The Benefits of Hybrid Caching in Gauss-Poisson D2D Networks

Na Deng, *Member, IEEE* and Martin Haenggi, *Fellow, IEEE*

***Abstract*—Device caching has recently been proposed as an efficient way to offload traffic from congested cellular networks. However, previous works usually ignore the fact that in practice the user device may not be willing to help others due to the limited battery capacity. In this paper, we introduce cooperation among the D2D transmitters and propose two novel hybrid caching strategies—single-point caching combined with two-point cooperative caching with joint transmission (SPC-CCJT) or multi-stream transmission (SPC-CCMT)—aiming at saving the energy cost of content deliverers. Using tools from stochastic geometry, we propose an analytical framework of the hybrid caching strategies by modeling the locations of the D2D transmitters as a Gauss-Poisson process (GPP) to accurately capture the clustering and cooperative behaviors. Firstly, we consider a probabilistic caching placement and optimize the caching distribution to maximize the cache hit probability. Secondly, to compare the performance between different content delivery strategies, we derive the success probability and per-user capacity for SPC, CCJT, and CCMT, respectively. These results are then applied to evaluate the offloading gain and the distribution of the content retrieval delay for SPC-CCJT and SPC-CCMT in the GPP-based D2D networks. It turns out that significant offloading gain and delay improvement can be achieved by hybrid caching with cooperation while the energy cost of each cooperator is kept low.**

***Index Terms*—Stochastic geometry; Gauss-Poisson process; device caching; D2D networks; hybrid caching; offloading gain.**

I. Introduction

A. Motivation

The recent proliferation of new mobile devices (e.g., smartphones) has led to a fundamental shift of mobile traffic from voices and messages to rich multi-media contents, such as video streaming and application downloads, exacerbating the growth of traffic demand [1]. To cope with this growth, techniques such as millimeter wave communication (new spectral resources), massive multiple-input multiple-output (MIMO) (improving spectral efficiency), and cell densification (exploiting spatial reuse), are developed. Despite the benefits from these techniques, the deployment costs of many RF chains or high-speed backhaul installation are prohibitively expensive. Driven by the interesting observation that a large amount of content requests are asynchronous but redundant, wireless edge caching, i.e., prefetching popular contents during off-peak times at the edge of wireless networks, e.g., base stations, helper nodes, and user devices, has drawn much attention as a promising technique to alleviate the network congestion and improve the users' quality of experience [2–4].

As one of the promising caching approaches, caching files on the user devices with short-range device-to-device (D2D) communications has been shown to provide increased spectral reuse and throughput gain in D2D networks [5]. This way, multiple devices form a common virtual cache space where a large number of files can be stored, and the storage capacity of each individual device does not necessarily need to be very large. Different from small cell caching, device caching does not require special infrastructure and has the further advantage of reducing the demand on radio resources for cellular transmissions. Interestingly, device caching has the unique feature that the number of caching devices is inherently concentrated in regions with large demand [2], and accordingly, the cache capacity of the virtual cache space grows with the user density. However, an important issue challenging the implementation of device caching is whether a user is willing to be a helper. Although the users can be incentivized by the network operators, a key question that can not be neglected is the limited battery capacity of wireless devices. In practice, the energy consumption of a caching content delivery largely determines the willingness of the helper user. In other words, a user may only be willing to use a fraction of its remaining battery energy for file transfer. To cope with this issue, the battery consumption allowed by the helping users should be seeded into the device caching design, and efficient regimes that can reduce the transmit power or duration of each helper user should be exploited in order to encourage more users to be the "helper". Exploring this issue quantitatively is the goal of this paper.

B. Related Work

Recently, device caching has received significant attention as a means to offload traffic from congested cellular networks and improve the throughput and latency without requiring additional infrastructure. Notable progress has been made on investigating the benefits of device caching in the offloading and the throughput performance with various caching strategies proposed, see [2, 6–9] and references therein. Although several key insights were provided on the design of the cache placement and content delivery, most previous works made

Manuscript received Dec. 1, 2017; revised Apr. 6, 2018; accepted Apr. 19, 2018. This work was supported by the National Natural Science Foundation of China under Grant 61701071, the China Postdoctoral Science Foundation (2017M621129), the Xinghai Scholars Program, the Fundamental Research Funds for the Central Universities (DUT16RC(3)119), and the US NSF grant CCF 1525904. *(Corresponding author: Na Deng.)*

Na Deng is with the School of Information and Communication Engineering, Dalian University of Technology (DLUT), Dalian, 116024, China (e-mail: dengna@dlut.edu.cn).

Martin Haenggi is with the Dept. of Electrical Engineering, University of Notre Dame, Notre Dame 46556, USA (e-mail: mhaenggi@nd.edu).

several ideal assumptions which would have a great influence on the real implementation of device caching. First of all, they implicitly assumed that users acting as the transmitters are by definition willing to provide content delivery service without considering the fact that users are not obligated to help. Secondly, they assumed the battery capacity of the helper user to be infinite so that the file will be definitely delivered completely as long as the D2D link is established. To address this concern, authors in [10] recently quantified the offloading gain of a cache-enabled D2D communication system considering a maximal permissible battery consumption. Due to the battery limitation, some of the helper users can only transmit a fraction of the file and the remaining part is left for the base station (BS). Consequently, the offloading gain benefit from device caching would be reduced, and a large number of signaling between the BS and users would be required to guarantee the seamless transmission, thus causing extra delay, overhead, and system complexity. Different from [10], in this paper, we aim to overcome this limitation through cooperation to reduce the energy cost of each individual helper, expecting to encourage more users to participate in the device caching while taking full advantage of device caching in terms of the traffic offloading from the BSs. The very recent work [9] also investigated the benefit of the cooperation in device caching, however, the adopted grid model does not capture the randomness and the clustering features of the spatial distribution of the wireless devices.

Due to the analytical tractability, prior works based on stochastic-geometry modeling mostly used the Poisson point process (PPP) to model the spatial distribution of nodes and quantitatively analyzed metrics like success probability, mean achievable rate, offloading gain, content retrieval delay, etc., for the respective caching strategies [8, 10–12]. Although the PPP model has several convenient features for the analysis and these PPP-based works have provided useful design insights, this model is inadequate for those scenarios where the node locations exhibit correlations, especially for D2D users who are likely to be clustered in reality. On the one hand, the content-centric nature of D2D communication is primarily driven by the spatiotemporal correlation in the content demand; on the other hand, the spatial distribution of smart devices is mostly determined by the uneven population distributions due to some social activities and hotspots. Thus, compared with the PPP, clustered point processes are more suitable for capturing the clustering feature of devices. Few prior studies adopted clustered point processes to model the caching devices, most notably, the well-known Thomas cluster process was used for modeling and analyzing cache-enable D2D networks in [13], however, the performance metrics such as coverage probability and area spectral efficiency derived therein are in complex form involving multiple integrals. Considering the accuracy, tractability, and practicability tradeoffs, in this paper, we propose the Gauss-Poisson process (GPP) [14], which is also a Poisson cluster process but not a doubly Poisson or Cox process as the Thomas cluster process, as a model for cache-enabled D2D networks when devices exhibit clustering. Since the GPP includes the PPP as a special case, even if in some cases the PPP was a sufficiently accurate model, it is included in our analysis. More importantly, the GPP constitutes a simple yet effective network model to analyze wireless networks that apply cooperative techniques [15], which is at the center of the paper.

C. Contributions

The main objective of this paper is to introduce and promote two hybrid caching strategies for cache-enable D2D networks, where the helper user locations exhibit correlation. The limited battery capacity of wireless devices motivates the use of cooperative techniques to save energy. The contributions of the paper are:

- **Novel hybrid caching schemes**. We propose two hybrid caching schemes for a D2D network where some of the helper users transmit the file completely on their own, while some need to finish the content delivery jointly with an other helper user according to the users' willingness. To reduce the energy cost of each helper user, we introduce two new cooperative device caching strategies, namely cooperative caching with joint transmission (CCJT) and cooperative caching with multi-stream transmission (CCMT), aiming at reducing the transmit power and transmission duration of each helper user, respectively. Thus, the two new hybrid caching schemes are composed of single-point caching (SPC) and CCJT or CCMT, called SPC-CCJT and SPC-CCMT, respectively.
- **Gauss-Poisson process-based placement**. We consider a GPP-based cache placement to capture the clustering feature of devices. The GPP is well suited to our proposed hybrid caching strategies and belongs to the family of Poisson cluster processes, with the number of points in each cluster restricted to one or two. It describes the scenario where traditional device caching (SPC) and cooperative device caching (CCJT or CCMT) coexist in the same network. Based on this model, we introduce a user-centric probabilistic caching placement where only the users within a caching radius of the requesting user can serve as helpers, and optimize the caching placement to maximize the *cache hit probability* [7, 8], defined as the probability to find the requested file in the local cache.
- **Performance derivation**. We first derive the success probabilities (or, equivalently, the SIR distributions) and the per-user capacity for the three (non-hybrid) caching strategies: SPC, CCJT, and CCMT. Then, we consider hybrid caching, where the node in each single-point cluster of the GPP is the SPC helper while the paired nodes in each two-point cluster of the GPP are the cooperating helpers, and evaluate the performance of two hybrid caching schemes in terms of an offloading metric and the distribution of the file retrieval delay.
- **Comparisons and design insights**. We first provide a detailed numerical study to compare the success probabilities and per-user capacities of different content delivery strategies. Then we investigate the benefits of hybrid caching in terms of the offloading gain and delay performance in a GPP-based network, which gives fundamental insights into the benefits of applying cooperation

techniques in wireless device caching in large networks, where the interference from all transmitting nodes and the limited battery capacity are properly accounted for.

II. System Model

In this section, we first introduce the components as well as the assumptions of the network and describe how each component is modeled using stochastic geometry. Then we describe the models for the two phases of cache placement and content delivery.

A. Network Model

We consider a hybrid cache-enabled D2D communication network with two types of device caching: non-cooperative caching (i.e., SPC) and cooperative caching (i.e., CCJT or CCMT). The devices that can perform proactive caching and provide content delivery are called *content providers* while the ones requesting files are called *content clients*. We model the locations of the content providers as a GPP on $\mathbb{R}^2$, which is denoted by Φ and defined as follows.

Definition 1. *(Gauss-Poisson process, GPP [16, Example 3.8]). The planar GPP is a Poisson cluster process on $\mathbb{R}^2$ where λ_p denotes the intensity of the parent process and each cluster has one or two points, with probabilities p and $1-p$, respectively. If a cluster consists of one point, that point is at the parent's location. If it has two points, one is at the parent's position, and the other is uniformly distributed on the circle with radius u centered at the parent.*

According to this definition, the density of Φ, denoted by λ, is $\lambda = \lambda_\mathrm{p}(2-p)$, and we have $\Phi = \Phi^{(1)} \bigcup \Phi^{(2)}$, where $\Phi^{(1)}$ and $\Phi^{(2)}$ are the unions of the one-point clusters and the two-point clusters in Φ, respectively. Note that in this GPP-based network model, the one-point clusters correspond to non-cooperative caching, while the two-point clusters correspond to cooperative caching. A single point representing a content client is placed at the origin. We call it the typical client, because upon expectation w.r.t. the GPP, this client becomes the typical client for any stationary point process model of clients.

There is also a tier of base stations (BSs) in the network, which are connected to the core network via backhaul links and communicate with the content clients only when their requests cannot be satisfied by D2D links. The locations of the BSs follow a homogeneous PPP Φ_b with density λ_b independent of Φ. We assume that the BSs and content providers operate over non-overlapping frequency bands to avoid cross-tier interference. We adopt a power path loss law $\ell(x) = \|x\|^{-\alpha}$, where $x \in \mathbb{R}^2$ and $\alpha > 2$, and independent Rayleigh fading where the channel power gain, denoted by h_x, is exponentially distributed with unit mean. We set all transmit powers to unity and focus on the interference-limited regime, thus omitting the thermal noise.

B. Content Popularity and Caching Model

We consider a static content catalog consisting of N_f files, denoted by $\mathcal{W} \triangleq \{W_1, W_2, \ldots, W_{N_\mathrm{f}}\}$, which are indexed in descending order of popularity. Each file has the same size of F bits. The popularity of W_k, denoted as $p_\mathrm{r}(k)$, for $k \in [N_\mathrm{f}] \triangleq \{1, 2, \ldots, N_\mathrm{f}\}$, follows the Zipf distribution

$$p_\mathrm{r}(k) = \frac{k^{-\gamma}}{\sum_{i=1}^{N_\mathrm{f}} i^{-\gamma}}, \tag{1}$$

where $\sum_{k=1}^{N_\mathrm{f}} p_\mathrm{r}(k) = 1$, and $\gamma > 0$ is the Zipf exponent that determines the skewness of the distribution. The larger γ, the fewer files that are responsible for the majority of requests [17].

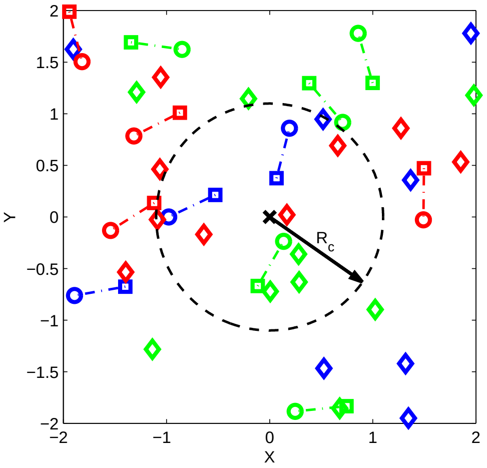


Fig. 1. A realization of the GPP with $\lambda_\mathrm{p} = 3$, $p = 0.5$, $u = 0.5$, $R_\mathrm{c} = 1.1$, $N_\mathrm{f} = 3$ and $\boldsymbol{p}_\mathrm{c} = (0.5, 0.3, 0.2)$, where points in green, red and blue constitute 1st, 2nd and 3th D2D caching tier, respectively; '◇' represents content providers belonging to one-point clusters while '○' and '□' represent cooperative content providers belonging to two-point clusters with a dashed-dot line indicating their paired relationship; and the dashed circle is the caching region of the typical client at the origin denoted by '×'.

We consider a cluster-based caching strategy, i.e., the cache placement and content delivery decisions are performed by clusters, and each cluster is considered as a *caching entity*. To be specific, for a one-point cluster, the caching entity refers to that point; while for a two-point cluster, once a content is assigned, both points in this entity store either the same content or different partitions of the same content depending on the transmission (delivery) scheme adopted[1]. Considering the relatively short communication distance and the limited battery capacity, we consider a client-centric protocol, i.e., a content provider will send a cached file to the client only if their distance is smaller than a given value, called *caching radius*, denoted by R_c. Thus, clusters with all their points located within the caching radius of a content client are potential caching entities and called *adjacent clusters*.

Furthermore, we adopt a probabilistic caching model, where a file is independently cached at different clusters according to a *caching distribution*, which is assumed the same at all clusters. The caching distribution is defined as $\boldsymbol{p}_\mathrm{c} \triangleq \big(p_\mathrm{c}(1), p_\mathrm{c}(2), \ldots, p_\mathrm{c}(N_\mathrm{f})\big)$, where $p_\mathrm{c}(k)$, $k \in [N_\mathrm{f}]$ is the probability that the k-th file is cached at a cluster. Each caching

[1] This may involve a quick handshake between the points in the same caching entity.

entity is assumed to store one file[2] in their local cache, as in [10, 19]. All clusters that have cached the k-th file constitute a tier, called the k-th D2D caching tier. An example for the D2D content distribution network with $N_\mathrm{f} = 3$ is illustrated in Fig. 1. In the next section, we will optimize this probabilistic caching policy with known demand statistics and find the optimal caching distribution.

C. Content Request and Delivery Model

A content client requests a file with a probability according to the file popularity distribution, i.e., the Zipf distribution. If it can find the requested file in the local caches of its adjacent clusters, this request will be fulfilled by a uniformly randomly chosen adjacent cluster via D2D links. Otherwise, the client will be served by its nearest BS, which is assumed to access all files through its backhaul link. Our analysis focuses on the typical client located at the origin.

In the content delivery phase, all content providers are assumed to be the active transmitters and the typical client becomes the typical receiver. Conditioning on there being at least one adjacent cluster, two cases that may occur for the established D2D links are considered:

1) Non-cooperative Case: a transmitter x_0 of a one-point adjacent cluster sends the complete file by itself, which is called *single-point caching* (SPC), with the SIR at the receiver (client) given by

$$\mathrm{SIR} = \frac{h_0 \ell(x_0)}{\sum_{x\in\Phi\setminus\{x_0\}} h_x \ell(x)}, \tag{2}$$

where $I_1 = \sum_{x\in\Phi\setminus\{x_0\}} h_x\ell(x)$ is the total interference power from other transmitters.

2) Cooperative Case: two transmitters x_1, x_2 of a two-point adjacent cluster Φ_0 cooperatively fulfill the complete file transmission. The cooperative case is further divided into two sub-cases according to the transmission scheme adopted:

- **Cooperative Caching with Joint Transmission (CCJT)**: a non-coherent joint transmission scheme is adopted, with the SIR at the receiver given by

$$\mathrm{SIR} = \frac{\sum_{x\in\Phi_0} h_x\ell(x)}{\sum_{x\in\Phi\setminus\Phi_0} h_x\ell(x)}, \tag{3}$$

  where $I_2 = \sum_{x\in\Phi\setminus\Phi_0} h_x\ell(x)$ is the total interference power from other transmitters.
- **Cooperative Caching with Multi-stream Transmission (CCMT)**: a parallel transmission scheme is adopted, i.e., two transmitters deliver different parts of the file concurrently over the same resource block, similar to the caching scheme in [20]. At the typical receiver, the successive interference cancellation (SIC) technique is adopted to decode the signals from two transmitters successively in the descending order of the average received signal strength as in [21]. Specifically, suppose that the elements of Φ_0 are indexed as $x^{(1)}$ and $x^{(2)}$ with $\ell(x^{(1)}) > \ell(x^{(2)})$, the signal from $x^{(1)}$ is decoded first, and if it is decoded successfully, it will be subtracted from the received signal, then the residual received signal (the signal from $x^{(2)}$) is decoded. The SIR expressions for the two data streams at the receiver for this caching strategy are given by

$$\mathrm{SIR}_1 = \frac{h_1\ell(x^{(1)})}{h_2\ell(x^{(2)}) + I_2}, \quad \mathrm{SIR}_2 = \frac{h_2\ell(x^{(2)})}{I_2}. \tag{4}$$

III. Optimal Caching Policy

In this section, we derive the caching distribution $\boldsymbol{p}_\mathrm{c}$ that maximizes the *cache hit probability*, defined as the probability that the desired file of a client can be found in the local caches of its adjacent clusters. As is described in Sec. II-B, due to the limited communication distance, only if all the points of the cluster are located within $b(o, R_\mathrm{c})$, they can provide the content for the client at o. Under such constraint, the optimal caching distribution is given by the following theorem. We use the notation $[x]^+ \triangleq \max\{0, x\}$ for $x \in \mathbb{R}$.

Theorem 1. *Let*

$$\chi \triangleq \frac{2}{\pi R_\mathrm{c}^2}\int_{\min\{|R_\mathrm{c}-u|,R_\mathrm{c}\}}^{R_\mathrm{c}} r \arccos\frac{r^2+u^2-R_\mathrm{c}^2}{2ru}\mathrm{d}r. \tag{5}$$

The cache hit probability of hybrid caching in the GPP-based D2D network is

$$p_\mathrm{hit} = \sum_{k\in[N_\mathrm{f}]} p_\mathrm{r}(k)\Big(1 - e^{-\lambda_\mathrm{p} p_\mathrm{c}(k)(p+(1-p)\epsilon)\pi R_\mathrm{c}^2}\Big), \tag{6}$$

where $\epsilon = \chi + \big([R_\mathrm{c}-u]^+\big)^2/R_\mathrm{c}^2$ *and the optimal caching distribution that maximizes* p_hit *is obtained from the waterfilling policy, given as*

$$p_\mathrm{c}^*(k) = \frac{1}{\lambda_\mathrm{p}(p+(1-p)\epsilon)\pi R_\mathrm{c}^2}\Big[\nu - \log(1/p_\mathrm{r}(k))\Big]^+, \tag{7}$$

where ν *is chosen such that* $\sum_{k\in[N_\mathrm{f}]} p_\mathrm{c}^*(k) = 1$.

Proof: See Appendix A.

Note that (6) and (7) are closed-form results since χ in (5) can be expressed in closed (but rather unwieldy) form.

When $p = 1$, the GPP reduces to a PPP, and the following corollary gives a consistent result with that in [10, Eq. (4)].

Corollary 1. *The cache hit probability in a PPP-based D2D network is*

$$p_\mathrm{hit} = \sum_{k\in[N_\mathrm{f}]} p_\mathrm{r}(k)\Big(1 - e^{-\lambda_\mathrm{p} p_\mathrm{c}(k)\pi R_\mathrm{c}^2}\Big), \tag{8}$$

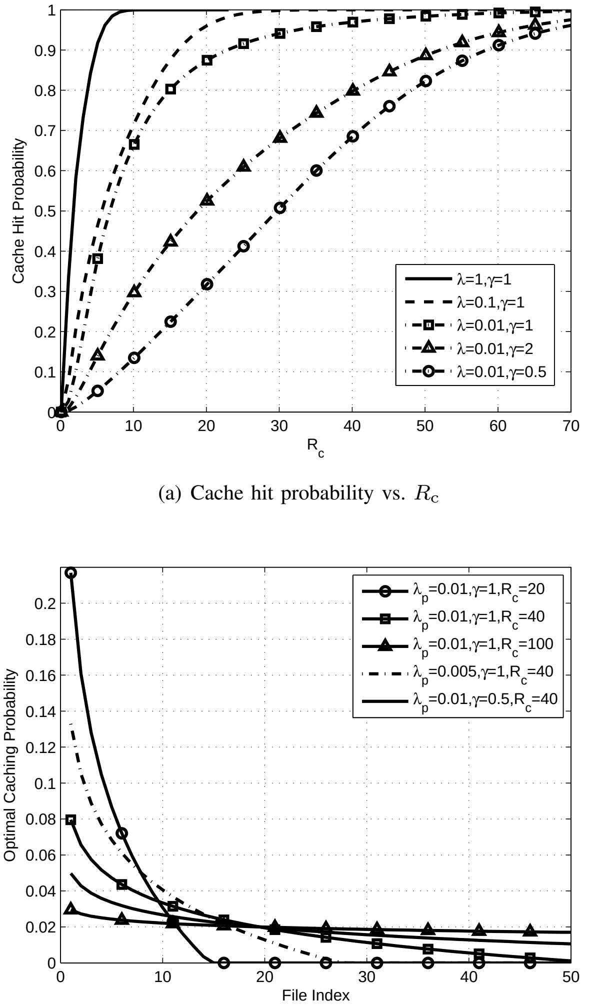
and the waterfilling policy yields the optimal caching distribution, given as

$$p_\mathrm{c}^*(k) = \frac{1}{\lambda_\mathrm{p}\pi R_\mathrm{c}^2}\Big[\nu - \log(1/p_\mathrm{r}(k))\Big]^+, \tag{9}$$

where ν *is chosen such that* $\sum_{k\in[N_\mathrm{f}]} p_\mathrm{c}^*(k) = 1$.

Corollary 2. *When* $\frac{\lambda_\mathrm{p}\pi R_\mathrm{c}^2}{\frac{\gamma}{p}\big(2N_\mathrm{f}\log(N_\mathrm{f}) - \log(N_\mathrm{f}!)\big)} \to \infty$, *the optimal caching distribution tends to the uniform distribution, i.e.,* $p_\mathrm{c}^*(k) \to \frac{1}{N_\mathrm{f}}$, $k \in [N_\mathrm{f}]$.

[2] The generalization to the storage of multiple files is possible following similar steps as in [18].



(a) Cache hit probability vs. R_c

(b) Optimal caching distribution vs. file index

Fig. 2. Cache hit probability and optimal caching distribution for $p = 0.5$, $u = 1$, $N_\mathrm{f} = 50$.

Proof: According to the waterfilling policy, the values $\log(1/p_\mathrm{r}(k))$ are regarded as the bottoms of a vessel for different files. If $\zeta = \lambda_\mathrm{p}(p + (1-p)\epsilon)\pi R_\mathrm{c}^2$ units of water are poured into the vessel, the caching probability of file k is the depth of the water normalized by ζ, and ν is the height of the water surface. When ζ increases to the level where all bottoms are below the water surface, the maximum depth gap between the files is $\log(p_\mathrm{r}(1)/p_\mathrm{r}(N_\mathrm{f}))$. A sufficiently large ζ (dominated by $\lambda_\mathrm{p}\pi R_\mathrm{c}^2$) makes the depth gap negligible and thus the caching distribution tends to the uniform distribution, which follows from

$$\frac{\zeta - \sum_{k\in[K]} \log(p_\mathrm{r}(k)/p_\mathrm{r}(N_\mathrm{f}))}{N_\mathrm{f}\log(p_\mathrm{r}(1)/p_\mathrm{r}(N_\mathrm{f}))} \to \infty. \tag{10}$$

Substituting the expressions of $p_\mathrm{r}(k)$, we have $\frac{\lambda_\mathrm{p}\pi R_\mathrm{c}^2}{\frac{\gamma}{p}\left(2N_\mathrm{f}\log(N_\mathrm{f}) - \log(N_\mathrm{f}!)\right)} \to \infty$, which indicates that the optimal caching distribution tends to the uniform distribution when the mean number of clusters in the caching region grows large. ■

Fig. 2(a) shows the cache hit probability versus the caching radius for different λ_p and γ. With the increase of λ_p, R_c and γ, the cache hit probability increases, which indicates that: (1) as the number of content providers in the networks grows, the possibility that a requested file can be found in the local cache grows accordingly. This point highlights that content caching is especially suitable to D2D networks where the number of content providers is inherently concentrated in regions with large demand; (2) as the content popularity distribution becomes more skewed (i.e., a larger value of γ), fewer popular contents constitute a majority of the content requests, which also makes the possibility of device caching larger. This is perfectly consistent with the original intention of introducing local caching to exploit the inherent mass reuse of a few popular contents while coping with the asynchronous requests.

Fig. 2(b) investigates the optimal caching distribution for different R_c, γ and λ_p. With the decrease of R_c and λ_p or increase of γ, the probability of caching popular files increases, which makes the distribution more concentrated in the fewer popular files. Intuitively, when the number of content providers within the local cache area decreases, it is efficient to let more providers cache the more popular files, and vice versa. When R_c is large enough, see the curve of $R_\mathrm{c} = 100$, the caching distribution is quite close to the uniform distribution. In this regime, the skewness of the popularity distribution has little effect on the cache placement since the number of providers is large enough to pre-fetch each file in the content catalog with almost the same possibility.

In Fig. 3, we compare the optimal caching policy in Theorem 1 with the *popularity-based caching policy* (i.e., each provider stores a file from the catalog according to the content popularity distribution) and the *uniform caching policy* (i.e., each provider stores a file from the catalog uniformly), which are commonly used in previous works, for different N_f. It is obvious that the optimal caching policy has the highest cache hit probability, i.e., the optimal caching policy has more opportunity to offload traffic from cellular networks. In addition, for small R_c, i.e., $R_\mathrm{c} \leq 30$, the gap between the optimal caching policy and the uniform caching policy is obviously larger than that between the optimal caching policy and the popularity-based caching policy. As R_c increases, the cache hit probability of the uniform caching policy tends to that of the optimal caching policy at a faster rate than that of the popularity-based caching policy. This is consistent with the case of $R_\mathrm{c} = 100$ in Fig. 2(b), where the optimal caching distribution reduces to a uniform distribution.

IV. Performance Analysis

A. Success Probability

We assume that the receiver can decode successfully if its SIR exceeds a threshold θ, and the success probability is defined as the probability of a transmission that is successfully decoded. In this section, we derive the complementary cumulative distribution function (CCDF) of the SIR (or, equivalently, the link success probability) at the receiver for SPC, CCJT, and CCMT, respectively. For simplicity, considering a point x of a two-point cluster, we

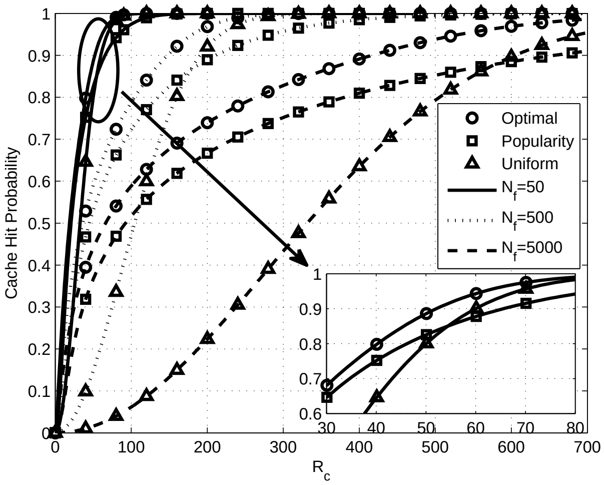


Fig. 3. Comparison of different caching policies in terms of cache hit probability for $\lambda_{\mathrm{p}} = 0.01$, $\gamma = 1$, $p = 0.5$ $u = 1$.

let $c(\|x\|, \Psi) \triangleq \sqrt{\|x\|^2 + u^2 + 2\|x\|u\cos\Psi}$ be the distance between its cooperator and the origin, where Ψ is the angle between the ray from x to its cooperator and the radial direction of x.

1) SPC: In this case, the typical receiver attempts to receive the caching file from a one-point adjacent cluster with transmitter x_0. Thus the link success probability is

$$P_{\mathrm{SPC}}(\theta) = \mathbb{P}\left(\frac{h_0\|x_0\|^{-\alpha}}{I_1} \geq \theta\right). \tag{11}$$

The following theorem gives the success probability of SPC for the typical receiver.

Theorem 2. *Conditioned on $x_0 \in b(o, R_{\mathrm{c}})$ being a point in a one-point cluster of Φ, the link success probability of SPC is*

$$P_{\mathrm{SPC}}(\theta) = \frac{1}{R_{\mathrm{c}}^2}\int_0^{R_{\mathrm{c}}^2} \mathcal{L}_{I_1}(\theta t^{\alpha/2})\mathrm{d}t, \tag{12}$$

where

$$\mathcal{L}_{I_1}(s) = \exp\Bigg(2\pi\lambda_{\mathrm{p}}\int_0^{\infty}\bigg(\frac{p}{1+sr^{-\alpha}} + \frac{1-p}{1+sr^{-\alpha}} \times \int_0^{\pi}\frac{1}{\pi}\frac{\mathrm{d}\psi}{1+s(c(r,\psi))^{-\alpha}} - 1\bigg) r\mathrm{d}r\Bigg). \tag{13}$$

Proof: We have

$$P_{\mathrm{SPC}}(\theta) = \mathbb{P}\bigg(\frac{h_0\|x_0\|^{-\alpha}}{I_1} > \theta\bigg) \overset{(a)}{=} \mathbb{E}_{\|x_0\|}\Big(\mathcal{L}_{I_1}(\theta\|x_0\|^{\alpha})\Big), \tag{14}$$

where (a) follows with Rayleigh fading of h_0 and expectation over $\|x_0\|$. From Slivnyak's theorem [16, Thm. 8.10], conditioning on x_0 does not change the distribution of the other clusters, and the distribution of the points excluding x_0 remains the same as in the original GPP Φ. Therefore, the Laplace transform of I_1 is derived as

$$\mathcal{L}_{I_1}(s) = \mathbb{E}\Bigg(\prod_{x\in\Phi}\frac{1}{1+s\|x\|^{-\alpha}}\Bigg)$$

$$\begin{aligned}
&\overset{(b)}{=} \exp\Bigg(\lambda_{\mathrm{p}}\int_{\mathbb{R}^2}\bigg(\mathbb{E}\prod_{y\in\Phi_x}\frac{1}{1+s\|y\|^{-\alpha}} - 1\bigg)\mathrm{d}x\Bigg)\\
&\overset{(c)}{=} \exp\Bigg(\lambda_{\mathrm{p}}\int_{\mathbb{R}^2}\bigg(\frac{p}{1+s\|x\|^{-\alpha}} + \frac{1-p}{1+s\|x\|^{-\alpha}}\\
&\qquad\times\mathbb{E}_{\Psi}\frac{1}{1+s(c(\|x\|,\Psi))^{-\alpha}} - 1\bigg)\mathrm{d}x\Bigg)\\
&= \exp\Bigg(2\pi\lambda_{\mathrm{p}}\int_0^{\infty}\bigg(\frac{p}{1+sr^{-\alpha}} + \frac{1-p}{1+sr^{-\alpha}}\\
&\qquad\times\int_0^{\pi}\frac{1}{\pi}\frac{\mathrm{d}\psi}{1+s(c(r,\psi))^{-\alpha}} - 1\bigg) r\mathrm{d}r\Bigg),
\end{aligned} \tag{15}$$

where $\{\Phi_x, x \in \Phi_{\mathrm{p}}\}$ are the clusters of Φ, (b) follows from the probability generating functional (PGFL) of Poisson cluster processes [16, Cor. 4.12] and (c) follows by the definition of the GPP. Since x_0 is uniformly distributed in $b(o, R_{\mathrm{c}})$, its probability density function (PDF) is $f_{\|x_0\|}(t) = \frac{2t}{R_{\mathrm{c}}^2}$. Ψ is uniformly distributed in $[0,\ 2\pi]$ according to the definition of the GPP. ■

2) CCJT: In this case, the typical receiver attempts to receive the caching file from a two-point adjacent cluster Φ_0 and the two transmitters x_1, x_2 deliver the same content jointly. Thus, the received power is the sum of the received signal powers from the two transmitters, and the link success probability is

$$P_{\mathrm{CCJT}}(\theta) = \mathbb{P}\left(\frac{h_1\ell(x_1) + h_2\ell(x_2)}{I_2} \geq \theta\right). \tag{16}$$

The following theorem gives the success probability of CCJT for the typical receiver.

Theorem 3. *Conditioned on Φ_0 being a two-point cluster of $\Phi \cap b(o, R_{\mathrm{c}})$, the link success probability of CCJT is*

$$\begin{aligned}
P_{\mathrm{CCJT}}(\theta) =& \frac{1}{R_{\mathrm{c}}^2}\int_0^{R_{\mathrm{c}}}\frac{2t}{\pi-\varphi(t)}\int_{\varphi(t)}^{\pi}\Bigg(\frac{\mathcal{L}_{I_2}\big(\theta t^{\alpha}\big)}{1-(t/c(t,\psi))^{\alpha}}\\
&-\frac{\mathcal{L}_{I_2}\big(\theta(c(t,\psi))^{\alpha}\big)}{(c(t,\psi)/t)^{\alpha}-1}\Bigg)\mathrm{d}\psi\mathrm{d}t,
\end{aligned} \tag{17}$$

where $\mathcal{L}_{I_2}(s) = \mathcal{L}_{I_1}(s)$ and

$$\varphi(t) = \begin{cases}\pi - \arccos\frac{t^2+u^2-R_{\mathrm{c}}^2}{2tu}, & \text{for } t \geq |R_{\mathrm{c}} - u|\\ \pi - \pi[R_{\mathrm{c}}-u]^{+}/(R_{\mathrm{c}}- u), & \text{for } t < |R_{\mathrm{c}} - u|.\end{cases} \tag{18}$$

Proof: See Appendix B.

Compared with SPC, CCJT is expected to provide a higher success probability due to the summation of two desired signals, which, in turn, implies that the content providers in CCJT can save their energy by reducing their transmit power while keeping the success probability no less than that of SPC. This way, some providers whose remaining energy is insufficient for delivering the complete file can still contribute to the local caching through CCJT. The following corollary investigates the relationship between the permitted fraction of the transmit power and the link success probability.

Corollary 3. *If the transmitters in CCJT only use a fraction η of their transmit power, the link success probability $P_{\mathrm{CCJT}}(\eta,\theta)$ is an increasing function of η for an arbitrary fixed θ.*

Proof: When the transmitters in CCJT only use a fraction η of their transmit power, the SIR is expressed as

$$\begin{aligned}\mathrm{SIR}(\eta) &= \frac{\eta h_1\ell(x_1)+\eta h_2\ell(x_2)}{\sum\limits_{x\in\Phi^{(1)}} h_x\ell(x) + \sum\limits_{x\in\Phi^{(2)}} \eta h_x\ell(x)} \\ &= \frac{h_1\ell(x_1)+h_2\ell(x_2)}{\sum\limits_{x\in\Phi^{(1)}} h_x\ell(x)/\eta + \sum\limits_{x\in\Phi^{(2)}} h_x\ell(x)}. \end{aligned} \tag{19}$$

For $0 \le \eta_1 < \eta_2 \le 1$, we have $\mathrm{SIR}(\eta_1) < \mathrm{SIR}(\eta_2)$ and thus $\mathbb{P}(\mathrm{SIR}(\eta_1) \ge \theta) \le \mathbb{P}(\mathrm{SIR}(\eta_2) \ge \theta)$. As a result, the link success probability increases with η for an arbitrary fixed θ. ■

3) CCMT: In this case, the typical receiver attempts to receive the caching file from a two-point adjacent cluster Φ_0 and the two transmitters x_1, x_2 send different data streams to the receiver simultaneously. The link success probability is then defined as the probability that both streams are decoded successfully, i.e.,

$$P_{\mathrm{CCMT}}(\theta) = \mathbb{P}\left(\mathrm{SIR}_1 \ge \theta, \mathrm{SIR}_2 \ge \theta\right). \tag{20}$$

The following theorem gives the success probability of CCMT for the typical receiver.

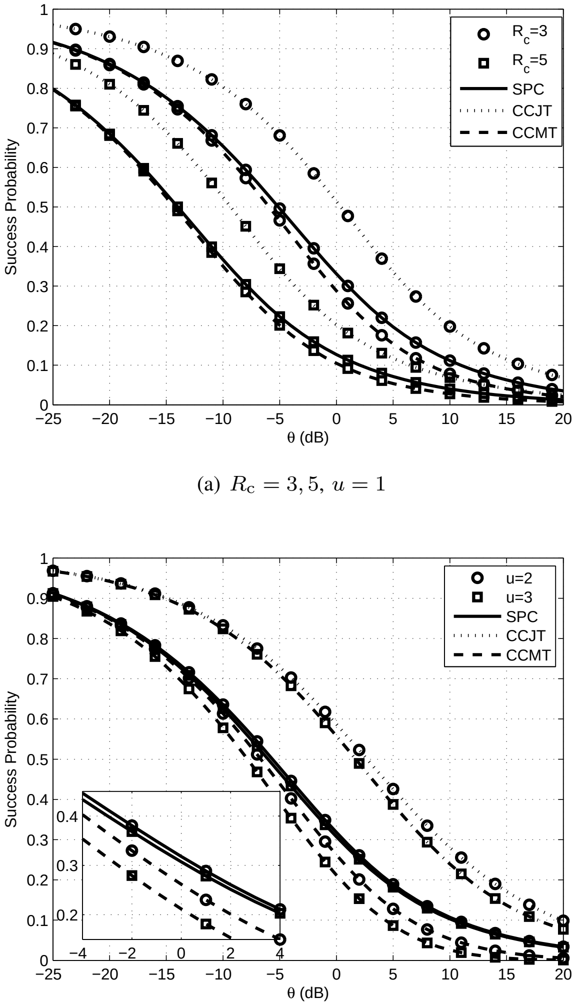
Theorem 4. *Conditioned on Φ_0 being a two-point cluster of $\Phi \cap b(o, R_{\mathrm{c}})$, the link success probability of CCMT using SIC is*

$$\begin{aligned}P_{\mathrm{CCMT}}(\theta) = &\frac{1}{R_{\mathrm{c}}^2}\int_0^{R_{\mathrm{c}}} \frac{2t}{\pi-\varphi(t)}\left(\int_{\varphi(t)}^{\vartheta(t)} \frac{\mathcal{L}_{I_2}\big(\theta(\theta+1)t^\alpha+\theta(c(t,\psi))^\alpha\big)}{1+\theta(t/c(t,\psi))^\alpha}\mathrm{d}\psi\right. \\ &\left.+\int_{\vartheta(t)}^{\pi} \frac{\mathcal{L}_{I_2}\big(\theta(\theta+1)(c(t,\psi))^\alpha\big)+\theta t^\alpha\big)}{1+\theta(c(t,\psi)/t)^\alpha}\mathrm{d}\psi\right)\mathrm{d}t, \end{aligned} \tag{21}$$

where $\vartheta(t) = \pi$ if $t \le u/2$, otherwise $\vartheta(t) = \pi - \arccos\frac{u}{2t}$.

Proof: See Appendix C.

Fig. 4 presents the success probabilities of SPC, CCJT and CCMT as a function of R_{c} (Fig. 4(a)), and u (Fig. 4(b)), respectively. Since CCMT delivers two data streams simultaneously, it would be unfair to compare the success probability with the same θ as SPC and CCJT. Given an SIR threshold θ for SPC and CCJT, the rate is (approximately) $R = \log(1+\theta)$, and the required rate for CCMT is only $R/2$ per stream, which means the SIR threshold for CCMT should be $\tilde{\theta} = \sqrt{1+\theta} - 1$. From Fig. 4(a), we observe that CCJT always achieves a higher success probability than SPC and CCMT, demonstrating the benefit of CCJT in terms of higher transmission reliability. In addition, SPC and CCMT have quite similar success probabilities and SPC results in a slightly higher success probability than CCMT for a larger θ. This is because: (1) having two transmitters in CCMT causes inter-stream interference, which reduces the SIRs of the two streams; (2) in the CCMT case, the transmission efficiency is improved through parallel transmission, i.e., the transmission duration is reduced in half, at the expense of the reliability[3], and its performance relies heavily on the SIC receiver, which was shown to be especially beneficial if very low-rate codes are used [22].



(a) $R_{\mathrm{c}} = 3, 5,\ u = 1$

(b) $R_{\mathrm{c}} = 3,\ u = 2, 3$

Fig. 4. The success probabilities of SPC, CCJT, and CCMT for $\lambda_{\mathrm{P}} = 0.05$, $p = 0.5$, $\alpha = 4$. (For CCMT, $\tilde{\theta} = \sqrt{1+\theta} - 1$ is used.)

Fig. 4(b) shows how the distance u affects the success probability for SPC, CCJT, and CCMT. Since u is the distance between the two cooperative transmitters, it has a more pronounced effect on the performance of CCJT and CCMT than that of SPC. A larger u means one of the two transmitters is more likely to be far from the receiver, and the desired signal is attenuated accordingly.

B. Per-User (Per-Client) Capacity

In addition to the transmission reliability, the per-user (per-client) capacity is another important performance metric since

[3] Intuitively, although the potential benefit of coded caching is strongly limited by the reduced transmission reliability when a high transmission rate is required, the performance can be improved significantly if a proper interference mitigation technique is adopted, which is left for the future work.

it directly reflects the transmission efficiency. Under a fixed-rate transmission based on the SIR threshold, the per-user capacities of SPC and CCJT are given as

$$\begin{aligned}\tau_{\mathrm{SPC}} &= \kappa_1 \log(1+\theta)P_{\mathrm{SPC}}(\theta),\\ \tau_{\mathrm{CCJT}} &= \kappa_2 \log(1+\theta)P_{\mathrm{CCJT}}(\theta),\end{aligned} \tag{22}$$

where $\kappa_1 = 1 - e^{-\lambda_{\mathrm{p}} p \pi R_{\mathrm{c}}^2}$ and $\kappa_2 = 1 - e^{-\lambda_{\mathrm{p}}(1-p)\epsilon\pi R_{\mathrm{c}}^2}$ are the probabilities that there is at least one adjacent cluster for the content delivery.

In the CCMT case, the per-user capacity depends on the number N of data streams successfully decoded by the receiver. Accordingly, the per-user capacity in this case is defined as the total information rate received at the receiver [22], expressed by

$$\tau_{\mathrm{CCMT}} = \kappa_2 \log(1+\theta)\mathbb{E}[N], \tag{23}$$

where $\mathbb{E}[N]$ is the mean number of successively decoded streams. The following corollary gives an exact expression of $\mathbb{E}[N]$ and the corresponding per-user capacity.

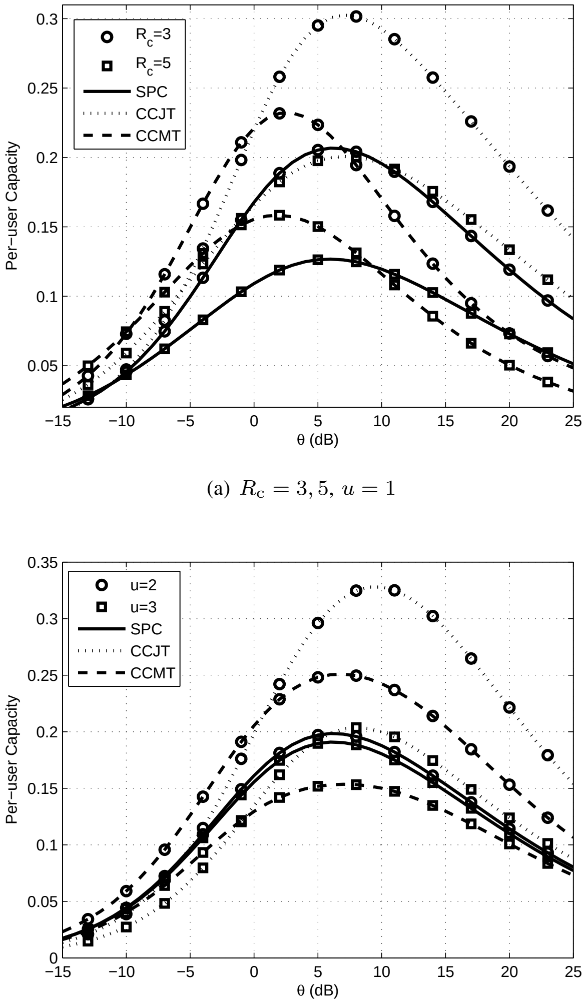
Corollary 4. *The mean number of successfully decoded streams with CCMT is*

$$\begin{aligned}\bar{N}_{\mathrm{CCMT}}(\theta) = &\frac{1}{R_{\mathrm{c}}^2}\int_0^{R_{\mathrm{c}}} \frac{2t}{\pi - \varphi(t)} \left(\int_{\varphi(t)}^{\vartheta(t)} \frac{1}{1+\theta(t/c(t,\psi))^{\alpha}} \Big[\mathcal{L}_{I_2}\big(\theta t^{\alpha}\big)\right.\\ &+\mathcal{L}_{I_2}\big(\theta(\theta+1)t^{\alpha}+\theta(c(t,\psi))^{\alpha}\big)\Big]\mathrm{d}\psi\\ &+\int_{\vartheta(t)}^{\pi} \frac{1}{1+\theta(c(t,\psi)/t)^{\alpha}} \Big[\mathcal{L}_{I_2}\big(\theta(c(t,\psi)^{\alpha}\big)\\ &\left.+\mathcal{L}_{I_2}\big(\theta(\theta+1)((c(t,\psi))^{\alpha}+\theta t^{\alpha}\big)\Big]\mathrm{d}\psi \right)\mathrm{d}t,\end{aligned} \tag{24}$$

and thus the per-client capacity is $\tau_{\mathrm{CCMT}} = \kappa_2 \log(1+\theta)\bar{N}_{\mathrm{CCMT}}(\theta)$.

Proof: See Appendix D.

Fig. 5 compares the per-user capacity curves of SPC, CCJT, and CCMT as a function of R_{c} (Fig. 5(a)), and u (Fig. 5(b)). Since the per-user capacity necessarily tends to zero for both $\theta \to 0$ and $\theta \to \infty$, it assumes a maximum at a finite value of θ. Fig. 5(a) shows that for $R_{\mathrm{c}} = 3$, the per-user capacity of SPC, CCJT, and CCMT is maximized quite exactly at $\theta = 5$ dB, 8 dB, and 2 dB, respectively. As R_{c} increases, the maximum per-user capacity of each caching scheme decreases due to the attenuation of the desired signal. Moreover, as expected, we observe that the performance benefit in terms of higher transmission efficiency of CCMT lies in the regime where θ is relatively small (i.e., $\theta < 0$ dB for $R_{\mathrm{c}} = 3$) while the benefit of CCJT lies in the opposite regime where θ is relatively large (i.e., $\theta > 0$ dB for $R_{\mathrm{c}} = 3$). Since $p = 0.5$, we have $\kappa_1 \approx \kappa_2$ from the parameter setting in Fig. 5(a), and under this condition, irrespective of θ, there is always a cooperation scheme that achieves higher per-user capacity than the non-cooperative transmission, thus highlighting the significant benefit of cooperation in device caching.



(a) $R_{\mathrm{c}} = 3, 5$, $u = 1$

(b) $R_{\mathrm{c}} = 3$, $u = 2, 3$

Fig. 5. The per-user capacity of SPC, CCJT, and CCMT for $\lambda_{\mathrm{p}} = 0.05$, $p = 0.5$, $\alpha = 4$.

Fig. 5(b) shows the impact of u on the per-user capacities of SPC, CCJT, and CCMT. It is seen that the per-user capacities of both CCJT and CCMT are more susceptible to the distance between the two cooperators than that of SPC. For all the three schemes, different u results in almost the same optimal value of θ but rather different maximum per-user capacities. As u increases, the number of adjacent two-point clusters decreases, reducing κ_2 significantly and, in turn, the per-user capacity.

Fig. 6 presents how the parameter p affects the per-user capacities of SPC, CCJT, and CCMT. It is seen that for both the two cooperative caching schemes: CCJT and CCMT, the per-user capacity is negatively correlated to p, while for the SPC, the situation is reversed. This indicates that κ_1 and κ_2 have obvious impacts on the per-user capacity. Since p represents the probability of a cluster having one point, as p increases, κ_1 increases while κ_2 decreases significantly, which means the more providers are available (either in cooperative mode or non-cooperative mode), the higher the per-user capacity. This figure reveals that both the non-cooperative and the cooperative strategies have their operating regimes where they achieve better performance. However, if we hope to have good performance in all cases, a hybrid caching strategy combining the non-cooperative scheme with the cooperative

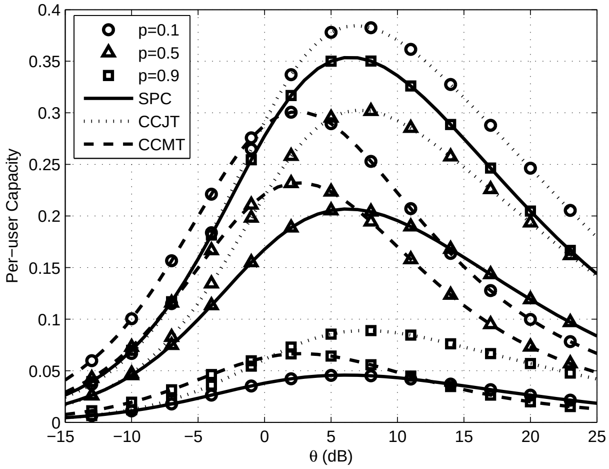


Fig. 6. The per-user capacity of SPC, CCJT, and CCMT with different p for $\lambda_{\mathrm{p}} = 0.05$, $u = 1$, $R_{\mathrm{c}} = 3$, $\alpha = 4$.

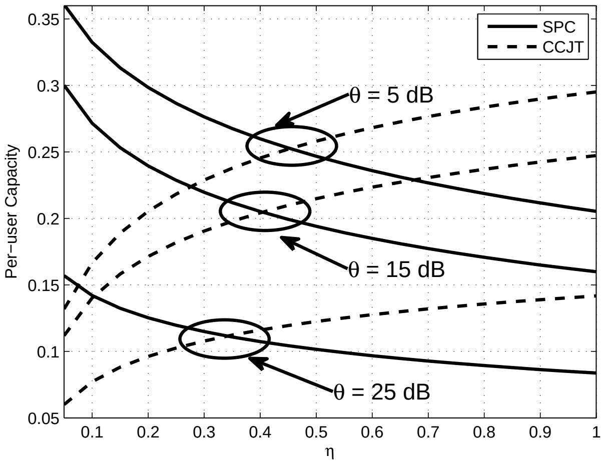


Fig. 7. The per-user capacity of CCJT vs. power ratio coefficient η for $\lambda_{\mathrm{p}} = 0.05$, $p = 0.5$, $u = 1$, $R_{\mathrm{c}} = 3$, $\alpha = 4$.

scheme should be exploited. Fig. 7 compares the per-user capacities of CCJT and SPC with different transmit power assignment for CCJT providers. An interesting observation in this figure is that for each θ, there is always an intersection point corresponding to a value of η, denoted by η_{m}. Then, CCJT achieves a higher per-user capacity than SPC as long as the CCJT provider permits at least η_{m} of its maximum transmit power. It is shown that for the given three values of θ, η_{m} is less than 0.5, which means that a CCJT provider uses less than half of the energy consumed by a SPC provider for the same per-user capacity. In other words, once the available energy of a provider is larger than the minimum transmit energy requirement, this provider can still transmit the file jointly with a cooperator. Thus, by cooperation in the content delivery phase, the energy consumption of each transmitter can be significantly reduced while maintaining the same performance as in the traditional non-cooperative case.

Intuitively, the significant benefit of cooperation plus proper reciprocity and incentives from operators will definitely encourage more users to participate in the device caching and thus offload more traffic from the congested cellular networks, providing a much better user experience. Motivated by this, we propose two novel hybrid caching policies for cache-enabled D2D networks, i.e., the combination of SPC and CCJT as well as that of SPC and CCMT, where SPC is modeling those providers who have enough battery energy to complete the content delivery individually while CCJT or CCMT is modeling providers who have limited energy and may not be willing to deliver the file by themselves. Two key performance indicators of local caching, i.e., offloading gain and delay, of the proposed hybrid caching policies are analyzed in the following.

C. Offloading Gain

In this subsection, we measure the offloading gain of the hybrid caching schemes, which is defined as the average fraction of each file that can be successfully delivered by the cache-enabled D2D network and hence offloaded from the cellular network at a given target SIR threshold. Two hybrid caching strategies are considered, namely SPC-CCJT and SPC-CCMT. For comparison, we first consider a conventional strategy as a baseline where only SPC is available for content delivery and providers of the two-point clusters will not contribute to any offloaded traffic due to limited battery energy. Thus, in this case, the offloading gain is given by

$$\rho_{\mathrm{SPC}} = \sum_{k=1}^{N_{\mathrm{f}}} p_{\mathrm{r}}(k)\tilde{p}_{\mathrm{f}}(k)P_{\mathrm{SPC}}(\theta), \tag{25}$$

where $\tilde{p}_{\mathrm{f}}(k) = 1 - \exp\big(-\lambda_{\mathrm{p}} p_{\mathrm{c}}(k) p\pi R_{\mathrm{c}}^2\big)$ is the probability that there exists at least one adjacent one-point cluster storing file k.

For hybrid caching, points of $\Phi^{(1)}$ operate in SPC mode while those of $\Phi^{(2)}$ operate in CCJT or CCMT mode. From the proof of Theorem 1, when the typical receiver requests file k and can be served by its adjacent clusters, it is easily obtained that the file is transmitted by the single-point cluster with probability $w_1 = \frac{p}{p+(1-p)\epsilon}$ and cooperatively transmitted by the two-point cluster with probability $w_2 = \frac{(1-p)\epsilon}{p+(1-p)\epsilon}$. Thus, the offloading gain of SPC-CCJT is expressed as

$$\rho_{\mathrm{SPC-CCJT}} = \sum_{k=1}^{N_{\mathrm{f}}} p_{\mathrm{r}}(k) p_{\mathrm{f}}(k)\big(w_1 P_{\mathrm{SPC}}(\theta) + w_2 P_{\mathrm{CCJT}}(\theta)\big), \tag{26}$$

where $p_{\mathrm{f}}(k) = 1 - \exp\big(-\lambda_{\mathrm{p}} p_{\mathrm{c}}(k)(p + (1-p)\epsilon)\pi R_{\mathrm{c}}^2\big)$ is the probability that there exists at least one adjacent cluster storing file k. When CCMT is adopted, since the data stream from each transmitter only accounts for half of the file, the average fraction of the file that can be successfully delivered for such a caching strategy is $\bar{N}_{\mathrm{CCMT}}(\theta)/2$. Thus, the offloading gain of SPC-CCMT is given by

$$\rho_{\mathrm{SPC-CCMT}} = \sum_{k=1}^{N_{\mathrm{f}}} p_{\mathrm{r}}(k) p_{\mathrm{f}}(k)\big(w_1 P_{\mathrm{SPC}}(\theta) + w_2 \bar{N}_{\mathrm{CCMT}}(\theta)/2\big). \tag{27}$$

Fig. 8 illustrates the offloading gains of SPC-CCJT and SPC-CCMT, in comparison with the conventional method

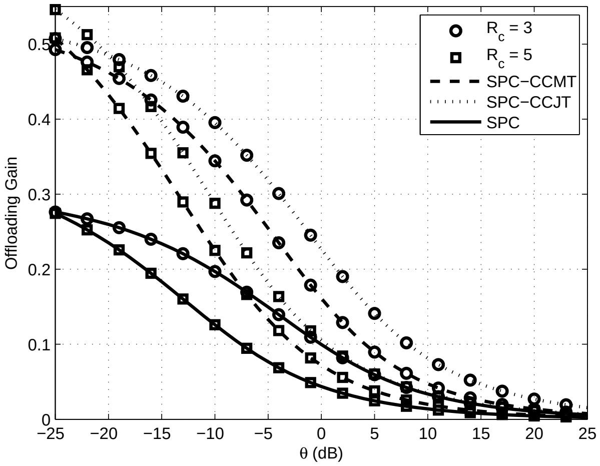


Fig. 8. Offloading gain versus θ for $\lambda_\mathrm{p}=0.05$, $p=0.5$, $u=1$, $\alpha=4$, $N_\mathrm{f}=10$, $\gamma=1$, $\eta=0.5$.

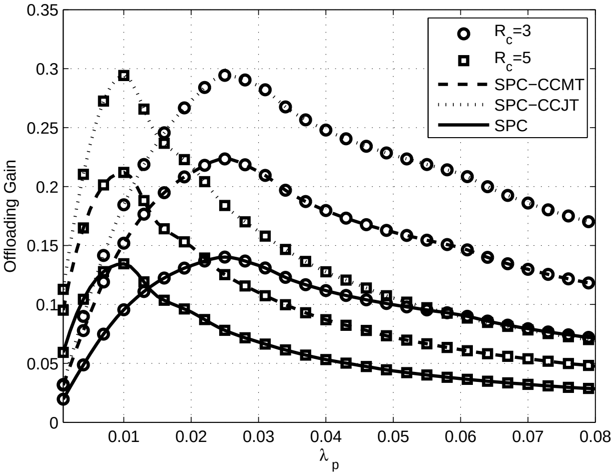


Fig. 9. Offloading gain versus λ_p for $p=0.5$, $u=1$, $\alpha=4$, $N_\mathrm{f}=10$, $\theta=0$ dB, $\gamma=1$, $\eta=0.5$.

with only SPC at different SIR requirements θ. As expected, we observe that both hybrid policies achieve a much higher offloading gain than conventional caching since in hybrid caching, more devices join the caching and make contributions to the traffic offloaded from cellular networks. The performance difference between the two hybrid caching policies lies in the volume of the traffic offloaded. However, from the definition of the offloading gain, this performance metric does not take the transmission time into account. Actually, since each stream in CCMT only accounts for half the size of the file, it is transmitted in half of the time that SPC or CCJT spends. Thus, from the perspective of transmission efficiency, CCMT is better; while from the perspective of transmission reliability, CCJT is better. The two cooperative schemes are complementary in performance. In addition, it is shown that the offloading performance decreases with the increase of R_c. Although the increase of R_c increases the probability of finding a cluster within $b(o,R_\mathrm{c})$, its direct effect on the desired signal strength is more prominent.

Fig. 9 shows the relationship between the offloading gain and the network density as a function of λ_p for different R_c. We observe that for the given system parameters, there is always a maximum offloading performance at some finite value of λ_p, highlighting the inherent trade-off between the cache hit probability and the success probability: on the one hand, increasing the density of providers will increase the probability to find the requested file in the local cache; on the other hand, increasing the density will cause more interference and hence reduce the success probability. Interestingly, changing R_c has a negligible impact on the maximum offloading gain but results in quite different node densities that maximize the offloading gain. This phenomenon indicates the inherent relationship between the optimal network density and the caching radius. Thus, considering the trade-off between the cache hit probability and the desired signal strength, the caching radius R_c should be set judiciously.

Fig. 10 plots the offloading performance of SPC-CCJT, SPC-CCMT, and only SPC as a function of the popularity exponent γ for different N_f. With the increase of γ or decrease of N_f, the offloading gain increases, which means the more skewed the popularity distribution, the better offloading performance can be achieved by device caching, validating the effectiveness of local caching to cope with the asynchronous but redundant pattern of content requests. Considering the limited storage capacity of devices, the cache-enabled D2D network is especially suitable for the case when most of the user requests are concentrated in few very popular files.

D. Delay Analysis

We characterize the delay performance of different caching strategies through the *content retrieval delay*, denoted by $\mathcal{D}$ and defined as the delay experienced by a client when retrieving a requested content from any available source. Specifically, if a client can find the requested file in the local caches of its adjacent clusters, this file request will be fulfilled via D2D links. Otherwise, this client will be served by its nearest BS in cellular networks, and the coverage probability is given in [23] via the Gaussian hypergeometric function ${}_2F_1$ as

$$P_\mathrm{cell}(\theta)=\frac{1}{{}_2F_1(1,-2/\alpha,1-2/\alpha,-\theta)}. \tag{28}$$

The transmission delay for the BS consists of the backhaul delay and the BS transmission delay because the BS needs to download the file from data servers first through the backhaul link. The backhaul delay D_B is defined as the file size divided by the data rate R_b of the backhaul link, i.e., $D_\mathrm{B}=F/R_\mathrm{b}$. From [24], $R_\mathrm{b}=\frac{C_1(1+p_\mathrm{hit})}{\lambda_\mathrm{b}}+C_2$, which is related to the density of BSs and associated traffic load. Given a delay requirement D for a requested file, the equivalence between the SIR and delay requirements is given as

$$\mathbb{P}(\mathcal{D}\le D)=\mathbb{P}(B\log(1+\mathrm{SIR})\ge F/D)=\mathbb{P}(\mathrm{SIR}\ge 2^{\frac{F}{BD}}-1), \tag{29}$$

where B is the system bandwidth. Assuming that the cellular bandwidth is equal to that of the D2D network [25, Chap. 9.2.1], the corresponding delay requirement for cellular

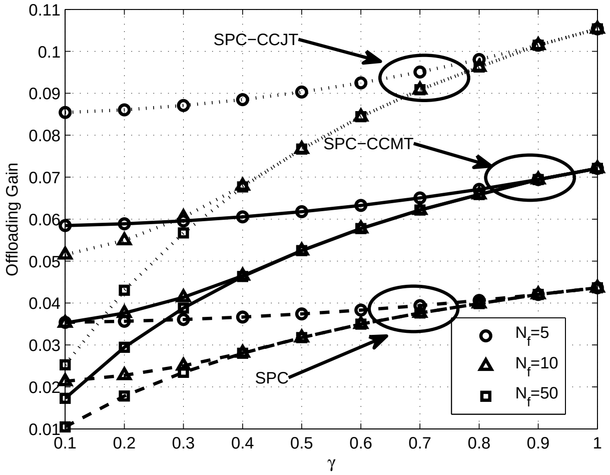


Fig. 10. Offloading gain versus γ for $\lambda_{\mathrm{p}} = 0.05$, $p = 0.5$, $u = 1$, $R_{\mathrm{c}} = 5$, $\theta = 0$ dB, $\alpha = 4$, $\eta = 0.5$.

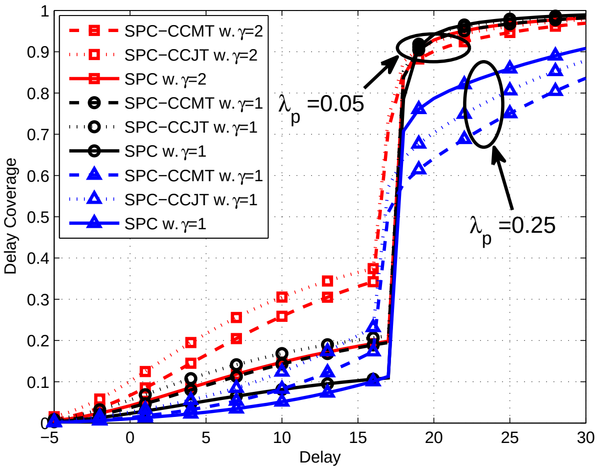


Fig. 11. Delay performance of SPC-CCJT, SPC-CCMT and SPC for $R_{\mathrm{c}} = 3$, $\alpha = 4$, $N_{\mathrm{f}} = 10$, $p = 0.5$, $u = 1$, $B = 10$MHz, $F = 25$Mb, $\lambda_{\mathrm{b}} = 10^{-4}$, $C_1 = 30$, $C_2 = 10^5$.

networks is modified to $D - D_{\mathrm{B}}$. Letting $\bar{\theta} = 2^{\frac{F}{BD}} - 1$, $\bar{\theta}_{\mathrm{cell}} = 2^{\frac{F}{B(D-D_{\mathrm{B}})}} - 1$, we have

$$\mathbb{P}_{\mathrm{SPC}}(\mathcal{D} < D) = \sum_{k=1}^{N_{\mathrm{f}}} p_{\mathrm{r}}(k) \Big[\tilde{p}_{\mathrm{f}}(k) P_{\mathrm{SPC}}(\bar{\theta}) + \big(1 - \tilde{p}_{\mathrm{f}}(k)\big) P_{\mathrm{cell}}(\bar{\theta}_{\mathrm{cell}}) \Big]. \tag{30}$$

Similarly, for SPC-CCJT, we have

$$\mathbb{P}_{\mathrm{SPC-CCJT}}(\mathcal{D} < D) = \sum_{k=1}^{N_{\mathrm{f}}} p_{\mathrm{r}}(k) \Big[p_{\mathrm{f}}(k) \big(w_1 P_{\mathrm{SPC}}(\bar{\theta}) + w_2 P_{\mathrm{CCJT}}(\bar{\theta}) \big) + (1 - p_{\mathrm{f}}(k)) P_{\mathrm{cell}}(\bar{\theta}_{\mathrm{cell}}) \Big]. \tag{31}$$

When CCMT is adopted, one data stream only accounts for half the file and the SIR threshold in this strategy is $\bar{\theta}_{\mathrm{MT}} = 2^{\frac{F}{2BD}} - 1$. Thus, we have

$$\mathbb{P}_{\mathrm{SPC-CCMT}}(\mathcal{D} < D) = \sum_{k=1}^{N_{\mathrm{f}}} p_{\mathrm{r}}(k) \Big[p_{\mathrm{f}}(k) \big(w_1 P_{\mathrm{SPC}}(\bar{\theta}) + w_2 P_{\mathrm{CCMT}}(\bar{\theta}_{\mathrm{MT}}) \big) + (1 - p_{\mathrm{f}}(k)) P_{\mathrm{cell}}(\bar{\theta}_{\mathrm{cell}}) \Big]. \tag{32}$$

In Fig. 11, the two hybrid caching policies are compared in terms of the delay performance, where the conventional caching policy only with SPC serves as the baseline. It is observed that both hybrid caching policies outperform the baseline for lower delay requirements, i.e., $D < 16$ dB in this figure, which demonstrates the benefits of the proposed caching policies for delay-sensitive services. The reason why SPC-CCMT achieves a lower performance than SPC-CCJT is that the inter-stream interference causes lower success probability even with smaller SIR requirement, and the CDF of the delay follows directly from the success probability which requires both transmission streams to satisfy the SIR requirement. In addition, the singularity of the curves is due to the backhaul delay from the cellular network, which can be interpreted as follows. When the delay requirement is less than the backhaul delay, the delay performance is dominated by the transmission delay of the local caching since most of the BS transmissions can not satisfy the delay requirement. Conversely, when the required delay is larger than the backhaul delay, see $D > 17$ dB, the delay performance curves of the three caching strategies tend to coincide with each other for $\lambda_{\mathrm{p}} = 0.05$, since at this time, the cellular transmission assumes the dominant role. As for the case of $\lambda_{\mathrm{p}} = 0.25$, the performance of the two hybrid caching schemes for large delay requirement is worse than that of the SPC as a result of two aspects: on the one hand, a large node density causes severe interference and hence leads to very poor D2D link performance; on the other hand, since in the SPC case, the cache hit probability is lower than that in the hybrid case, most content requests are actually served by the cellular network which provides better performance than D2D links.

V. Conclusions

In this paper, we proposed a hybrid device caching strategy that combines the conventional single-device caching with two-device cooperative caching, where the latter is designed for users unwilling to help transmit the complete file by themselves due to limited battery energy. To quantify the performance of the hybrid caching strategy, we adopted the GPP as a model for the cache-enabled D2D networks where caching strategies with and without cooperation coexist and provided a general and tractable framework for the performance analysis.

We first considered a probabilistic caching placement and derived the optimal caching distribution that maximizes the cache hit probability. It turns out that there are close connections between the cache hit probability and parameters such as the node density, caching radius, user request statistics and the number of files in the catalog. Furthermore, using this framework, we derived the success probability and per-user capacity of SPC, CCJT, and CCMT. The results indicate that CCJT is superior to SPC and CCMT in terms of higher transmission reliability (i.e., the success probability) while

CCMT is superior to SPC and CCJT in terms of higher transmission efficiency (i.e., the per-user capacity) and particularly beneficial in the low-SIR regime. These results were then applied to quantify the offloading gain and delay performance of two hybrid caching schemes: SPC-CCJT and SPC-CCMT. It is shown that both performance metrics can be improved significantly compared with the conventional non-cooperative caching strategy, since cooperation reduces the energy cost of each content provider and thus effectively copes with the limited user allowed battery consumption in practice.

In summary, the hybrid caching strategy is expected to bring substantial benefits for future wireless edge caching. However, since SPC-CCMT with SIC is vulnerable to the harsh interference environment, its reliability can hardly be guaranteed despite the high complexity. Thus, SPC-CCJT is preferred, due to its higher transmission reliability and lower system complexity.

Appendix A
Proof of Theorem 1

Proof: Since each cluster caches the same file independently, the locations of content providers caching the k-th file constitute a GPP Φ_k where the intensity of its parent point process $\Phi_{\mathrm{p},k}$ is $\lambda_k = \lambda_{\mathrm{p}} p_{\mathrm{c}}(k)$. According to Def. 1, each cluster has one or two points independently, thus we have $\Phi_k = \Phi_k^{(1)} \cup \Phi_k^{(2)}$, where $\Phi_k^{(1)}$ and $\Phi_k^{(2)}$ represent the point sets composed by the one-point clusters and two-point clusters in Φ_k, respectively. Let $b(o, R_{\mathrm{c}})$ be the caching region for the typical client, and $\#\mathcal{A}$ denotes the cardinality of the set $\mathcal{A}$. Further, $\mathcal{A}_k = \{\Phi_k^{(1)} \cap b(o, R_{\mathrm{c}}) = \emptyset\}$ denotes the event that no point of $\Phi_k^{(1)}$ is located in $b(o, R_{\mathrm{c}})$, and $\mathcal{B}_k = \{\#\{\Phi_x \cap b(o, R_{\mathrm{c}})\} \leq 1 : \forall \Phi_x \subseteq \Phi_k^{(2)}\}$ denotes the event that at most one point in each cluster of $\Phi_k^{(2)}$ within $b(o, R_{\mathrm{c}})$. Therefore, the cache hit probability of requesting the k-th file is $1 - \mathbb{P}(\mathcal{A}_k)\mathbb{P}(\mathcal{B}_k)$. Since $\Phi_k^{(1)}$ is a homogeneous PPP with density $\lambda_k p$, it is easy to see that $\mathbb{P}(\mathcal{A}_k) = \exp(-\lambda_k p \pi R_{\mathrm{c}}^2)$. It is not straightforward to obtain the expression of $\mathbb{P}(\mathcal{B}_k)$, however, $\mathcal{B}_k$ can be decomposed into disjoint events depending on the number of parent points from $\Phi_k^{(2)}$ located in $b(o, R_{\mathrm{c}})$. Letting $C_n = \{\#\{\Phi_{C_n} = \Phi_{\mathrm{p},k} \cap \Phi_k^{(2)} \cap b(o, R_{\mathrm{c}})\} = n\}$ denote the event that there only exist n parent points from $\Phi_k^{(2)}$ located in $b(o, R_{\mathrm{c}})$, we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{B}_k) &= \sum_{n=0}^{\infty} \mathbb{P}(C_n)\mathbb{P}(\mathcal{B}_k \mid C_n) \\ &= \sum_{n=0}^{\infty} e^{-\lambda_k(1-p)\pi R_{\mathrm{c}}^2} \frac{(\lambda_k(1-p)\pi R_{\mathrm{c}}^2)^n}{n!} \mathbb{P}(\mathcal{B}_k \mid C_n). \quad (33)\end{aligned}$$

Obviously, we have $\mathbb{P}(\mathcal{B}_k \mid C_0) = 1$. Since Φ_{C_n} is also a PPP, it suffices to derive the probability $\bar{\epsilon}$ that $x + V_x$ is outside of $b(o, R_{\mathrm{c}})$ for an arbitrary $x \in \Phi_{C_n}$ and then $\mathbb{P}(\mathcal{B}_k \mid C_n) = \bar{\epsilon}^n$, where V_x is the vector from x to its cooperator and $\{x, x+V_x\}$ forms a two-point cluster. It is known that the PDF of the distance from the point x to the origin is $f_{\|x\|}(r) = 2r/R_{\mathrm{c}}^2$, and given $\|x\| = r$, we have that $x + V_x$ is outside $b(o, R_{\mathrm{c}})$ with probability $\frac{2\varphi_x}{2\pi}$ from Fig. 12, where

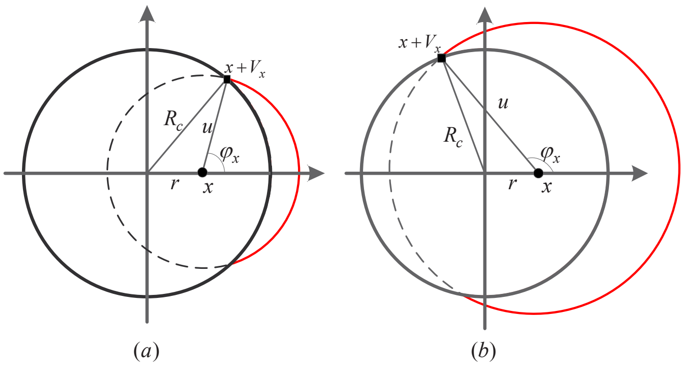


Fig. 12. Illustration for the proof of Thm. 1.

$$\varphi_x = \begin{cases} \pi - \arccos \frac{r^2+u^2-R_{\mathrm{c}}^2}{2ru}, & \text{for } r \geq |R_{\mathrm{c}}-u| \\ \pi - \pi[R_{\mathrm{c}}-u]^+/(R_{\mathrm{c}}-u), & \text{for } r < |R_{\mathrm{c}}-u|. \end{cases} \quad (34)$$

Averaging over $f_{\|x\|}(r)$ for φ_x/π, we have

$$\bar{\epsilon} = 1 - \frac{\left([R_{\mathrm{c}}-u]^+\right)^2}{R_{\mathrm{c}}^2} - \frac{2}{\pi R_{\mathrm{c}}^2} \int_{\zeta}^{R_{\mathrm{c}}} r \arccos \frac{r^2+u^2-R_{\mathrm{c}}^2}{2ru} \mathrm{d}r, \quad (35)$$

$\zeta = \min\{|R_{\mathrm{c}} - u|, R_{\mathrm{c}}\}$ and $\mathbb{P}(\mathcal{B}_k) = \exp(-\lambda_k(1-p)(1-\bar{\epsilon})\pi R_{\mathrm{c}}^2)$. Letting $\epsilon = 1 - \bar{\epsilon}$, and following from the total probability law over the content popularity distribution (1), we have the cache hit probability

$$p_{\mathrm{hit}} = \sum_{k \in [N_{\mathrm{f}}]} p_{\mathrm{r}}(k) \left(1 - e^{-\lambda_{\mathrm{p}} p_{\mathrm{c}}(k)(p+(1-p)\epsilon)\pi R_{\mathrm{c}}^2}\right). \quad (36)$$

The optimal caching distribution that maximizes the cache hit probability can be found from the following problem

$$\max_{\boldsymbol{p}_{\mathrm{c}}} \sum_{k \in [N_{\mathrm{f}}]} p_{\mathrm{r}}(k) \left(1 - e^{-\lambda_{\mathrm{p}} p_{\mathrm{c}}(k)(p+(1-p)\epsilon)\pi R_{\mathrm{c}}^2}\right),$$
$$\text{subject to} \sum_{k \in [N_{\mathrm{f}}]} p_{\mathrm{c}}(k) = 1, \; p_{\mathrm{c}}(k) \geq 0, k \in [N_{\mathrm{f}}]. \quad (37)$$

Since the objective function is the sum of N_{f} exponential functions and the constraints are linear, this problem is convex. From the Lagrange multiplier method, the optimal caching distribution should satisfy the following conditions,

$$p_{\mathrm{c}}^*(k) = \frac{1}{\lambda_{\mathrm{p}}(p+(1-p)\epsilon)\pi R_{\mathrm{c}}^2} \Big[\nu - \log(1/p_{\mathrm{r}}(k))\Big]^+, \; k \in [N_{\mathrm{f}}], \quad (38)$$

where $\sum_{k \in [N_{\mathrm{f}}]} p_{\mathrm{c}}^*(k) = 1$ and ν is related to the Lagrange multiplier. In essence, this optimal distribution can be easily obtained by common *waterfilling* methods, where ν is the height of the water surface. ■

Appendix B
Proof of Theorem 3

Proof: In the CCJT case, the received power, denoted by μ_{sum}, is

$$\mu_{\mathrm{sum}} = h_1 \|x_1\|^{-\alpha} + h_2 (c(\|x_1\|, \Psi))^{-\alpha}. \quad (39)$$

Since both transmitters should be located within the caching radius, Ψ is uniformly distributed in $[\varphi(\|x_1\|), \pi] \cup$

$[-\pi, -\varphi(\|x_1\|)]$, where $\varphi(\|x_1\|)$ is obtained through Fig. 12 as

$$\varphi(\|x_1\|)=\begin{cases}\pi- \arccos \frac{\|x_1\|^2+u^2-R_{\rm c}^2}{2\|x_1\|u}, & \text{for } \|x_1\| \geq |R_{\rm c}-u|,\\ \pi- \pi[R_{\rm c}-u]^+/(R_{\rm c}-u), & \text{for } \|x_1\| < |R_{\rm c}-u|.\end{cases} \tag{40}$$

Since h_1 and h_2 are independent exponentially distributed random variables, $\mu_{\rm sum}$ follows a hypoexponential or Erlang distribution, and the corresponding complementary cumulative distribution function is given by

$$\bar{F}_{\mu_{\rm sum}}(x) = \begin{cases} \frac{\xi_2}{\xi_2-\xi_1}e^{-\xi_1 x} - \frac{\xi_1}{\xi_2-\xi_1}e^{-\xi_2 x}, & \text{if } \xi_1 \neq \xi_2,\\ \xi_1^2 x e^{-\xi_1 x}, & \text{if } \xi_1 = \xi_2,\end{cases} \tag{41}$$

where $\xi_1 = \|x_1\|^\alpha$ and $\xi_2 = (c(\|x_1\|, \Psi))^\alpha$. The success probability is given by

$$P_{\rm CCJT}(\theta)=\mathbb{E}\left(\frac{\mu_{\rm sum}}{I_2}\geq\theta\right)=\mathbb{E}_{\|x_1\|,\Psi,I_2}\Big(\bar{F}_{\mu_{\rm sum}}(I_2\theta)\Big). \tag{42}$$

Since Ψ has a continuous uniform distribution, the event $\xi_1 = \xi_2$ occurs with probability 0. Thus, we have

$$P_{\rm CCJT}(\theta)=\mathbb{E}_{\|x_1\|,\Psi}\Big(\frac{\xi_2}{\xi_2-\xi_1}\mathcal{L}_{I_2}(\xi_1\theta)-\frac{\xi_1}{\xi_2-\xi_1}\mathcal{L}_{I_2}(\xi_2\theta)\Big). \tag{43}$$

Due to Slivnyak's theorem, the distribution of the points excluding Φ_0 also remains the same as the original GPP Φ and thus $\mathcal{L}_{I_2}(s) = \mathcal{L}_{I_1}(s)$. The final result follows by substituting ξ_1, ξ_2, with the PDF of Ψ by $\frac{1}{2(\pi-\varphi(t))}$ and the PDF of $\|x_1\|$ by $f_{\|x_1\|}(t) = 2t/R_{\rm c}^2$. ■

Appendix C
Proof of Theorem 4

Proof: The receiver decodes the signals from the desired two-point cluster successively in the descending order of the average received signal strength, where the order is in fact determined by Ψ with its valid range given in (40). Due to the symmetry, it suffices to analyze the decoding order for $\Psi \in [\varphi(\|x_1\|),\ \pi]$. As Ψ increases from $\varphi(\|x_1\|)$ to π, see Fig. 12 (a), $\|x_2\|$ decreases monotonously and thus there is another critical angle of $\vartheta(\|x_1\|)$ to make $\|x_1\|^{-\alpha} \geq \|x_2\|^{-\alpha}$ when $\Psi \leq \vartheta(\|x_1\|)$ (similar to Fig. 12 (b)). Through the geometric relationship, we obtain

$$\vartheta(\|x_1\|) = \begin{cases} \pi, & \text{for } \|x_1\| \leq u/2,\\ \pi - \arccos\frac{u}{2\|x_1\|}, & \text{otherwise.}\end{cases} \tag{44}$$

When $\Psi \in [\varphi(\|x_1\|),\ \vartheta(\|x_1\|)] \cup [-\vartheta(\|x_1\|),\ -\varphi(\|x_1\|)]$, we have $\|x_1\|^{-\alpha} \geq \|x_2\|^{-\alpha}$ and

$$\begin{aligned}p_1(\theta) &= \mathbb{P}\Big(\frac{\|x_1\|^{-\alpha}h_1}{I_2+\|x_2\|^{-\alpha}h_2}\geq\theta, \frac{\|x_2\|^{-\alpha}h_2}{I_2}\geq\theta\Big)\\ &=\mathbb{E}_{I_2,\|x_1\|,\Psi}\bigg(e^{-\|x_1\|^\alpha\theta I_2}\mathbb{E}_{h_2}\bigg(e^{-\big(\|x_1\|/c(\|x_1\|,\Psi)\big)^\alpha\theta h_2}\\ &\quad\times\mathbf{1}\big(c(\|x_1\|,\Psi))^{-\alpha}h_2\geq I_2\theta\big)\bigg)\bigg)\\ &=\mathbb{E}_{\|x_1\|,\Psi}\bigg(\frac{\mathcal{L}_{I_2}\big(\theta\|x_1\|^\alpha+\theta(c(\|x_1\|,\Psi))^\alpha+\theta^2\|x_1\|^\alpha\big)}{1+\theta(\|x_1\|/c(\|x_1\|,\Psi))^\alpha}\bigg)\\ &=\frac{1}{R_{\rm c}^2}\int_0^{R_{\rm c}}\frac{2t}{\pi-\varphi(t)}\int_{\varphi(t)}^{\vartheta(t)}\frac{1}{1+\theta(t/c(t,\psi))^\alpha}\\ &\quad\times\mathcal{L}_{I_2}\big(\theta t^\alpha+\theta(c(t,\psi))^\alpha)+\theta^2t^\alpha\big){\rm d}\psi{\rm d}t.\end{aligned} \tag{45}$$

When $\Psi\in[\vartheta(\|x_1\|),\pi]\cup[-\pi,\ -\vartheta(\|x_1\|)]$, we have $\|x_1\|^{-\alpha}< \|x_2\|^{-\alpha}$ and

$$\begin{aligned}p_2(\theta) &= \mathbb{P}\Big(\frac{\|x_2\|^{-\alpha}h_2}{I_2+\|x_1\|^{-\alpha}h_1}\geq\theta, \frac{\|x_1\|^{-\alpha}h_1}{I_2}\geq\theta\Big)\\ &=\frac{1}{R_{\rm c}^2}\int_0^{R_{\rm c}}\frac{2t}{\pi-\varphi(t)}\int_{\vartheta(t)}^{\pi}\frac{1}{1+\theta(c(t,\psi)/t)^\alpha}\\ &\quad\times\mathcal{L}_{I_2}\big(\theta t^\alpha+\theta(c(t,\psi))^\alpha)+\theta^2(c(t,\psi))^\alpha\big){\rm d}\psi{\rm d}t.\end{aligned} \tag{46}$$

By summing the two cases, the final link success probability for CCMT is obtained. ■

Appendix D
Proof of Corollary 4

Proof: In the CCMT case, the mean number of the successfully decoded data streams is

$$\begin{aligned}\bar{N}_{\rm CCMT}(\theta) &= 2\mathbb{P}({\rm SIR}_1\geq\theta, {\rm SIR}_2\geq\theta)\\ &\quad+\mathbb{P}({\rm SIR}_1\geq\theta, {\rm SIR}_2<\theta),\end{aligned} \tag{47}$$

where $\mathbb{P}({\rm SIR}_1 \geq \theta, {\rm SIR}_2 \geq \theta)$ is obtained by Theorem 4. As for the probability that only one data stream is decoded successfully, i.e., $\mathbb{P}({\rm SIR}_1 \geq \theta, {\rm SIR}_2 < \theta)$, we derive as follows. When $\Psi \in [\varphi(\|x_1\|),\ \vartheta(\|x_1\|)] \cup [-\vartheta(\|x_1\|),\ -\varphi(\|x_1\|)]$, we have

$$\begin{aligned}\tilde{p}_1(\theta) &= \mathbb{P}\Big(\frac{\|x_1\|^{-\alpha}h_1}{I_2+\|x_2\|^{-\alpha}h_2}\geq\theta, \frac{\|x_2\|^{-\alpha}h_2}{I_2}<\theta\Big)\\ &= \mathbb{E}_{I_2,\|x_1\|,\Psi}\bigg(e^{-\|x_1\|^\alpha\theta I_2}\mathbb{E}_{h_2}\Big(e^{-(\|x_1\|/(\|x_1\|,\Psi))^\alpha h_2}\\ &\quad\times\mathbf{1}\big((c(\|x_1\|,\Psi))^{-\alpha}h_2<I_2\theta\big)\Big)\bigg)\\ &= \mathbb{E}_{\|x_1\|,\Psi}\bigg(\frac{1}{1+\theta(\|x_1\|/c(\|x_1\|,\Psi))^\alpha}\Big(\mathcal{L}_{I_2}\big(\theta\|x_1\|^\alpha\big)\\ &\quad-\mathcal{L}_{I_2}\big(\theta\|x_1\|^\alpha+\theta(c(\|x_1\|,\Psi))^\alpha+\theta^2\|x_1\|^\alpha\big)\Big)\bigg)\\ &= \frac{1}{R_{\rm c}^2}\int_0^{R_{\rm c}}\frac{2t}{\pi-\varphi(t)}\int_{\varphi(t)}^{\vartheta(t)}\frac{1}{1+\theta(t/c(t,\psi))^\alpha}\\ &\quad\times\Big(\mathcal{L}_{I_2}\big(\theta t^\alpha\big)-\mathcal{L}_{I_2}\big(\theta t^\alpha+\theta c(t,\psi))^\alpha\big)+\theta^2t^\alpha\big)\Big){\rm d}\psi{\rm d}t.\end{aligned} \tag{48}$$

When $\Psi \in [\vartheta(\|x_1\|),\pi] \cup [-\pi,\ -\vartheta(\|x_1\|)]$, we have

$$\begin{aligned}\tilde{p}_2(\theta) &=\mathbb{P}\Big(\frac{\|x_2\|^{-\alpha}h_2}{I+\|x_1\|^{-\alpha}h_1}\geq\theta, \frac{\|x_1\|^{-\alpha}h_1}{I_2}<\theta\Big)\\ &=\frac{1}{R_{\rm c}^2}\int_0^{R_{\rm c}}\frac{2t}{\pi-\varphi(t)}\int_{\vartheta(t)}^{\pi}\frac{1}{1+\theta(c(t,\psi)/t)^\alpha}\Big(\mathcal{L}_{I_2}\big(\theta(c(t,\psi))^\alpha\big))\\ &\quad-\mathcal{L}_{I_2}\big(\theta t^\alpha+\theta(\theta+1)(c(t,\psi))^\alpha\big))\Big){\rm d}\psi{\rm d}t.\end{aligned} \tag{49}$$

By summing the two cases, we have $\mathbb{P}({\rm SIR}_1 \geq \theta, {\rm SIR}_2 < \theta) = \tilde{p}_1(\theta)+\tilde{p}_2(\theta)$. Thus, the mean number of the successfully decoded data streams can be obtained. ■

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021," Feb. 2017.

[2] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

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

[4] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: design aspects, challenges, and future directions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 22–28, Sept. 2016.

[5] N. Naderializadeh, D. T. H. Kao, and A. S. Avestimehr, "How to utilize caching to improve spectral efficiency in device-to-device wireless networks," in *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept. 2014, pp. 415–422.

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

[7] 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 Communications*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[8] N. Giatsoglou, K. Ntontin, E. Kartsakli, A. Antonopoulos, and C. Verikoukis, "D2D-aware device caching in mmwave-cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2025–2037, Sept. 2017.

[9] A. Liu, V. Lau, and G. Caire, "Cache-induced hierarchical cooperation in wireless device-to-device caching networks," 2018, accepted at *IEEE Transactions on Information Theory*. Available on IEEE Xplore Early Access.

[10] B. Chen, C. Yang, and A. F. Molisch, "Cache-enabled device-to-device communications: Offloading gain and energy cost," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4519–4536, Jul. 2017.

[11] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4365–4380, Oct. 2016.

[12] S. Krishnan and H. S. Dhillon, "Distributed caching in device-to-device networks: A stochastic geometry perspective," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, Nov. 2015, pp. 1280–1284.

[13] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Fundamentals of cluster-centric content placement in cache-enabled device-to-device networks," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2511–2526, Jun. 2016.

[14] D. S. Newman, "A new family of point processes which are characterized by their second moment properties," *Journal of Applied Probability*, vol. 7, no. 2, pp. 338–358, 1970.

[15] A. Guo, Y. Zhong, W. Zhang, and M. Haenggi, "The Gauss-Poisson process for wireless networks and the benefits of cooperation," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 1916–1929, May 2016.

[16] M. Haenggi, *Stochastic geometry for wireless networks*. Cambridge University Press, 2012.

[17] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, Mar. 1999, pp. 126–134.

[18] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *2015 IEEE International Conference on Communications (ICC)*, Jun. 2015, pp. 3358–3363.

[19] 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 Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

[20] V. Bioglio, F. Gabry, and I. Land, "Optimizing MDS codes for caching at the edge," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2015, pp. 1–6.

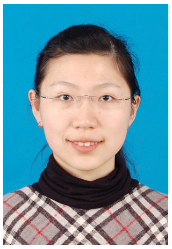
[21] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.

[22] X. Zhang and M. Haenggi, "The performance of successive interference cancellation in random wireless networks," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6368–6388, Oct. 2014.

[23] M. Haenggi, "The meta distribution of the SIR in Poisson bipolar and cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2577–2589, Apr. 2016.

[24] L. Wu and W. Zhang, "Caching-based scalable video transmission over cellular networks," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1156–1159, Jun. 2016.

[25] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.211, Dec. 2017, version 15.0.0.



Na Deng (S'12-M'17) received the Ph.D. and B.S. degrees in electronic engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2015 and 2010, respectively. From 2013 to 2014, she was a Visiting Student in Prof. Martin Haenggi's Group at the University of Notre Dame, Notre Dame, IN, USA. From June 2015 to November 2016, she was a Senior Engineer at Huawei Technologies Co., Ltd., Shanghai, China. Since then, she has been a lecturer with the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. Her scientific interests include networking and wireless communications, green communications, and network design based on wireless big data.

Martin Haenggi (S'95-M'99-SM'04-F'14) received the Dipl.-Ing. (M.Sc.) and Dr.sc.techn. (Ph.D.) degrees in electrical engineering from the Swiss Federal Institute of Technology in Zurich (ETH) in 1995 and 1999, respectively. Currently he is the Freimann Professor of Electrical Engineering and a Concurrent Professor of Applied and Computational Mathematics and Statistics at the University of Notre Dame, Indiana, USA. In 2007-2008, he was a visiting professor at the University of California at San Diego, and in 2014-2015 he was an Invited Professor at EPFL, Switzerland. He is a co-author of the monograph "Interference in Large Wireless Networks" (NOW Publishers, 2009) and "Stochastic Geometry Analysis of Cellular Networks" (Cambridge University Press, 2018) and the author of the textbook "Stochastic Geometry for Wireless Networks" (Cambridge, 2012), and he published 14 single-author journal articles. His scientific interests lie in networking and wireless communications, with an emphasis on cellular, amorphous, ad hoc (including D2D and M2M), cognitive, and vehicular networks. He served as an Associate Editor of the Elsevier Journal of Ad Hoc Networks, the IEEE Transactions on Mobile Computing (TMC), the ACM Transactions on Sensor Networks, as a Guest Editor for the IEEE Journal on Selected Areas in Communications, the IEEE Transactions on Vehicular Technology, and the EURASIP Journal on Wireless Communications and Networking, as a Steering Committee member of the TMC, and as the Chair of the Executive Editorial Committee of the IEEE Transactions on Wireless Communications (TWC). Currently he is the Editor-in-Chief of the TWC. He also served as a Distinguished Lecturer for the IEEE Circuits and Systems Society, as a TPC Co-chair of the Communication Theory Symposium of the 2012 IEEE International Conference on Communications (ICC'12), of the 2014 International Conference on Wireless Communications and Signal Processing (WCSP'14), and the 2016 International Symposium on Wireless Personal Multimedia Communications (WPMC'16). For both his M.Sc. and Ph.D. theses, he was awarded the ETH medal. He also received a CAREER award from the U.S. National Science Foundation in 2005 and three awards from the IEEE Communications Society, the 2010 Best Tutorial Paper award, the 2017 Stephen O. Rice Prize paper award, and the 2017 Best Survey paper award, and he is a 2017 Clarivate Analytics Highly Cited Researcher.