transmission is [125]

$$\mathsf{SINR}(\mathcal{K}_{\mathsf{M}}) \overset{(a)}{=} \frac{\rho\Big(\sum\limits_{m\in\mathcal{M}} h_m\Big)^2}{\sum\limits_{m\in\mathcal{M}} h_m\left(1+I_{k_m}\right)} \overset{(b)}{=} \frac{M^2\rho h_M}{M+\sum\limits_{m\in\mathcal{M}} I_{k_m}}, \tag{5.11}$$

where $h_m, g_{i,m} \sim \exp(\mu)$, $m \in \mathcal{M}$ are independent and identically distributed (i.i.d.) channel power distributions of the desired device and the interferers, respectively, where $i_m \in \{1, \ldots, k_m - 1\}$ is the interferer index at retransmission attempt m, (a) follows from letting $I_{k_m} = \sum_{i=1}^{k_m-1} \rho g_{i_m}$ denote the total interference seen at transmission attempt m, and (b) is based on the assumption that h_M is unchanged within a time slot[4].

We mathematically write the hit rate –as characterized by Poisson arrival rate– optimization problem as

$$\begin{aligned} \mathsf{R}_{\mathsf{Hit}} = \max_{B,\, M\in\mathbb{Z}^+} \quad & 1 - \mathrm{P}_{\mathsf{Fail}}(M) \\ \text{s.t.} \quad & \mathrm{P}_{\mathsf{Fail}}(M) \leq \delta, \quad \Gamma = 2^{\frac{L}{n}} - 1, \\ & C(\mathsf{SINR}(\mathcal{K}_{\mathsf{m}})) \geq \frac{L}{n},\ m \in \mathcal{M},\ n \leq \frac{TW}{MB}, \end{aligned} \tag{5.12}$$

[3]Assuming perfect channel inversion power control such that the average received SNR per device is fixed, the locations of the devices do not play a role in the system performance. This assumption can be relaxed using fractional or no power control.

[4]We use the Rayleigh block fading model [128] in which the power fading coefficients remain static over each time slot, and are temporally (and spatially) independent with exponential distribution of mean $\mu = 1$.

where $\mathrm{P}_{\mathsf{Fail}}(M)$ is given in (5.8), and $C(\mathsf{SINR}(\mathcal{K}_{\mathsf{m}}))$ is the Shannon capacity as a function of the Chase combiner output SINR given in (5.11) conditioned on the set of arrivals $\mathcal{K}_m$.

The probability of failure $\mathrm{P}_{\mathsf{Fail}}(M)$ can be computed as a function of λ_M using a similar methodology as in Theorem 2 of Chapter 2 as

$$\mathrm{P}_{\mathsf{Fail}} = \mathrm{P}_{\mathsf{Fail}}(T, \lambda_M, p_c, p_r, \alpha). \tag{5.13}$$

From (5.10), we can observe that λ_M is also a function of $\mathrm{P}_{\mathsf{Fail}}(m)$'s, where $m \in \{1, \ldots, M\}$. Therefore, to optimize the performance to compute the maximum cache hit rate, it is required to solve a fixed point equation that comes from (5.10) and (5.13) and determine the values of B and M under low SNR and high SNR regimes for optimization of resources.

As discussed in this dissertation, content caching has been studied using different tools ranging from stochastic geometry to game theory. However, these models have certain limitations. It is hard to model the geographical locality of the content, the request distributions might be non-stationary, and the SINR coverage of the network varies due to locality of the content. Despite previous research efforts, to the best of our knowledge, there has been no study focusing on the investigation of the spatial-temporal dynamics of content caching with realistic interference and fading models in the context of cellular networks.

We propose to build a Matlab simulator to test and compare these caching algorithms on increasingly large networks and file sets (may require

developing new, more efficient suboptimal algorithms). Test recent theoretical results covered in Chapters 2, 3, 4, including other well known caching models and approaches. We hope and expect that doing the above will lead to insights and new, improved caching algorithms, and accompanying theoretical models and analysis.

Our broad objective is to analyze and design cellular networks in order to optimize the performance of caching, and best support broadband data and short packets for the development of future 5G networks with heterogeneous QoS constraints.

Appendices

Appendix A

Appendix to Chapter 2

A.1 Proof of Lemma 6

We investigate the general solution of (2.7). Using the Lagrange multiplier method [129], we define

$$\Lambda(\mathbf{p_c}, \eta) = \sum\nolimits_{i=1}^{M} \lambda_t p_r(i)\, \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha) + \eta\Big(\sum\nolimits_{i=1}^{M} p_c(i) - 1\Big).$$

The partial derivatives of $\Lambda(\mathbf{p_c}, \eta)$ with respect to $p_c(i)$ for $i = 1, \ldots, M$ give M equations.

$$\begin{aligned}
\frac{\partial \Lambda(\mathbf{p_c}, \eta)}{\partial p_c(i)} &= \lambda_t p_r(i) \frac{\partial\, \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha)}{\partial p_c(i)} + \eta \\
&= \lambda_t p_r(i) \frac{\partial \Big[\pi \lambda_t p_c(i) \int\limits_0^\infty e^{-\pi \lambda_t p_c(i) r \beta(\mathrm{T},\alpha) - \mu\, \mathrm{T}\, \sigma^2 r^{\alpha/2}}\, \mathrm{d}r\Big]}{\partial p_c(i)} + \eta \\
&= \lambda_t p_r(i) \Big[\pi \lambda_t \int_0^\infty e^{-\pi \lambda_t p_c(i) r \beta(\mathrm{T},\alpha) - \mu \mathrm{T} \sigma^2 r^{\alpha/2}}\, \mathrm{d}r \\
&\quad - (\pi \lambda_t)^2 \beta(\mathrm{T}, \alpha) p_c(i) \int_0^\infty e^{-\pi \lambda_t p_c(i) r \beta(\mathrm{T},\alpha) - \mu \mathrm{T} \sigma^2 r^{\alpha/2}} r\, \mathrm{d}r\Big] + \eta. \quad \text{(A.1)}
\end{aligned}$$

To maximize $\Lambda(\mathbf{p_c}, \eta)$, equate the RHS of (A.1) to 0 and obtain

$$\int_0^\infty [1 - \pi \lambda_t \beta(\mathrm{T}, \alpha) p_c(i) r] e^{-\pi \lambda_t p_c(i) r \beta(\mathrm{T},\alpha) - \mu \mathrm{T} \sigma^2 r^{\alpha/2}}\, \mathrm{d}r = -\frac{\eta}{p_r(i) \pi \lambda_t^2}. \quad \text{(A.2)}$$

The partial derivative $\frac{\partial \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha)}{\partial \lambda_t}$ is given as

$$\frac{\partial \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha)}{\partial \lambda_t} = \pi p_c(i) \int_0^\infty e^{-\pi \lambda_t p_c(i) r \beta(\mathrm{T},\alpha) - \mu \mathrm{T} \sigma^2 r^{\alpha/2}}\, \mathrm{d}r$$

$$- (\pi p_c(i))^2 \beta(\mathrm{T},\alpha)\lambda_t \int_0^\infty e^{-\pi\lambda_t p_c(i) r\beta(\mathrm{T},\alpha) - \mu \mathrm{T}\sigma^2 r^{\alpha/2}} r \,\mathrm{d}r$$
$$= \frac{p_c(i)}{\lambda_t} \frac{\partial\, \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha)}{\partial p_c(i)}. \tag{A.3}$$

Combining the relations (A.2) and (A.3) results in $\frac{\partial\, \mathrm{p_{cov}}(\mathrm{T},\lambda_t p_c(i),\alpha)}{\partial \lambda_t} = -\eta \frac{p_c(i)}{\lambda_t^2 p_r(i)}$. Using the definition of $\mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha)$, we note that $\mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha) = \mathrm{p_{cov}}(\mathrm{T}\,(p_c(j)/p_c(i))^{\alpha/2}, \lambda_t p_c(j), \alpha)$. Taking the derivative of this expression with respect to λ_t, we have

$$\frac{\partial\, \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(j), \alpha)}{\partial \lambda_t} = -\eta \frac{p_c(j)}{\lambda_t^2 p_r(j)} = \frac{\partial\, \mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha)}{\partial \lambda_t} \frac{p_r(i)/p_c(i)}{p_r(j)/p_c(j)}. \tag{A.4}$$

We can rewrite (A.4) using the expression for $\mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(i), \alpha)$ as follows

$$= \frac{\partial\, \mathrm{p_{cov}}(\mathrm{T}\left(\frac{p_c(j)}{p_c(i)}\right)^{\alpha/2}, \lambda_t p_c(j), \alpha)}{\partial \lambda_t} \frac{p_r(i)/p_c(i)}{p_r(j)/p_c(j)}. \tag{A.5}$$

Next, by employing a change of variables $v = r\beta(\mathrm{T}, \alpha)$, we can rewrite (2.1) in Definition 2.1 as

$$\mathrm{p_{cov}}(\mathrm{T}, \lambda_t, \alpha) = \frac{\pi\lambda_t}{\beta(\mathrm{T},\alpha)} \int_0^\infty e^{-\pi\lambda_t v - \mu\left[\frac{\mathrm{T}}{\beta(\mathrm{T},\alpha)^{\frac{\alpha}{2}}}\right]\sigma^2 v^{\frac{\alpha}{2}}} \,\mathrm{d}v. \tag{A.6}$$

We investigate the relation between $\beta(\mathrm{T},\alpha)^{\alpha/2}$ and T in Fig. 2.4, for practical α and μ values, and observe the linear dependence, where the slope is mainly determined by α, and changes only slightly by varying μ. Based on these simulations, since $\beta(\mathrm{T},\alpha)^{\alpha/2}/\mathrm{T}$ is invariant to T and using the relation in (A.6), it is reasonable to write $\mathrm{p_{cov}}(\mathrm{T}, \lambda_t p_c(j), \alpha)$ as a separable function which is the form $f(\lambda_t p_c(j), \alpha)g(\mathrm{T})$. By taking its derivative with respect to λ_t, we can then rewrite (A.5) as

$$g(\mathrm{T}) = g\Big(\mathrm{T}\Big(\frac{p_c(j)}{p_c(i)}\Big)^{\alpha/2}\Big) \frac{p_c(j)}{p_c(i)} \Big(\frac{j}{i}\Big)^{\gamma_r}. \tag{A.7}$$

Taking the derivative of both sides with respect to T, we obtain

$$\frac{dg(\mathrm{T})}{d\,\mathrm{T}} = \Big(\frac{p_c(j)}{p_c(i)}\Big)^{\alpha/2} \frac{dg(\mathrm{T})}{d\,\mathrm{T}} \frac{p_c(j)}{p_c(i)} \Big(\frac{j}{i}\Big)^{\gamma_r},$$

implying that $p_c(j)/p_c(i) = (i/j)^{\frac{\gamma_r}{\alpha/2+1}}$. Then, $p_c(\cdot)$ is also Zipf(γ_c) distributed with parameter $\gamma_c = \frac{\gamma_r}{\alpha/2+1}$.

Appendix B

Appendix to Chapter 3

B.1 Proof of Proposition 3

We evaluate L and K by approximating $\mathrm{p}_{\mathrm{c,G}_c}^{\mathrm{Lin}}(m)$ using the expression $\mathrm{p}_{\mathrm{c,G}}^{*}(m)$ in (3.5). Incorporating the finite cache size constraint to (3.8) and solving $\sum_{m=1}^{M} \mathrm{p}_{\mathrm{c,G}}^{\mathrm{Lin}}(m) = N$, we obtain:

$$K + L - 1 = 2N. \tag{B.1}$$

Using the optimal cache placement probabilities for independent placement as given in (3.5) and the relation $\frac{p_r(m)}{p_r(K)} = (K/m)^{\gamma_r}$, we have

$$\frac{\mathrm{p}_{\mathrm{c,G}}^{\mathrm{Lin}}(m)}{p_c^{\mathrm{Lin}}(n)} \approx \frac{\mathrm{p}_{\mathrm{c,G}}(m)}{\mathrm{p}_{\mathrm{c,G}}(n)} = \frac{\log\left(\frac{\lambda_{\mathrm{t}}\,\pi \mathrm{R}_{\mathsf{D2D}}^2}{\mu^*}\right) - \log\left(\sum_{i=1}^{M} \frac{1}{i^{\gamma_r}}\right) - \gamma_r \log(m)}{\log\left(\frac{\lambda_{\mathrm{t}}\,\pi \mathrm{R}_{\mathsf{D2D}}^2}{\mu^*}\right) - \log\left(\sum_{i=1}^{M} \frac{1}{i^{\gamma_r}}\right) - \gamma_r \log(n)}, \; L < m, n < K, \tag{B.2}$$

which yields the following approximation for K:

$$K \approx \frac{1}{\gamma_r} \log\Big(\frac{\lambda_{\mathrm{t}}\,\pi \mathrm{R}_{\mathsf{D2D}}^2}{\mu^*}\Big) - \frac{1}{\gamma_r} \log\Big(\sum_{i=1}^{M} \frac{1}{i^{\gamma_r}}\Big). \tag{B.3}$$

Using the boundary conditions in (3.5), the optimal value μ^* is such that $p_r(L+1)\mathbb{P}(\mathcal{N}_P = 1) \leq \mu^* \leq p_r(K-1)\mathbb{E}[\mathcal{N}_P]$. Equivalently, $p_r(L+1)\,\lambda_{\mathrm{t}}\,\pi \mathrm{R}_{\mathsf{D2D}}^2 e^{-\lambda_{\mathrm{t}}\,\pi \mathrm{R}_{\mathsf{D2D}}^2} \leq \mu^* \leq p_r(K-1)\,\lambda_{\mathrm{t}}\,\pi \mathrm{R}_{\mathsf{D2D}}^2$. We determine the best pair (L, K), given the relations (B.1) and (B.3) and the optimal value μ^*.

B.2 Proof of Proposition 6

We first consider the case $r_m \geq \mathrm{R}_{\mathsf{D2D}}$, where the user can be covered by at most one transmitter that has file m. The probability that the user is covered is given by the probability that there exists a transmitter of the HCP-A process of file m at the origin as determined by [61, Ch. 2.1]

$$\begin{aligned}\mathbb{P}(\tilde{C}_m = 1 | r_m \geq \mathrm{R}_{\mathsf{D2D}}) &= \mathbb{E}[\tilde{C}_m | r_m \geq \mathrm{R}_{\mathsf{D2D}}] \\ &= \lambda_{\mathsf{HCP\text{-}A}}(m) \pi \, \mathrm{R}_{\mathsf{D2D}}^2 \\ &= [1 - e^{-\bar{C}_m}] \Big(\frac{\mathrm{R}_{\mathsf{D2D}}}{r_m} \Big)^2. \end{aligned} \tag{B.4}$$

For the case where $r_m < \mathrm{R}_{\mathsf{D2D}}$, we can estimate $\mathbb{P}(\tilde{C}_m \geq 1 | r_m < \mathrm{R}_{\mathsf{D2D}})$ using the second-order product density of the MHC model. However, we use a simpler approximation for tractability. The probability that a transmitter is eliminated in the HCP-A with exclusion radius r_m is equal to $1 - \frac{\lambda_{\mathsf{HCP\text{-}A}}(m)}{\lambda_{\mathrm{t}}}$. For the case of $r_m < \mathrm{R}_{\mathsf{D2D}}$, let the number of points in $B(r_m)$ from the original PPP satisfy $\Phi(B_0(\mathrm{R}_{\mathsf{D2D}})) = k$. Since HCP-A is negatively correlated, from Definition 2, we can exploit the PPP approximation for the MHC in [130] to calculate the following upper bound for the probability that k points are eliminated in HCP-A Φ_M:

$$\begin{aligned}&\mathbb{P}(k \text{ points eliminated in } \Phi_M \text{ given exclusion radius } = r_m | \mathcal{N}_P = k) \\ &\qquad \leq \Big(1 - \frac{\lambda_{\mathsf{HCP\text{-}A}}(m)}{\lambda_{\mathrm{t}}} \Big)^k. \end{aligned} \tag{B.5}$$

Using (B.5), the void probability of the HCP-A is approximated as

$$\mathbb{P}(\tilde{C}_m = 0 | r_m < \mathrm{R}_{\mathsf{D2D}}) \leq \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{N}_P = k) \Big(1 - \frac{\lambda_{\mathsf{HCP\text{-}A}}(m)}{\lambda_{\mathrm{t}}} \Big)^k$$

$$= \exp\left(-\lambda_{\mathsf{HCP\text{-}A}}(m)\pi \mathrm{R}_{\mathsf{D2D}}^2\right). \tag{B.6}$$

The relations (B.4) and (B.6) yield the final result.

B.3 Proof of Proposition 9

Using the hit probabilities given in (3.4) and (3.19), respectively for the independent and HCP-A content placements, a necessary condition for the HCP-A to perform better than the optimal independent placement model in [60] in terms of hit probability is given by

$$\begin{aligned}\mathrm{P}_{\mathsf{Hit,HCP\text{-}A}} &= \sum_{m=1}^{M} p_r(m)\mathbb{P}(\tilde{C}_m \geq 1 | r_m) \\ &\geq \mathrm{P}_{\mathsf{Hit,G}} = \sum_{m=1}^{M} p_r(m)[1 - \exp(-\lambda_{\mathrm{t}}\, \mathrm{p}^*_{\mathrm{c,G}}(m)\pi \mathrm{R}_{\mathsf{D2D}}^2)]. \end{aligned} \tag{B.7}$$

A sufficient condition for (B.7) to be valid is given by $\mathbb{P}(\tilde{C}_m \geq 1 | r_m) \geq 1 - \exp\left(-\lambda_{\mathrm{t}}\, \mathrm{p}^*_{\mathrm{c,G}}(m)\pi \mathrm{R}_{\mathsf{D2D}}^2\right)$. For files with very high popularity, from (B.6):

$$\begin{aligned}\mathbb{P}(\tilde{C}_m \geq 1 | r_m < \mathrm{R}_{\mathsf{D2D}}) &\geq 1 - \exp(-\lambda_{\mathsf{HCP\text{-}A}}(m)\pi \mathrm{R}_{\mathsf{D2D}}^2) \\ &\geq 1 - \exp(-\lambda_{\mathrm{t}}\, \mathrm{p}^*_{\mathrm{c,G}}(m)\pi \mathrm{R}_{\mathsf{D2D}}^2). \end{aligned} \tag{B.8}$$

For files with very low popularity, r_m tends to be very high, i.e., $r_m \geq \mathrm{R}_{\mathsf{D2D}}$, and from (B.4),

$$\begin{aligned}\mathbb{P}(\tilde{C}_m = 1 | r_m \geq \mathrm{R}_{\mathsf{D2D}}) &= \lambda_{\mathsf{HCP\text{-}A}}(m)\pi\, \mathrm{R}_{\mathsf{D2D}}^2 \\ &\geq 1 - \exp(-\lambda_{\mathrm{t}}\, \mathrm{p}^*_{\mathrm{c,G}}(m)\pi \mathrm{R}_{\mathsf{D2D}}^2). \end{aligned} \tag{B.9}$$

Solving (B.8) and (B.9), the final result is obtained.

The following relation is established from (B.8) and (B.9):

$$\sum_{m=1}^{M} \lambda_{\mathsf{HCP\text{-}A}}(m) \;\; \geq \sum_{m=1}^{\mathrm{m_c}} \lambda_{\mathrm{t}}\, \mathrm{p}^{*}_{\mathrm{c,G}}(m) + \sum_{m=\mathrm{m_c}+1}^{M} \frac{1-\exp(-\lambda_{\mathrm{t}}\, \mathrm{p}^{*}_{\mathrm{c,G}}(m)\pi \mathrm{R}^2_{\mathsf{D2D}})}{\pi\, \mathrm{R}^2_{\mathsf{D2D}}}, \quad \text{(B.10)}$$

where using $1 - e^{-x} \leq x$ for $x \geq 0$, the RHS of (B.10) can be shown to satisfy:

$$\leq \sum_{m=1}^{\mathrm{m_c}} \lambda_{\mathrm{t}}\, \mathrm{p}^{*}_{\mathrm{c,G}}(m) + \sum_{m=\mathrm{m_c}+1}^{M} \lambda_{\mathrm{t}}\, \mathrm{p}^{*}_{\mathrm{c,G}}(m) = \lambda_{\mathrm{t}} \sum_{m=1}^{M} \mathrm{p}^{*}_{\mathrm{c,G}}(m) = N\, \lambda_{\mathrm{t}}\,.$$

For a feasible cache placement strategy, we also require that $\sum_{m=1}^{M} \lambda_{\mathsf{HCP\text{-}A}}(m) \leq N\, \lambda_{\mathrm{t}}$. Hence, it is possible to set $\lambda_{\mathsf{HCP\text{-}A}}(m)$'s as in (3.29) and satisfy the feasible placement condition.

Appendix C

Appendix to Chapter 4

C.1 Proof of Theorem 6

Letting $f(r_{xu}) = p_r^x(c_u) \exp\left(-\gamma/l(r_{xu}) - \beta r_{xu}^2\right) = p_r^x(c_u)\left(1 - \gamma l(r_{xu}) - \beta r_{xu}^2\right)$, where $\gamma = \mu T \sigma^2$ and $\beta = \pi\lambda\rho(T,\alpha)$, the MGF of $\mathcal{B}_\phi(x)$, i.e., $M_{\mathcal{B}_\phi(x)}(t) = \mathbb{E}[e^{t\,\mathcal{B}_\phi(x)}]$, is given as follows:

$$\begin{aligned} M_{\mathcal{B}_\phi(x)}(t) &= \mathbb{E}_{|\mathcal{U}_x|}\Big[\mathbb{E}\Big[\exp\left(t\,\mathcal{B}_\phi(x)\right)\Big||\mathcal{U}_x|\Big]\Big] \\ &= \mathbb{E}_{|\mathcal{U}_x|}\Big[\mathbb{E}\Big[\exp\Big(t\sum_{u=1}^{|\mathcal{U}_x|} f(r_{xu})\Big)\Big||\mathcal{U}_x|\Big]\Big] \\ &= \mathbb{E}_{|\mathcal{U}_x|}\Big[\mathbb{E}\Big[\prod_{u=1}^{|\mathcal{U}_x|} \exp\left(tf(r_{xu})\right)\Big||\mathcal{U}_x|\Big]\Big] \\ &\overset{(a)}{=} \mathbb{E}_{|\mathcal{U}_x|}\Big[\mathbb{E}\Big[\exp\left(tf(r_{xu})\right)\Big]^{|\mathcal{U}_x|}\Big], \end{aligned} \tag{C.1}$$

where (a) is due to that conditional on having $|\mathcal{U}_x|$ receivers in $B_x(\mathrm{R}_{\mathsf{D2D}})$, via Poisson property, u's are i.i.d. in $B_x(\mathrm{R}_{\mathsf{D2D}})$.

Letting $a(t) = \mathbb{E}\Big[\exp\left(tf(r_{xu})\right)\Big]$, we have

$$\begin{aligned} a(t) &\overset{(a)}{=} \frac{1}{\pi\,\mathrm{R}_{\mathsf{D2D}}^2}\int_{u\in B_x(\mathrm{R}_{\mathsf{D2D}})} \exp\left(tf(|x-u|)\right)\mathrm{d}u \\ &= \frac{2}{\mathrm{R}_{\mathsf{D2D}}^2}\int_0^{\mathrm{R}_{\mathsf{D2D}}} \exp\left(tf(r)\right) r\mathrm{d}r \end{aligned}$$

$$\stackrel{(b)}{=} \frac{1}{\mathrm{R}_{\mathsf{D2D}}^2} \int_0^{\mathrm{R}_{\mathsf{D2D}}^2} \exp\left(t p_r^x(c_u)\left(1-\gamma/l(v^{1/2})-\beta v\right)\right) \mathrm{d}v, \tag{C.2}$$

where (a) is due to that u's are i.i.d. and uniformly distributed inside $B_x(\mathrm{R}_{\mathsf{D2D}})$, and (b) is obtained by employing a change of variables $v = r^2$. Therefore,

$$M_{\mathcal{B}_\phi(x)}(t) = \mathbb{E}_{|\mathcal{U}_x|}\left[a(t)^{|\mathcal{U}_x|}\right],$$

where we used a similar approach to the definition of the MGF of the Poisson distribution, which yields $E[e^{t|\mathcal{U}_x|}] = \exp(\lambda_r^x \pi \, \mathrm{R}_{\mathsf{D2D}}^2 (e^t - 1))$, where $|\mathcal{U}_x| \sim \mathrm{Poisson}(\lambda_r^x \pi \, \mathrm{R}_{\mathsf{D2D}}^2)$. We can obtain the final result by using the fact that $p_r^x(c_u)$'s are i.i.d. and $a(t)$ does not depend on $\mathcal{U}_x$.

Bibliography

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2016 - 2021," white paper, 2017.

[2] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication," *IEEE Commun. Mag.*, pp. 36–43, May 2014.

[3] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and R. Baeza-Yates, "A machine learning approach for result caching in web search engines," *Information Processing and Management*, vol. 53, no. 4, pp. 834 – 850, Jul. 2017.

[4] D. Freitag, "Greedy attribute selection," in *Machine Learning Proceedings 1994: Proceedings of the Eighth International Conference*. Morgan Kaufmann, Jan. 2017, p. 28.

[5] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, pp. 131–139, Feb. 2014.

[6] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "FlashLinQ: A synchronous distributed scheduler for Peer-

to-Peer ad hoc networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1215–1228, Aug. 2013.

[7] 3GPP, "Study on LTE Device to Device proximity services," 3GPP, Tech. Rep., Dec. 2012.

[8] 3GPP TR 22.803 V1.0.0, "3rd Generation Partnership Project; technical specification group SA; feasibility study for proximity services (ProSe) (release 12)," 3GPP, Tech. Rep., Aug. 2012.

[9] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014.

[10] ——, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, Apr. 2014.

[11] Firechat. [Online]. Available: https://www.opengarden.com/firechat.html

[12] T. Simonite, "FireChat could be the first in a wave of mesh networking apps," Mar. 2014.

[13] N. Naderializadeh, D. T. Kao, and A. S. Avestimehr, "How to utilize caching to improve spectral efficiency in device-to-device wireless networks," in *Proc., Annu. Allerton Conf.*, Illinois, USA, Oct. 2014.

[14] Y.-D. Lin and Y.-C. Hsu, "Multihop cellular: A new architecture for wireless communications," in *Proc., IEEE Infocom*, vol. 3, Mar. 2000, pp. 1273–1282.

[15] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated cellular and ad hoc relaying systems: iCAR," *IEEE Journ. on Sel. Areas in Comm.*, vol. 19, no. 10, pp. 2105–2115, Oct. 2001.

[16] H. Luo, R. Ramjee, P. Sinha, L. E. Li, and S. Lu, "UCAN: A unified cellular and ad-hoc network architecture," in *Proc., MobiCom*, Sep. 2003, pp. 353–367.

[17] B. Liu, Z. Liu, and D. Towsley, "On the capacity of hybrid wireless networks," in *Proc., IEEE Infocom*, vol. 2, Mar. 2003, pp. 1543–1552.

[18] U. C. Kozat and L. Tassiulas, "Throughput capacity of random ad hoc networks with infrastructure support," in *Proc., ACM MobiCom*, Sep. 2003, pp. 55–65.

[19] H.-Y. Hsieh and R. Sivakumar, "On using Peer-to-Peer communication in cellular wireless data networks," *IEEE Trans. Mobile Comput.*, vol. 3, no. 1, pp. 57–72, Jan.-Feb. 2004.

[20] A. Zemlianov and G. de Veciana, "Capacity of ad hoc wireless networks with infrastructure support," *IEEE Journ. on Sel. Areas in Comm.*, vol. 23, no. 3, pp. 657–667, Mar. 2005.

[21] F. H. Fitzek, M. Katz, and Q. Zhang, "Cellular controlled short-range communication for cooperative P2P networking," *Wireless Personal Communications*, vol. 48, no. 1, pp. 141–155, Jan. 2009.

[22] A. Berl and H. De Meer, "Integration of mobile devices into popular Peer-to-Peer networks," in *Proc., Next Generation Internet Networks*, Jul. 2009, pp. 1–9.

[23] E. Yaacoub and O. Kubbar, "Energy-efficient Device-to-Device communications in LTE public safety networks," in *Proc., IEEE Globecom Workshops*, Dec. 2012, pp. 391–395.

[24] Q. Ye, C. Caramanis, and J. G. Andrews, "Device-to-Device communications: a survey," University of Texas at Austin, Austin, TX, Tech. Rep., 2013.

[25] X. Lin, "Integrated cellular and Device-to-Device networks," Ph.D. dissertation, The University of Texas at Austin, Dec. 2014.

[26] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Resource optimization in Device-to-Device cellular systems using time-frequency hopping," *IEEE Trans. Wireless Comm.*, vol. 13, no. 10, pp. 5467–5480, Oct. 2014.

[27] Q. Ye, B. Rong, Y. Chen, M. A.-S. M, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–16, Jun. 2013.

[28] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. Wireless Comm.*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014.

[29] X. Lin and J. G. Andrews, "Optimal spectrum partition and mode selection in Device-to-Device overlaid cellular networks," in *Proc. IEEE Globecom*, Dec. 2013, pp. 1837–1842.

[30] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "A tractable model for optimizing Device-to-Device communications in downlink cellular networks," in *Proc., IEEE ICC*, Jun. 2014, pp. 2039–2044.

[31] F. Baccelli, N. Khude, R. Laroia, J. Li, T. Richardson, S. Shakkottai, and X. Wu, "On the design of Device-to-Device autonomous discovery," in *Proc., IEEE International Conference on Communication Systems and Networks (COMSNETS)*, Jan. 2012, pp. 1–9.

[32] Y. Du and G. de Veciana, "Wireless networks without edges: Dynamic radio resource clustering and user scheduling," in *Proc., IEEE Infocom*, Apr. 2014, pp. 1321–1329.

[33] V. Shah and G. de Veciana, "Performance evaluation and asymptotics for content delivery networks," in *Proc., IEEE Infocom*, Apr. 2014, pp. 2607–2615.

[34] N. Naderializadeh and A. S. Avestimehr, "ITLinQ: A new approach for spectrum sharing in Device-to-Device communication systems," *IEEE*

Journ. on Sel. Areas in Comm., vol. 32, no. 6, pp. 1139–1151, Jun. 2014.

[35] R. K. Mungara, X. Zhang, A. Lozano, and R. W. Heath, "On the spatial spectral efficiency of ITLinQ," in *Proc. Annual Asilomar Conf. Signals, Syst., Comp.*, Nov. 2014, pp. 1806 – 1810.

[36] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive resource allocation: Harnessing the diversity and multicast gains," *IEEE Trans. Inf. Theory*, vol. 59, no. 8, pp. 4833–4854, Aug. 2013.

[37] J. Tadrous and A. Eryilmaz, "On optimal proactive caching for mobile networks with demand uncertainties," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2715–2727, Oct. 2016.

[38] O. Shoukry, M. A. ElMohsen, J. Tadrous, H. E. Gamal, T. ElBatt, N. Wanas, Y. Elnakieb, and M. Khairy, "Proactive scheduling for content pre-fetching in mobile networks," in *Proc., IEEE ICC*, Jun. 2014, pp. 2848 – 2854.

[39] O. Gungor, O. O. Koyluoglu, H. E. Gamal, and C. E. Koksal, "Proactive source coding," in *Proc., IEEE ISIT*, Jul. 2011, pp. 2213 – 2217.

[40] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive data download and demand shaping," Ohio State University, Tech. Rep., 2013. [Online]. Available: http://www2.ece.ohio-state.edu/~tadrousj/ProactiveTechReport.pdf

[41] ——, "Proactive content download and user demand shaping for data networks," *IEEE/ACM Trans. Netw.*, vol. 23, no. 6, pp. 1917–1930, Dec. 2015.

[42] J. Tadrous, H. E. Gamal, and A. Eryilmaz, "Can carriers make more profit while users save money?" in *Proc., IEEE ISIT*, Jun. 2014, pp. 1757 – 1761.

[43] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Joint pricing and proactive caching for data services: Global and user-centric approaches," in *Proc., IEEE Infocom Workshops*, Apr. 2014.

[44] F. Alotaibi, S. Hosny, H. E. Gamal, and A. Eryilmaz, "Dynamic pricing and proactive caching for P2P mobile networks," in *Proc., Information Theory and Applications Workshop*, Feb. 2015.

[45] ——, "A game theoretic approach to content trading in proactive wireless networks," in *Proc., IEEE ISIT*, Jun. 2015, pp. 2216 – 2220.

[46] E. Bastug, K. Hamidouche, W. Saad, and M. Debbah, "Centrality-based caching for mobile wireless networks," in *Proc., 1st KuVS Workshop on Anticipatory Networks*, Stuttgart, Germany, Sep. 2014.

[47] Y. Chen and B. Tang, "Data caching in ad hoc networks using game-theoretic analysis," in *Proc., IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*, 2010.

[48] M. X. Goemans, L. E. Li, V. S. Mirrokni, and M. Thottan, "Market sharing games applied to content distribution in ad hoc networks," *IEEE Journ. on Sel. Areas in Comm.*, vol. 24, no. 5, pp. 1020–1033, May 2006.

[49] M. K. Hanawal, E. Altman, and F. Baccelli, "Stochastic geometry based medium access games in wireless ad hoc networks," *IEEE Journ. on Sel. Areas in Comm.*, vol. 30, no. 11, pp. 2146–2157, Dec. 2012.

[50] R. Zhang, J. Zhang, Y. Zhang, J. Sun, and G. Yan, "Privacy-preserving profile matching for proximity-based mobile social networking," *IEEE Journ. on Sel. Areas in Comm.*, vol. 31, no. 9, pp. 656–668, Sep. 2013.

[51] Y. Zhang, E. Pan, L. Song, W. Saad, Z. Dawy, and Z. Han, "Social network enhanced Device-to-Device communication underlaying cellular networks," in *Proc., IEEE/CIC Intl. Conf. on Communications in China - Workshops (CIC/ICCC), Intl. Workshop on Device-to-Device Communications and Networks (D2D)*, Aug. 2013, pp. 182–186.

[52] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4365–4380, Oct. 2016.

[53] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *Proc., IEEE Infocom*, Mar. 2010, pp. 1478 – 1486.

[54] Y. Birk and T. Kol, "Coding on demand by an informed source (ISCOD) for efficient broadcast of different supplemental data to caching clients," *IEEE/ACM Trans. Netw.*, vol. 14, no. SI, pp. 2825 – 2830, Jun. 2006.

[55] Z. Bar-Yossef, Y. Birk, T. S. Jayram, and T. Kol, "Index coding with side information," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1479 – 1494, Mar. 2011.

[56] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–67, May 2014.

[57] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[58] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Comm.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

[59] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Info. Theory*, vol. 59, no. 12, pp. 8402–13, Dec. 2013.

[60] B. Błaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc., IEEE ICC*, UK, Jun. 2015, pp. 3358–3363.

[61] F. Baccelli and B. Błaszczyszyn, *Stochastic Geometry and Wireless Networks*. NOW: Found. Trends. Network., 2010.

[62] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Temporal locality in today's content caching: Why it matters and how to model it," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 5, pp. 5–12, Nov. 2013.

[63] J. Elias and B. Błaszczyszyn, "Optimal geographic caching in cellular networks with linear content coding," *arXiv preprint arXiv:1704.08625*, Apr. 2017.

[64] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd ed. John Wiley and Sons, 1996.

[65] T. Liggett, *Interacting particle systems*. Springer Science and Business Media, 2012, vol. 276.

[66] P. L. Krapivsky, S. Redner, and E. Ben-Naim, *A kinetic view of statistical physics*. Cambridge University Press, 2010.

[67] S. Zhou, D. Lee, B. Leng, X. Zhou, H. Zhang, and Z. Niu, "On the spatial distribution of base stations and its relation to the traffic density in cellular networks," *IEEE Access*, vol. 3, pp. 998 – 1010, Jul. 2015.

[68] H. Sarkissian, "The business case for caching in 4G LTE networks," 2012, lSI-Wireless Technical Report.

[69] D. Malak and M. Al-Shalash, "Optimal caching for Device-to-Device content distribution in 5G networks," in *Proc., IEEE Globecom Workshops*, Austin, TX, Dec. 2014.

[70] ——, "Device-to-Device content distribution: Optimal caching strategies and performance bounds," in *Proc., IEEE ICC Workshops*, London, UK, Jun. 2015.

[71] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833–6859, Dec. 2015.

[72] S. Gitzenis, G. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Info. Theory*, vol. 59, no. 5, pp. 2760–2776, May 2013.

[73] S.-W. Jeon, S.-N. Hong, M. Ji, and G. Caire, "Caching in wireless multihop device-to-device networks," in *Proc., IEEE ICC*, Jun. 2015, pp. 6732–6737.

[74] R. L. Cruz, "Ad-hoc networks at global scale," in *Proc., Intl. Conf. on Computing, Networking, and Communications (ICNC)*, Jan. 2013, pp. 813–817.

[75] N. Golrezaei, M. Alexandros G. Dimakis, and A. F. Molisch, "Device-to-Device collaboration through distributed storage," in *Proc., IEEE Globecom*, Dec. 2012, pp. 2397–2402.

[76] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, Jul. 2014.

[77] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[78] M. Ji, G. Caire, and A. F. Molisch, "Wireless Device-to-Device caching networks: Basic principles and system performance," *IEEE Journ. on Sel. Areas in Comm.*, vol. 34, no. 1, pp. 176–89, Jan. 2016.

[79] E. Altman, K. Avrachenkov, and J. Goseling, "Coding for caches in the plane," *arXiv preprint arXiv:1309.0604*, Sep. 2013.

[80] H. P. Keeler, B. Błaszczyszyn, and M. Karray, "SINR-based k-coverage probability in cellular networks with arbitrary shadowing," in *Proc., IEEE ISIT*, Istanbul, Jul. 2013, pp. 1167 – 1171.

[81] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Comm.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

[82] S. Vanichpun, "Comparing strength of locality of reference: Popularity, temporal correlations, and some folk theorems for the miss rates and outputs of caches," Ph.D. dissertation, University of Maryland, 2005.

[83] A. M. Makowski and S. Vanichpun, *Comparing Locality of Reference-Some Folk Theorems for the Miss Rates and the Output of Caches*. Springer US, 2005, ch. 13, pp. 333–365.

[84] L. Pietronero, E. Tosatti, V. Tosatti, and A. Vespignani, "Explaining the uneven distribution of numbers in nature: The laws of Benford and Zipf," *Physica A: Statistical Mechanics and its Applications*, vol. 293, no. 1-2, pp. 297–304, Apr. 2001.

[85] M. Carter, *Foundations of mathematical economics*. MIT Press, 2001.

[86] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing the spatial content caching distribution for device-to-device communications," in *Proc., IEEE ISIT*, Barcelona, Spain, Jul. 2016, pp. 280–284.

[87] ——, "Spatially correlated content caching for device-to-device communications," *arXiv preprint arXiv:1609.00419, under revision, IEEE Transactions on Wireless Communications*, Jun. 2017.

[88] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Hit optimal vs. throughput optimal," *IEEE Communications Letters*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[89] U. Niesen, P. Gupta, and D. Shah, "On capacity scaling in arbitrary wireless networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 3959 – 3982, Sep. 2009.

[90] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–107, May 2017.

[91] A. Liu and V. K. N. Lau, "How much cache is needed to achieve linear capacity scaling in backhaul-limited dense wireless networks?" *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 179–88, Feb. 2017.

[92] K. Shanmugam, M. Ji, A. M.Tulino, J. Llorca, and A. G. Dimakis, "Finite length analysis of caching-aided coded multicasting," *IEEE Trans. Info. Theory*, vol. 62, no. 10, pp. 5524–37, Oct. 2016.

[93] G. Vettigli, M. Ji, A. M. Tulino, J. Llorca, and P. Festa, "An efficient coded multicasting scheme preserving the multiplicative caching gain," in *Proc., IEEE Infocom Wkshps*, Apr., 2015, pp. 251–256.

[94] B. Błaszczyszyn, P. Keeler, and P. Muhlethaler, "Optimizing spatial throughput in device-to-device networks," *arXiv preprint arXiv:1612.09198*, Dec. 2016.

[95] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Modeling and performance analysis of clustered device-to-device networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4957–72, Jul. 2016.

[96] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," in *Proc., IEEE Infocom*, Apr. 2016.

[97] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1305–1314, Sep. 2002.

[98] A. Giovanidis and A. Avranas, "Spatial multi-LRU caching for wireless networks with coverage overlaps," in *Proc., ACM Sigmetrics/IFIP Performance*, Antibes, France, Jun. 2016, pp. 403–405.

[99] M. Gerasimov, V. Kruglov, and A. Volodin, "On negatively associated random variables," *Lobachevskii Journal of Mathematics*, vol. 33, no. 1, pp. 47–55, Jan. 2012.

[100] D. J. Aldous, *Exchangeability and related topics.* Springer Berlin Heidelberg, 1985, vol. 1117.

[101] J. Pitman, "Exchangeable and partially exchangeable random partitions," *Probability theory and related fields*, vol. 102, no. 2, pp. 145–158, Jun. 1995.

[102] D. Aldous. (2013) Exchangeability and related topics. [Online]. Available: www.stat.berkeley.edu/~aldous/206-Exch/

[103] J. Møller, M. L. Huber, and R. L. Wolpert, "Perfect simulation and moment properties for the Matérn type III process," *Stoch. Process. Appl.*, vol. 120, no. 11, pp. 2142–58, Nov. 2010.

[104] M. Hörig and C. Redenbach, "The maximum volume hard subset model for Poisson processes: simulation aspects," *J. Statist. Comput. Simul.*, vol. 82, no. 1, pp. 107–121, Jan. 2012.

[105] D. Malak, M. Al-Shalash, and J. G. Andrews, "A distributed auction policy for user association in device-to-device caching networks," in *Proc., IEEE PIMRC*, Montreal, QC, Canada, Oct. 2017.

[106] ——, "Resource allocation for content caching in D2D-enabled cellular networks," *journal preprint*, 2017.

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

[108] N. Lee, X. Lin, J. G. Andrews, and R. W. Heath, "Power control for D2D underlaid cellular networks: Modeling, algorithms, and analysis," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 1, pp. 1–13, Jan. 2015.

[109] A. Busson, G. Chelius, and J.-M. Gorce, "Interference modeling in CSMA multi-hop wireless networks," Ph.D. Thesis, INRIA, Paris, France, Feb. 2009.

[110] B. Kauffmann, F. Baccelli, A. Chaintreau, V. Mhatre, K. Papagiannaki, and C. Diot, "Measurement-based self organization of interfering 802.11

wireless access networks," in *Proc., IEEE International Conference on Computer Communications (INFOCOM)*, May 2007, pp. 1451–1459.

[111] A. Chattopadhyay and B. Błaszczyszyn, "Gibbsian on-line distributed content caching strategy for cellular networks," *arXiv preprint arXiv:1610.02318*, Oct. 2016.

[112] B. Chen, C. Yang, and Z. Xiong, "Optimal caching and scheduling for cache-enabled D2D communications," *arXiv preprint arXiv:1701.04596*, Jan. 2017.

[113] J. Krolikowski, A. Giovanidis, and M. D. Renzo, "Fair distributed user-traffic association in cache equipped cellular networks," in *Proc., IEEE WiOpt*, May 2017, pp. 1 – 6.

[114] A. Pourmiri, M. J. Siavoshani, and S. P. Shariatpanahi, "Proximity-aware balanced allocations in cache networks," in *Proc., IEEE Int. Parallel and Distributed Proc. Sym.*, May 2017, pp. 1068 – 1077.

[115] R. K. Ganti and M. Haenggi, "Spatial and temporal correlation of the interference in ALOHA ad hoc networks," *IEEE Commun. Lett.*, vol. 13, no. 9, pp. 631–633, Sep. 2009.

[116] M. Haenggi, "Mean interference in hard-core wireless networks," *IEEE Commun. Lett.*, vol. 15, no. 8, pp. 792–794, Aug. 2011.

[117] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Device-to-device modeling and analysis with a modified Matern hardcore BS location model," in *Proc., IEEE Globecom*, Dec. 2013, pp. 1825–1830.

[118] F. Baccelli and B. Błaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume II — Applications*, ser. Foundations and Trends in Networking. Now Publishers, 2009, vol. 4, no. 1–2.

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

[120] P. Brémaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues.* Springer Science & Business Media, 2013, vol. 31.

[121] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times.* American Mathematical Society, 2006.

[122] B. A. Cipra, "An introduction to the Ising model," *American Mathematical Monthly*, vol. 94, no. 10, pp. 937–959, Dec. 1987.

[123] J. Abate and W. Whitt, "Numerical inversion of laplace transforms of probability distributions," *ORSA Journal on Computing*, vol. 7, no. 1, pp. 36 – 43, Feb. 1995.

[124] H. S. Dhillon, M. Kountouris, and J. G. Andrews, "Downlink MIMO HetNets: Modeling, ordering results and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 5208–5222, Oct. 2013.

[125] D. Malak, H. C. Huang, and J. G. Andrews, "Fundamental limits of random access communication with retransmissions," in *Proc., IEEE ICC*, May 2017.

[126] H. S. Dhillon, H. C. Huang, H. Viswanathan, and R. A. Valenzuela, "Fundamentals of throughput maximization with random arrivals for M2M communications," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4094–4109, Nov. 2014.

[127] N. Abramson, "The ALOHA system - another alternative for computer communications," in *Proc, Fall AFIPS Computer Conference*, Nov. 1970, pp. 281–285.

[128] Y. Zhong, M. Haenggi, T. Q. Quek, and W. Zhang, "On the stability of static Poisson networks under random access," *IEEE Trans. Commun.*, vol. 64, no. 7, pp. 2985–2998, Jul. 2016.

[129] D. P. Bertsekas, *Nonlinear programming.* Athena Scientific, 1999.

[130] A. M. Ibrahim, T. ElBatt, and A. El-Keyi, "Coverage probability analysis for wireless networks using repulsive point processes," in *Proc., IEEE PIMRC*, Sep. 2013, pp. 1002 – 1007.

Vita

Derya Malak received the B.S. in Electrical and Electronics Engineering with minor in Physics at Middle East Technical University, Ankara, Turkey, in Jun 2010. She received the M.S. in Electrical and Electronics Engineering at Koc University, Istanbul, Turkey, in Feb 2013. Currently, she is a Ph.D. candidate in the Department of Electrical and Computer Engineering at Wireless Networking & Communications Group at the University of Texas at Austin. She has held summer internships at Huawei Technologies, Plano, TX, and Nokia Bell Labs, Murray Hill, NJ. She is broadly interested in the area of communication theory and networks, wireless, and information theory. Her doctoral research has focused primarily on stochastic modeling, design, analysis of content caching using device-to-device communications, and machine-type communications with high reliability, low-latency and high energy efficiency. During her M.S. studies, she has focused on molecular communications and human nervous system using information theory and coding. She was awarded the Graduate School (UTGS) fellowship by UT Austin between 2013-2017.

Permanent email: deryamalak@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.