

Mobility Increases the Data Offloading Ratio in D2D Caching Networks

Rui Wang*, Jun Zhang*, S.H. Song* and K. B. Letaief*[†], *Fellow, IEEE*

*Dept. of ECE, The Hong Kong University of Science and Technology, [†]Hamad Bin Khalifa University, Doha, Qatar

Email: *{rwangae, eejzhang, eeshsong, eekhaled}@ust.hk, [†]kletaief@hbku.edu.qa

Abstract—Caching at mobile devices, accompanied by device-to-device (D2D) communications, is one promising technique to accommodate the exponentially increasing mobile data traffic. While most previous works ignored user mobility, there are some recent works taking it into account. However, the duration of user contact times has been ignored, making it difficult to explicitly characterize the effect of mobility. In this paper, we adopt the alternating renewal process to model the duration of both the contact and inter-contact times, and investigate how the caching performance is affected by mobility. The *data offloading ratio*, i.e., the proportion of requested data that can be delivered via D2D links, is taken as the performance metric. We first approximate the distribution of the *communication time* for a given user by beta distribution through moment matching. With this approximation, an accurate expression of the data offloading ratio is derived. For the homogeneous case where the average contact and inter-contact times of different user pairs are identical, we prove that the data offloading ratio increases with the user moving speed, assuming that the transmission rate remains the same. Simulation results are provided to show the accuracy of the approximate result, and also validate the effect of user mobility.

I. INTRODUCTION

The mobile data traffic is growing at an exponential rate, among which mobile video accounts for more than a half [1]. Caching popular contents at helper nodes or user devices is a promising approach to reduce the data traffic on the backhaul links, as well as improving the user experience of video streaming applications [2], [3]. In comparison with the commonly considered femto-caching system, caching at devices enjoys a unique advantage, i.e., the devices' aggregate caching capacity grows with the number of devices [2]. Moreover, device caching can promote device-to-device (D2D) communications, where nearby mobile devices may communicate directly rather than being forced to communicate through the base station (BS)[4].

Recently, caching in D2D networks has attracted lots of attentions. In [5], the scaling behavior of the number of D2D collaborating links was identified. Three concentration regimes, classified by the concentration of the file popularity, were investigated. The outage-throughput tradeoff and optimal scaling laws of both the throughput and outage probability were studied in [6]. Two coded caching schemes, i.e., centralized and decentralized, were proposed in [7], where the contents are delivered via broadcasting.

So far, an important characteristic of mobile users, i.e., the user mobility, has been ignored in previous studies of D2D caching networks. There are some works starting to consider the effect of user mobility. Effective methodologies to utilize the user mobility information in caching design were discussed in [8]. In [9], the effect of mobility was evaluated in D2D networks with coded caching, with the conclusion that mobility can improve the scaling law of throughput. This result was based on the assumption that the user locations are random and independent in each time slot, which failed to take into account the temporal correlation.

The inter-contact model, which considers the temporal correlation of the user mobility, has been widely applied [10], where the timeline for an arbitrary pair of mobile users are divided into *contact times* and *inter-contact times*. Specifically, the *contact times* denote the time intervals when the mobile users are located within the transmission range. Correspondingly, the *inter-contact times* denote the time intervals between contact times [11]. This model has been used to develop device caching schemes to exploit the user mobility pattern in [12]. The throughput-delay scaling law was developed by characterizing the inter-contact pattern of the random walk model [13]. In these works, it was assumed that a fixed amount of data can be delivered within one contact time, while the duration of the contact times was not considered. However, as the user moving speed will affect the durations of both the contact and inter-contact times, it is critical to account for their effects when investigating the impact of user mobility on caching performance.

In this paper, we shall analytically evaluate the effect of mobility in D2D caching networks, by adopting an alternating renewal process to model the mobility pattern so that both the contact and inter-contact times are accounted for. The *data offloading ratio*, which is defined as the proportion of data that can be obtained via D2D links, is adopted as the performance metric. The main contribution is an approximate expression for the data offloading ratio, for which the main difficulty is to deal with multiple alternating renewal processes. We tackle it by first deriving the expectation and variance of the *communication time* of a given user, and then use a beta random variable to approximate it by moment matching. Furthermore, we investigate the effect of mobility in a homogeneous case, where the average contact and inter-contact times for all the user pairs are the same. In the low-to-medium mobility scenario, by assuming that the transmission rate is irrelevant to the user speed, it is proved that the data offloading ratio increases with the user speed for any caching strategy that

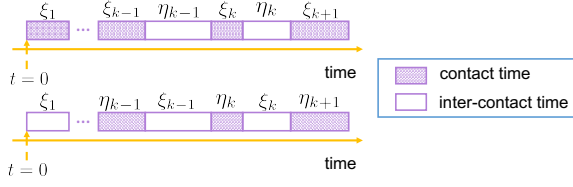


Fig. 1. The timeline for an arbitrary pair of mobile users.

does not cache the same contents at all devices. Simulation results validates the accuracy of the derived expression, as well as the effect of the user mobility.

II. SYSTEM MODEL AND PERFORMANCE METRIC

In this section, we will first introduce the alternating renewal process to model the user mobility pattern, and discuss the caching and file delivery models. Then, the performance metric, i.e., the data offloading ratio, will be defined.

A. User Mobility Model

The inter-contact model, which captures the temporal correlation of the user mobility [10], is used to model the user mobility pattern. Specifically, the timeline of each pair of users is divided into *contact times*, i.e., the times when the users are in the transmission range, and *inter-contact times*, i.e., the times between consecutive contact times. Considering that contact times and inter-contact times appear alternatively in the timeline of a pair of users, similar to [14], an alternating renewal process is applied to model the pairwise contact pattern, as defined below [15].

Definition 1. Consider a stochastic process with state space $\{A, B\}$, and the successive durations for the system to be in states A and B are denoted as $\xi_k, k = 1, 2, \dots$ and $\eta_k, k = 1, 2, \dots$, respectively, which are i.i.d.. Specifically, the system starts at state A and remains for ξ_1 , then switches to state B for η_1 , then backs to state A for ξ_2 , and so forth. Let $\psi_k = \xi_k + \eta_k$. The counting process of ψ_k is called as an *alternating renewal process*.

As shown in Fig. 1, if the pair of users is in contact at $t = 0$, ξ_k and η_k represent the contact times and inter-contact times, respectively; otherwise, ξ_k and η_k represent the inter-contact times and contact times, respectively. It was shown in [16] that exponential curves well fit the distribution of inter-contact times, while in [17], it was identified that exponential distribution is a good approximation for the distribution of the contact times. Thus, same as [14], we assume that the contact times and inter-contact times follows independent exponential distributions. For simplicity, the timelines of different user pairs are assumed to be independent. Specifically, we consider N_u users in a network, and the index set of the users is denoted as $\mathcal{S} = \{1, 2, \dots, N_u\}$. The contact times and inter-contact times of users $i \in \mathcal{S}$ and $j \in \mathcal{S} \setminus \{i\}$ follow independent exponential distributions with parameters $\lambda_{i,j}^C$ and $\lambda_{i,j}^I$, respectively.

B. Caching and File Transmission Model

There is a library with N_f files, whose index set is denoted as $\mathcal{F} = \{1, 2, \dots, N_f\}$, each with size C . Each user device

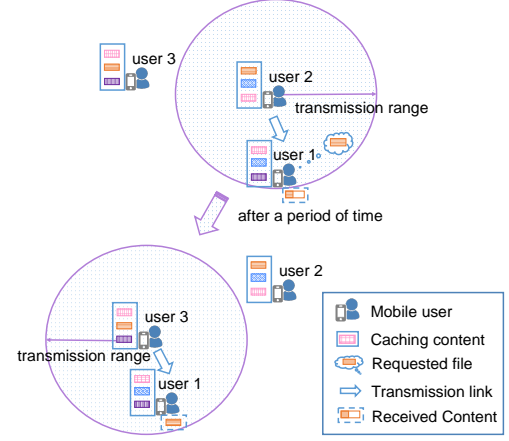


Fig. 2. A sample network with three mobile users.

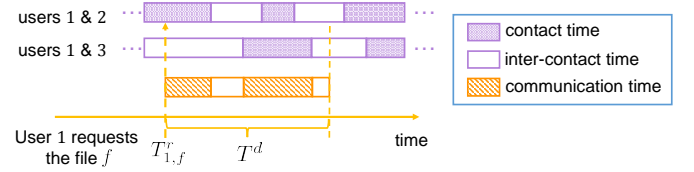


Fig. 3. An illustration of the communication time.

has a limited storage capacity, and each file can be completely cached or not cached at all at each user device. Specifically, the caching placement is denoted as

$$x_{j,f} = \begin{cases} 1, & \text{if user } j \text{ caches file } f, \\ 0, & \text{if user } j \text{ does not cache file } f, \end{cases} \quad (1)$$

where $j \in \mathcal{S}$ and $f \in \mathcal{F}$. User $i \in \mathcal{S}$, is assumed to request a file $f \in \mathcal{F}$ with probability $p_{i,f}^r$, where $\sum_{f \in \mathcal{F}} p_{i,f}^r = 1$. When a user requests a file f , it will first check its own cache, and then download the file from the users that are in contact and store file f , with a fixed transmission rate, denoted as R . If the user cannot get the whole file within a certain delay threshold, denoted as T^d , it will download the remaining part from the BS. We assume that the delay threshold is larger than the time required to download each content, i.e., $T^d > \frac{C}{R}$. Fig. 2 shows a sample network, where user 1 gets part of the requested file during the contact time with user 2, then gets the whole file after the contact time with user 3.

C. Performance metric

The *data offloading ratio*, which is defined as the percentage of requested content that can be obtained via D2D links rather than downloading from the BS, is used as the performance metric. Specifically, the data offloading ratio for user $i \in \mathcal{S}$ is defined as

$$\mathcal{P}_i = \sum_{f \in \mathcal{F}} p_{i,f}^r \left\{ x_{i,f} + \frac{(1 - x_{i,f}) \mathbb{E}_{D_{i,f}} [\min(D_{i,f}, C)]}{C} \right\}, \quad (2)$$

where $D_{i,f}$ denotes the amount of requested data that can be delivered via D2D links when user i requests file f . Since a fixed transmission rate is assumed, $D_{i,f}$ can be written as

$D_{i,f} = RT_{i,f}^c$, where $T_{i,f}^c$ is the *communication time* for user i to download file f from other users caching file f within time T^d . We assume that user i can download file f while at least one user caching file f is in contact, where the handover time is ignored. Fig. 3 shows the communication time of user 1 in Fig. 2. Then, the average data offloading ratio is

$$\mathcal{P} = \frac{1}{N_u} \sum_{i \in \mathcal{S}} \sum_{f \in \mathcal{F}} p_{i,f}^r \left\{ x_{i,f} + \frac{(1 - x_{i,f}) \mathbb{E}_{T_{i,f}^c} [\min(RT_{i,f}^c, C)]}{C} \right\}. \quad (3)$$

In the following, we will evaluate the data offloading ratio given in (3) for any given caching strategy, and investigate the effect of user mobility on caching performance.

III. DATA OFFLOADING RATIO ANALYSIS

The main difficulty of evaluating the data offloading ratio is to find the distribution of the communication time. As this distribution is highly complicated, instead of deriving it directly, we will develop an accurate approximation. In this section, we will first approximate the distribution of the communication time by a beta distribution, and then, an approximation of the data offloading ratio will be obtained.

A. Communication time analysis

To help analyze the communication time, we first define some stochastic processes.

Definition 2. Define $\mathbf{H}_{i,j}$, where $i \in \mathcal{S}$ and $j \in \mathcal{S} \setminus \{i\}$, as the continuous-time random process, i.e., $\mathbf{H}_{i,j} = \{H_{i,j}(t), t \in (0, \infty)\}$ with state space $\{0, 1\}$, where $H_{i,j}(t) = 1$ means that users i and j are in contact at the time instant t ; otherwise $H_{i,j}(t) = 0$. The durations of staying in states 1 and 0 follow i.i.d. exponential distributions with parameter $\lambda_{i,j}^C$ and $\lambda_{i,j}^I$, respectively.

Definition 3. Define \mathbf{H}_i^f , where $i \in \mathcal{S}$ and $f \in \mathcal{F}$, as the continuous-time random process, i.e., $\mathbf{H}_i^f = \{H_i^f(t), t \in (0, \infty)\}$ with state space $\{0, 1\}$, where $H_i^f(t) = 1$ means that users i can download file f from any other user caching file f at time instant t ; otherwise $H_i^f(t) = 0$.

At time t , since user i can download file f when at least one user caching file f is in contact, we get $H_i^f(t) = 1 - \prod_{j \in \mathcal{S} \setminus \{i\}, x_{j,f}=1} [1 - H_{i,j}(t)]$. Similar to [14], it is reasonable to assume that when a user requests a file, the alternating process between each pair of users has been running for a long time. Thus, denote $T_{i,f}^r$, $i \in \mathcal{S}$ and $f \in \mathcal{F}$, as the time of user i requests file f , and the communication time $T_{i,f}^c$ can be derived as $T_{i,f}^c = \lim_{T_{i,f}^r \rightarrow \infty} \int_{T_{i,f}^r}^{T_{i,f}^r + T^d} H_i^f(t) dt$. In the following, we will derive the expectation and variance of the communication time.

Lemma 1. When user $i \in \mathcal{S}$ requests file $f \in \mathcal{F}$, which is not stored at its own cache, the expectation and variance of its communication time is

$$\mathbb{E}[T_{i,f}^c] = T^d \left(1 - \prod_{j \in \mathcal{S}, x_{j,f}=1} \frac{\lambda_{i,j}^C}{\lambda_{i,j}^C + \lambda_{i,j}^I} \right). \quad (4)$$

and

$$\begin{aligned} \text{Var}[T_{i,f}^c] = & 2 \int_0^{T^d} (T^d - u) \prod_{j \in \mathcal{S}, x_{j,f}=1} \frac{\lambda_{i,j}^C}{(\lambda_{i,j}^I + \lambda_{i,j}^C)^2} \\ & \times [\lambda_{i,j}^C + \lambda_{i,j}^I e^{-u(\lambda_{i,j}^C + \lambda_{i,j}^I)}] du \\ & - (T^d)^2 \prod_{j \in \mathcal{S}, x_{j,f}=1} \left(\frac{\lambda_{i,j}^C}{\lambda_{i,j}^C + \lambda_{i,j}^I} \right)^2. \end{aligned} \quad (5)$$

Proof: See Appendix A. \blacksquare

Since the communication time $T_{i,f}^c$ is a bounded random variable, we propose to approximate its distribution by a beta distribution, which is widely used to model the random variables limited to finite ranges. Specifically, we consider $T_{i,f}^c \approx T^d Y_{i,f}$, where $Y_{i,f} \sim B(\alpha_{i,f}, \beta_{i,f})$, $i \in \mathcal{S}$ and $f \in \mathcal{F}$, if $\sum_{j \in \mathcal{S} \setminus \{i\}} x_{j,f} > 0$, which means that user i may download file f from at least one user; otherwise, $T_{i,f}^c = 0$. Let $\mathbb{E}[T^d Y_{i,f}] = \mathbb{E}[T_{i,f}^c]$ and $\text{Var}[T^d Y_{i,f}] = \text{Var}[T_{i,f}^c]$, and the parameters of the beta distribution to match the first two moments can be derived as¹

$$\begin{cases} \alpha_{i,f} = \frac{\mathbb{E}[T_{i,f}^c]^2 (T^d - \mathbb{E}[T_{i,f}^c])}{\text{Var}[T_{i,f}^c] T^d} - \frac{\mathbb{E}[T_{i,f}^c]}{T^d} \\ \beta_{i,f} = \frac{T^d - \mathbb{E}[T_{i,f}^c]}{\mathbb{E}[T_{i,f}^c]} \alpha_{i,f} \end{cases} \quad (6)$$

B. Data offloading ratio approximation

Based on the above approximation, we get an approximate expression of the data offloading ratio in Proposition 1. Simulations will show that the approximation is quite accurate.

Proposition 1. The data offloading ratio is approximated as

$$\mathcal{P} = \frac{1}{N_u} \sum_{i \in \mathcal{S}} \sum_{f \in \mathcal{F}} p_{i,f}^r [x_{i,f} + (1 - x_{i,f}) \mathcal{P}_{i,f}], \quad (7)$$

where $\mathcal{P}_{i,f}$ is the data offloading ratio when user i requests file f , which is not in its own cache, approximated by

$$\begin{aligned} \mathcal{P}_{i,f} \approx & 1 - I_{\frac{C}{T^d R}}(\alpha_{i,f}, \frac{T^d - \mathbb{E}[T_{i,f}^c]}{\mathbb{E}[T_{i,f}^c]} \alpha_{i,f}) \\ & + \frac{\mathbb{E}[T_{i,f}^c] R}{C} I_{\frac{C}{T^d R}}(\alpha_{i,f} + 1, \frac{T^d - \mathbb{E}[T_{i,f}^c]}{\mathbb{E}[T_{i,f}^c]} \alpha_{i,f}) \end{aligned} \quad (8)$$

if $\sum_{j \in \mathcal{S} \setminus \{i\}} x_{j,f} > 0$ and 0 elsewhere, where $I_r(\cdot, \cdot)$ is the incomplete beta function, and $\alpha_{i,f}$ is given in (6).

Proof: Following (3), (4), (5), and (6), the expression in (7) can be obtained. Due to the space limitation, the detail is omitted. \blacksquare

IV. EFFECT OF MOBILE USER SPEED

In this section, we will consider a homogeneous case, where the contact and inter-contact parameters among all pairs of users are the same, i.e., $\lambda^C = \lambda_{i,j}^C > 0$ and $\lambda^I = \lambda_{i,j}^I > 0$, where $i \in \mathcal{S}$ and $j \in \mathcal{S} \setminus \{i\}$. We will investigate how the

¹The parameters of the beta distribution should be positive, and it can be proved that $\alpha_{i,f} > 0$ and $\beta_{i,f} > 0$, by $e^{-u(\lambda_{i,j}^I + \lambda_{i,j}^C)} \leq 1$. The detail is omitted due to the space limitation.

user moving speed affects the data offloading ratio for a given caching strategy. If all users cache the same contents, the D2D communications will not help the content delivery. Thus, in the following, we assume that the contents cached at different users are not all the same. This investigation will be based on the approximate expression in (7), and simulations will be provided later to verify the results.

A. Communication time analysis

Under the above assumptions, the expectation and variance of the communication time can be simplified, as in the following corollary.

Corollary 1. When $\lambda^C = \lambda_{i,j}^C$ and $\lambda^I = \lambda_{i,j}^I$, where $i \in \mathcal{S}$ and $j \in \mathcal{S} \setminus \{i\}$, the expectation and variance of a user requests file f , which is not stored at its own cache, are given by

$$\mathbb{E}[T_{i,f}^c] = T^d \left[1 - \left(\frac{\lambda^C}{\lambda^C + \lambda^I} \right)^{n_f} \right], \quad (9)$$

$$\begin{aligned} \text{Var}[T_{i,f}^c] &= \left[\frac{\lambda^C}{(\lambda^C + \lambda^I)^2} \right]^{n_f} \sum_{l=1}^{n_f} \binom{n_f}{l} \frac{(\lambda^C)^{n_f-l} (\lambda^I)^l}{l(\lambda^C + \lambda^I)} \\ &\times \left[T^d - \frac{1}{l(\lambda^C + \lambda^I)} + \frac{e^{-l(\lambda^C + \lambda^I)T}}{l(\lambda^C + \lambda^I)} \right], \end{aligned} \quad (10)$$

where $i \in \mathcal{S}$ and $n_f = \sum_{j \in \mathcal{S}} x_{j,f}$ denotes the number of users caching file f .

Proof: See Appendix A. ■

B. Mobile user speed

We first characterize the relationship between the user speed and the parameters λ^C and λ^I in Lemma 2.

Lemma 2. When all the user speeds change by s times, the contact and inter-contact parameters will also change by s times, i.e., from λ^C and λ^I to $s\lambda^C$ and $s\lambda^I$, respectively.

Proof: The time for user i to move along a certain path L_i can be given as a curve integral $\int_{L_i} \frac{dz}{v_i(z)}$, where $v_i(z)$ is the speed of user i when passing by a point z on the path L_i . When the speed of user i changes by s times, the time for moving along the path L_i changes to $\int_{L_i} \frac{dz}{sv_i(z)} = \frac{1}{s} \int_{L_i} \frac{dz}{v_i(z)}$, which scales by $\frac{1}{s}$ times. During each contact or inter-contact time, users i and j move along certain paths. When all the user speeds change by s times, each contact or inter-contact time changes by $\frac{1}{s}$ times, and thus, the average ones change by $\frac{1}{s}$ times. Since the contact and inter-contact times are assumed to be exponential distributed with mean $\frac{1}{\lambda^C}$ and $\frac{1}{\lambda^I}$, respectively, the parameters λ^C and λ^I scale by s times. ■

Considering that a larger s means that users are moving faster, in the following, we will investigate how changing s will affect the data offloading ratio. For simplicity, we assume that the transmission rate is a constant, and will not change with the user speed. This is reasonable in the low-to-medium mobility regime. Firstly, the effect of user speed on the communication time is shown in Lemma 3.

Lemma 3. When s increases, which is equivalent to increasing the user speed, the expectation of the communication time

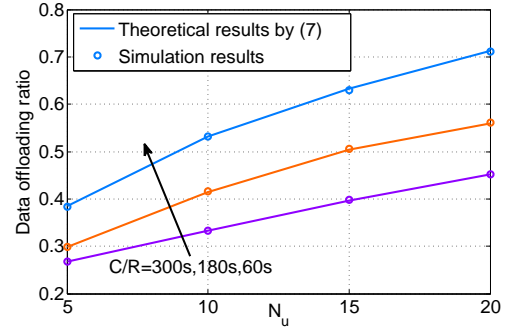


Fig. 4. Data offloading ratio with $N_f = 100$, $T^d = 300s$ and $\gamma_r = 0.6$.

when a user $i \in \mathcal{S}$ requests file $f \in \mathcal{F}$ that is not in its own cache, i.e., $\mathbb{E}[T_{i,f}^c]$, keeps the same, and the corresponding variance, i.e., $\text{Var}[T_{i,f}^c]$, decreases, if the number of users caching file f is larger than 0, i.e., $n_f > 0$. Accordingly, the parameter $\alpha_{i,f}$ of the beta distribution increases.

Proof: See Appendix B. ■

Then, we evaluate the relationship between $\alpha_{i,f}$ and the data offloading ratio when user i requests file f that is not in its own cache, i.e., $\mathcal{P}_{i,f}$ in (8), in Lemma 4.

Lemma 4. When user $i \in \mathcal{S}$ requests file $f \in \mathcal{F}$ and cannot find it in its own cache, the data offloading ratio, i.e., $\mathcal{P}_{i,f}$, increases with $\alpha_{i,f}$.

Proof: See Appendix C. ■

Base on Lemmas 3 and 4, we can specify the effect of user speed on the data offloading ratio in Proposition 2.

Proposition 2. If the transmission rate does not change with the user speed, and the average contact and inter-contact times among all the pairs are the same, the data offloading ratio increases with the user moving speed.

Proof: See Appendix D. ■

Remark. The result in Proposition 2 is valid for any caching strategy, only excluding the case that all the users have the same cache contents.

V. SIMULATION RESULTS

In the simulation, the content request probability follows a Zipf distribution with parameter γ_r , i.e., $p_f = \frac{f^{-\gamma_r}}{\sum_{i \in \mathcal{F}} i^{-\gamma_r}}$, $f \in \mathcal{F}$ [2]. Meanwhile, each user caches 5 contents, and a random caching strategy is applied [18], where the probabilities of the contents cached at each user are proportional to the file request probabilities.

Fig. 4 validates the accuracy of the approximation in (7). The inter-contact parameters $\lambda_{i,j}^I$, $i \in \mathcal{S}$, $j \in \mathcal{S} \setminus \{i\}$ are generated according to a gamma distribution as $\Gamma(4.43, 1/1088)$ [19]. Similar as [14], we assume the average of the contact parameters are 5 times larger than the inter-contact parameters. Thus, the contact parameters are generated according to $\Gamma(4.43 \times 25, 1/1088/5)$. It is shown from Fig. 4 that the theoretical results are very close to the simulation results,

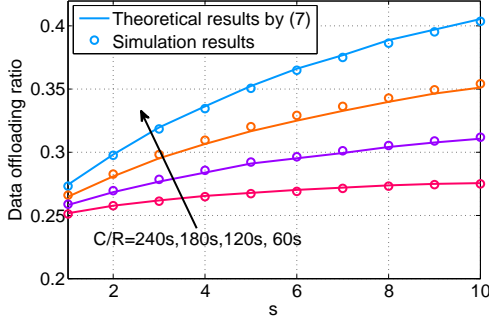


Fig. 5. Data offloading ratio with $N_u = 15$, $\lambda^C = 0.001s$, $\lambda^I = 0.0002s$, $N_f = 100$, $T^d = 300s$ and $\gamma_r = 0.6$.

which means the approximate expression (7) is quit accurate. Furthermore, the data offloading ratio increases with the number of users, which is brought by the increasing aggregate caching capacity and the content sharing via D2D links.

In Fig. 5, the effect of s is demonstrated, where increasing s is equivalent to increasing the user speed. Firstly, the small gap between the theoretical and simulation results again verifies the accuracy of the approximate expression in (7). It is also shown that the data offloading ratio increases with s , which confirms the conclusion in Proposition 2. Moreover, from Fig. 5, the increasing rate of the data offloading ratio is decreases with the user moving speed.

VI. CONCLUSIONS

In this paper, we investigated the effect of user mobility on the caching performance in a D2D caching network. The communication time of a given user was firstly approximated by a beta distribution, through matching the first two moments. Then, an approximate expression of the data offloading ratio was derived. For a homogeneous case, where the average contact and inter-contact times are the same for all the user pairs, we evaluated how the user moving speed affects the data offloading ratio. Specifically, it was proved that the data offloading ratio increases with the user speed, assuming that the transmission rate is irrelevant to the user speed. Simulation results validated the accuracy of the approximate expression of the data offloading ratio, and demonstrated that the data offloading ratio increases with the user speed, while the increasing rate decreases with the user speed.

APPENDIX

A. Proof of Lemma 1 and Corollary 1

As the timeline of different user pairs are independent, the expectation of the communication time when user i requests file f , which is not in its own cache, can be written as

$$\mathbb{E}[T_{i,f}^c] = \lim_{T_{i,f}^r \rightarrow \infty} \int_{T_{i,f}^r}^{T_{i,f}^r + T^d} \left[1 - \prod_{j \in \mathcal{S}_{x_j, f=1}} (1 - \mathbb{E}H_{i,j}(t)) \right] dt. \quad (11)$$

Since the timeline between each pair of users is modeled as an alternating renewal process, according to [15], we have $\lim_{t \rightarrow \infty} \Pr[H_{i,j}(t) = 1] = \frac{\lambda_{i,j}^I}{\lambda_{i,j}^C + \lambda_{i,j}^I}$. Thus, $\lim_{t \rightarrow \infty} \mathbb{E}[H_{i,j}(t)] =$

$\frac{\lambda_{i,j}^I}{\lambda_{i,j}^C + \lambda_{i,j}^I}$, and then, the expectation in (4) can be obtained. Let $\lambda^C = \lambda_{i,j}^C$ and $\lambda^I = \lambda_{i,j}^I$, and we can get the expression in (9). The variance of the communication time is

$$\begin{aligned} \text{Var}[T_{i,f}^c] &= \\ 2 \lim_{T_{i,f}^r \rightarrow \infty} &\int_{T_{i,f}^r}^{T_{i,f}^r + T^d} \int_{T_{i,f}^r}^{\tau} \Pr[H_i^f(t) = 1, H_i^f(\tau) = 1] dt d\tau \\ &- (\mathbb{E}[T_{i,f}^c])^2 \end{aligned} \quad (12)$$

According to [15], $\Pr[H_{i,j}(\tau) = 0 | H_{i,j}(t) = 0] = \frac{\lambda_{i,j}^C}{\lambda_{i,j}^C + \lambda_{i,j}^I} + \frac{\lambda_{i,j}^I}{\lambda_{i,j}^C + \lambda_{i,j}^I} e^{-(\lambda_{i,j}^C + \lambda_{i,j}^I)(\tau - t)}$. Then, when $T_{i,f}^r \rightarrow \infty$, we can get

$$\begin{aligned} \Pr[H_i^f(\tau) = 1, H_i^f(t) = 1] &= 1 - 2 \prod_{j \in \mathcal{S}_{x_j, f=1}} \frac{\lambda_{i,j}^C}{\lambda_{i,j}^C + \lambda_{i,j}^I} \\ &+ \prod_{j \in \mathcal{S}_{x_j, f=1}} \frac{\lambda_{i,j}^C}{(\lambda_{i,j}^I + \lambda_{i,j}^C)^2} \left[\lambda_{i,j}^C + \lambda_{i,j}^I e^{-(\lambda_{i,j}^C + \lambda_{i,j}^I)(\tau - t)} \right] \end{aligned} \quad (13)$$

Let $u = \tau - t$ and substitute (13) into (12), and we can get (5). Let $\lambda^C = \lambda_{i,j}^C$ and $\lambda^I = \lambda_{i,j}^I$, and we can get (10) with the binomial theorem.

B. Proof of Lemma 3

When the user speed changes by s times, the expectation of the communication time in (9) keeps the same, while the variance changes to

$$\begin{aligned} \text{Var}[T_{i,f}^c] &= \left[\frac{\lambda^C}{(\lambda^C + \lambda^I)^2} \right]^{n_f} \sum_{l=1}^{n_f} \binom{n_f}{l} \frac{(\lambda^C)^{n_f-l} (\lambda^I)^l}{sl(\lambda^C + \lambda^I)} \\ &\times \left[T^d - \frac{1}{sl(\lambda^C + \lambda^I)} + \frac{e^{-sl(\lambda^C + \lambda^I)T^d}}{sl(\lambda^C + \lambda^I)} \right], \end{aligned} \quad (14)$$

To prove that $\text{Var}[T_{i,f}^c]$ decreases with s , we will prove that $\frac{\partial \text{Var}[T_{i,f}^c]}{\partial s} < 0$. The partial derivation of $\text{Var}[T_{i,f}^c]$ is

$$\begin{aligned} \frac{\partial \text{Var}[T_{i,f}^c]}{\partial s} &= \\ \left[\frac{\lambda^C}{(\lambda^C + \lambda^I)^2} \right]^{n_f} &\sum_{l=1}^{n_f} \binom{n_f}{l} \frac{(\lambda^C)^{n_f-l} (\lambda^I)^l}{s^3 l^2 (\lambda^C + \lambda^I)^2} \mathcal{A}_1(x), \end{aligned} \quad (15)$$

where $\mathcal{A}_1(x) = -x - xe^{-x} - 2(e^{-x} - 1)$ and $x = sl(\lambda^C + \lambda^I)T^d > 0$. Since $\mathcal{A}_1'(x) = -1 + (1+x)e^{-x} < -1 + (1+x)\frac{1}{1+x} = 0$, $\mathcal{A}_1(x)$ is a decreasing function of x . Thus, $\mathcal{A}_1(x) < \mathcal{A}_1(0) = 0$. According to (15), when $n_f > 0$, we have $\frac{\partial \text{Var}[T_{i,f}^c]}{\partial s} < 0$. The parameter $\alpha_{i,f}$ given in (6) is a decreasing function of $\text{Var}[T_{i,f}^c]$, and thus increases with s .

C. Proof of Lemma 4

To simplify the expression in (8), denote $r \triangleq \frac{C}{T^d R} \in (0, 1)$, $y \triangleq \frac{T^d - \mathbb{E}[T_{i,f}^c]}{\mathbb{E}[T_{i,f}^c]} \geq 0$, and $\alpha \triangleq \alpha_{i,f}$. The expression in (8) can be rewritten as a function of α , given as

$$\mathcal{P}_{i,f} = 1 - \frac{\int_0^r (1 - \frac{u}{r}) u^{\alpha-1} (1-u)^{y\alpha-1} du}{B(\alpha, y\alpha)}. \quad (16)$$

Let $g(\alpha) = 1 - \mathcal{P}_{i,f}$, the derivation of $g(\alpha)$ is

$$g'(\alpha) = \frac{1}{B(\alpha, y\alpha)} \left\{ \int_0^r \left(1 - \frac{u}{r}\right) u^{\alpha-1} (1-u)^{y\alpha-1} [\ln u + y \ln(1-u)] du - \int_0^r \left(1 - \frac{u}{r}\right) u^{\alpha-1} (1-u)^{y\alpha-1} du D(y, \alpha) \right\}, \quad (17)$$

where $D(y, \alpha) = \psi(\alpha) + y\psi(y\alpha) - (1+y)\psi[(1+y)\alpha]$ and $\psi(\cdot)$ is the digamma function. If $r = 1$, $g'(\alpha) = \frac{\partial[y/(1+y)]}{\partial\alpha} = 0$. Denote $\mathcal{A}_2(r) = \frac{B(\alpha, y\alpha)}{r} g'(\alpha)$, $\mathcal{A}_2(1) = 0$ and

$$\lim_{r \rightarrow 0^+} \mathcal{A}_2(r) = \lim_{r \rightarrow 0^+} \int_0^r (r-u) u^{\alpha-1} (1-u)^{y\alpha-1} [\ln u + y \ln(1-u)] du \quad (18)$$

Since $r \geq u \geq 0$ and $y \geq 0$, $(r-u)u^{\alpha-1}(1-u)^{y\alpha-1} \geq 0$ and $\ln u + y \ln(1-u) \leq 0$, thus, $\lim_{r \rightarrow 0^+} \mathcal{A}_2(r) \leq 0$. The derivation of $\mathcal{A}_2(r)$ is

$$\mathcal{A}_2'(r) = \int_0^r u^{\alpha-1} (1-u)^{y\alpha-1} [\ln u + y \ln(1-u)] du - \int_0^r u^{\alpha-1} (1-u)^{y\alpha-1} du D(y, \alpha). \quad (19)$$

Thus, $\mathcal{A}_2'(1) = \frac{\partial B(\alpha, y\alpha)}{\partial\alpha} - \frac{\partial B(\alpha, y\alpha)}{\partial\alpha} = 0$ and $\lim_{r \rightarrow 0^+} \mathcal{A}_2'(r) \leq 0$.

Then, we can get $\mathcal{A}_2''(r) = r^{\alpha-1}(1-r)^{y\alpha-1} [\ln r + y \ln(1-r) - D(y, \alpha)]$. Let $\mathcal{A}_3(r) = r^{1-\alpha}(1-r)^{1-y\alpha} \mathcal{A}_2'(r)$, then, there is one zero point of $\mathcal{A}_3'(r) = \frac{1-(1+y)x}{x(1-x)}$ in $(0, 1]$. Thus, there is one inflection point of $\mathcal{A}_3(r)$. Considering that $\lim_{r \rightarrow 0^+} \mathcal{A}_3(r) = \lim_{r \rightarrow 1^-} \mathcal{A}_3(r) = -\infty$, the sign of $\mathcal{A}_3(r)$ may be negative, or first negative, then positive, and then negative, while r increases in $(0, 1)$. If $\mathcal{A}_3(r) < 0$, then $\mathcal{A}_2''(r) < 0$ when $r \in (0, 1)$. However, we have $\lim_{r \rightarrow 0^+} \mathcal{A}_2'(r) \leq \mathcal{A}_2'(1)$,

which means that $\mathcal{A}_2'(r)$ can not be a decreasing function in $(0, 1)$. Thus, the sign of $\mathcal{A}_3(r)$ is first negative, then positive, and then negative, while r increases in $(0, 1)$. Since $\mathcal{A}_2''(r)$ has the same sign with $\mathcal{A}_3(r)$ in $(0, 1)$, $\mathcal{A}_2'(r)$ first decreases, then increases, and then decreases while r increases in $(0, 1)$. Considering that $\lim_{r \rightarrow 0^+} \mathcal{A}_2'(r) \leq 0$ and $\mathcal{A}_2'(1) = 0$, the sign of $\mathcal{A}_2'(r)$ must be first negative, and then positive in $(0, 1)$. Therefore, while r increases in $(0, 1)$, $\mathcal{A}_2(r)$ first decreases, and then increases. Considering that $\lim_{r \rightarrow 0^+} \mathcal{A}_2(r) \leq 0$ and $\mathcal{A}_2(1) = 0$, we have $\mathcal{A}_2(r) < 0$ in $(0, 1)$ and $\mathcal{A}_2(r) = 0$ when $r = 1$. Since $g'(\alpha) = \frac{r}{B(\alpha, y\alpha)} \mathcal{A}_2(r)$, we get $g'(\alpha) < 0$ in $(0, 1)$. Thus, $g(\alpha)$ decreases with α , and $\mathcal{P}_{i,f} = 1 - g(\alpha)$ increases with α .

D. Proof of Proposition 2

The data offloading ratio in (7) increases with the increasing of $\mathcal{P}_{i,f}$ if $x_{i,f} = 0$, $i \in \mathcal{S}$, $f \in \mathcal{F}$. Then, based on Lemmas 3 and 4, we can get that the data offloading ratio when user i requests file f from other users, i.e., $\mathcal{P}_{i,f}$, decreases with the user speed when $n_f > 0$, otherwise $\mathcal{P}_{i,f} = 0$. Accordingly, the data offloading ratio when user i requests file f , i.e., $x_{i,f} + (1 - x_{i,f})\mathcal{P}_{i,f}$, increases with the user speed when

$x_{i,f} = 0$ and $n_f > 0$; otherwise, it keeps the same, where $i \in \mathcal{S}$, $f \in \mathcal{F}$. Since we consider that not all the users cache the same contents, there must exists $i' \in \mathcal{S}$, $j' \in \mathcal{S}$ and $f' \in \mathcal{F}$, where $x_{i',f'} = 0$ and $x_{j',f'} = 1$, i.e., $n_{f'} > 0$. Thus, the data offloading ratio increases with the user speed.

REFERENCES

- [1] Cisco Systems Inc., "Cisco visual networking index: Global mobile data traffic forecast update, 2015c2020," *White Paper*, Feb. 2016.
- [2] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 99, pp. 176–189, Jul. 2015.
- [3] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Content caching at the wireless network edge: A distributed algorithm via brief propagation," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016.
- [4] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklós, and Z. Turányi, "Design aspects of network assisted device-to-device communications," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 170–177, Mar. 2012.
- [5] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, Apr. 2014.
- [6] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Istanbul, Jul. 2013.
- [7] M. Ji, G. Caire, and A. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849 – 869, Feb. 2016.
- [8] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77 – 83, Aug. 2016.
- [9] A. Shabani, S. P. Shariatpanahi, V. Shah-Mansouri, and A. Khonsari, "Mobility increases throughput of wireless device-to-device networks with coded caching," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016.
- [10] V. Conan, J. Leguay, and T. Friedman, "Fixed point opportunistic routing in delay tolerant networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 5, pp. 773–782, Jun. 2008.
- [11] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot, "Pocket switched networks and human mobility in conference environments," in *Proc. ACM Special Interest Group on Data Commun. (SIGCOMM) Workshop*, Philadelphia, PA, Aug. 2005.
- [12] R. Wang, J. Zhang, S. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," *arXiv preprint arXiv:1606.05282*, 2016.
- [13] G. Alfano, M. Garetto, and E. Leonardi, "Content-centric wireless networks with limited buffers: when mobility hurts," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 299–311, Feb. 2016.
- [14] Y. Li and W. Wang, "Can mobile cloudlets support mobile applications?" in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM)*, Toronto, Canada, Apr. 2014.
- [15] M. Rausand and A. Høyland, *System reliability theory: models, statistical methods, and applications*. John Wiley & Sons, 2004.
- [16] V. Conan, J. Leguay, and T. Friedman, "Characterizing pairwise inter-contact patterns in delay tolerant networks," in *Proc. Int. Conf. on Autonomic Computing and Commun. syst.*, Rome, Italy, 2007.
- [17] Z. Ming, L. Yujin, and W. Wenye, "Modeling and analytical study of link properties in multihop wireless networks," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 445 – 455, Feb. 2012.
- [18] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Int. Conf. on Commun. (ICC)*, London, UK, Jun. 2015.
- [19] A. Passarella and M. Conti, "Analysis of individual pair and aggregate intercontact times in heterogeneous opportunistic networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 12, pp. 2483–2495, Oct. 2013.