Li, L., Zhao, G. and Blum, R. S. (2018) A survey of caching techniques in cellular networks: research issues and challenges in content placement and delivery strategies. *IEEE Communications Surveys and Tutorials*, 20(3), pp. 1710-1732. (doi: [10.1109/COMST.2018.2820021](https://doi.org/10.1109/COMST.2018.2820021))

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/238672/>

Deposited on 04 June 2021

A Survey of Caching Techniques in Wireless Communication Networks: Research Issues and Challenges

Liying Li, *Member, IEEE*, Guodong Zhao, *Senior Member, IEEE*,
and Rick S. Blum, *Fellow, IEEE*

Abstract

Mobile data traffic is currently growing exponentially and these rapid increases have caused the backhaul data rate requirements to become the major bottleneck to reducing costs and raising revenue for operators. To address this problem, caching techniques have attracted significant attention since they can effectively reduce the backhaul traffic by eliminating duplicate data transmissions that carry popular content. In addition, other system performance metrics can also be improved through caching techniques, e.g., spectrum efficiency, energy efficiency, and transmission delay. In this paper, we provide a systematical review of the state-of-the-art caching techniques that were recently developed in different wireless communication networks, including cellular networks, heterogeneous networks (HetNets), device-to-device (D2D) networks, cloud-radio access networks (C-RANs), and fog-radio access networks (C-RANs). We discuss content placement, content delivery, and joint placement and delivery algorithms. We also compare the different proposed approaches in terms of different performance metrics, including throughput, backhaul cost, power consumption, and network delay. In addition, we summarize the main research achievements in different networks, and highlight some challenges and promising potential research topics.

Index Terms

Caching, cache-enabled networks, cellular networks, content-centric wireless networks, wireless communication networks.

The work of L. Li and G. Zhao was supported by the National Nature Science Foundation of China under Grants No. 61601094. The work of R. S. Blum was supported by the National Science Foundation under Grant No. ECCS-1405579.

L. Li is with the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, 611731 China (e-mail: liyingli0815@gmail.com).

G. Zhao is with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, 611731 China (e-mail: gdngzhao@gmail.com).

R. S. Blum is with the Department of Electrical and Computer Engineering, Lehigh University, Bethlehem, PA 18015 USA (e-mail: rblum@eecs.lehigh.edu).

A Survey of Caching Techniques in Cellular Network: Research Issues and Challenges

Abstract—Mobile data traffic is currently growing exponentially and these rapid increases have caused the backhaul data rate requirements to become the major bottleneck to reducing costs and raising revenue for operators. To address this problem, caching techniques have attracted significant attention since they can effectively reduce the backhaul traffic by eliminating duplicate data transmissions that carry popular content. In addition, other system performance metrics can also be improved through caching techniques, e.g., spectrum efficiency, energy efficiency, and transmission delay. In this paper, we provide a systematical review of the state-of-the-art caching techniques that were recently developed in different wireless communication networks, including cellular networks, heterogeneous networks (HetNets), device-to-device (D2D) networks, cloud-radio access networks (C-RANs), and fog-radio access networks (F-RANs). We discuss content placement, content delivery, and joint placement and delivery algorithms. We also compare the different proposed approaches in terms of different performance metrics, including throughput, backhaul cost, power consumption, and network delay. In addition, we summarize the main research achievements in different networks, and highlight some challenges and promising potential research topics.

Index Terms—Caching, cache-enabled networks, cellular networks, content-centric wireless networks, wireless communication networks.

I. INTRODUCTION

In the past decade, due to the fast development of smart phone and tablet technologies, mobile data traffic has been growing exponentially to reach a volume of 7.2 exabytes* per month by the end of 2016. In the near future, the *internet-of-things* (IoTs) and other new applications are expected to generate even more mobile traffic, such that the total traffic is predicted to reach 49 exabytes per month by the end of 2021 [1]. These predictions justify a need for new wireless networks, e.g., the *fifth generation* (5G) networks[†] [2], to meet the predicted traffic requirement [3]. Currently, ultra-wide-band communication [4], massive *multiple-input multiple-output* (MIMO) communication [5], millimeter-wave communication [6], and *heterogeneous networks* (HetNets) [7] are among several promising techniques proposed for a 5G communication standard. However, all these techniques require the existence of expensive backhaul links between *base stations* (BSs) and the core internet (or among BSs) [7]–[9], which motivates reducing backhaul traffic.

As demonstrated recently, caching techniques are good candidates to reduce backhaul traffic. A large portion of

backhaul traffic is contributed by transmitting *duplicate popular data* to multiple users. The duplicate data transmission is mainly caused by transmitting popular content (including audio, video, and webpages). When the popular content is requested by a large number of users at different times, the wireless network needs to repeatedly send the same content to each user. As a result, the duplicate data passes through the backhaul links over and over again, which causes significant backhaul traffic. If BSs can cache the popular content, they are able to directly serve each user. Then, the backhaul links only need to send the popular content once to the BSs rather than multiple times, which can effectively reduce the duplicate data transmitted. Even better, if BSs can predict the user requests and cache the requested content in advance [10], [11], they can directly serve the users without requiring use of the backhaul link. As a result, the BS can provide simultaneous transmission with low backhaul requirements, in particular in peak hours. In addition, the cost of installing memory storage is much lower compared with that of improving the backhaul capacity, e.g., for a memory storage of 2 – 3 TB is about 100 US\$ at retail prices [12]. Therefore, due to the above advantages, caching techniques have attracted significant attention in wireless communication networks. This paper aims at providing a survey of key research progress in this young field. We shall also present a research outlook containing ensemble of promising research directions for caching techniques in wireless communication networks.

A. The Development of Wireless Caching

Caching technique has been comprehensively studied in the areas of web caching, *content distribution networks* (CDNs) [13] and *information centric networks* (ICNs) [14]–[17] to solve network congestion in wired networks. In 1990s, web caching technique was developed to reduce the transmission delay and traffic by caching the web documents, including HTML pages and images, on the computers. Then in CDNs, caching technique was proposed to further reduce the traffic by caching popular content in many geographical areas and saves bandwidth by avoiding unnecessary multi-hop retransmissions caused by the client-server model of connectivity. In 2000s and 2010s, caching in ICNs are proposed to reduce the network traffic, delay and power consumption by caching content in the routers of core networks.

Fig. 1 provides an architecture of a wired network with caching, which consists of subscribers, routers, and servers [16]. Subscribers issue content requests by sending a packet which carries the name of the desired content. The routers on the path that cache the desired content can serve the users with low latency. Otherwise, the servers need to send the request content to subscribers. [18].

*One exabyte is equivalent to one billion gigabytes, and one thousand petabytes.

[†]*Fifth generation* (5G) networks are the proposed next telecommunication standards beyond the current *fourth-generation* (4G)/IMT-Advanced standards. They are designed to enable the mobile communication with higher capacity than current 4G, allow a higher density of mobile broadband users, and support device-to-device, ultra reliable, and massive machine communications.

Figure 1. Architecture of a typical wired network with caching.

Figure 2. Architecture of a typical cellular network with caching.

Fig. 2 provide a communication architecture of cellular networks with caching, which consist of the wireless communication systems, wired core networks, and internet. When users request content (videos or webpages), the neighboring users or relays, which cache the requested content, will serve the request locally. Otherwise, the content needs to be served by the BSs cooperatively with the cached content or by connecting the core networks.

Since the architecture of cellular networks is different from wired networks, the caching techniques regarding to the content placement and delivery are also different. Even more, the caching in cellular networks is more challenging [29]. This is because, the wireless edges have a smaller service coverage and caching space compared with the wired edges in the core network or internet. Thus, the content placement strategies designed for wired network are not effective for the wireless networks. In addition, in wireless networks, the content placement strategies also need to consider the flexible communication topology is flexible and wireless transmission fading and interference. In wireless communication networks, the content delivery strategies need to consider not only the transmission routing between the cache and users, but also the wireless transmission fading and interferences. In particular, some transmission techniques are used for content delivery, such as multicast and coordination multipoint transmission techniques.

As mobile traffic continues to grow significantly, since 2009,

wireless caching techniques have been widely investigated to improve the transmission efficiency and reduce the traffic congestion in cellular networks. They were first investigated in [19] from the perspective of information theory by using the caching capacity region. A communication scheme that is able to achieve the inner bound of the caching capacity region was proposed. The communication scheme consists of three layers: routing, cooperative, and physical layer. In 2010, the architecture of a cache-enabled femtocell was first proposed in [20] where the main idea is to add a storage resource at each femtocell. By designing the caching and user association schemes, the authors demonstrated that the caching technique was able to increase the system throughput by up to 500%, which shows the great potential of caching techniques. After 2010, caching techniques have been widely applied in many types of cellular networks, producing *cache-enabled macro-cellular networks* [21], [22], *cache-enabled HetNets* [23], [24], *cache-enabled device-to-device (D2D) cellular networks* [25], [26], *cache-enabled cloud-radio access networks (C-RANs)* and *fog-radio access networks (F-RANs)* [27], [28]. In these papers, the authors mainly design the content placement and content delivery algorithms, where the former will cache the content to the storage that is located at either BSs or users, and the latter decides how to retrieve a content and convey the content to a user that requests it. Typically, the objective is to reduce backhaul cost and network delay, or increase the spectral/energy efficiency.

B. Paper Motivation and Outline

Since caching is an effect way to reduce backhaul cost and can significantly reduce the transmission delay for cellular networks, a lot of contribution has been done in this area recently. There are some published papers have surveyed caching techniques in wireless communication networks [30]–[32]. In [30], the authors review only their own caching approach in HetNets and D2D networks. In particular, the surveyed approach mainly maximizes the system throughput and is used for video transmission. In [31], the authors focus on introducing and comparing three caching techniques that are suggested for 5G networks: evolved packet core caching, radio access network caching, and content-centric networking caching. They demonstrate that the caching techniques can effectively reduce mobile traffic. In [32], the authors discuss the throughput scaling law of wireless networks with caching capability, where a D2D approach, a conventional unicasting approach, a harmonic broadcasting approach, and a coded multicasting approach are compared. However, in view of prior works, there still lacks a survey article providing a systematic and concrete overview on the caching techniques in different wireless communication networks, which motivates the current work.

In this paper, we systematically review recent caching techniques in cellular networks, including macro-cellular networks, HetNets, D2D networks, C-RANs, and F-RANs. For each network, we further compare the caching algorithms in terms of different performance metrics, including throughput, backhaul cost, power consumption, and network delay. In addition, we summarize the main research achievements, identify the algorithm designing pitfalls in different networks, and highlight some challenges and potential research topics.

The paper is organized as follows. In Section II, we introduce the taxonomy of the caching techniques in cellular networks. In Section III, we review the existing works on cache-enabled macro cellular networks. In Sections IV and V, we discuss the recent advances and contributions on cache-enabled HetNets and D2D networks, respectively. In Section VI, we focus on cache-enabled C-RAN and F-RAN, where cloud and edge computing is also considered. In Section VII, we discuss the challenges and potential research topics. Finally, Section VIII concludes the paper.

II. TAXONOMY

In this section, we provide a taxonomy for research on the topic of caching techniques in wireless communication networks to systematically survey the existing research. The taxonomy also serves as a framework for this paper that accounts for all the design aspects. Fig. 3 illustrates the taxonomy structure which includes five topics: (1) network type; (2) caching strategy; (3) performance metrics; (4) control structure; and (5) mathematical tools. Each topic is further subdivided into different categories.

A. Network Type

As shown in Fig. 4, there are four main types of wireless communication networks: cache-enabled macro-cellular networks, cache-enabled HetNets, cache-enabled D2D networks,

and cache-enabled C-RANs/F-RANs, where the content is usually cached at BSs, *small-cell base stations*[‡] (SBSs), *user equipment* (UE), and *remote radio heads* (RRHs), respectively. In different types of the networks, a caching resource may provide very different gains. Specifically, in cache-enabled cellular networks and HetNets, caching mainly reduces the backhaul cost. In contrast, in cache-enabled D2D networks and C-RANs/F-RANs, caching mainly improves the spectral efficiency and energy efficiency, respectively. Thus, to make full use of the caching resource, we need to design different caching strategies for different types of networks. Next, we provide the details of the different networks.

• Cache-Enabled Macro-Cellular Networks

In cache-enabled macro-cellular networks [21], as shown in Fig. 4(a), each BS has large caching space. This allows the users to obtain the required content directly from the BS, rather than from the internet through backhaul links. In this category, the typical transmission range of a BS in urban areas is less than 800 m and the typical transmission power ranges from 43 dBm to 48 dBm [34]. In addition, there is no overlap among the coverage of the different BSs.

• Cache-Enabled HetNets

As shown in 4(b), the cache-enabled HetNets consist of macro-cells, small cells (femtocells, picocells), and relay nodes, where the coverage of macrocells overlaps that of the small cells and relays, and they share the same spectrum [7], [23]. The macro-cell BSs typically transmit at high power level, i.e., $5W - 40W$. The small cell BSs and relay nodes transmit at substantially lower power levels, i.e., $100mW - 2W$ [35]. In cache-enabled HetNets, both *macro-cell base stations* (MBSs) and SBSs have large caching space. The users can obtain the requested files directly from the closest SBS or MBS. There are two main types of topologies in cache-enabled HetNets. One type is called a grid topology, as shown in Fig. 4(b). As in the figure, the MBS is usually located at the center of each cell and several SBSs are also deployed inside the cell [36]. The other type is called a random spatial topology as shown in Fig. 5. In this topology, the MBSs and SBSs can be considered, for modeling purpose, to be randomly deployed in each cell. Recent approaches, called stochastic geometry, model the MBS and SBS locations in order to analyze system performance as independent homogeneous *Poisson point processes* (PPPs) [37], [38].

• Cache-Enabled D2D Networks

Cache-enabled D2D networks [25], [39] allow users with limited caching space to communicate with each other through wireless links. Here, the users are often organized into clusters that can be controlled in a centralized manner or in a distributed way. In this setting, a user's content request can be satisfied by other users in the same cluster. Most of the recent work considers one-hop D2D

[‡]Small-cell base stations are the radio access base stations with low transmission power and the transmission range of 10–2000 meters. They are “small” compared to macro cell base stations, which have the transmission range of a few tens of kilometers [33].

Figure 3. Taxonomy structure

Figure 5. Random topology of cache-enabled HetNets.

communications.

- **Cache-Enabled C-RAN/F-RAN**

C-RAN [40] is a centralized cloud computing-based radio access network that supports 4G and future wireless communication standards [41]. As shown in Fig. 4(d), the *baseband units* (BBUs) with large caching space are usually clustered as a BBU pool and responsible for providing large-scale collaborative processing, co-operative radio resource allocation, and intelligent networking. The distributed RRHs connect to the BBU pool through fronthaul links, which are usually transport through cheap fiber resources. In C-RAN, UEs are usually served by the BBU pool without connecting the core network through backhaul links. In contrast, F-RAN [42], [43] is a distributed edge computing-based radio access network, where the memory storage and signal processing functionalities can be installed at the edge of the network,

i.e., RRHs. In an F-RAN, most of transmission task are processed directly by local RRHs with lower transmission cost and delay. Only the tasks those cannot be well processed by edges (RRHs) are handled by the could (the BBU pool) [44], [45].

B. Caching Strategy

In cache-enabled wireless networks, the network operates in two phases: the content placement phase and the content delivery phase. The content placement phase usually works in off-peak traffic periods. In this phase, the network prefetches the content to the storage and one needs to design a *caching placement strategy* to decide which file should be cached in which BS or UE, and whether coded or uncoded files should be cached. The content delivery phase occurs after the content placement phase and usually works in peak traffic periods. In this phase, the network needs to send the requested files to the users and one needs to design a *caching delivery strategy* to decide how to send the files, by determining power control, channel allocation, and the other required transmission details.

- **Caching Placement Strategy**

In general, there are two kinds of placement strategies, one is the uncoded placement strategy [46], [47], and the other is the coded one [48]–[50]. Both are designed to optimize some performance metrics, including hitting rate, area spectral efficiency, and backhaul cost. In addition, a content replacement strategy that determines how to replace the old content with new content also needs to be designed. The two most famous replacement strategies are the recency-based [51] and the frequency-based [52] strategies, which remove the *least recently used* (LRU)

Figure 4. Four types of wireless communication networks.

objects and *least frequent used* (LFU) objects, and then replace them with updated content, respectively [53]. When we design a caching placement strategy, the popularity of content needs to be considered. Often a few popular files are requested by a large number of users at different times. In particular, the request probability of the videos is typically modeled using a Zipf distribution [54], [55]. Specifically, the request probability of the i -th most popular file, $P_r(f)$, can be modeled as

$$P_r(i) = \frac{\frac{1}{i^{r_r}}}{\sum_{j=1}^m \frac{1}{j^{r_r}}}, \quad 1 \leq i \leq m, \quad (1)$$

where r_r is the Zipf exponent which describes the probability of content reuse. The larger the value of r_r , the smaller the number of files requested on average. In (1), m is the total number of the files requested.

• Caching Delivery Strategy

In the content delivery phase, a delivery strategy is designed to decide how to transmit the requested files to the users, e.g., where should the files be transmitted from, what the transmission power is, and on which frequency band should the files be transmitted. Different performance metrics and physical transmission conditions

need to be considered when we design the delivery strategies.

C. Performance Metrics

In cache-enabled wireless communication networks, different metrics are adopted in designing the placement and delivery strategies, including throughput, backhaul cost, network cost, network delay, hitting rate, area spectral efficiency, and outage/coverage probability. Next we describe each metric in more details.

- **Throughput:** Since most of the content is made up of video files in cache-enabled wireless communication networks, throughput is critical in system design. In particular, the user experience in video service mainly depends on throughput. When we consider the fairness among different users, the minimum per-user average throughput under the constraint of a network outage probability is the metric used in designing the caching strategies [25].
- **Backhaul Cost:** In wireless networks, backhaul refers to the communication links connecting the BSs to each other or to the core network, which usually employs expensive optical fiber. Future high data rate wireless

services require extremely high backhaul bandwidth and result in high backhaul cost. Thus, backhaul cost needs to be considered in system design.

- **Power Consumption:** In general, the power consumption is dominated by that from the BSs and the user devices. The former determines the electricity bill at the operators. The latter affects the battery life time of user devices. To satisfy both the operators and users, power consumption needs to be considered.
- **Network Delay:** Network delay represents the time duration between the time when the file is requested until delivery. It affects the user's experience and is critical for delay-sensitive content services. Specifically, the network delay can be divided into three parts: 1) Processing delay: the time that routers take to process the data; 2) Queuing delay: the time that the data spends in routing queues; 3) Propagation delay: the time it takes for a signal to reach its destination.
- **Hitting Rate:** Hitting rate is the ratio between the number of cached files requested by the users and the overall number of files in the caching space. This criteria indicates the percentage of cached files used, which is used as a performance metric to evaluate the caching placement algorithm performance [38].
- **Area Spectral Efficiency:** Area spectral efficiency is defined as the maximum data rate that a certain bandwidth can support in a given area, with the unit of $\text{bps}/(\text{Hz}\cdot\text{m}^2)$ [56]. It is usually used to evaluate the spectral utilization efficiency of a cellular system. The higher the area spectral efficiency is, the more users a certain area can accommodate.
- **Outage/Coverage Probability:** Outage/coverage probability is the probability that the *signal-to-interference-plus-noise ratio* (SINR) is below or above a certain threshold, respectively. These measures describe the percentage of time that a user experiences loss of service. Thus, they are critical for a placement algorithm design that aims at user experience [37], [38].

D. Control Structure

The caching placement and delivery strategies can be controlled by one entity in a centralized structure, or by multiple entities in a distributed structure. In the following, we provide more details on the two control structures.

- **Centralized:** In a centralized structure, a single central entity decides the content placement and delivery strategies. The central entity collects information, such as the file's popularity, user requests, and relevant channel information. Based on this information, the central entity decides which file should be placed in which BS or UE, and where should the files be transmitted from [24], [57], [58]. In general, the centralized content placement and delivery strategies can provide the optimal solution for the entire network, but may require high complexity.
- **Distributed:** In a distributed structure, each BS or UE will decide which file should be cached and how to send the requested files. Since distributed content placement

and delivery strategies do not need to collect the information, they have lower complexity compared to centralized algorithms. However, distributed content placement and delivery strategies can not guarantee global optimality of the algorithms, since the BSs or UEs obtain the solutions based on their own information.

E. Mathematical Tools

Different mathematical tools can be used to formulate and solve problems. Considering different network architectures and performance metrics, caching problems are usually formulated using four mathematical tools: optimization, game theory, machine learning, and stochastic geometry.

- **Optimization:** Optimization is one of the most popular tools for designing the placement and delivery strategies. The optimization problem is usually formulated to maximize a performance metric under the resource constraints and *quality of service* (QoS) requirements.
- **Game Theory:** Game theory is a mathematical modeling tool, which attempts to model a situation where players compete to try to maximize their payoffs. In cache-enabled wireless communication networks, the players can be the BSs [53], [59] or the UEs [60]. A Nash Equilibrium is reached when none of the players can increase their payoff by changing only its own strategy [61].
- **Network Coding** Network coding [22], [48], [62] is a mathematical technique, in which transmitted data is encoded and decoded to increase network throughput, reduce delays, and make the network more robust. In network coding, algebraic algorithms are applied to the data to accumulate the various transmissions. The received transmissions are decoded at their destinations. This means that fewer transmissions are required to transmit all the data, but this requires more processing at intermediary and terminal nodes.
- **Machine Learning:** Machine learning is a subject where measured data is often used to model the actions a system may take. It can be used in designing caching strategies in wireless networks, since it allows prediction of the content popularity or rating by exploiting the available data from social networks, localization systems, user velocity measurements or other available data [10], [46].
- **Stochastic Geometry:** Stochastic geometry is a statistical modeling tool that can be used for analyzing and designing caching techniques in wireless communication networks. The model assumes a random topology [63] determined by a random process, e.g., the widely used PPP and *Poisson cluster process* (PCP). Based on the specific properties of the selected point process, an analytical expression can be derived for the outage probability [37] and coverage probability [38] of cache-enabled wireless communication networks.

III. CACHE-ENABLED MACRO-CELLULAR NETWORKS

The main purpose of introducing cache to a cellular network is to reduce backhaul transmission cost. By installing large

caching storage at BSs, cellular users can obtain the requested content directly from the BSs rather than from the core network using a backhaul link. In this section, we review recent research on cache-enabled cellular networks from three aspects: (1) placement strategy; (2) delivery strategy; (3) joint placement and delivery strategy. In particular, we provide Table II to compare the existing work. From the table, most of caching placement and delivery strategies either minimize the backhaul cost or network delay or peak data rate over a shared link[§], or maximize the successful transmission probability. They cache and deliver coded data, where a segment of the coded data is cached at each BS. There are also some strategies caching and delivering uncoded data, where the whole content is cached at each BS and delivered to UEs. Regarding the mathematic tools, optimization is employed in both distributed and centralized manners. In Table II, the mark “-” means the corresponding algorithm did not consider this metric. The abbreviations used in Table II are summarized in Table I.

A. Placement Strategy

In cache-enabled macro-cellular networks, BSs usually co-operatively cache a set of files. Then, a user can obtain the requested content from either its local BS or from a set of neighbor BSs [21], [64], [65]. In general, different placement strategies are designed under different assumptions, e.g., the assumed size of a BS’s cache and whether the popularity of the content is known to the BSs. In [21], the authors propose a content placement method to minimize the download cost, where the BSs are installed with finite cache storage and the content popularity is known to the BSs. In [64], the authors propose a content placement method to minimize the summation of download cost and caching cost, where the BSs are assumed to have infinite cache storage and they do not have information on content popularity. Note that both works [21], [64] consider uncoded as well as coded data. For uncoded data, the problem was formulated as an integer optimization problem, which is an NP-complete problem. For coded data, the problem was formulated into a linear programming optimization problem, which can be solved in polynomial-time. In [65], the authors propose a content placement method to minimize the average download delay, including both processing and propagation delays. They find that obtaining the optimal placement strategy using mixed integer programming is difficult. Then they propose a low-complexity sub-optimal algorithm by relaxing the integer constraint. In [66], the authors consider user mobility and propose coded and uncoded content placement strategies to minimize the cache fail probability that mobile users cannot obtain the requested files from the cached content at BSs. In [67], the authors propose an optimal content placement to maximize the hitting probability when a UE can be served by multiple BSs. They find that the optimal content placement policy is related to the *signal-to-interference-and-noise ratio* (SINR).

[§]The shared link is defined in Fig. 1 of [48]. It is a transmission link between a BS and multiple users, occupying the same time and frequency resource.

Table I
SUMMARY OF ABBREVIATIONS USED IN TABLES II-V

ASE	Area Spectral Efficiency
BC	Backhaul Cost
BL	backhaul Load
CC	Caching Cost
CFP	Cache Failure Probability
CP	Coverage Probability
CU	Caching Utility
DC	Download Cost
DD	Download Delay
DR	Delivery Rate
DoF	Degree of Freedom
ECPC	Electronic Content Provisioning Cost
EC	Energy Consumption
EE	Energy Efficiency
ENAL	Expected Number of Active D2D Links
FL	Fronthaul Loading
HP	Hit Probability
LC	Link Capacity
LTAR	Long-Term Average Rate
NC	Network Cost
ND	Network Delay
NUS	Number of Users can be Served
OL	Offloading Loss
PC	Power Consumption
PR	Peak Rate
PUT	Per-User Throughput
SE	Spectral Efficiency
SP	Service Probability
STP	Successful Transmission Probability
SOP	Successful Offloading Probability
TC	Transmission Cost
TP	Transmission Power
TL	Traffic Load
WCNR	Worst-Case Normalized Rate

The popularity of the content varies with time and new content is being created all the time. Thus the cached content needs to be updated, which is called content replacement. In [53], a distributed caching replacement strategy is investigated, where the BSs update the content in their caches without knowing the cached content at other BSs. Then, the BSs can cache the latest popular content to increase the probability that a user is served by its BS’s cache. In algorithm design, the authors in [53] employ a *Markov decision process* (MDP) to model the caching replacement problem and propose an optimal replacement policy, where network coding [73] is used to encode the content. The results show that the transmission cost goes to zero when the storage is large enough.

B. Delivery Strategy

In contrast to placement strategies that move the content to the edges of the cellular networks, e.g., BSs, delivery strategies serve the UEs using the cached content. Most of the existing work discusses delivery strategies under the systems using

Table II
COMPARISON OF DIFFERENT PLACEMENT AND DELIVERY STRATEGIES IN CACHE-ENABLED CELLULAR NETWORKS

Ref.	Strategy	Objective	Data Type	Mathematical Tools	Control Structure	Backhaul Cost	Throughput	Power Consumption
[21]	Placement	max-Profit	Coded & uncoded	Optimization	Distributed	Reduced	Large	Reduced
[64]		min-(CC+DC)	Coded & uncoded	Optimization	Centralized	Reduced	Large	-
[65]		min-DD	Coded	Optimization	Centralized	Reduced	Large	-
[66]		min-CFP	Coded & uncoded	Optimization	Centralized	Reduced	Moderate	-
[67]		max-HP	Uncoded	Stochastic Geometry	Centralized	Reduced	-	Moderate
[53]		min-BC	Coded	Game Theory	Distributed	Reduced	-	-
[68]	Delivery	min-(DD+TC)	Uncoded	Optimization	Centralized	Reduced	-	Reduced
[69]		max-STP	Coded	Stochastic Geometry	Centralized	-	Large	-
[70]		min-TP	Coded	Optimization	Centralized	Reduced	Moderate	Low
[48]	Placement & Delivery	min-PR	Coded	Optimization	Centralized	-	Small	-
[49]		min-WCNR	Coded	Optimization	Distributed	-	Small	-
[71]		min-TL	Coded	Optimization	Distributed	-	Small	-
[72]		min-NC	Coded	Optimization	Centralized	-	Large	-
[50]		min-LTAR	Coded	Optimization	Centralized	-	Small	-

multicast and *coordinated multi-point* (CoMP) transmission techniques, which can obtain high spectral efficiency.

The multicast transmission technique achieves high spectral efficiency by serving multiple users simultaneously when those users require the same content in the same cell. However, this requires BS to delay some of the requests until a group of requests are collected so they can serve multiple users that request a common content in a certain time window. Therefore, the main task in designing a multicast delivery strategy is to tradeoff the delay and spectral efficiency. If the time window used to collect requests is large, the BS can collect a large number of requests. Then the system achieves high spectral efficiency, but has large time delay for some of the users. In contrast, a short time window has small delay, but lowers the spectral efficiency. In [68], an optimal dynamic multicast method is proposed to balance the delay and spectral efficiency. Delivery cost is also considered in this paper. This delivery cost includes the average network delay, backhaul cost, and transmission power. Practical conditions, including the geographic locations of BSs and UEs, *channel state information* (CSI) and interference, are considered in [69]. The authors propose a multicast delivery strategy to maximize the successful transmission probability, when a geometric statistical tool is used to model the locations of the BSs and UEs.

The CoMP transmission technique achieves high spectral efficiency by sharing the realtime CSI and data among BSs. However, it requires high backhaul capacity for CSI and data sharing, which is a bottleneck to implementing CoMP in practical systems. Since the data sharing dominates the backhaul traffic, we can precache the data at BSs to reduce the backhaul requirement for CoMP transmission, i.e., the BSs are able to use the cached data and the shared CSI to conduct CoMP transmission. In [70], a content delivery method is proposed to minimize the transmit power for CoMP transmission, where the same data is cached at different BSs.

C. Joint Placement and Delivery Strategy

When we jointly design the content placement and delivery using coding techniques, the traffic between BSs and UEs can be significantly offloaded. In [48], the authors propose a centralized coded placement and delivery strategy, where a BS connects to users through a shared and error free link. In their model, the BS controls the content placement and delivery for all UEs. The BS precaches all content that might be requested by the UEs and each UE caches a part of each file based on an encoding function. Then, different requests can be satisfied by decoding the content from multicast streams. In this way, the traffic between BS and UEs can be effectively offloaded. Here, we provide two examples in Fig. 6 to explain the basic idea of the coded caching method. We consider four users U_1 , U_2 , U_3 , and U_4 , and each of them requires a file. Specifically, the users U_1 , U_2 , U_3 , and U_4 request the files A, B, C, and D, respectively. In the first example in Fig. 6(a), we assume that the cache of each user is 1 unit in size and for simplicity we only consider files of size one unit. Thus, each user can cache only one file. In the placement phase, each file is evenly split into four subfiles, e.g., $A = (A_1, A_2, A_3, A_4)$, then the size of each subfile becomes 0.25. This allows each user to cache one fourth of each file, e.g., the user U_1 caches the files (A_1, B_1, C_1, D_1) , which satisfies the caching capacity constraint of each user. In the delivery phase, the BS broadcasts the linearly processed coded data to all the users, i.e., $(A_2 \oplus B_1, A_3 \oplus C_1, A_4 \oplus D_1, B_3 \oplus C_2, B_4 \oplus D_2, C_4 \oplus D_3)$, where $\oplus$ denotes bitwise XOR operation. Then, U_1 can recover A_2 , A_3 , and A_4 from $A_2 \oplus B_1$, $A_3 \oplus C_1$, and $A_4 \oplus D_1$, respectively, since it already has B_1 , C_1 , and D_1 . Similarly, other users are able to obtain their requested files by receiving the coded data transmitted in the shared link. The required data rate in the delivery phase becomes 1.5 units/broadcast time since the size of each processed component in the broadcast transmission is 0.25 units so multiplying by 6 components yields 1.5 units. Therefore, the coded caching method significantly outperforms the uncoded one, in which the required

data rate is 3 units/broadcast time. In the second example as shown in 6(b), we assume that the caching size of each user is 3 units. Using the same coded caching method, we can further reduce the required data rate rate in the delivery phase from 1.5 units/broadcast time to 0.25 units/broadcast time.

From above two examples, it seems that the coded caching can significantly save transmission bandwidth. However, the computational complexity of the coded method grows dramatically as the number of users increases. To reduce the complexity, the authors in [49] propose a decentralized coded method. In the placement phase, each user caches a part of the coded content. In the delivery phase, the BS first obtains some information about the active users, including their cache content and their requests. Then the BS sends the requested content to the active users using a multicast technique. The results indicate that the data rate of the decentralized algorithm is close to that of the centralized algorithm in [48]. There are other recent works discussing the joint placement and delivery method. In [71], an erasure packet channel is considered as a model for the communication links. The optimal data rate region is obtained when using the decentralized coded placement method in [49] and delivery method in [74].

Some other joint placement and delivery methods simultaneously conduct the content placement and delivery. For example, when a BS delivers a requested content to a UE, other UEs may simultaneously receive the same content and put it into their cache. Then, the users that did not request the file need to decide whether to cache the file. As an example of such an approach, a coded joint placement and delivery method is studied [72], where an integer non-linear programming problem is formulated to minimize the network cost and sub-optimal solutions are provided.

Since the popularity of content varies with time, the BS and UEs need to replace the unpopular content with the currently popular ones. In this situation, the replacement strategy actually affects the delivery strategy. When a BS delivers a file, it prefers to delivery the one that can simultaneously serve a user's request and replace other users' cached files. This saves spectrum and related delay. In [50], an online coded joint replacement and delivery method is proposed, which is called a coded *least-recently sent* (LRS) algorithm. It minimizes the long-term average data rate and allows the users to evict the least-recently sent file from the BS.

D. Summary and Discussion

The just given discussion illustrates that coding and multicast are important techniques to obtain advantages from caching. Most of the existing work uses these two techniques to design the caching methods. In particular, the work in [48]–[50] investigates coded caching and delivery from the perspective of information theory. They consider error-free channel conditions between BS and UEs. In the future, more practical channel conditions and user demand distributions need to be considered when we design the content placement and delivery strategies.

IV. CACHE-ENABLED HETNETS

HetNets improve area spectral efficiency through densely deployed SBSs. However, the benefit relies on high-speed backhaul links to each SBS, which can be expensive. Thus, reducing the backhaul cost is critical for HetNets. Using massive storage at the SBSs is an effective way to reduce the backhaul cost and has been called cache-enabled HetNets. The basic idea is to use low bandwidth backhaul links to provide high data rate wireless services by adding cache. Specifically, before any content is requested, the SBSs first cache popular content using low bandwidth backhaul. When the content is requested, the cached content serves the user requests. As a result, the required backhaul bandwidth can be significantly reduced.

In this section, we introduce the existing work on cache-enabled HetNets from three aspects: (1) placement strategy; (2) delivery strategy; and (3) joint placement and delivery strategy. Table III shows a comparison of existing work. From the table, most strategies are designed to minimize the network delay, or transmission cost, or backhaul load. Most of them cache uncoded data and use optimization tools to formulate and solve the placement and delivery problems, where both centralized and distributed control are used. The abbreviations used in Table III are summarized in Table I.

A. Placement Strategy

Due to the huge volume of content and limited caching space at SBSs, only a part of the popular content can be cached at the SBSs. Therefore, we need to carefully design content placement strategies to determine which part of the content should be cached at each BS. Here, the distribution of content popularity, which describes the probability that each file is requested by a group of users, plays an important role in algorithm design. In the following, we first review the existing work when the distribution of content popularity is available to the SBSs. Then, we introduce the work considering the cases when the distribution is unknown to the SBSs. Finally, we discuss the content replacement in the SBSs.

When the distribution of content popularity is available to SBSs, different content placement strategies are designed for different network topologies, by using various metrics combining spectral efficiency, hitting probability, energy efficiency, and download delay. In [24], an optimal placement algorithm is proposed to minimize the download delay, when the network topology is known to each BS. In [89], the capacity of a multi-hop HetNet is studied under the uncoded and coded content placement methods. The results show that the coded method achieves higher capacity than the uncoded one. In [75] and [90], user mobility is considered in content placement, in which the users frequently move from one cell to another. In [59], a distributed content placement problem under ultra dense small cell networks is studied. The problem is formulated as a stochastic differential game to reduce the backhaul load. In [76], the caching capability at user side is considered, where a user can obtain its requested content either from the BSs or from itself. The content placement problem is formulated as an optimization problem to minimize the

Figure 6. Examples of coded caching methods: (a) the caching size of each user is 1; (b) the caching size of each user is 3.

Table III
COMPARISON OF DIFFERENT PLACEMENT AND DELIVERY STRATEGIES IN CACHE-ENABLED HETNETS

Ref.	Strategy	Algorithm	File Type	Model	Control	Backhaul Cost	Throughput	Delay
[24]	Placement	min-ND	Coded&Uncoded	Optimization	Centralized	Reduced	Moderate	Reduced
[75]		min-CU	Uncoded	Optimization	Centralized	Reduced	-	-
[59]		min-BL	Coded	Game Theory	Distributed	Low	-	Moderate
[76]		min-EC	Uncoded	Optimization	Centralized	-	Moderate	-
[77]		min-OL	Uncoded	Stochastic Geometry	Centralized	Reduced	Moderate	Reduced
[78]		max-TL	Uncoded	Optimization	Centralized	Reduced	Large	-
[46]		min-BL	Uncoded	Machine learning	Centralized	Low	Moderate	Reduced
[58]		max-SP	Uncoded	Stochastic Geometry	Centralized	Low	Moderate	Moderate
[79]		min-ND	Uncoded	Optimization	Distributed	Reduced	Moderate	Reduced
[38]		max-CP	Uncoded	Stochastic Geometry	Distributed	Low	-	-
[60]	Delivery	min-ND	Uncoded	Game Theory	Distributed	Low	-	Reduced
[80]		min-BL	Uncoded	Game Theory	Distributed	Reduced	-	-
[81]		min-TD	Uncoded	Optimization	Distributed	Reduced	-	Reduced
[82]		min-ND	Uncoded	Optimization	Centralized	Reduced	-	Reduced
[83]	Placement & Delivery	min-DD	Coded	Optimization	Centralized	Low	Moderate	Reduced
[84]		min-TL	Uncoded	Optimization	Centralized	Low	Moderate	-
[23]		max-TL	Uncoded	Stochastic Geometry	Centralized	Low	Large	Reduced
[85]		min-EC	Uncoded	Optimization	Centralized	Low	Moderate	Moderate
[86]		min-PR	Coded	Optimization	Centralized	Low	Moderate	-
[87]		min-TC	Coded	Optimization	Centralized	Low	Large	-
[88]		max-DoF	Coded	Optimization	Centralized	Low	Large	-

energy consumption. A low complexity suboptimal method is proposed to properly place the content at the relay station and user devices.

In practice, the SBSs may not know the content popularity distribution. Then, they need to estimate the distribution before conducting placement strategies. In [77], the authors propose a learning-based approach to estimate the content popularity by using the instantaneous user demands within a specified time interval, called the training time. The minimum training time that achieves the desired estimation accuracy is obtained in the paper, which scales with the total number of cached files. In addition, the estimated popularity is also used to design

the content placement strategy. In [78], the authors propose to estimate the distribution of content popularity based on the previous cached files. They also propose an optimized content placement strategy to maximize the throughput provided by the cache of SBSs. In [46], the authors propose a proactive content placement method using big data analysis tools. They collect a huge amount of real data from a telecom operator in Turkey, process them on a big data platform, and predict the distribution of content popularity. They also propose an optimal placement strategy to minimize the download delay. In [58], the authors use local user interest to estimate content popularity. In their proposed method, each SBS samples the

content requests of all users in a certain region, and then estimates the content popularity for the region. Based on the estimated popularity, the authors propose a placement strategy to maximize the cache hitting probability. In [79], the authors proposed a cluster-based placement strategy, where the users with similar content interests are grouped into a cluster. The proposed method first estimates the distribution of content popularity for each cluster and optimizes the placement parameters.

Since new content is added all the time and the caching space of each SBS is limited, the content placement control needs to replace the old content by the new content. In [38], the authors propose a content replacement method to maximize the coverage probability. They model the locations of the SBSs and those of the macro BSs as two independent homogeneous PPPs. Their results provide the relationships between the coverage probability, the cache space of each BS, and the required data rate at each user.

B. Delivery Strategy

In cache-enabled HetNets, the requested content can be delivered from different BSs. Thus, we need to design content delivery strategies to choose which BSs will deliver the content. In addition, the corresponding transmission power also needs to be optimized. In the following, we introduce the main delivery strategies for unicast and multicast transmission, where the unicast allows a BS to use different time-frequency resources to serve different users while the multicast allows a BS to use a same time-frequency resource to serve a group of users simultaneously.

In unicast transmission, content delivery strategies are designed to reduce network delay and backhaul cost. In [60], a distributed delivery strategy is designed to decide whether a user should access the cache-enabled small cell network or the macro network. The authors formulate the delivery problem as a congestion game to minimize the network delay, where a unique Nash equilibrium is achieved for all the active users. In [80], backhaul usage at each SBS is considered when designing the content delivery strategy. The authors first use collaborating filtering to estimate the required backhaul usage at each SBS based on the previous users' request. Then, they design a content delivery strategy to find the best SBS to serve the requested UE. In particular, the content delivery problem is formulated as a matching game to minimize the backhaul usage at each SBS. In [81], a content placement strategy is designed to reduce both content transmission delay and backhaul load. The strategy is obtained through re-association of requested UEs and SBSs.

In multicast transmission, the content delivery strategies are designed to reduce the network delay. In [82], the authors propose an optimal content delivery strategy by using multicast at each SBS. They formulate the multicast content delivery problem as a Markov decision process, considering the dynamic feature of the users' requests. The objective is to jointly minimize the average network delay and power consumption under a multiple access constraint.

C. Joint Placement and Delivery Strategies

The content placement and delivery strategies actually interact with each other. The placement strategy decides the quantity and location of cached content, based on which the delivery strategy can determine how to select the transmission links between the requested users and the cached content. In contrast, the delivery strategy may adopt different transmission techniques, including unicast and multicast transmission, which affects the placement strategy as well. For example, the unicast transmission is usually adopted when small cells have little overlap. In this case, it is reasonable to cache the same content at different cells, which provides the same hitting probability in different cells. On the other hand, the multicast transmission is usually adopted when small cells overlap heavily. In this case, it prefers to cache different content at different small cells, which achieves a high hitting probability.

In unicast transmission, joint content placement and delivery strategies are designed under different environments. In [83], the authors consider a two-tier cache-enabled HetNet and investigate a joint content placement and delivery strategy to minimize the delivery time, where the SBSs are capable of caching content. Here, the joint placement and delivery problem is formulated as a two-stage optimization problem. In the first stage, the content placement is decided based on the historical profile of user request and CSI. In the second stage, the proper BS associated with the requested user can be found according to the proposed delivery strategy. In [84], the authors consider user mobility and propose a joint content placement and routing method to minimize the load of the MBS. In [23], the authors consider a three-tier HetNet that consists of macro BS, relays, and D2D pairs. They model the locations of nodes in the three tiers as mutually independent PPPs, and derive the traffic throughput of the network. By maximizing the traffic throughput, they obtain a joint content placement and delivery strategy.

In multicast transmission, the joint content placement and delivery is usually designed to reduce the transmission power. In [85], the authors aim to minimize the expected energy consumption within a multicast period. They formulate the problem as an integer programming problem, which is NP-hard. Then, they propose a low complexity heuristic algorithm to provide a feasible solution. In [86], the authors study the joint placement and delivery problem in a two tier system, where both the SBSs and users have caching capability. They propose three methods to place the content in different layers and deliver the content to the requested users using multicast transmission. In the first method, the SBSs and users cache a part of the coded content. Then the request content is delivered by the SBSs directly from their storages. In the second method, the users cache a part of the coded content. The rest of requested content is forwarded by the SBSs from the MBS to the users. In the third method, the system is divided into two subsystems, one uses the first method and the other uses the second method. The authors compare the data rate in each method and show that the third method can minimize the required data rate under the QoS constraints.

Some research also considers multi-point cooperative transmission, where macro BS and cache-enabled SBSs or relays cooperatively serve users. In [87], a joint optimization of content placement and power control is discussed to maximize the probability of conducting the cooperative transmission, where the limited caching storage at each relay station is considered. In [88], the degree of freedom in cooperative transmission is discussed for the content placement strategy. The research shows that the upper bound is achievable even with small caching space at each relay station.

D. Summary and Discussion

The main purpose of introducing caching into HetNets is to reduce the backhaul bandwidth. Most of existing contributions use PPPs to model the locations of the users and BSs. The algorithm design is mainly based on the distribution of content popularity, which is treated as either prior information or estimated. Optimization and game theory are usually used in the algorithm designs. The results indicate system performance is sensitive to the caching space at each BS or user.

V. CACHE-ENABLED D2D NETWORKS

The goal of introducing caching techniques into D2D communication networks is to improve the area spectral efficiency while providing low backhaul cost. Conventionally, the D2D communication networks allow each user to directly communicate with its neighboring users. However, the neighboring users still need to connect to the BS to obtain the content, which requires high speed backhaul as well as suitable data rate wireless links between neighboring users and BSs. In contrast, when we introduce caching techniques into D2D networks, each user can precache the content with low speed and low cost backhaul links. When the cached content is requested, the users can establish high speed D2D communication without involving BSs. As a result, the spectral efficiency can be significantly improved with low backhaul cost.

In algorithm design, existing contributions are mainly based on three types of system models that describe the locations of users: (1) PPP-based models [39], [91]–[94]; (2) Poisson-based clustering models [95]; and (3) grid-based clustering models [25], [96]–[99]. In PPP-based models, the locations of all devices are modeled as a homogeneous PPP. In Poisson-based clustering models, the users are grouped into different clusters and each cluster has a cluster center and some cluster members. The locations of the cluster centers follow a PPP and the locations of the cluster members in each cluster follow a normal distribution. In grid-based models, the locations of all devices are at the intersections of a regular grid.

In the following, we introduce the existing works from three aspects: (1) placement strategy; (2) delivery strategy; and (3) joint placement and delivery strategy. Table IV shows a comparison of existing work. From the table, most strategies maximize the area spectral efficiency or the coverage probability or per-user throughput. They consider uncoded data and use optimization tools to formulate and solve the placement and delivery problems. The abbreviations used in Table IV are listed in Table I.

A. Placement Strategy

The content placement strategy is very important in cache-enabled D2D communication networks since their performance determines the portion of the overall traffic that can be offloaded by D2D communication. If the content is well placed, each user can easily find its requested content at a neighboring user. Then a large portion of the overall traffic can go through the short range D2D communication and the system can achieve high area spectral efficiency. Otherwise, if the content is not well placed, most of the users may not find their requested content at a neighboring user. Then, the system cannot benefit from D2D communication. Therefore, designing a good placement strategy is crucial for the performance of cache-enabled D2D networks.

In [91], the authors consider a PPP-based model and propose a content placement algorithm, where co-channel interference, noise, and channel fading are considered. The content placement algorithm is described by the caching probability, which is the probability a given content is stored in a cache. In the algorithm, each user independently caches the content based on a caching probability. By maximizing the coverage probability, they obtain the optimal caching distribution, which indicates the caching probabilities over different content. The results show that the optimal caching distribution follows a Zipf distribution. In [95], the authors consider a Poisson-based clustering model and propose a cluster-centric content placement strategy to maximize the coverage probability of the whole cluster. The results indicate that to obtain a large coverage probability, the popular content should be placed at the users that are close to cluster centers.

From the above discussion, the D2D caching technique benefits service providers by increasing the number of users that can be supported. However, users may not be willing to participate in caching because caching content will inevitably consume the users' storage and energy resources. Thus, we need to introduce an incentive to encourage users to participate in D2D caching. In [98], the authors consider the incentive issue and propose a cooperative content placement method. They formulate an economic problem to minimize the total economic cost for the service provider. They show that the proposed method can reduce the total cost, compared with the traditional ones without incentive. In [100], the authors design an incentive mechanism based on the amount of content that the users provide to their neighboring users. They formulate the content placement problem as a Stackelberg game, where the BS serves as a leader announcing the reward to minimize the BS's economic cost and users act as followers determining the caching strategy to maximize their utility. Then, the authors propose an iterative gradient algorithm to obtain the Stackelberg Equilibrium. The results show that the proposed scheme outperforms other content placement schemes which do not employ an incentive in terms of lower BS cost and higher user utilities.

B. Delivery Strategy

In cache-enabled D2D networks, each user may find its requested content is available at multiple neighboring users,

Table IV
COMPARISON OF DIFFERENT PLACEMENT AND DELIVERY STRATEGIES IN CACHE-ENABLED D2D NETWORKS

Ref.	Strategy	Algorithm	File Type	Model	Control	Throughput	Interference
[91]	Placement	max-CP	Uncoded	Stochastic Geometry	Distributed	Large	Considered
[95]		max-ASE	Uncoded	Optimization	Centralized	Large	-
[98]		min-ECPC	Uncoded	Optimization	Centralized	Large	-
[100]		min-DC	Uncoded	Game Theory	Distributed	Moderate	-
[101]	Delivery	max-NUS	Uncoded	Optimization	Centralized	Moderate	Considered
[99]	Placement & Delivery	min-LC	Uncoded	Optimization	Centralized	Small	Considered
[39]		max-SOP	Uncoded	Optimization	Distributed	Moderate	Considered
[93]		min-ENAL	Uncoded	Optimization	Distributed	Moderate	Considered
[97]		max-PUT	Coded & uncoded	Optimization	Distributed	Moderate	Considered
[32]		max-PUT	Coded & uncoded	Optimization	Distributed	Large	-
[25]		max-PUT	Coded & uncoded	Optimization	Distributed	Large	-
[94]		min-DD	Uncoded	Optimization	Centralized	Large	-

and each cached content may also be requested by multiple users. To efficiently use our limited resources over time and frequency to delivery the content, we need to carefully design the content delivery strategy to determine which cached file serves which user, i.e., *link scheduling*. Here, designing good delivery strategies is quite challenging since the topology of the network and the interference among different links are complicated.

In [101], the authors propose a heuristic method to schedule the D2D communication links based on the geographic distance between the users who request content, i.e., receivers, and the users who can potentially provide the content, i.e., transmitters. In the proposed method, they randomly choose a receiver and link it to the closest transmitter. Then, they repeat the same process for the rest of the receivers until all receivers are linked to a transmitter. Their results show that the proposed method can improve the spectral reuse compared with traditional interference avoidance schemes [97].

C. Joint Placement and Delivery Strategy

When designing joint placement and delivery strategies, we usually consider two kinds of methods, deterministic-caching-based methods and random-caching-based methods. In the former method, the BS or the cluster center determines how to cache content while in the latter method, each user determines how to cache content (usually based on a caching distribution).

For deterministic-caching-based methods, the existing contributions usually assume that the location and CSI of the users are known to the BS, and they are constant in both placement and delivery phases. In [99], the authors consider multi-hop wireless D2D communication networks with a deterministic square grid topology, where each user can only communicate with its four neighboring users. They propose two joint content placement and multi-hop routing schemes to minimize the average data rate of all links. With the proposed methods, the authors obtain the scaling law of link capacity when the number of users goes to infinity.

For random-caching-based methods, the existing contributions usually consider the PPP-based models [39], [92], [93] or grid-based models [96], [97]. In [39], the authors consider

interference among users and jointly optimize the caching distribution and placement policy to maximize the probability that a user can obtain its requested file in own cache or its neighboring users. In [92], [93], the whole communication area is divided into multiple square clusters of equal areas. The authors propose a joint placement and delivery strategy to maximize the number of active D2D links. They find the optimal caching distribution as well as the optimal transmission distance for each user. In [96], [97], the authors adopt the grid-based clustering model and propose a probability-based content placement strategy as well as a transmission policy. They can maximize the minimum average user throughput under an outage probability constraint. The results show that the average user throughput is mainly related to the size of the caching space at each user. The authors in [32] adopt more practical channel models, i.e., 4G cellular channel model and microwave D2D channel model. They compare the per-user throughput of the caching D2D method in [97] and some conventional D2D methods, including the D2D broadcasting and coded multicasting methods. The results show that the caching D2D method outperforms the conventional methods in many scenarios with practical channels.

There are also some contributions discussing both deterministic- and random-caching-based methods. In [25], the authors investigate the information theoretic bounds of the average per-user throughput under the grid-based clustering model. They propose two content placement and delivery strategies. The results show that both strategies can achieve the information theoretic upper bound under certain conditions. In [94], the authors divide a macro-cell into multiple virtual clusters and assume that only a one-hop D2D link can be activated in each cluster. They optimize the cluster size under both deterministic- and random-caching-based methods in terms of download delay. The results show that the proposed method can improve the spectral efficiency by one to two orders of magnitude, compared with the method without caching. They also show that the random-caching-based method outperforms the deterministic-caching-based method in robustness, and that it is more suitable to the dynamic scenario where users are moving among different cells.

D. Summary and Discussion

The existing contributions are mainly based on three system models: (1) the grid-based model is suitable to conduct analysis from the prospective of information theory, (2) the PPP-based model and (3) Poisson-based clustering model are suitable to design practical algorithms. In particular, the clustering model usually ignores the interference between different clusters, which simplifies the algorithm design. The PPP-based model considers the co-channel interference among different links, which is more practical. In addition, caching strategies can be divided into deterministic- and random-caching-based methods, where the latter is more suitable for the dynamic scenario with user mobility. Other issues, e.g., time delay constraint, caching space constraint, and the incentive to encourage user to participate in caching, are also considered in existing literature.

VI. CACHE-ENABLED C-RAN & F-RAN

In cache-enabled C-RANs and F-RANs (see Fig. 4(d)), content is usually cached at either the cloud center or the RRHs. If the content is cached at the cloud center, the goal is to reduce backhaul communication between the BBU and core network. If the content is cached at the RRHs, the main purpose is to reduce the fronthaul communication between BBU and RRHs. As a result, users can obtain their requested content directly from the cloud center or the RRHs, which is expected to significantly reduce either the backhaul or fronthaul cost in deploying C-RANs and F-RANs.

In this section, we introduce existing work from three aspects: (1) placement strategy; (2) delivery strategy; and (3) joint placement and delivery strategy. In addition, we provide a comparison of existing work in Table V. From the table, most of the strategies are designed to minimize backhaul or fronthaul traffic load, or RRH power consumption. They consider uncoded data and use optimization tools to solve the placement and delivery problems. The abbreviations used in Table V are listed in Table I.

A. Placement Strategy

In cache-enabled C-RANs and F-RANs, it is important to decide where to cache the content. If all the content is cached at the cloud center, users can obtain the required content with low backhaul cost. However, this raises the signaling overhead in the cloud center. On the other hand, if content is cached at local RRHs, the users obtain the required content directly from RRHs with low fronthaul cost. However, this weakens the processing capability of the cloud center, since most of the signal processing is conducted by the RRHs. Thus, when designing content placement strategies, we need to consider both the signal processing capability at the cloud center and the backhaul/fronthaul costs.

In [102], the authors propose a virtual caching framework, where content is cached at multiple *virtual machines* (VMs)[¶]

[¶]According to [109], virtual machine is an emulation of a computer system, which is based on a computer architecture and provides functionality of a physical computer.

inside the cloud center. In the proposed framework, VMs connecting with several RRHs can share their caching information, which enables high efficient cooperation among the VMs. The authors design an optimized content placement strategy to minimize the expected sum traffic load of the backhaul links between the cloud center and the core networks. They formulate an optimization problem to maximize the user demands satisfied by the cache of VMs. The results show that the proposed method can effectively reduce the expected sum traffic load compared with traditional schemes without caching. At the same time, the processing overhead at the cloud center is reduced since they use multiple VMs to communicate with different RRHs. In [103], the authors further propose a cluster content caching scheme, where they divide the RRHs into different clusters and the RRHs in the same cluster share a common local cluster storage. In the scheme, the popular content is cached in each local cluster storage. The authors use stochastic geometry tools to analyze the capacity and backhaul cost. The results show that the proposed scheme outperforms the centralized caching scheme that performs the caching of popular content only in the cloud center, in terms of system capacity and backhaul cost.

B. Delivery Strategy

To improve the performance of C-RANs and F-RANs, each user is usually cooperatively served by a group of RRHs. In this way, the requested data only needs to be cooperatively processed by the RRHs in the group rather than all RRHs under the control of a center processor. Thus, the clustering technique is very important when we design the delivery strategies. In addition, a multicasting technique is also considered when multiple users request the same content, which can further improve the spectrum efficiency.

In [104] and [110], the authors investigate the content delivery problem by jointly designing RRH clustering and multicasting in cache-enabled C-RANs, where content is cached at the RRHs. They formulate the problem as an optimization problem to minimize the total network cost subject to the *quality-of-service* (QoS) constraint. Then they reformulate an equivalent sparse multicasting beamforming design problem and propose a convex-concave algorithm to find the solution. The results show that the proposed method can significantly reduce the network cost compared with the conventional unicasting beamforming and clustering methods. They also find that the proper BS clustering is related to not only the channel conditions between the users and RRHs, but also the RRHs' cached content. In [105], the authors consider the fronthaul cost and design the clustering and multicasting beamforming in cache-enabled F-RAN. They formulate an optimization problem to minimize the transmission power consumption of all RRHs in a cluster under the constraints of each user's QoS and fronthaul capacity. They propose an iterative optimization algorithm to solve the problem and show its advantages compared with [104] in terms of power consumption.

Some delivery strategy research focuses on the cooperation between the local RRHs and cloud center. In [43], the authors propose a delivery strategy based on fronthaul quantization and

Table V
COMPARISON OF DIFFERENT PLACEMENT AND DELIVERY STRATEGIES IN CACHE-ENABLED C-RAN/F-RAN

Ref.	Strategy	Algorithm	File Type	Model	Control	Backhaul /Fronthaul Cost	Power Consumption
[102]	Placement	min-TL	Uncoded	Optimization	Centralized	Reduced	-
[103]		min-TL	Uncoded	Stochastic Geometry	Centralized	Reduced	-
[104]	Delivery	min-(TC+BC)	Uncoded	Optimization	Centralized	Moderate	Reduced
[105]		min-PC	Uncoded	Optimization	Centralized	Considered	Reduced
[43]		max-DR	Coded	Optimization	Centralized	Reduced	Moderate
[106]		min-NC	Coded	Optimization	Centralized	Reduced	Reduced
[107]	Placement & Delivery	min-FL	Coded	Optimization	Centralized	Reduced	-
[108]		max-Utility	Uncoded	Stochastic Geometry	Distributed	Moderate	Low
[27]		min-DT	Coded	Optimization	Centralized	-	Moderate

superposition coding in cache-enabled F-RANs, where a user's requested content is cooperatively delivered by the local RRHs and cloud center. In this method, they consider two fronthaul modes, a hard-transfer mode and a soft-transfer mode. In the hard-transfer mode, the fronthaul links are used to transfer the uncoded files that are not cached in the local RRHs. In the soft-transfer mode, the fronthaul links are used to transfer the precoded files that are not cached in the local RRHs. They propose to maximize the delivery rate under fronthaul capacity and power constraints on each RRH. The results show that the soft-transfer mode uses the fronthaul resources more efficiently than the hard-transfer mode. In [106], the authors investigate a beamforming method in C-RAN, where each RRH is installed with multiple antennas and cached coded files. They optimize the beamforming vector of all RRHs by minimizing the total network cost, including transmission power and backhaul cost. The results show that the backhaul cost can be significantly reduced even with small caching space in each RRH.

C. Joint Placement and Delivery Strategy

Some research has considered joint placement and delivery strategy to further enhance the system performance. In [107], the authors propose a joint content placement and user-centric RRH clustering method in C-RANs. The purpose is to obtain a tradeoff between the caching space at the RRHs and the fronthaul cost. In the algorithm, the user-centric RRH clustering is modeled as a multi-timescale problem to optimize the average weighted sum rate. In addition, the content placement at RRHs is exploited to reduce the fronthaul load. In [108], the authors propose a cluster-based content placement structure in C-RANs and further design the delivery strategy considering resource allocation and the association between the RRH and users. They analyze the capacity and energy efficiency of the proposed structure. Based on that, they jointly design the radio resource allocation and RRH association using a nested game. The results show that the proposed structure and strategy can effectively improve the capacity as well as the energy efficiency. In [27], the authors jointly optimize the content placement, fronthaul cost, and RRH resource allocation in F-RANs from an information theoretic view. The purpose is to minimize the normalized delivery time measured under an ideal interference-free channel condition, i.e., the delivery time

per bit normalized by power consumption. From this work, they provide a latency-centric understanding of the F-RAN system, which focuses on the high SNR regions.

D. Summary and Discussion

The existing contributions focus on designing placement and delivery strategies that can exploit the caching ability at the RRHs as well as the signal processing ability at the cloud center simultaneously. They propose different network architectures for C-RANs or F-RANs in term of energy efficiency and backhaul/fronthaul costs. Some consider coded content placement methods, while frequently using clustering and multicast beamforming techniques to design the transmission strategies.

VII. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

In this section, we identify some of the challenges and discuss some directions that have not been satisfactorily addressed regarding cache-enabled wireless networks, including mobility-aware caching, coded caching, and big data processing.

A. Mobility-Aware Caching

In wireless mobile communication networks, the performance of caching is highly affected by user mobility. In dynamic scenarios the networks may need to cache multiple copies of content to different BSs, so that they can use the cached content to serve the moving users. However, under finite caching capacity at each BS, the BSs cannot cache all the content. Thus, only a small amount of requested content can be cached, which severely limits the performance of the networks. If the networks know user mobility, e.g., the path of all users, the networks can precisely cache the requested content at the right BSs along the users' path, which allows the maximum amount of requested content to be served by caching. In practice, it is difficult to precisely predict the exact path a user may take before they take it. We need to design caching algorithms under imperfect user path knowledge, which raises many challenges.

In addition, there is no tractable mathematical model available to describe user mobility, which impedes the understanding of the user mobility in cache-enabled wireless communication networks. Current research, e.g., [12], [66], [75], mainly proposed a framework for considering the user mobility in cache-enabled wireless communication networks. Therefore, there are lots of research opportunities, e.g., from measurement and modeling to algorithm design and analysis.

B. Coded Caching

Even though coded caching has impressive performance improvement, it faces significant challenges in practical scenarios. The first one is the strong condition that requires the content is well placed at a group of users, e.g., in D2D networks, each user in a group needs to pre-cache a part of the files that other users request. The second challenge is imperfect channels. Currently, most existing contributions consider a perfect error-free channel, i.e., the message is always successfully delivered from base station to each user [48], [86]. However, in practice, the randomness of wireless channel causes some failures of message delivery. Therefore, it is important to design practical coded caching algorithms to consider suboptimally placed content among users and channel imperfection.

On the other hand, most existing work on coded caching is designed for the networks with one tier, i.e., a BS serves multiple mobile users. However, they are not suitable for the networks with two tiers, i.e., a BS serves SBSs and the SBSs further serve mobile users. Thus, it is important to extend the coded caching methods to two tier networks, which is very challenging because of the complicated network structure.

C. Big Data Processing Techniques for Wireless Caching

In cellular systems, there are large amounts of data generated every minute which include user locations, transmission parameters, and resource allocation parameters [111]. When we further consider the caching capability, there is even more data, e.g., the feature and the popularity of the content. Therefore, it is reasonable to use big data processing techniques to improve the performance of the cache-enabled wireless communication networks. For example, big data processing may predict the future location of the each user, what content each user requests, when it is requested [11], and even the social relationship among users [112]. As a result, the network can better handle user mobility.

Another advantage that big data processing can offer is the capability for interaction between users and cellular network current configuration. Conventionally, we design and optimize the cellular network configuration according to users' requests, i.e., the network adapts its parameters to meet the users' requests. If we consider big data processing techniques which tend to recommend some content that is very close to each user [113], the network resources and users' requests are able to interact with each other to encourage requests more suited to the current configuration. Then, the communication performance can be significantly improved. For example, based on user preference, big data processing may recommend the

content that is partly pre-cached at its neighbor user. As a result, we can create more opportunities to meet the condition of coded caching, so that the system can benefit from coded caching.

D. Others

There are many others challenges that may open new research directions. For example, the energy efficiency of the system needs to be considered, in particular, when we consider the caching capability at the user side [114] [115]. The capacity of the caching storage also needs to be considered since each device usually has finite storage [116]. In some situations, the available capacity for caching even varies with time, which further complicate caching strategies.

In addition, the network resources that are used in the caching phase should be further considered. Note that the cost of delivery can be highly related to how long the content is pre-cached in advance. For example, if the content is pre-cached via WiFi around midnight, the network resource that is used for caching is usually inexpensive since WiFi is usually free. On the other hand, if the content is pre-cached via cellular resource only several minutes in advance, the network resource can be costly.

VIII. CONCLUSIONS

This paper surveyed the existing contributions on caching techniques in different wireless communication networks, including cache-enabled cellular networks, cache-enabled HetNets, cache-enabled D2D networks, and cache-enabled C-RAN/F-RAN. In particular, we first provided a taxonomy to outline the network types, mathematic tools, control structures, performance metrics, and caching strategies. Then, we introduced the contributions of each network type from three aspects: caching placement, caching delivery, and joint placement and delivery. Finally, we discussed the main research challenges and future research directions, e.g., mobility-aware caching, coded caching, and big data processing techniques.

Currently, the development of caching techniques in wireless communication networks is at a very preliminary stage and most of the existing work focuses on theoretical study, in particular from the prospectives of information theory and algorithm design. The performance is demonstrated by numerical and simulation results under strong assumptions. The primary goal of these work is to show the potential of cache-enabled wireless communication networks compared with conventional system without caching capability. From published materials, there are some experimental demonstrations from academic [11], [117], [118], but no demonstration from industry.

From existing academic research results, the following aspects have been clarified: The primary goals of introducing caching into cellular systems are to reduce the backhaul cost, download delay, and maximize per-user throughput. For cache-enabled cellular networks and HetNets, caching can effectively reduce the backhaul cost when it combines with coding and multicasting techniques. In particular, for

cache-enabled HetNets, improving the accuracy of the estimated content popularity can enhance the system performance. For cache-enabled D2D networks, the network topology and caching capacity of each user highly affect the efficiency of the caching. For cache-enabled C-RANs and F-RANs, the clustering methods have significant impact on the caching strategies, which further affect the backhaul/fronthaul costs. In the future, wireless caching techniques are expected to attract significant attention in both academy and industry, in particular, in 5G development. Many challenges need to be addressed, which opens broad research opportunities.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their invaluable comments that improve the paper.

REFERENCES

- [1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021 white paper [OL]," <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html>, 2017.
- [2] A. Gupta and R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 52, pp. 1206–1232, Aug. 2015.
- [3] J. G. Andrews, S. Buzzi, W. Choi *et al.*, "What will 5G be?" *IEEE Commun. Mag.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
- [4] M.-G. D. Benedetto and B. R. Vojcic, "Ultra wide band wireless communications : A tutorial," *Journ. Commun. and Networks*, vol. 5, no. 4, pp. 290–302, Dec. 2003.
- [5] Z. Xiang, M. Tao, and X. Wang, "Massive MIMO multicasting in noncooperative cellular networks," *IEEE Journal on Selected Areas in Commun.*, vol. 32, no. 6, pp. 1180–1193, Jun. 2014.
- [6] R. Taori and A. Sridharan, "Point-to-multipoint in-band mmwave backhaul for 5G networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 195–201, Jan. 2015.
- [7] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 136–144, Mar. 2013.
- [8] S. Annapureddy, A. Barbieri, S. Geirhofer, S. Mallik, and A. Gorokhov, "Coordinated joint transmission in WWAN," in *IEEE Commun. Theory Workshop*, Dublin, Ireland, Aug. 2010.
- [9] R. Irmer, H. Droste, P. Marsch, M. Grieger *et al.*, "Coordinated multi-point: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.
- [10] E. Bastug, M. Bennis *et al.*, "Big data meets telcos: A proactive caching perspective," *Journ. Commun. and Networks, Special Issue on Big Data Networking Challenges and Applications*, vol. 17, no. 6, pp. 549–557, Dec. 2015.
- [11] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader *et al.*, "Big data caching for networking: Moving from cloud to edge," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 36–42, Sep. 2016.
- [12] N. Golrezaei, A. F. Molisch, A. G. Dimakis *et al.*, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.
- [13] G. Zhang, Y. Li, and T. Lin, "Caching in information centric networking: A survey," *Computer Networks*, vol. 57, pp. 3128–3141, 2013.
- [14] M. Zhang, H. Luo, and H. Zhang, "A survey of caching mechanisms in information-centric networking," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 3, pp. 1473–1499, Aug. 2015.
- [15] A. Ioannou and S. Weber, "A survey of caching policies and forwarding mechanisms in information-centric networking," *IEEE Commun. Surveys & Tutorials*, vol. pp, no. 99, pp. 1–41, 2016.
- [16] G. Xylomenos, C. N. Ververidis, V. A. Siris, N. Fotiou, C. Tsilopoulos *et al.*, "A survey of information-centric networking research," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 2, pp. 1024–1049, Apr. 2014.
- [17] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, "A survey of energy efficient caching in information-centric networking," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 122–129, Nov. 2014.
- [18] —, "A survey of green information-centric networking: Research issues and challenges," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 3, pp. 1455–1469, Oct. 2015.
- [19] U. Niesen, D. Shah, and G. Wornell, "Caching in wireless networks," in *2009 IEEE International Symposium on Information Theory (ISIT)*, Seoul, Korea, Jun. 2009, pp. 2111–2115.
- [20] <https://software.intel.com/en-us/articles/video-aware-wireless-networks>.
- [21] A. Khreishah and J. Chakareski, "Collaborative caching for multicell-coordinated systems," in *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Hong Kong, Apr. 2015, pp. 257–262.
- [22] M. A. Maddah-Ali and U. Niesen, "Coding for caching: Fundamental limits and practical challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 23–29, Aug. 2016.
- [23] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.
- [24] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inform. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.
- [25] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inform. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.
- [26] M. Gregori, J. Gomez-Vilardebo, J. Matamoros, and D. Gunduz, "Wireless content caching for small cell and D2D networks," *IEEE Journal on Selected Areas in Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.
- [27] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *IEEE International Symposium on Information Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 2029–2033.
- [28] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "Mobile edge computing: Survey and research outlook," [Online]. Available: <https://arxiv.org/pdf/1701.01090.pdf>, 2017.
- [29] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.
- [30] A. F. Molisch, G. Caire, D. Ott, J. R. Foerster, D. Bethanabhotla, and M. Ji, "Caching eliminates the wireless bottleneck in video-aware wireless networks," *Hindawi Advances in Electrical Engineering*, vol. 2014, pp. 1–13, Oct. 2014.
- [31] 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.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.
- [32] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE Journal on Selected Areas in Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.
- [33] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.
- [34] "LTE radio link budgeting and RF planning," <https://sites.google.com/site/teencyclopedia/lte-radio-link-budgeting-and-rf-planning>.
- [35] "Lte advanced: Heterogeneous networks."
- [36] N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Wireless video content delivery through distributed caching and peer-to-peer gossiping," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, US, Nov. 2011, pp. 1177–1180.
- [37] E. Bastug, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *11th International Symposium on Wireless Communications Systems (ISWCS)*, Barcelona, Spain, Aug. 2014, pp. 649–653.
- [38] S. Zaidi, M. Ghogho, and D. C. McLernon, "Information centric modeling for two-tier cache enabled cellular networks," in *IEEE Intern. Conference on Commun. Workshop (ICCW)*, London, UK, Jun. 2015, pp. 80–86.
- [39] B. Chen, C. Yang, and Z. Xiong, "Optimal caching and scheduling for cache-enabled D2D communications," *IEEE Commun. Letters*, vol. pp, no. 99, pp. 1–1, Jan. 2017.
- [40] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari *et al.*, "Cloud RAN for mobile networks: A technology overview," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, pp. 405–426, Jan. 2015.

- [41] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 3, pp. 2282–2308, Mar. 2016.
- [42] S.-C. Hung, H. Hsu, S.-Y. Lien, and K.-C. Chen, "Architecture harmonization between cloud radio access networks and fog networks," *IEEE Access*, vol. 3, pp. 3019–3034, Dec. 2015.
- [43] S.-H. Park, O. Simeone, and S. Shamai, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7621–7632, Nov. 2016.
- [44] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Network*, vol. 30, no. 4, pp. 46–53, Jul. 2016.
- [45] R. Tandon and O. Simeone, "Harnessing cloud and edge synergies: Toward an information theory of fog radio access networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 45–50, Aug. 2016.
- [46] M. A. Kader, E. Bastug, M. Bennis, E. Zeydan, A. Karatepe, A. S. Er, and M. Debbah, "Leveraging big data analytics for cache-enabled wireless networks," in *2015 IEEE Global Communications Conf. (Globecom)*, San Diego, CA, US, Dec. 2015, pp. 1–6.
- [47] X. Li, X. Wang, and V. C. M. Leung, "Weighted network traffic offloading in cache-enabled heterogeneous networks," in *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
- [48] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
- [49] —, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Networking*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.
- [50] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Trans. Networking*, vol. 24, no. 2, pp. 836–1040, Apr. 2016.
- [51] M. Abrams, C. R. Standridge, G. Abdulla, S. Williams, and E. A. Fox, "Caching proxies: Limitations and potentials," in *Intern. WWW-4 Conference*, Boston, US, Jul. 1995.
- [52] M. Arlitt, L. Cherkasova, J. Dilley, R. Friedrich, and T. Jin, "Evaluating content management techniques for web proxy caches," *ACM SIGMETRICS Performance Evaluation Review*, vol. 27, no. 4, pp. 3–11, Mar. 2000.
- [53] J. Gu, W. Wang, A. Huang, H. Shan, and Z. Zhang, "Distributed cache replacement for caching-enable base stations in cellular networks," in *2014 IEEE Intern. Conference on Commun. (ICC)*, Sydney, Jun. 2014, pp. 2648–2653.
- [54] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *IEEE INFOCOM*, New York, US, Apr. 1999, pp. 126–134.
- [55] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the worlds largest user generated content video system," in *7th ACM SIGCOMM Conference on Internet Measurement*, San Diego, US, Oct. 2007, pp. 1–14.
- [56] L. Zhang, H.-C. Yang, and M. O. Hasna, "Generalized area spectral efficiency: An effective performance metric for green wireless communications," *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 747–757, Feb. 2014.
- [57] N. Golrezaei, K. Shanmugam, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *IEEE INFOCOM*, Orlando, Florida USA, Mar. 2012, pp. 1107–1115.
- [58] Z. Chen and M. Kountouris, "Cache-enabled small cell networks with local user interest correlation," in *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Stockholm, Sweden, Jun. 2015, pp. 680–684.
- [59] K. Hamidouche, W. Saad, M. Debbah, and H. V. Poor, "Mean-field games for distributed caching in ultra-dense small cell networks," in *American Control Conference (ACC)*, Boston, US, Jul. 2016, pp. 4699–4704.
- [60] V. S. Varma and T. Q. S. Quek, "Congestion games in caching enabled heterogeneous cellular networks," in *2015 IFIP Networking Conference (IFIP Networking)*, Toulouse, France, May 2015, pp. 1–6.
- [61] Z. Hu, Z. Zheng, T. Wang, L. Song, and X. Li, "Game theoretic approaches for wireless proactive caching," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 37–43, Aug. 2016.
- [62] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inform. Theory*, vol. 46, no. 4, pp. 1204–1216, Jul. 2000.
- [63] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 3, pp. 996–1019, Jun. 2013.
- [64] P. Ostovari, J. Wu, and A. Khreishah, "Efficient online collaborative caching in cellular networks with multiple base stations," in *13th IEEE International Conference on Mobile Ad hoc and Sensor Systems (IEEE MASS)*, Brasilia, Brazil, Oct. 2016, pp. 1–9.
- [65] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," in *2015 IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, US, Dec. 2015, pp. 1–6.
- [66] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: Modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, Aug. 2016.
- [67] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *IEEE International Conference on Communications (ICC)*, London, UK, Jun. 2015, pp. 3358–3363.
- [68] B. Zhou, Y. Cui, and M. Tao, "Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks," in *2015 IEEE International Symposium on Information Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 1412–1416.
- [69] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, Jul. 2016.
- [70] A. Liu and V. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, Dec. 2013.
- [71] A. Ghorbel, M. Kobayashi, and S. Yang, "Cache-enabled broadcast packet erasure channels with state feedback," in *Fifty-third Annual Allerton Conference*, UIUC, IL, US, Sep. 2015, pp. 1446–1453.
- [72] J. Llorca, A. M. Tulino, K. Guan, and D. C. Kilper, "Network-coded caching-aided multicast for efficient content delivery," in *IEEE Intern. Conference on Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 3557–3562.
- [73] A. Dimakis, P. Godfrey, Y. Wu, M. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4539–4551, Sep. 2010.
- [74] M. Gatzianas, L. Georgiadis, and L. Tassiulas, "Multiuser broadcast erasure channel with feedback-capacity and algorithms," *IEEE Trans. Inform. Theory*, vol. 59, no. 9, pp. 5779–5804, Sep. 2013.
- [75] Y. Guan, Y. Xiao, H. Feng, C.-C. Shen, and L. J. Cimini, "Mobicacher: Mobility-aware content caching in small-cell networks," in *2014 IEEE Global Communications Conference (GLOBECOM)*, Austin, TX, US, Dec. 2014, pp. 4537–4542.
- [76] C. Yang, Z. Chen, Y. Yao, B. Xia, and H. Liu, "Energy efficiency in wireless cooperative caching networks," in *2014 IEEE Intern. Conference on Commun. (ICC)*, Sydney, Jun. 2014, pp. 4975–4980.
- [77] B. N. Bharath, K. G. Nagananda, and H. V. Poor, "A learning-based approach to caching in heterogeneous small cell networks," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1674–1686, Apr. 2016.
- [78] P. Blasco and D. Gunduz, "Learning-based optimization of cache content in a small cell base station," in *2014 IEEE Intern. Conference on Commun. (ICC)*, Sydney, Jun. 2014, pp. 1897–1903.
- [79] M. S. ElBamby, M. Bennis, W. Saad, and M. Latva-aho, "Content-aware user clustering and caching in wireless small cell networks," in *IEEE International Symposium on Wireless Communications Systems (IEEE ISWCS 2014)*, Barcelona, Spain, Aug. 2014, pp. 945–949.
- [80] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Match to cache: Joint user association and backhaul allocation in cache-aware small cell networks," in *2015 IEEE International Conference on Communications (ICC)*, London, UK, Jun. 2015, pp. 3082–3087.
- [81] H. Li, Z. Wang, and D. Hu, "Joint wireless and backhaul load balancing in cooperative caches enabled small-cell networks," in *2015 IEEE 26th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Hong Kong, Sep. 2015, pp. 1889–1894.
- [82] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6284–6297, Aug. 2016.
- [83] L. Xiang, D. W. K. Ng, T. Islam, R. Schober, and V. W. S. Wong, "Cross-layer optimization of fast video delivery in cache-enabled relaying networks," in *2015 IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, US, Dec. 2015, pp. 1–7.
- [84] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation caching and routing algorithms for massive mobile data delivery," in *IEEE Global Commun. Conference*, Atlanta, USA, Dec. 2013, pp. 3534–3539.
- [85] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

- [86] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, "Hierarchical coded caching," *IEEE Trans. Inform. Theory*, vol. 62, no. 6, pp. 3212–3229, Jun. 2016.
- [87] A. Liu and V. Lau, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 390–402, Jan. 2014.
- [88] W. Han, A. Liu, and V. K. N. Lau, "Degrees of freedom in cached MIMO relay networks," *IEEE Trans. Signal Processing*, vol. 63, no. 15, pp. 3986–3997, Aug. 2015.
- [89] M. K. Kiskani and H. R. Sadjadpour, "Capacity of cellular networks with femtocache," in *IEEE INFOCOM*, San Francisco, CA, US, Apr. 2016, pp. 9–14.
- [90] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," in *IEEE Intern. Symposium on Information Theory (ISIT)*, Istanbul, Turkey, Jul. 2013, pp. 1017–1021.
- [91] D. Malak and M. Al-Shalash, "Optimal caching for device-to-device content distribution in 5G networks," in *2014 IEEE Globecom Workshops (GC Wkshps)*, Austin, TX, US, Dec. 2014, pp. 863–868.
- [92] N. Golrezaei, A. G. Dimakis, and A. Molisch, "Wireless device-to-device communications with distributed caching," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, Cambridge, MA, US, Jul. 2012, pp. 2781–2785.
- [93] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inform. Theory*, vol. 60, no. 7, pp. 4286–4298, Jul. 2014.
- [94] 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 Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.
- [95] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Fundamentals of cluster-centric content placement in cache-enabled device-to-device networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2511–2526, Jun. 2016.
- [96] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *IEEE Intern. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013, pp. 1461–1465.
- [97] —, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inform. Theory*, vol. 61, no. 12, pp. 6833–6859, Dec. 2015.
- [98] M. Taghizadeh, K. Micinski, S. Biswas, E. Torng, and S. Biswas, "Distributed cooperative caching in social wireless networks," *IEEE Trans. Mobile Computing*, vol. 12, no. 6, pp. 1037–1053, Jun. 2013.
- [99] S. Gitis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inform. Theory*, vol. 59, no. 5, pp. 2760–2776, May 2013.
- [100] Z. Chen, Y. Liu, B. Zhou, and M. Tao, "Caching incentive design in wireless D2D networks: A stackelberg game approach," in *IEEE Intern. Conference on Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
- [101] N. Naderializadeh, D. T. Kao, and A. S. Avestimehr, "How to utilize caching to improve spectral efficiency in device-to-device wireless networks," in *Fifty-second Annual Allerton Conference*, UIUC, IL, US, Oct. 2014, pp. 415–422.
- [102] X. Li, X. Wang, C. Zhu, W. Cai, and V. C. M. Leung, "Caching-as-a-service: Virtual caching framework in the cloud-based mobile networks," in *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Hong Kong, Apr. 2015, pp. 372–377.
- [103] Z. Zhao, S. Jia, Y. Li, M. Peng, and C. Wang, "Performance analysis of cluster content caching in cloud-radio access networks," in *2015 IEEE Globecom Workshops (GC Wkshps)*, San Diego, CA, US, Dec. 2015, pp. 1–6.
- [104] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Aug. 2016.
- [105] D. Chen, S. Schedler, and V. Kuehn, "Backhaul traffic balancing and dynamic content-centric clustering for the downlink of fog radio access network," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Edinburgh, UK, Jul. 2016, pp. 1–5.
- [106] Y. Ugur, Z. H. Awan, and A. Sezgin, "Cloud radio access networks with coded caching," in *IEEE 20th Intern. ITG Workshop on Smart Antennas*, Munich, Germany, Mar. 2016, pp. 1–5.
- [107] A. Liu, V. Lau, and W. Han, "Randomized user-centric clustering for cloud radio access network with PHY caching," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Orlando, FL, US, Dec. 2015, pp. 948–952.
- [108] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE Journal on Selected Areas in Commun.*, vol. 34, no. 5, pp. 1207–1221, May 2016.
- [109] https://en.wikipedia.org/wiki/Virtual_machine.
- [110] H. Zhou, M. Tao, E. Chen, and W. Yu, "Content-centric multicast beamforming in cache-enabled cloud radio access networks," in *2015 IEEE Global Communications Conf. (Globecom)*, San Diego, CA, US, Dec. 2015, pp. 1–6.
- [111] S. Bi, R. Zhang, Z. Ding, and S. Cui, "Wireless communications in the era of big data," *IEEE Commun. Magazine*, vol. 53, no. 10, pp. 190–199, Oct. 2015.
- [112] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Magazine*, vol. 52, no. 3, pp. 82–89, Aug. 2014.
- [113] H. Yin, X. Zhou, B. Cui *et al.*, "Adapting to user interest drift for POI recommendation," *IEEE Trans. Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2566–2581, Oct. 2016.
- [114] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE Journal on Selected Areas in Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.
- [115] R. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.
- [116] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.
- [117] M. Sheng, C. Xu, J. Liu *et al.*, "Enhancement for content delivery with proximity communications in caching enabled wireless networks: Architecture and challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 70–76, Aug. 2016.
- [118] U. Niesen and M. A. Maddah-Ali, "Coded caching for delay-sensitive content," in *IEEE Intern. Conference on Commun. (ICC)*, Kuala Lumpur, Malaysia, Jun. 2015, pp. 5559–5564.