# Optimizing Joint Probabilistic Caching and Channel Access for Clustered D2D Networks

Ramy Amer, Mohamed Baza, Tara Salman, M. Majid Butt, Ahmad Alhindi, and Nicola Marchetti

*Abstract:* **Caching at mobile devices and leveraging device-to-device (D2D) communication are two promising approaches to support massive content delivery over wireless networks. Analysis of such D2D caching networks based on a physical interference model is usually carried out by assuming uniformly distributed devices. However, this approach does not capture the notion of device clustering. In this regard, this paper proposes a joint communication and caching optimization framework for clustered D2D networks. Devices are spatially distributed into disjoint clusters and are assumed to have a surplus memory that is utilized to proactively cache files, following a random probabilistic caching scheme. The cache offloading gain is maximized by jointly optimizing channel access and caching scheme. A closed-form caching solution is obtained and bisection search method is adopted to heuristically obtain the optimal channel access probability. Results show significant improvement in the offloading gain reaching up to $10\%$ compared to the Zipf caching baseline.**

*Index Terms:* **Caching, channel access, D2D communication, offloading gain.**

## I. INTRODUCTION

CACHING at mobile devices significantly improves system performance by facilitating device-to-device (D2D) communications, which enhances the spectrum efficiency and alleviate the heavy burden on backhaul links [1]–[3]. There are two main approaches for content placement in the literature, deterministic and probabilistic. For deterministic placement, files are cached and optimized for specific networks in a deterministic manner [3]–[5]. However, in practice, the wireless channels and the geographic distribution of devices are time-variant. This triggers the optimal content placement strategy to be frequently updated, which makes the content placement quite complex. To cope with this problem, probabilistic content placement is proposed whereby each device randomly caches a subset of content with a certain caching probability in stochastic networks [6], [7]. In this paper, we focus on the probabilistic caching model. In essence, this model is shown to be powerful and tractable for the analysis of random networks and possibly yields a convex content caching problem, which can be effectively solved [6], [7].

Modeling of wireless caching networks also follows two main directions in the current state-of-art. The first line of work focuses on the fundamental scaling results by assuming a simple protocol channel model [3]–[5], known as the protocol model. This model assumes that two devices can always communicate if they are within a certain distance. The second line of work, which is similar to the one adopted in this paper, considers a more realistic model for the underlying physical and medium access control (MAC) layers [8]–[16]. This is commonly defined as the physical interference model. Driven by this, the physical interference model allows us to study joint caching and communications for D2D networks and design efficient channel access-aware caching scheme.

The analysis of wireless caching networks that underlies a physical interference model, is commonly conducted by means of stochastic point processes. For instance, modeling device locations as a Poisson point process (PPP) is a widely adopted approach in the wireless caching area [8]–[10], [17]. However, while the PPP model is tractable, a realistic model for D2D caching networks needs to capture the notion of clustering. In particular, in clustered D2D networks, each device has multiple proximate devices, where any of them can act as a serving device. Such deployments can be effectively characterized by cluster processes [18].

Performance of clustered D2D caching networks is studied in [12]–[16], [19], [20]. For instance, the authors in [12] discussed different strategies of content placement in a Poisson cluster process (PCP) deployment. Meanwhile, the authors extended their work in [13], to optimize the collective performance of all the devices in each cluster. Moreover, the authors in [14] proposed cooperation among the D2D transmitters and hybrid caching strategies to save the energy cost of content providers, where the location of these providers is modeled by a Gauss-Poisson process. In [15], we jointly optimized content caching and frequency partitioning between D2D cellular communications for clustered cache-enabled networks. Moreover, we studied the role of cooperative communication for clustered D2D networks in [16]. We particularly showed that cooperative communication becomes more appealing in denser D2D caching networks and adverse interference conditions. Meanwhile, in [20], we formu-

Manuscript received October 27, 2020; revised April 10, 2021; approved for publication by Abbas Jamalipour, Division II Editor, May 23, 2021.

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 13/RC/2077P2 and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077.

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for supporting this work by Grant Code: 19-COM-1-01-0017

M. Baza is with the Department of Computer Science, College of Charleston, SC, USA, email: bazam@cofc.edu.

T. Salman is with the Department of Computer Science & Engineering, Washington University, St. Louis, MO, USA, email: tara.salman@wustl.edu.

A. Alhindi is with Department of Computer Science, Umm Al-Qura University, Makkah, KSA, email: ahhindi@uqu.edu.sa.

R. Amer and N. Marchetti are with CONNECT-2 Centre for Future Networks, Trinity College Dublin, Ireland, emails: {ramyr, nicola.marchetti}@tcd.ie.

M. Majid Butt is with Nokia Bell Labs, France, and Trinity College Dublin, Ireland, email: Majid.Butt@tcd.ie.

R. Amer is the corresponding author.

Digital Object Identifier: 10.23919/JCN.2021.000019

lated and solved the energy minimization problem for clustered D2D caching networks.

While the prior works in [12]–[16], [19], [20] studied cache-enabled D2D networks from various perspectives, the joint optimization of caching and channel access for clustered D2D networks has not been addressed yet in the literature. Such a study is vital for proper quantifying and optimizing the achievable performance of D2D cache-enabled networks under practical MAC scheduling protocols.

Compared with this prior art, the main contributions of this paper are as follows:

- We study the content placement and delivery for a network wherein cache-enabled devices are spatially distributed into disjoint clusters. We conduct a performance analysis and joint optimization of channel access and probabilistic content placement with the aim to maximize the cache offloading gain.
- We characterize the optimal content placement as a function of the system parameters, namely, density of clusters, displacement distance between devices, and required signal-to-interference ratio (SIR) threshold. We then propose a heuristic approach to obtain the optimal channel access probability.
- Our results reveal that the optimal caching scheme heavily depends on the channel access probability and the geometry of the network. Overall, joint optimization of content placement and communication, e.g., channel access, is shown to be vital to enhance the performance of wireless caching networks.

The rest of this paper is organized as follows. Section II and Section III introduce the system model and the rate coverage analysis, respectively. The offloading gain maximization problem is discussed in Section IV. Numerical results are presented in Section V and conclusions are drawn in Section VI.

## II. SYSTEM MODEL

### A. System Setup

We model the location of mobile devices with a Thomas cluster process (TCP). The use of TCP allows us to factor in the notion of clustering for D2D caching networks, which is commonly ignored in the literature. The TCP is composed of the parent points, which are drawn from a PPP $\Phi_p$ with density $\lambda_p$, and the daughter points that are drawn from a Gaussian PPP around each parent point [18]. In particular, the daughter points are normally scattered with variance $\sigma^2 \in \mathbb{R}$ around each parent point. The parent points and offspring are referred to as cluster centers and cluster members, respectively. By the TCP definition, the number of devices per cluster is a Poisson random variable (RV) with mean $\overline{n}$. Therefore, the density function of a cluster member location relative to its cluster center is

$$f_Y(y) = \frac{1}{2\pi\sigma^2}\exp\Big(-\frac{\|y\|^2}{2\sigma^2}\Big), \qquad y \in \mathbb{R}^2, \tag{1}$$

where $\|.\|$ is the Euclidean norm. The intensity function of a cluster is given by $\lambda_c(y) = \overline{n}/(2\pi\sigma^2)\exp\big(-\frac{\|y\|^2}{2\sigma^2}\big)$, and therefore, the intensity of the entire process is given by $\lambda = \overline{n}\lambda_p$.

We assume that the D2D communication is operating as out-of-band D2D under flat Rayleigh fading channels. D2D communication is enabled within each cluster to deliver popular content. It is assumed that the devices adopt a slotted-ALOHA medium access protocol, where each transmitter during each time slot, independently and randomly accesses the channel with the same probability $q$. One can alternatively assume that each device makes a coin flip at each time about whether or not it accesses a shared-channel. This allows us to define a Bernoulli process $N_y$ with the probability that a device located at $y$ accesses a channel being $\mathbb{P}(N_y) = q$. The key advantage of adopting slotted-ALOHA is that it is a simple yet fundamental MAC protocol, where there is no need for a central controller to schedule the users' transmissions. Moreover, despite the vast amount of existing studies on MAC protocols, only variations of ALOHA and CSMA are still used in the majority of technologies adopted in the Internet of Things [21]. According to this access model, multiple active D2D links might coexist within a cluster. Therefore, $q$ is a design parameter that directly controls intra- as well as inter-cluster interference, as described later.

If a requesting device caches the desired content, the device directly retrieves the content. However, if the content is not locally cached, it can be downloaded from a randomly selected neighboring device that caches the file within the same cluster, henceforth called *catering device*. This catering device is, in turn, admitted to access the channel according to the proposed slotted-ALOHA protocol. Finally, the device attaches to the nearest base station (BS) as a last resort to download the content, in the case it is not cached within the device cluster. Since there are memory and battery consumption costs borne by a catering device, the geographically closest device may not want to participate in the content caching and/or delivery. Hence, randomizing the catering device reflects the possibility of being served by a distant device that is willing to participate in the content delivery, which is not necessarily the nearest one. Note that this assumption is commonly adopted in the literature [12], [14].

### B. Content Popularity and Caching

We assume that each device has a surplus memory of size $M$ designated for caching files. The total number of files is $N_f > M$, and the set (library) of content indices is denoted as $\mathcal{F} = \{1, 2, \cdots, N_f\}$. These files represent the content catalog that all devices in a cluster may request, which are indexed in a descending order of popularity. The probability that the $i$th file is requested follows a Zipf's distribution given by [22],

$$p_i = \frac{i^{-\beta}}{\sum_{k=1}^{N_f} k^{-\beta}}, \tag{2}$$

where $\beta$ is a parameter that reflects how skewed the popularity distribution is. For example, if $\beta = 0$, the popularity of the files has a uniform distribution. Increasing $\beta$ increases the disparity among the files popularity such that lower indexed files have higher popularity. By definition, $\sum_{i=1}^{N_f} p_i = 1$. We use the Zipf's distribution to model the popularity of files per cluster.

We adopt a random content placement where each device independently selects a file to cache according to a specific prob-

ability function $\boldsymbol{b} = \{b_1, b_2, \cdots, b_{N_f}\}$, where $b_i$ is the probability that a device caches the $i$th file, $0 \leq b_i \leq 1$ for all $i = \{1, \cdots, N_f\}$. To avoid duplicate caching of the same content within the memory of the same device, we follow a probabilistic caching approach proposed in [7], which implies that $\sum_{i=1}^{N_f} b_i = M$.[1]

Next, we proceed with the rate coverage analysis to obtain the offloading gain, which is a key performance metric for D2D caching networks [9]. Particularly, the offloading gain is defined as the probability of obtaining a requested file from the local cluster, either via self-cache or from a neighboring device in the same cluster, with a received SIR higher than a required threshold $\vartheta$.

## III. RATE COVERAGE ANALYSIS

We conduct the next analysis for a cluster whose center is assumed at $x_0 \in \Phi_p$, referred to as representative cluster. The device requesting a content in this cluster, henceforth called typical device, is located at the origin. We denote the location of the catering device by $y_0$ relative to $x_0$, where $\{x_0, y_0\} \in \mathbb{R}^{\not\in}$. The distance between the typical and catering devices is denoted as $r = \|x_0 + y_0\|$, which is a realization of a RV $R$ whose distribution is described later. Having explained the channel access and the random selection of catering devices, the offloading gain can be expressed as

$$\mathbb{P}_o(q, \boldsymbol{b}) = \sum_{i=1}^{N_f} p_i b_i + p_i(1 - b_i)(1 - e^{-b_i \overline{n}}) \times \underbrace{\int_{r=0}^{\infty} f_R(r) \mathbb{P}(\mathrm{SIR}_{|r} > \vartheta)\, \mathrm{d}r}_{\Upsilon}, \tag{3}$$

where $\mathrm{SIR}_{|\mathrm{r}}$ is the received SIR at the typical device when downloading a content from a catering device $r$ apart from the origin, and $\Upsilon$ represents the rate coverage probability. The first term in (3) is the probability of requesting a locally cached file (self-cache). The second term is the probability that a requested file $i$ is cached among at least one cluster member and being downloadable with an SIR greater than $\vartheta$, given that it was not self-cached. More precisely, since the number of devices per cluster has a Poisson distribution, the probability that there are $k$ devices per cluster is equal to $\overline{n}^k e^{-\overline{n}}/k!$. Accordingly, the probability that there are $k$ devices caching content $i$ is $(b_i\overline{n})^k e^{-b_i\overline{n}}/k!$. Hence, the probability that at least one device caches content $i$ is $1 - e^{-b_i\overline{n}}$.

For the serving distance distribution $f_R(r)$, since both the typical device and catering device have their locations drawn from a normal distribution with variance $\sigma^2$, then by definition, the serving distance has a Rayleigh distribution of scale parameter $\sqrt{2}\sigma$, i.e., $f_R(r) = r/(2\sigma^2)\mathrm{e}^{\frac{-r^2}{4\sigma^2}}$. It is worth noting that the serving distance is independent of the caching probability.

[1]It is clear that the benefits of content caching at devices is prominent only if these devices have sufficient memory to store files of interest. The availability of unused memory on these devices can not be always maintained. In other words, it might happen that the devices run out of sufficient memory to cache popular files. The analysis of device caching networks with such variations of unused memory to store popular files is an important extension for our future work.

To clarify, from the thinning theorem [18], the set of devices caching content $i$ in a given cluster forms a Gaussian PPP $\Phi_{ci}$ whose intensity is $\lambda_{ci} = b_i \lambda_c(y)$. The probability distribution function (PDF) of the distance between a randomly selected caching device from $\Phi_{ci}$ and the typical device is $f_R(r)$, which is again independent of $b_i$.

The received power at the typical device from a catering device located at $y_0$ relative to the cluster center is given by

$$P = P_d g_0 \|x_0 + y_0\|^{-\alpha} = P_d g_0 r^{-\alpha}, \tag{4}$$

where $P_d$ denotes the D2D transmission power, $g_0 \sim \exp(1)$ is the complex Gaussian fading channel coefficient, and $\alpha > 2$ is the path loss exponent. Under this setup, the typical device sees two types of interference, namely, the intra- and inter-cluster interference. We first describe the inter-cluster interference, then the intra-cluster interference is characterized. The set of active devices in any remote cluster is denoted as $\mathcal{B}^{\amalg}$, where $q$ refers to the access probability. Similarly, the set of active devices in the local cluster is denoted as $\mathcal{A}^{\amalg}$. The received interference at the typical device from simultaneously active D2D transmitters within the remote clusters is

$$I_{\Phi_p^!} = \sum_{x \in \Phi_p^!} \sum_{y \in \mathcal{B}^q} P_d g_{y_x} \|x + y\|^{-\alpha} = \sum_{x \in \Phi_p^!} \sum_{y \in \mathcal{B}^q} P_d g_u u^{-\alpha},$$

where $\Phi_p^! = \Phi_p \setminus x_0$ for ease of notation, $y$ is the marginal distance between a potential interfering device and its cluster center at $x \in \Phi_p$ , $u = \|x + y\|$ is a realization of a RV $U$ that models the inter-cluster interfering distance, $g_{y_x} \sim \exp(1)$, and $g_u = g_{y_x}$. The intra-cluster interference is then given by

$$I_{\Phi_c} = \sum_{y \in \mathcal{A}^p} P_d g_{y_{x_0}} \|x_0 + y\|^{-\alpha} = \sum_{y \in \mathcal{A}^p} P_d g_h h^{-\alpha},$$

where $y$ is the marginal distance between the intra-cluster interfering devices and the cluster center at $x_0 \in \Phi_p$, $h = \|x_0 + y\|$ is a realization of a RV $H$, which models the intra-cluster interfering distance, $g_{y_{x_0}} \sim \exp(1)$, and $g_h = g_{y_{x_0}}$. From the thinning theorem [18], the set of active transmitters based on the slotted-ALOHA medium access forms a Gaussian PPP $\Phi_{cq}$ whose intensity is given by

$$\lambda_{cq} = q\lambda_c(y) = q\overline{n} f_Y(y) = \frac{q\overline{n}}{2\pi\sigma^2} \exp\Big( - \frac{\|y\|^2}{2\sigma^2}\Big), \quad y \in \mathbb{R}^2.$$

Assuming that the thermal noise is neglected as compared to the aggregate interference, the received SIR at the typical device can be written as

$$\mathrm{SIR}_{|\mathrm{r}} = \mathbf{1}\{N_r = 1\} \frac{P}{I_{\Phi_p^!} + I_{\Phi_c}} = \mathbf{1}\{N_r = 1\} \frac{P_d g_0 r^{-\alpha}}{I_{\Phi_p^!} + I_{\Phi_c}}, \tag{5}$$

where $\mathbf{1}\{.\}$ is the indicator function, and, for ease of exposition, $N_r = N_{y_0}$ is a Bernoulli RV that takes the value one with probability $q$. Thus, the event $\{N_r = 1\}$ captures the incident

when the serving device is admitted to access the channel. Then, the probability that the received SIR is higher than the required threshold $\vartheta$ is derived as follows:

$$\begin{aligned}\Upsilon_{|r} &= \mathbb{P}(\mathrm{SIR}_{|\mathrm{r}} > \vartheta) \\ &= \mathbb{P}\Big(\mathbf{1}\{\mathrm{N_r}=1\}\frac{\mathrm{P_d g_0 r^{-\alpha}}}{\mathrm{I}_{\Phi_\mathrm{p}^!}+\mathrm{I}_{\Phi_\mathrm{c}}} > \vartheta\Big) \\ &\overset{(a)}{=} q\mathbb{P}\Big(\frac{\mathrm{P_d g_0 r^{-\alpha}}}{I_{\Phi_\mathrm{p}^!}+\mathrm{I}_{\Phi_\mathrm{c}}} > \vartheta\Big), \end{aligned} \tag{6}$$

where (a) follows from the assumption of a Bernoulli's RV with mean $q$. Rearranging the right-hand side, we get

$$\begin{aligned}\Upsilon_{|r} &\overset{(b)}{=} q\mathbb{E}_{\mathrm{I}_{\Phi_\mathrm{p}^!},\mathrm{I}_{\Phi_\mathrm{c}}}\left[\exp\Big(\frac{-\vartheta r^\alpha}{P_d}[I_{\Phi_p^!}+I_{\Phi_c}]\Big)\right] \\ &\overset{(c)}{=} q\mathscr{L}_{I_{\Phi_p^!}}(s)\mathscr{L}_{I_{\Phi_c}}(s), \end{aligned} \tag{7}$$

where (b) follows from the assumption $g_0 \sim \mathcal{CN}(0,1)$, and (c) follows from the independence of the intra- and inter-cluster interference and calculating the Laplace transform of them, with $s = \vartheta r^\alpha / P_d$. The classical tradeoff between frequency reuse and higher interference power is depicted in (7). In other words, increasing the access probability $q$ allows more opportunities to access the channel, however, this channel access would be accompanied with higher interference power.

Next, we first derive the Laplace transform of interference to obtain the rate coverage probability $\Upsilon$. Then, we formulate the offloading gain maximization problem.

**Lemma 1:** Laplace transform of the inter-cluster aggregate interference $I_{\Phi_p^!}$ is given by

$$\mathscr{L}_{I_{\Phi_p^!}}(s) = \exp\Big(-2\pi\lambda_p \int_{v=0}^{\infty}\Big(1-\mathrm{e}^{-q\overline{n}\varphi(s,v)}\Big)v\,\mathrm{dv}\Big), \tag{8}$$

where $s = \frac{\vartheta r^\alpha}{P_d}$, $\varphi(s,v) = \int_{u=0}^{\infty}\frac{s}{s+u^\alpha}f_U(u|v)\,\mathrm{du}$, and $f_U(u|v) = \mathrm{Rice(u|v,}\sigma)$ represents Rice's PDF of parameter $\sigma$, and $v = \|x\|$.

*Proof:* Please see Appendix A. □

**Lemma 2:** Laplace transform of the intra-cluster aggregate interference $I_{\Phi_c}$ is approximated as

$$\mathscr{L}_{\mathscr{I}}() \approx \exp\Big(-\int_{=}^{\infty}\frac{}{+\alpha}\mathscr{H}()\,\mathrm{dh}\Big), \tag{9}$$

where $f_H(h) = \mathrm{Rayleigh(h,}\sqrt{2}\sigma)$ represents Rayleigh's PDF with a scale parameter $\sqrt{2}\sigma$.

The proof of Lemma 2 proceeds in a similar way to the proof of Lemma 1, and the approximation follows from neglecting the correlation among intra-cluster serving distances, i.e., the common part $x_0$ in $\|x_0+y\|$ with the detailed proof omitted.

To validate the approximation in Lemma 2, in Fig. 1, we plot the rate coverage probability $\Upsilon$, computed from (3), against the displacement standard deviation $\sigma$. Fig. 1 verifies that the adopted approximation is accurate. It is intuitive to see that the $\Upsilon$ decreases as both $\sigma$ and $\lambda_p$ increase. This is attributed to the fact that the desired signal level increases as $\sigma$ decreases, meanwhile, the interference power increases with $\lambda_p$ and $\sigma$. This is attributed to the higher density of clusters (larger $\lambda_p$) and statistically shorter distance between interfering devices and the typical device (larger $\sigma$).

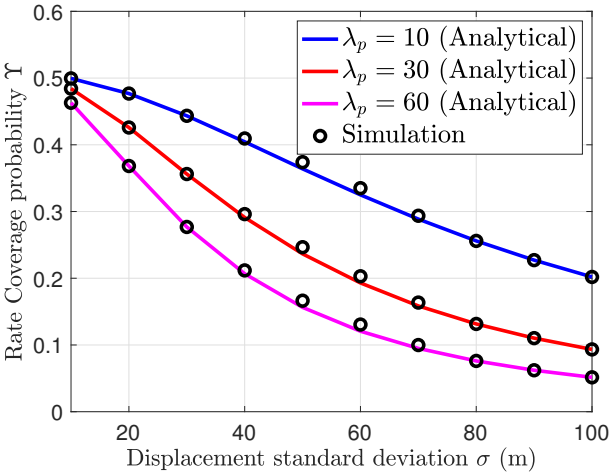



Fig. 1. The rate coverage probability $\Upsilon$ versus the displacement standard deviation $\sigma$ ($\overline{n}=5$, $\vartheta = 0\,\mathrm{dB}$, $p=0.3$).

From (3), (8), and (9), we get

$$\begin{aligned}\mathbb{P}_o(q,\boldsymbol{b}) = &\sum_{i=1}^{N_f} p_i b_i + p_i(1-b_i)(1-e^{-b_i\overline{n}}) \\ &\times \underbrace{\int_{r=0}^{\infty}\frac{r}{2\sigma^2}\mathrm{e}^{\frac{-r^2}{4\sigma^2}}p\mathscr{L}_{I_{\Phi_p^!}}(s)\mathscr{L}_{I_{\Phi_c}}(s)\,\mathrm{dr}}_{\Upsilon}. \end{aligned} \tag{10}$$

Having characterized the offloading gain, we next formulate the joint channel access and caching optimization problem.

## IV. MAXIMIZING OFFLOADING GAIN

The offloading gain maximization problem is formulated as

$$\mathbf{P1:} \quad \max_{q,\boldsymbol{b}} \quad \mathbb{P}_o(q,\boldsymbol{b}) \tag{11}$$

$$\text{s.t.} \quad \sum_{i=1}^{N_f} b_i = M, \tag{12}$$

$$b_i \in [0,1], \tag{13}$$

$$q \in [0,1], \tag{14}$$

where (12) is the device cache size constraint. Since the offloading gain depends on the caching probability $\boldsymbol{b}$ and access probability $q$, and since $q$ exists as a complicated exponential term in $\Upsilon$ (see (7) and (9)), it is difficult to analytically characterize the objective function, e.g., show concavity or find a tractable expression for the optimal access probability. In order to tackle this, we propose to find the optimal access probability $q^*$ that maximizes $\Upsilon$ via the bisection search method in its feasible range $q \in [0,1]$. Then, the obtained $q^*$ is used to solve for

Table 1. Simulation parameters.

| Description | Parameter | Value |
|---|---|---|
| Displacement standard deviation | $\sigma$ | 10 m |
| Popularity index | $\beta$ | 0.5 |
| Path loss exponent | $\alpha$ | 4 |
| Library size and cache size per device | $N_f$, $M$ | 100, 8 files |
| Average number of devices per cluster | $\overline{n}$ | 4 |
| Density of clusters | $\lambda_p$ | 10 clusters/km$^2$ |
| SIR threshold | $\vartheta$ | 0 dB |

the caching probability $\boldsymbol{b}$ in the optimization problem below.

$$\textbf{P2:} \quad \max_{\boldsymbol{b}} \quad \mathbb{P}_o(q^*, \boldsymbol{b}) \qquad (15)$$
$$\text{s.t.} \quad (12), (13)$$

**Lemma 3:** For fixed $q^*$, $\mathbb{P}_o(q^*, \boldsymbol{b})$ is a concave function w.r.t. $\boldsymbol{b}$ and the optimal caching probability $\boldsymbol{b}^*$ that maximizes the offloading gain is given by

$$b_i^* = \begin{cases} 1 & , v^* < p_i - p_i(1-e^{-\overline{n}})\Upsilon \\ 0 & , v^* > p_i + \overline{n} p_i \Upsilon \\ \psi(v^*) & , \text{otherwise}, \end{cases}$$

where $\psi(v^*)$ is the solution of $v^* = p_i + p_i\big(\overline{n}(1-b_i^*)e^{-\overline{n}b_i^*} - (1-e^{-\overline{n}b_i^*})\big)\Upsilon$, that satisfies $\sum_{i=1}^{N_f} b_i^* = M$.

*Proof:* Please see Appendix B. □

Clearly, the optimal caching solution $\boldsymbol{b}^*$ depends on the scheduling of devices through channel access probability $q^*$ from $\Upsilon$, while $q^*$ is independent of $\boldsymbol{b}^*$. [9] shows that a PPP network exhibits the same property, i.e., the caching scheme is scheduling-dependent. To gain some insights, it is useful to consider a simple case when only one D2D link per cluster is allowed. In this case, the rate coverage probability of the proposed clustered model with one active D2D link within a cluster will be [20, Lemma 2]:

$$\Upsilon = \frac{1}{\big(4\sigma^2\pi\lambda_p\vartheta^{2/\alpha}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha)+1\big)}. \qquad (16)$$

Substituting in (10) for $\Upsilon$, we get the offloading gain as

$$\mathbb{P}_o(\boldsymbol{b}) = \sum_{i=1}^{N_f} p_i b_i + \frac{p_i(1-b_i)(1-e^{-b_i\overline{n}})}{4\sigma^2\pi\lambda_p\vartheta^{2/\alpha}\Gamma(1+2/\alpha)\Gamma(1-2/\alpha)+1}. \qquad (17)$$

**Remark 1:** From (17), it is clear that the offloading gain increases as $\sigma$ and $\lambda_p$ decrease. Particularly, the offloading gain is inversely proportional to the density of clusters $\lambda_p$ and the variance of the displacement $\sigma^2$. This is because smaller $\sigma$ results in higher levels of the desired signal, while lower $\lambda_p$ leads to smaller encountered interference at the typical device.

## V. NUMERICAL RESULTS

We first validate the developed mathematical model via Monte Carlo simulations. Then we benchmark the proposed caching scheme against conventional caching schemes. Unless otherwise stated, the network parameters are selected as shown in Table 1.

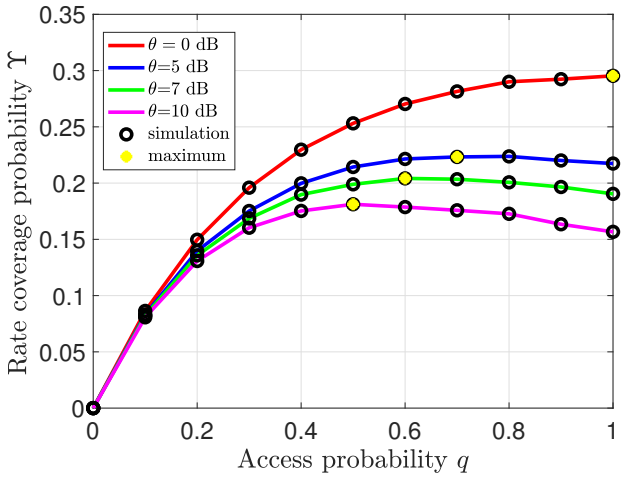



Fig. 2. The rate coverage probability $\Upsilon$ versus the access probability $q$.

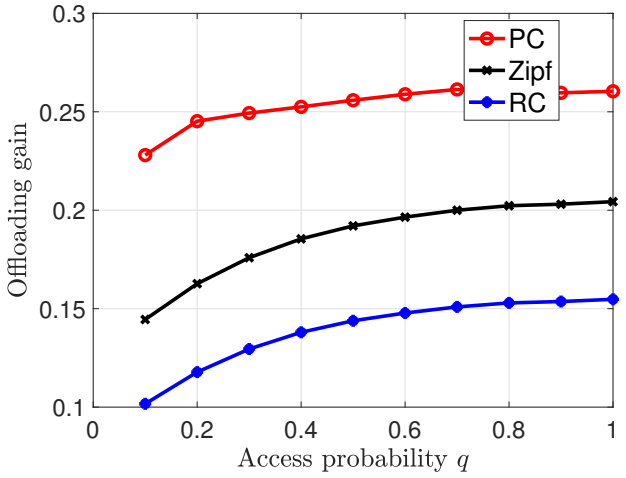
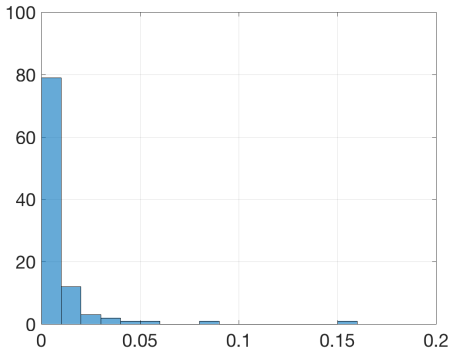


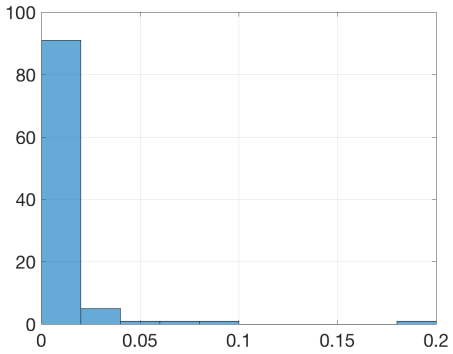

Fig. 3. The offloading gain versus the access probability $q$.

In Fig. 2, we plot the rate coverage probability $\Upsilon$ against the channel access probability $q$. The theoretical and simulated results are plotted together, and they are consistent. Clearly, there is an optimal $q^*$; before it, $\Upsilon$ tends to increase as the probability of accessing the channel increases, and beyond it, $\Upsilon$ tends to decrease due to the effect of aggressive interference. It is intuitive to observe that the optimal access probability $q^*$, which maximizes $\Upsilon$, decreases as $\vartheta$ increases. This reflects the fact the system becomes more sensitive to the effect of interference when a higher SIR threshold is required.

Fig. 3 manifests the effect of the access probability $q$ on the offloading gain. The offloading gain is plotted against $q$ for different caching schemes, namely, the proposed probabilistic caching (PC), Zipf caching (Zipf), and uniform random caching (RC). Fig. 3 is plotted for an SIR threshold $\vartheta = 0$ dB, hence, the optimal access probability $q^*$ is near one from Fig. 2. Clearly, the offloading gain for the different caching schemes improves as $q$ approaches its optimal value, which reveals the crucial impact of the device scheduling on the content placement and ac-




(a)




(b)

Fig. 4. Histogram of the optimal caching probability $\boldsymbol{b}^*$: (a) $q = q^*$ and (b) $q < q^*$.

cordingly, on the offloading gain. Moreover, the proposed PC is shown to attain the best performance as compared to other benchmark schemes.

To show the effect of $q$ on the caching probability, in Fig. 4, we plot the histogram of the optimal caching probability at different $q$ values. Specifically, $p = q^*$ in Fig. 4(a) and $q < q^*$ in Fig. 4(b). It is clear from the histograms that the optimal caching probability $\boldsymbol{b}^*$ tends to be more skewed when $q < q^*$, i.e., when $\Upsilon$ decreases. This shows that file sharing is more difficult when $q$ is not optimized. Broadly speaking, for $q < q^*$, the system is too conservative, while for $q > q^*$, the outage probability is high due to the aggressive interference. In such regimes, each device tends to cache the most popular files leading to fewer opportunities of content transfer.

Fig. 5 illustrates the prominent effect of the content popularity on the offloading gain, and compares the achievable gain of three different caching schemes. Clearly, the offloading gain of the proposed PC attains the best performance as compared to other schemes. Particularly, 10% improvement in the offloading gain is observed compared to the Zipf caching when $\beta = 1$. Moreover, we note that all caching schemes encompass the same offloading gain when $\beta = 0$ owing to the uniformity of content popularity.

To show the effect of network geometry, in Fig. 6, we plot the closed-form offloading gain in (17) against $\sigma$ at different $\lambda_p$. Fig. 6 shows that the offloading gain monotonically decreases with both $\sigma$ and $\lambda_p$. This is because content sharing between

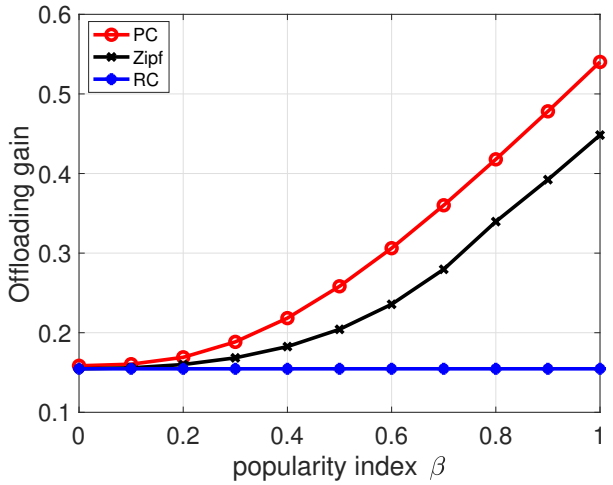



Fig. 5. The offloading gain versus the popularity of files $\beta$.

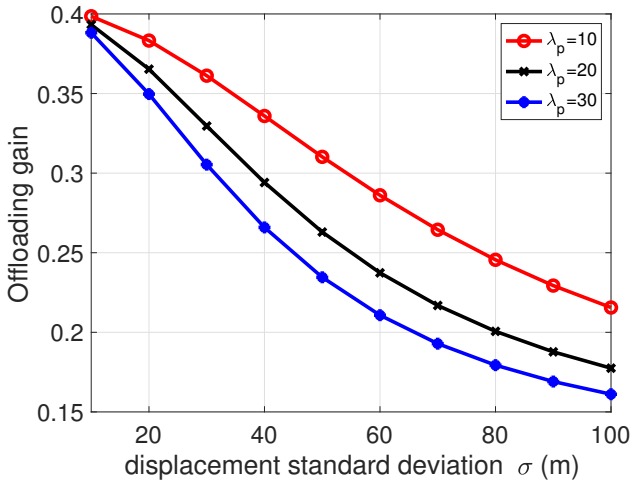



Fig. 6. The offloading gain versus the displacement standard deviation $\sigma$ at different density of clusters $\lambda_p$ (under the probabilistic caching scheme).

devices turns out to be less successful when the distance between devices is large, i.e., larger $\sigma$. Analogously, file sharing among the cluster devices is accompanied with higher interference when $\lambda_p$ and $\sigma$ are higher. Accordingly, this expected degradation prohibits successful content delivery via D2D communication.

Last, in Fig. 7, we plot the offloading gain versus the popularity index $\beta$ at different densities of cluster devices $\bar{n}$. Fig. 7 first shows that the proposed optimized probabilistic caching scheme achieves the best performance as compared to caching popular files (CPF) and Zipf caching. In addition, Fig. 7 shows that the attained offloading gain increases as the number of devices per cluster increases. This is attributed to the fact that the probability of having requested contents cached at a neighbor device within the same cluster increases when the number of cluster members is higher.

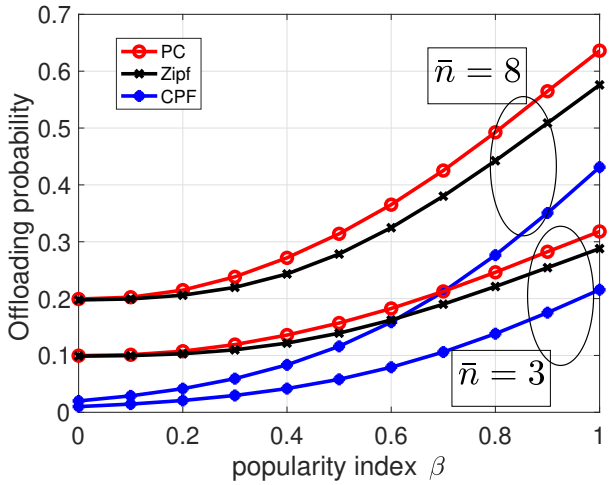



Fig. 7. The offloading gain versus the popularity index $\beta$ at different density of cluster devices $\bar{n}$.

## VI. CONCLUSION

In this paper, we have proposed a joint communication and caching optimization framework for clustered D2D networks. In particular, we have conducted joint optimization of channel access probability and content placement in order to maximize the offloading gain. We have characterized the optimal content caching scheme as a function of the system parameters, namely, density of clusters, average number of devices per cluster, content caching, placement and access probabilities. A bisection search method is also proposed to calculate the optimal channel access probability. We have demonstrated that deviating from the optimal access probability makes file sharing more difficult, i.e., the system is too conservative for small access probabilities, while the interference is too aggressive for larger access probabilities. Results showed up to $10\%$ enhancement in offloading gain compared to the Zipf caching technique.

## APPENDIX A
## PROOF OF LEMMA 1

Laplace transform of the inter-cluster aggregate interference $I_{\Phi_p^!}$ can be evaluated as

$$\mathscr{L}_{I_{\Phi_p^!}}(s) = \mathbb{E}\left[e^{-s\sum_{\Phi_p^!}\sum_{y\in\mathcal{B}^{\mathrm{II}}} g_{yx}\|x+y\|^{-\alpha}}\right]$$
$$\overset{(a)}{=} \mathbb{E}_{\Phi_p}\left[\prod_{\Phi_p^!}\mathbb{E}_{\Phi_{cq}}\prod_{y\in\mathcal{B}^{\mathrm{II}}}\frac{1}{1+s\|x+y\|^{-\alpha}}\right]$$
$$\overset{(b)}{=} \mathbb{E}_{\Phi_p}\prod_{\Phi_p^!} e^{-q\overline{n}\int_{\mathbb{R}^2}\left(1-\frac{1}{1+s\|x+y\|^{-\alpha}}\right)f_Y(y)\mathrm{dy}}$$
$$\overset{(c)}{=} e^{-\lambda_p\int_{\mathbb{R}^2}\left(1-e^{-q\overline{n}\int_{\mathbb{R}^2}\left(1-\frac{1}{1+s\|x+y\|^{-\alpha}}\right)f_Y(y)\mathrm{dy}}\right)\mathrm{dx}},$$

where (a) follows from the Rayleigh fading assumption, (b) follows from the probability generating functional (PGFL) of Gaussian PPP $\Phi_{cq}$, and (c) follows from the PGFL of the parent PPP $\Phi_p$. By using change of variables $z = x + y$ with $\mathrm{dz} = \mathrm{dy}$, we proceed as

$$\mathscr{L}_{I_{\Phi_p^!}}(s) = e^{-\lambda_p\int_{\mathbb{R}^2}\left(1-e^{-q\overline{n}\int_{\mathbb{R}^2}\left(1-\frac{1}{1+s\|z\|^{-\alpha}}\right)f_Y(z-x)\mathrm{dy}}\right)\mathrm{dx}} \quad (18)$$
$$\overset{(d)}{=} e^{-2\pi\lambda_p\int_{v=0}^{\infty}\left(1-e^{-q\overline{n}\int_{u=0}^{\infty}\left(1-\frac{1}{1+su^{-\alpha}}\right)f_U(u|v)\mathrm{du}}\right)v\mathrm{dv}}$$
$$= e^{-2\pi\lambda_p\int_{v=0}^{\infty}\left(1-e^{-q\overline{n}\int_{u=0}^{\infty}\frac{s}{s+u^{\alpha}}f_U(u|v)\mathrm{du}}\right)v\mathrm{dv}}, \quad (19)$$

where (d) follows from converting the cartesian coordinates to the polar coordinates with $u = \|z\|$. To clarify how in (d) the normal distribution $f_Y(z-x)$ is converted to the Rice distribution $f_U(u|v)$, consider a remote cluster centered at $x \in \Phi_p^!$, with a distance $v = \|x\|$ from the origin. Every interfering device belonging to the cluster centered at $x$ has its coordinates in $\mathbb{R}^2$ chosen independently from a Gaussian distribution with standard deviation $\sigma$. Then, by definition, the distance from such an interfering device to the origin, denoted as $u$, has a Rice distribution, denoted as $f_U(u|v) = u/\sigma^2\exp\big(-(\mathrm{u}^2+\mathrm{v}^2)/2\sigma^2\big)\mathrm{I}_0\big(\mathrm{uv}/\sigma^2\big)$, where $I_0$ is the modified Bessel function of the first kind with order zero and $\sigma$ is the scale parameter. Letting $\varphi(s,v) = \int_{u=0}^{\infty} s/(s+u^{\alpha})f_U(u|v)\,\mathrm{du}$, the proof is completed.

## APPENDIX B
## PROOF OF LEMMA 3

First, to prove concavity, we proceed as follows.

$$\frac{\partial\mathbb{P}_{\bowtie}}{\partial b_i} = q_i + q_i\big(\overline{n}(1-b_i)e^{-\overline{n}b_i} - (1-e^{-\overline{n}b_i})\big)\Upsilon$$
$$\frac{\partial^2\mathbb{P}_{\bowtie}}{\partial b_i\partial b_j} = -q_i\big(\overline{n}e^{-\overline{n}b_i} + \overline{n}^2(1-b_i)e^{-\overline{n}b_i} + \overline{n}e^{-\overline{n}b_i}\big)\Upsilon \quad (20)$$

It is clear that the second derivative $\frac{\partial^2\mathbb{P}_o}{\partial b_i\partial b_j}$ is negative. Hence, the Hessian matrix $\mathbf{H}_{i,j}$ of $\mathbb{P}_o(p^*,b_i)$ w.r.t. $b_i$ is negative semidefinite, and the function $\mathbb{P}_o(p^*,b_i)$ is concave with respect to $b_i$. Also, the constraints are linear, which implies that the necessity and sufficiency conditions for optimality exist. The dual Lagrangian function and the KKT conditions are then employed to solve **P2** [23]. The KKT Lagrangian function of the energy minimization problem is given by

$$\mathcal{L}(b_i,w_i,\mu_i,v) \quad (21)$$
$$= \sum_{i=1}^{N_f} q_ib_i + q_i(1-b_i)(1-e^{-b_i\overline{n}})\Upsilon$$
$$+ v(M-\sum_{i=1}^{N_f}b_i) + \sum_{i=1}^{N_f}w_i(b_i-1) - \sum_{i=1}^{N_f}\mu_ib_i, \quad (22)$$

where $v, w_i$, and $\mu_i$ are the dual equality and two inequality constraints, respectively. Now, the optimality conditions are written

as $\nabla_{b_i}\mathcal{L}(b_i^*, w_i^*, \mu_i^*, v^*) =$

$$q_i + q_i\big(\overline{n}(1-b_i)e^{-\overline{n}b_i} - (1-e^{-\overline{n}b_i})\big)\Upsilon - v^* + w_i^* - \mu_i^* = 0$$

$$w_i^* \geq 0 \tag{23}$$

$$\mu_i^* \leq 0 \tag{24}$$

$$w_i^*(b_i^* - 1) = 0 \tag{25}$$

$$\mu_i^* b_i^* = 0 \tag{26}$$

$$(M - \sum_{i=1}^{N_f} b_i^*) = 0. \tag{27}$$

The optimality conditions imply that:

1. $w_i^* > 0$: We have $b_i^* = 1$, $\mu_i^* = 0$, and

$$q_i - q_i(1-e^{-\overline{n}})\Upsilon = v^* - w_i^*,$$
$$v^* < q_i - q_i(1-e^{-\overline{n}})\Upsilon \tag{28}$$

2. $\mu_i^* < 0$: We have $b_i^* = 0$, and $w_i^* = 0$, and

$$q_i + \overline{n}q_i\Upsilon = v^* + \mu_i^*,$$
$$v^* > q_i + \overline{n}q_i\Upsilon \tag{29}$$

3. $0 < b_i^* < 1$: We have $w_i^* = \mu_i^* = 0$, and

$$v^* = q_i + q_i\big(\overline{n}(1-b_i^*)e^{-\overline{n}b_i} - (1-e^{-\overline{n}b_i})\big)\Upsilon. \tag{30}$$

By combining (28), (29), and (30), with the fact that $\sum_{i=1}^{N_f} b_i^* = M$, Lemma 3 is proven.

## REFERENCES


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

[2] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[3] K. Shanmugam *et al.*, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[4] N. Golrezaei *et al.*, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, July 2014.

[5] R. Amer, M. M. Butt, M. Bennis, and N. Marchetti, "Inter-cluster cooperation for wireless D2D caching networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6108–6121, Sept. 2018.

[6] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[7] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE ICC*, June 2015.

[8] S. Andreev *et al.*, "Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 67–80, 2015.

[9] B. Chen, C. Yang, and Z. Xiong, "Optimal caching and scheduling for cache-enabled D2D communications," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1155–1158, May 2017.

[10] S. H. Chae, T. Q. S. Quek, and W. Choi, "Content placement for wireless cooperative caching helpers: A tradeoff between cooperative gain and content diversity gain," *IEEE Trans.Wireless Commun.*, vol. 16, no. 10, pp. 6795–6807, Oct. 2017.

[11] M. Lee, A. F. Molisch, N. Sastry, and A. Raman, "Individual preference probability modeling and parameterization for video content in wireless caching networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 676–690, Apr. 2019.

[12] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Modeling and performance analysis of clustered device-to-device networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4957–4972, July 2016.

[13] ——, "Fundamentals of cluster-centric content placement in cache-enabled device-to-device networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2511–2526, June 2016.

[14] N. Deng and M. Haenggi, "The benefits of hybrid caching in gauss-poisson D2D networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1217–1230, June 2018.

[15] R. Amer *et al.*, "Optimized caching and spectrum partitioning for D2D enabled cellular systems with clustered devices," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4358–4374, July 2020.

[16] R. Amer, H. Elsawy, J. Kibiłda, M. M. Butt, and N. Marchetti, "Performance analysis and optimization of cache-assisted CoMP for clustered D2D networks," *IEEE Trans. Mobile Comput.*, 2020, early Access.

[17] R. Amer, W. Saad, H. ElSawy, M. Butt, and N. Marchetti, "Caching to the sky: Performance analysis of cache-assisted CoMP for cellular-connected UAVs," in *Proc. IEEE WCNC*, Apr. 2019.

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

[19] R. Amer, H. ElSawy, J. Kibiłda, M. M. Butt, and N. Marchetti, "Cooperative transmission and probabilistic caching for clustered D2D networks," in *Proc. IEEE WCNC*, Apr. 2019.

[20] R. Amer *et al.*, "On minimizing energy consumption for D2D clustered caching networks," in *Proc. IEEE GLOBECOM*, Dec. 2018.

[21] A. Rachedi, M. H. Rehmani, S. Cherkaoui, and J. J. P. C. Rodrigues, "The plethora of research in Internet of things (IoT)," *IEEE Access*, vol. 4, pp. 9575–9579, 2016.

[22] L. Breslau, Pei Cao, Li Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *Proc. IEEE INFOCOM*, Mar. 1999.

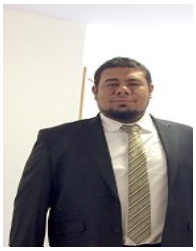
[23] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.


**Ramy Amer** received his Ph.D. in Electrical and Communications Engineering from Trinity College Dublin in November 2020. He was a Visiting Scholar at Virginia Tech, USA from September 2018 to March 2019. He holds one best paper award, IEEE student travel grant, and IEEE exemplary reviewer award. His research interests include edge caching and Intelligence, edge computing and IoT, machine learning, and UAVs. Ramy has been first author in one book chapter and six journals and eight IEEE conference papers.

**Mohamed Baza** is currently an Assistant Professor at the Department of Computer Science at College of Charleston, SC, USA. He received his Ph.D. degree in Electrical and Computer Engineering from Tennessee Tech University, Cookeville, Tennessee, United States in Dec. 2020 and B.S. and M.S. degrees in Electrical & Computer Engineering from Benha University, Egypt in 2012 and 2017 respectively. From August 2020 to May 2021, he worked as a Visiting Assistant Professor at the Department of Computer Science at Sam Houston State University, TX, USA. He also has more than two years of industry experience in information security in Apache-khalda petroleum company, Egypt. Dr. Baza is the Author of numerous papers published in major IEEE conferences and journals, such as IEEE Wireless Communications and Networking Conference (IEEE WCNC), IEEE International Conference on Communications (IEEE ICC), IEEE Vehicular Technology Conference (IEEE VTC), IEEE Transactions on Dependable and Secure Computing, IEEE Transactions of Vehicular Technology (TVT), IEEE Transactions on Network Science and Engineering (TNSE), and IEEE Systems Journal. He served as a Reviewer for several journals and conferences such as IEEE Transactions on Vehicular Technology, IEEE IoT Journal, and the journal of Peer-to-Peer Networking. His research interests include blockchains, cyber-security, machine learning, smart-grid, and vehicular ad-hoc networks. He also a recipient of best IEEE paper award in the International Conference on Smart Applications, Communications and Networking (SmartNets 2020).

**Tara Salman** is finishing a graduate Research Assistant at Washington University in St. Louis. She received her B.S. in Computer Engineering and M.S. degrees in Computer Networking from Qatar University Doha, Qatar in 2012 and 2015, respectively. She is currently pursuing a Ph.D. in Computer Science & Engineering at Washington University in St Louis, Missouri, USA. From 2012 -2015, she worked as a Research Assistant with Qatar University on an NPRP (National Priorities Research Program) funded project targeting physical layer security. Since 2015, she has been working as a Graduate Research Assistant at Washington University in St. Louis. Her research interests span network security, distributed systems, the Internet of things, and financial technology. She is an Author of 1 book chapter and numerous papers published at major IEEE conferences and journals. Tara Salman is an EECS Rising Star in UC Berkeley 2020, is a Recipient of the Cisco Certified Network Associate (CCNA) certification in 2012, and has completed all four levels of CCNA at Cisco academy-Qatar university branch.

**M. Majid Butt** received the M.Sc. degree in Digital Communications from Christian Albrechts University, Kiel, Germany, in 2005, and the Ph.D. degree in Telecommunications from the Norwegian University of Science and Technology, Trondheim, Norway, in 2011. He is currently a Senior Research Specialist 5G+ at Nokia Bell Labs, Paris-Saclay, France, and also an adjunct Research Professor at Trinity College Dublin, Dublin, Ireland. Prior to that, he has held various positions at the University of Glasgow, U.K., Trinity College Dublin, Ireland, Fraunhofer HHI, Germany, and the University of Luxembourg. His current research interests include communication techniques for wireless networks with a focus on radio resource allocation, scheduling algorithms, energy efficiency, and machine learning for RAN. He has Authored more than 70 peer-reviewed conference and journal publications, and filed some 10 patents in these areas. He frequently gives invited talks, as well as technical tutorial talks on various topics in IEEE conferences, including ICC, Globecom, PIMRC, VTC, ISWCS, etc. He was a Recipient of the Marie Curie Alain Bensoussan Post-Doctoral Fellowship from the European Research Consortium for Informatics and Mathematics. He has served as the Organizer/chair for technical workshops on various aspects of communication systems in conjunction with major IEEE conferences. He has been an Associate Editor of the IEEE ACCESS and the IEEE Communication Magazine since 2016.



**Ahmad Alhindi** received the B.Sc. degree in Computer Science from Umm Al-Qura University (UQU), Makkah, Saudi Arabia, in 2006, and the M.Sc. degree in Computer Science and the Ph.D. degree in Computing and Electronic Systems from the University of Essex, Colchester, U.K., in 2010 and 2015, respectively. He is currently an Assistant Professor in Artificial Intelligence (AI) with Computer Science Department, UQU. His current research interests include evolutionary multi-objective optimization and machine learning techniques. He is currently involved in AI algorithms, focusing particularly on machine learning and optimization with a willingness to implement them in a context of decision making and solving combinatorial problems in real-world projects.

**Nicola Marchetti** is Associate Professor in Wireless Communications at Trinity College Dublin, Ireland. He performs his research under the Irish Research Centre for Future Networks and Communications (CONNECT), where he leads the Wireless Engineering and Complexity Science (WhyCOM) lab. He received the Ph.D. in Wireless Communications from Aalborg University, Denmark in 2007, and the M.Sc. in Electronic Engineering from University of Ferrara, Italy in 2003. He also holds an M.Sc. in Mathematics which he received from Aalborg University in 2010. His research interests include Complex Systems Science, Self-Organizing Networks, Signal Processing for Communication, and Radio Resource Management. He has authored in excess of 140 journals and conference papers, 2 books and 8 book chapters, holds 4 patents, and received 4 best paper awards. He is a senior member of IEEE and serves as an associate editor for the IEEE Internet of Things Journal since 2018.