Performance Analysis of a Device-to-Device Offloading Scheme in a Vehicular Network Environment

Loreto Pescosolido, Marco Conti, Andrea Passarella

Italian National Research Council, Institute for Informatics and Telematics (CNR-IIT)
Via Giuseppe Moruzzi 1, 56124 Pisa, Italy
Email: {loreto.pescosolido, marco.conti, andrea.passarella}@iit.cnr.it

Abstract

We consider a scheme for offloading the delivery of contents to mobile devices in a vehicular networking scenario. Each content can be delivered to the requesting device either by a neighboring device or, at the expiration of a maximum delay, by the network infrastructure nodes. We compute the analytical expression of the probability that the content delivery is offloaded through a Device-to-Device (D2D) communication as a function of the maximum transmission range allowed for D2D communications, the content popularity, and the vehicles speed. We show that, using the model, it is possible to identify the optimal maximum transmission range, which minimizes the total energy consumption (of the infrastructure plus mobile devices).

Index Terms

D2D data offloading, vehicular networks, Poisson Point Process

I. Introduction

In many wireless network scenarios, where mobile devices retrieve contents by network infrastructure elements (e.g., the eNodeBs in an LTE or 5G network), Device-to-Device (D2D) communications can be exploited to obtain relevant network-level performance improvement in terms of reduction of congestion at the eNodeBs, reduction of the system-wise transmission energy consumption, and increase of the overall system spectral efficiency. Most existing system-level studies on D2D data offloading (see e.g., [1]–[5], and [6] for an extensive survey), focus on the congestion reduction of the cellular network as the main problem to solve through D2D offloading techniques, and aim to maximize offloading efficiency, defined as the percentage of contents delivered through D2D, as the major performance metric. However, in this work, we show that only maximizing offloading efficiency might result in a very significant increase of the energy spent by the overall system, i.e., by the eNodeBs plus the mobile devices, while it is possible to optimize energy efficiency with a modest reduction of offloading efficiency.

We show that, by properly setting system parameters at the physical layer, and in particular the maximum transmission range of the mobile devices for D2D communications (or, equivalently, the maximum transmit power for the devices), it is possible to minimize the energy consumption of the overall system. Intuition suggest that, by increasing the maximum transmission range of the devices, the probability of offloading keeps increasing since, at any instant, the number of neighbors of each node (i.e, the nodes within the transmission range) increases. However, beyond a certain range, the overall system energy consumption stops decreasing, since the power required to perform D2D transmissions starts to become comparable with the power that would be used, to deliver the contents, by the eNodeBs.

Our goal is to evaluate this effect in quantitative terms, in order to compute the optimal transmission range which minimizes the overall system energy consumption, and to understand the impact on offloading efficiency of operating the system at this optimal operating point. To achieve this goal, we present an analytical model which captures application- and scenario-dependent system parameters such as content popularity, mobility patterns, and vehicle speed, as well as physical aspects such as the propagation model.

Specifically, we consider a vehicular network scenario and analyze the performance of the offloading protocol recently proposed in [2]. In this protocol, a node requesting a content can obtain it (i) immediately through D2D, if a neighbor is caching a copy of that content; (ii) delayed through D2D, if, by a maximum deadline called content timeout, it encounters another node caching it; or (iii) delayed through an eNodeB, if such a node is not encountered within the content timeout.

To the best of our knowledge, in prior system-level studies, the link with physical parameters, paired with an analytical model which allows to identify the optimal maximum transmit power for the devices, has not been considered yet.

The model derived in this work takes into account system model parameters such as the content popularity and vehicles density and speed, as well as physical aspects such as the propagation model, and provides a set of analytical expressions to compute the probability of offloading a content request, and the average transmit power used to fulfill it. Both expressions are a function of the maximum D2D transmission distance, and of the considered system model and channel model. Building on this result, we observe that there exist an optimal value of the maximum D2D transmit power which allows to minimize the overall system energy consumption. The model is validated through simulations.

The paper is organized as follows. Section II positions and motivates our work with respect to the existing literature. Section III introduces our system model, the considered offloading protocol, and a Content Dissemination Management System (CDMS), operated by the network infrastructure in coordination with the mobile devices, to execute the offloading protocol. In Section IV we derive the model, and highlight the existence of the optimal operating point at the system level, in terms of the maximum transmission range of the devices. In Section V we validate the model by means of simulations. Finally, Section VI concludes the paper.

II. Related work and motivation

Techniques for offloading data delivery to D2D-communications have been extensively investigated in the literature. The interested reader may want to check, e.g., [6] for an extensive survey. From a system-level perspective, the the idea of jointly designing the offloading and caching/content-injection strategies at the network layer with information coming from the application layer, such as content popularity, synchronized or asynchronous requests, and content-sensitive delay tolerance, has been put forward by several works, under different assumptions on which contents need to be delivered to which users, and the absence or presence of delay tolerance. For instance, in [1], [4], assuming delay-tolerant applications, and a scenario in which content delivery mostly relies on D2D-offloading, a strategy for I2D re-injection of contents in the network is proposed to face temporal content starving in a certain area. In [2], a CDMS for contents originated from delay-tolerant applications, suited to vehicular network scenario, is proposed. In [3] the link between social connectivity and physical connectivity is exploited to select hub nodes in the network which may assist the cellular infrastructure in the data offloading process. In [5], in the framework of a content dissemination problem, the authors propose a mixed I2D-multicast and D2D-relaying reinforcement-learning-based strategy, which determines which users should receive the contents through a direct I2D transmission or through a D2D relaying from a neighboring device.

At the data-link layer, a class of works is related to organizing the local D2D topology in order to optimize performance metrics such as throughput, fairness, and energy/spectral efficiency. For instance, in [7] the authors devise an out-of-band D2D-clustering strategy, based on coalitional game-theory, aimed at improving these performance metrics, in a LTE-standard compliant way. In these works, information related to the application layer, such as the content popularity, is not considered. Finally, works like [8], [9] (amongst many others), aim at devising radio resource allocation strategies, and/or other physical layer parameters, like coding rates and transmit power levels, assuming that coexistence among D2D and/or I2D links are given as an input to the problem. Finally, design and fundamental limits of D2D caching-based content delivery protocols from an information-theoretic perspective, are investigated (for an infrastructureless scenario) in works like [10], [11], see also the references therein.

The designs proposed in the works considered above take typically into account system parameters of the sole referenced architectural level. For instance, [1]–[5] only consider system parameters at the application level. Moreover, in this class of works, an analytical model for the spatiotemporal stochastic process which controls the positions of the nodes and the content of their caches, and the related impact on system-level performance, is lacking, as the performance evaluation is based on simulations [2]–[5]. On the other hand, in works in the class of [7]–[9], targeting data-link/radio resource management issues, the impact of content popularity is not considered. Additionally, they typically assume fully-backlogged traffic and, for the modeling aspects, they take into account stochastic but static node distribution as in [8] or a fixed (and given) one [7], [9]. Finally, these works typically deal more with the coexistence of different D2D links, rather then dealing with the delivery of individual contents. For these works, the performance evaluation relies on simulations ([8], [9]) or experiments [7]. Finally, the results of works in the class of [10], [11], are often focusing on scaling laws and network throughput, assuming a fully backlogged traffic, but many details of the physical layer are necessarily abstracted out.

Differently from most of the existing literature, in this work, we take a full cross-layer approach, including aspects related to the application layer (such as the content popularity), aspects related the geographical distribution and mobility of nodes in the network, and physical layer aspects such as the radio propagation model[1]. Additionally, we derive an analytical model able to predict the system performance in terms of a high-level metric such as the offloading efficiency and metric related to physical parameters such as the system level average energy consumption. In this way, we can find the optimal maximum transmit power of the devices, which minimizes the energy consumption. A cross-layer approach in quite general terms similar to our one, i.e., which includes the content popularity-related aspects in the framework of an analytical model based on point process, is taken in [13], but in a static scenario and under completely different assumptions[2].

III. System model and Content Dissemination Management System

A. Vehicle arrival and content requests

We consider a Region of Interest (ROI) consisting of a street chunk. Vehicles enter, traverse, and exit the ROI. Each vehicle has onboard a mobile devices, which can be either a human hand-held device or part of the vehicle equipment[3].

We assume that vehicles enter the street from both ends, according to a homogeneous Temporal Poisson Point Process (TPPP), with vehicle arrival rate λ_t. The direction from which each new vehicle enters the street is randomly chosen with equal probability (equal to $1/2$). Accordingly, the arrival rate of vehicles entering at one end of the street is $\lambda_t/2$. The vehicles traverse the street at a constant speed. We assume that the speed of each vehicle, v, is a random variable with Probability Density Function (PDF) $\tilde{p}_V(v)$. However, for the derivations in Section IV, it is convenient to reformulate the above model by incorporating the vehicles' motion direction in their speed v. This can be done by including negative speed values. With this formulation, the PDF is given by

$$p_V(v) = \frac{1}{2} u_{(-\infty,0]}(v) \tilde{p}_V(-v) + \frac{1}{2} u_{[0,\infty)}(v) \tilde{p}_V(v), \quad (1)$$

where $u_{[x,y]}(\cdot)$ represents the indicator function equal to one for values of its argument in the interval $[x, y]$, and zero outside it (the interval is open if one the two extremes is infinite).

[1]The effect of the radio propagation model is explicitly considered in our model, but the impact of using different models, due to space reasons, is not analyzed in this work. A preliminary study of the impact of different channel models on the performance of the CDMS considered in this work, has been presented in [12].

[2]In that work, content popularity is related to clusters of users in the social domain, that are then mapped to physical clusters.

[3]The considered offloading protocol is particularly suited to the latter case since, in that case, the devices extract the energy required to send cached contents to their neighbors, from a "virtually" renewable source like the vehicle battery.

Most of the results obtained in this work are general with respect to the speed PDF $p_V(v)$. However, for the purpose of performing simulations to validate the results, it will be useful to consider a special case. Particularly we shall consider, as a special case, a uniform distribution of the vehicles speed between two values v_a and v_b, or $\tilde{p}_V(v) = \frac{1}{v_b - v_a} u_{[v_a, v_b]}(v)$. The two-side PDF (which incorporates the direction) is hence

$$p_V(v) = \frac{1}{2(v_b - v_a)} u_{[-v_b, -v_a]}(v) + \frac{1}{2(v_b - v_a)} u_{[v_a, v_b]}(v). \tag{2}$$

As vehicles enter the ROI, they start requesting contents according to a given content request process[4]. The content request process is characterized by a content-request *arrival* process, which defines the time instant at which the request is generated, and a content-interest probability distribution, which defines which content is requested. Particularly, we assume that the devices issue content requests according to a homogeneous TPPP of constant intensity (rate) λ_Z content requests per second, and that contents belong to a finite library $\mathcal{L}$ of size N_Z. Without loss of generality, we assign an index $z \in \{1, \ldots, N_Z\}$ to contents. We assume that content requests follow a given distribution with Probability Mass Function (PMF) $p_Z(z) = Pr\,(Z = z)$, with support $[1, \ldots, N_Z]$. We also assume that successive requests are independent and identically distributed (iid), and that requests from different nodes are also iid. Finally, we assume that, upon obtaining a content, a device keeps it cached for an amount of time τ_s called *sharing timeout*, so that the cache occupation is kept limited.

B. Content Dissemination Management

The ROI is served by a set of eNodeBs. During its path within the ROI, at each instant, each device is associated to an eNodeB, which is responsible of handling the process of delivering the contents requested by that device, during the time the devices is in its cell[5]. Similarly to [1], [2], we assume that content requests have some delay tolerance, i.e., that they must be served at most within a content timeout τ_c. Whenever possible, a device should obtain a desired content by neighboring or encountered devices. This is obviously possible if, at the time of content request, a neighboring device has the content cached locally. However, as devices are mobile, there is the chance that, in the event that no neighbor has the desired content in its cache at the time of request, the requesting device encounters, later on, another device which does have the content cached. Only after the content timeout has elapsed, if the requesting device has not yet obtained the content, it obtains it from the eNodeB to which it is associated at that time.

A pictorial representation of this basic idea is provided in Fig. 1, where the succession of six events at different time instants is represented. Two vehicles (V1 and V3) request two contents,

[4]The content request process is originated at the application layer. Here, it is of no importance whether the interest is generated by a human or by, for instance, an IoT application executed by the software on a vehicle.

[5]The content delivery handling can be also handed over to another eNodeB, if during the process the requesting device falls within another cell, see below.

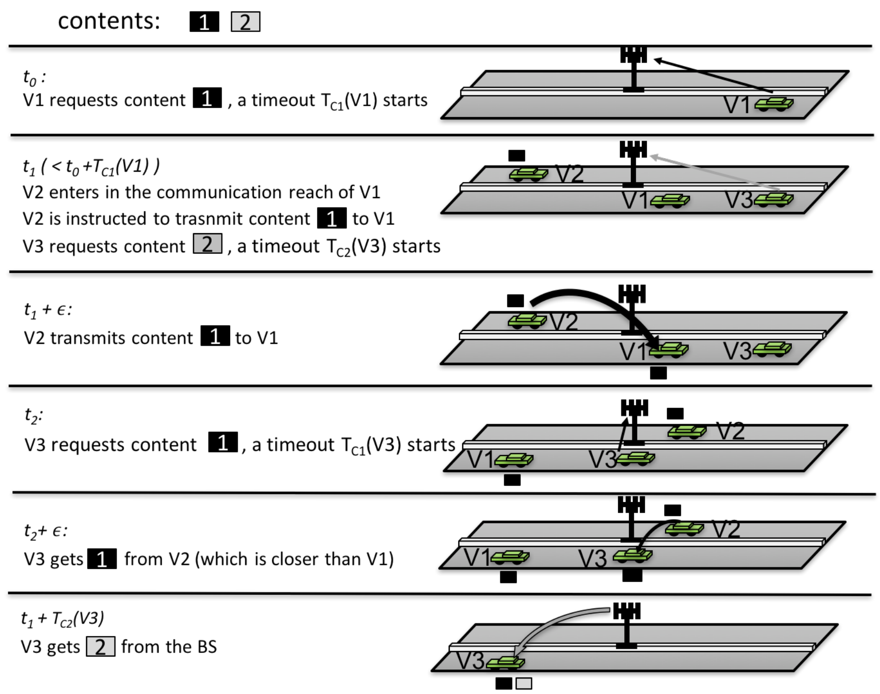

Fig. 1. System model and offloading protocol sketch

represented by the black and grey rectangles. V1 only requests the black content, and succeeds in obtaining it from the encountered V2 (which has it in its cache) before the associated content timeout $T_{c1}(V_1)$ elapses. V3 requests, in successive instants, first the grey content and then the black content. Similarly to V1, V3 obtains the black content from V2 before the associated content timeout $T_{c1}(V_3)$ elapses, while it obtains the grey content from the eNodeB at the end of the associated content timeout $T_{c2}(V_3)$ since it has encountered no device with the content available for D2D-offloading.

The CDMS, essentially, acts on a distributed database (residing at the eNodeBs) containing the up-to-date list of each node's position, the list of its neighbors, and the *nominal* channel gain (see below) between any two neighbors and between each device and the surrounding eNodeBs. For each device k, the list of neighbors, $\mathcal{N}_k$, held at the eNodeB to which device k is associated, is composed of pairs of the form (j, r_j^k). In this pair, j is the id of any device which is a neighbor of device k, and r_j^k is a ranking index of device j as "seen" by device k on the basis of a given criterion. In this work, the criterion to establish if two devices are neighbors, and the ranking of each node's neighbors, is based on a nominal indicator of the channel quality between the devices[6]. At each eNodeBs, the lists $\mathcal{N}_k$ are kept up-to-date on the basis of Hello messages sent periodically by the devices, containing a device unique identifier. Each device k has an internal content cache $\mathcal{C}_k$ populated with previously downloaded contents. At any time, the CDMS also has an index of the contents in each node's cache, although the CDMS does not necessarily hold a copy of the contents itself. Time is organized in Control Intervals (CIs). We assume that the duration of the CIs is much smaller than the content timeout.

Before providing details of the behavior of the CDMS in each CI, it is worth describing the tasks it performs on a coarse timescale. We do this through Algorithms 1 and 2, which describe at a high level[7] the actions taken on demand, i.e., as a consequence of content requests, by the devices and the CDMS. We briefly introduce the notation required for a correct interpretation of the algorithms: $\biguplus \{\mathcal{C}_j | \text{condition on } j\}$ is used to indicate the union of the caches of devices satisfying a given condition; $\hat{j}(k, z)$ is used to indicate the device j that has the best ranking r_j^k among the neighbors of device k which have content z in their caches; $j \xrightarrow{z} k$ indicates the transmission of content z from device j to device k. These transmissions are triggered by the CDMS. The remaining notation used in Algorithms 1-2 is self-explaining.

Upon the generation of a content request, a device (Algorithm 1) notifies the CDMS that it is interested in that content (step 3), and then waits for receiving it either from a BS or from a neighbor (step 4). The system guarantees that the content will be delivered within the predefined *content timeout*. After the reception of the content, the device makes it available for other devices that may request it, for a limited amount of time determined by the *sharing timeout* (steps 9, and 13-20).

Algorithm 2 describes the actions taken by the CDMS to handle a content request. Here, a key point, which effectively allows to increase the system energy efficiency, is that the CDMS selects the best device for delivering the content, on the basis of channel quality considerations, represented by the ranking of each node's neighbors (step 7). If, however the content cannot be delivered through a D2D communication within the content timeout, the CDMS uses the eNodeBs to deliver it (steps 15-20).

The behavior of the CDMS at the CI timescale is as follows. In every CI, the CDMS schedules which transmissions should be performed, based on the physical information contained in the above described lists, and on the ongoing content requests to be handled, including those whose content timeout has not expired and those whose content timeout has expired (which need to be fulfilled through I2D transmission). Specifically, if, at the time of a content request, a requesting device k has at least one neighboring device with the desired content, the CDMS selects, out of these neighboring devices, the device j with the best ranking r_j^k, and schedules it to transmit the content to the requesting device in the following CI. If there are no neighbors with the requested

[6] In general, the nominal channel quality may be computed, by the CDMS, on the basis of the positions of the devices, which the eNodeBs are assumed to know. In this work, we assume that the nominal channel gain can be computed using any deterministic channel model which relates the channel gain g to the distance d, i.e., a function $g(d)$, see Subsection IV-C and Section V.

[7] In the pseudocode, the temporal succession of Control Intervals is not appearing explicitly.

Algorithm 1 Actions taken by device k to request content z

1) **Upon** request for content z from the application layer
2) **Set** k_content_received = **false**
3) **Send** (k, z)_cont_req to CDMS
4) **while** k_content_received == **false do**
   ▷ Wait for receiving content z, from a BS or from a neighbor
5) **if** content z is received **then**
6) **Set** k_content_received = **true**
7) **Send** (k, z)_ACK to CDMS and/or the sending device
8) **Add** z to $\mathcal{C}_k$
9) **Set** (k, z)_sharing_timeout
10) **break**
11) **end if**
12) **end while**
13) **while** (k, z)_sharing_timeout is not expired **do**
    ▷ Available for opportunistic sharing of content z
14) **Upon** request from CDMS (step 7 of Algorithm 2)
15) **Send** z to device requesting it
16) **end while**
17) **Remove** content z from $\mathcal{C}_k$
18) **Cancel** (k, z)_sharing_timeout

content available, the first device encountered by the device k, with the desired content cached, is scheduled by the CDMS to transmit it. Finally, if no device is encountered within the content timeout, the CDMS schedules the transmission of the content from the infrastructure. To handle the handover of ongoing requests originated from a device that crosses a cell border during the content timeout, adjacent eNodeBs periodically exchange the up to date status of the ongoing request procedures (see below) of devices moving across cells[8].

Remark: The model introduced in Subsection III-A for the issuing of content requests from a device, does *not* account for the fact that, at the time of request, the content may be already present in the cache of the requesting device. If this is the case, the system is assumed to take the following actions: (i) the request is labelled as "repeated", and therefore the CDMS does not perform the transmission of the content (either through D2D or I2D); (ii) the sharing timeout related to that content is reinitialized to its initial value.

[8]This information exchange can be performed using high speed fiber connections, or dedicated radio channels forming a wireless backbone for the Radio Access Network (RAN).

Algorithm 2 Actions taken by CDMS for handling content request (k, z)

Upon receiving (k, z)_`cont_req`
Set (k, z)_`served` = **false**
Set (k, z)_`content_timeout`
while (k, z)_`content_timeout` is not expired **do**
  if $z \in \biguplus \{\mathcal{C}_j | j \in \mathcal{N}_k\}$ **then**
    Identify $\hat{j}(k, z)$
    Trigger transmission $\hat{j}(k, z) \xrightarrow{z} k$
    Wait for (k, z)_ACK
    Upon (k, z)_ACK reception
    Set (k, z)_`served` = **true**
    Remove (k, z) from $\mathcal{L}_{\text{req}}$
    break
  end if
end while
if (k, z)_`served` == **false**
  Send z to k
  Wait for ACK_(k, z)
  Upon reception of ACK_(k, z)
  Set (k, z)_`served` = **true**
end if
Cancel (k, z)_`content_timeout`

IV. OFFLOADING EFFICIENCY AND ENERGY CONSUMPTION

In this section, we compute the probability of offloading through D2D a non-repeated content request, and the associated transmit power used on average in each content transmission[9]. We shall derive expressions of these quantities as a function of the maximum transmission range allowed to the devices, and of the content request process and of the vehicles arrival and mobility models introduced in Subsection III-A. To derive our analytical results, we first present some preliminary results obtained by applying standard tools from the theory of temporal and Spatial Poisson Point Processes (SPPPs), [14], [15], (Subsection IV-A)[10], then we compute the probability of offloading (Subsection IV-B), and finally we compute the average transmit power (Subsection IV-C).

[9]Note that the average energy consumption of the entire system during a given time interval is simply given by the product of the average transmit power used for fulfilling a non repeated request, times the duration of each transmission, times the average number of non-repeated requests in the interval.

[10]It is likely that either of results in Subsection IV-A have appeared elsewhere. Nonetheless, for the sake of readability of the successive derivations, we deem it useful to collect them in a preliminary subsection.

A. Preliminary results

We start by proving the following result, which characterizes the spatial distribution of the vehicles as a function of a given (temporal) vehicles arrival process.

Lemma 1. *Under the assumptions in Subsection III-A, the following results hold true:*

1) At any instant, the vehicles are spatially distributed according to a homogeneous unidimensional[11] *SPPP with linear density*

$$\rho = \int_{-\infty}^{+\infty} \frac{1}{|v|} \lambda_t p_V(v) dv. \tag{3}$$

In the special case of uniformly distributed vehicles' speed ((2)),

$$\rho = \frac{\lambda_t (\ln v_b - \ln v_a)}{(v_b - v_a)}. \tag{4}$$

2) Considering a vehicle moving at a specific speed v^* on a straight line, the temporal process of the instants at which the vehicle encounters[12] *other vehicles, moving at any speed is a homogeneous TPPP with rate*

$$\lambda_e^{(v^*)} = \int_{-\infty}^{\infty} \lambda_t p_V(v) \frac{|v^* - v|}{|v|} dv. \tag{5}$$

In the special case (2)*,*

$$\lambda_e^{(v^*)} = \frac{\lambda_t}{(v_b - v_a)} \left(|v^*| \left(\ln |v^*| - \ln v_a - 1 \right) + v_b \right). \tag{6}$$

Proof: See Appendix A. ■

We focus now on the spatial point process of devices containing a specific content z in their caches at a given instant, and the temporal process of the instants at which a point which moves at constant speed v^* encounters vehicles moving at any speed, which have a specific content z in their caches.

Lemma 2. *Under the assumptions in Subsection III-A, the following results hold true:*

1) The process of requests for a specific content z issued by a given device is a homogeneous TPPP, with arrival rate

$$\lambda_z = p_Z(z) \lambda_Z. \tag{7}$$

[11] In our mathematical analysis, we only consider the horizontal coordinate of the vehicles positions, i.e., we do not take into account that vehicles moving in opposite directions are located on different lanes of the street. The comparison of the results of the simulations we performed to validate our analysis (in which vehicles moving in opposite direction are placed on different lanes) with the analytical results shows that the effect of this approximation on the computation of the offloading efficiency and energy consumption is negligible, see Section V.

[12] The "encountering" between two vehicles means that they fall within a range $d_{\max}$ off each other. The instant of the encountering is the instant at which their distance is exactly equal to $d_{\max}$.

2) At any instant, the probability $Pr(\mathcal{C} \ni z)$ that the cache of a generic device contains a specific content z is upper and lower bounded as follows

$$1 - e^{-\lambda_z(\tau_s - \tau_c)} \leq Pr(\mathcal{C} \ni z) \leq 1 - e^{-\lambda_z \tau_s}. \tag{8}$$

3) At a given instant, the spatial process of the position of the devices containing a specific content z in their caches can be very well approximated by a homogeneous SPPP, with linear density ρ_z tightly lower bounded as in

$$\rho_z \gtrsim \rho \left(1 - e^{-\lambda_z(\tau_s - \tau_c)}\right), \tag{9}$$

where ρ is given by (3) *or* (4)*.*

4) Consider a vehicle moving at speed v^*. The temporal process of devices, that have a specific content z in their caches, encountered by the vehicle, can be very well approximated by a a homogeneous TPPP with encountering rate tightly lower bounded as

$$\lambda_e^{(v^*,z)} \gtrsim \lambda_e^{(v^*)} \left(1 - e^{-\lambda_z(\tau_s - \tau_c)}\right), \tag{10}$$

where $\lambda_e^{(v^*)}$ is given by (5)*, or* (6) *in the special case* (2)*.*

Proof: See Appendix A. ■

Note that, since in practical scenarios we may reasonably assume that $\tau_s \gg \tau_c$, the bounds in (8) are very tight. As a result, in practice, the lower bounds (9) and (10) can be considered as very accurate approximations[13]. In the following, we will use the notation "$\simeq$" in all the Equations that stem from (9) and (10), as a convention to state that the approximation is quite accurate since it is supported by tight upper and lower bounds.

B. Probability of content delivery offloading

The results obtained in Subsection IV-A allow us to compute the probability that the fulfilling of a content request is offloaded to a D2D transmission among nearby devices. We shall compute this probability as a function of the maximum nominal transmission range of the devices, indicated in the following with $d_{\max}$. If two devices, at a given instant, are closer than $d_{\max}$, they are considered to be neighbors. Note that $d_{\max}$ is tightly related to physical layer parameters, such as transmit power and information rate, which play a major role in the determination of the system energy consumption, see Subsection IV-C.

In the following, in using the terminology "probability of offloading", *we always refer to the probability conditioned on the fact that the request is not repeated* (see the final remark in

[13]Both expressions (9) and (10) result from approximating $Pr(\mathcal{C} \ni z)$ with its lower bound in (8). Using the upper bound in (8) would still entail an accurate approximation (consisting in tight upper bounds instead of lower bounds) of the density and rate appearing in (9) and (10), respectively. For practical purposes, the impact of using either of the two bounds is negligible. Selecting the lower bound represents a conservative choice for the performance evaluation, since it tends to underestimate (in a negligible way) the probability of offloading the content requests.

Subsection III-B). To avoid using an excessively cumbersome notation, we omit this conditioning from the notation of this subsection up to Eq. (18).

To compute the probability of offloading (of a non-repeated request), we start computing the probability of offloading a non-repeated request, further conditioned on the fact that the requested content is a specific one, say z. We first compute the probability that the request is fulfilled *immediately*. We indicate the probability of this event with $Pr^{(d_{\max})}$ (off.imm $\mid z$). Now, the request can be fulfilled *immediately,* through a D2D transmission, if at least one neighbor of the requesting node has content z in its cache at the time of request. This is equivalent to say that the *closest neighbor*[14] has content z in its cache at the time of request. This event is determined by the SPPP of the devices containing content z, which is a homogeneous SPPP with intensity ρ_z given by (9), and by the maximum nominal transmission range $d_{\max}$. In fact, $Pr^{(d_{\max})}$ (off.imm $\mid z$) coincides with the probability that the closest point of the homogeneous SPPP of the devices containing content z, is at a distance less than $d_{\max}$ at the time of request. It is well known that the Cumulative Distribution Function (CDF) of the "closest neighbor distance" d_{cn} determined by a unidimensional homogeneous SPPP with (linear) density $\tilde{\rho}$, is given by $F_{\text{cn}}(d) \triangleq Pr\left(d_{\text{cn}} \leq d\right) = 1 - e^{-\tilde{\rho}2d}$, [14], [15]. Accordingly, we obtain

$$Pr^{(d_{\max})}\left(\text{off.imm} \mid z\right) = 1 - e^{-2d_{\max}\rho_z}. \tag{11}$$

Using (7) in the expressions of the upper and lower bounds in (8), plugging the lower bound in (8) to compute ρ_z from (9), and using (9) in (11), we obtain the following expression for the probability of immediate offloading of a non-repeated request of a specific content z:

$$Pr^{(d_{\max})}\left(\text{off.imm} \mid z\right) \backsimeq 1 - e^{-2d_{\max}\cdot\left(1-e^{-p_Z(z)\lambda_Z\cdot(\tau_s-\tau_c)}\right)\rho}, \tag{12}$$

where, in the special case of $p_V(v)$ given by (2), ρ can be replaced by $\lambda_t\left(\ln v_b - \ln v_a\right)/(v_b - v_a)$, see Eq. (6).

Next, we compute the probability that the request is still fulfilled through a D2D transmission, but the content is obtained by a device which, during the content timeout following the request, comes within a range $d_{\max}$ off the requesting device, i.e., it is encountered by it. We first compute such probability for a requesting device moving at a specific speed v.

First, consider a vehicle moving at speed v and an instant t_0. The probability that the first device with a given content z in its cache comes within a range $d_{\max}$ off the requesting device, starting from t_0, within an interval of duration equal to the content timeout τ_c, is given by

$$Pr\left(\text{enc} \mid z, v\right) = 1 - e^{-\lambda_e^{(v,z)}\tau_c} \backsimeq 1 - e^{-\lambda_e^{(v)}\left(1-e^{-\lambda_z(\tau_s-\tau_c)}\right)\tau_c}, \tag{13}$$

where $\lambda_e^{(v,z)}$ has the expression (10). In the special case (2), we have

$$Pr\left(\text{enc} \mid z, v\right) \backsimeq 1 - e^{-\frac{\lambda_t}{(v_b-v_a)}\left(|v|(\ln|v|-\ln v_a-1)+v_b\right)\left(1-e^{-\lambda_z(\tau_s-\tau_c)}\right)\tau_c}. \tag{14}$$

We observe that this probability does not depend on $d_{\max}$.

[14]Under the assumption that the neighbors ranking is performed on the basis of a distance-based criterion.

Now, the probability that a non-repeated request for a specific content z, issued by a vehicle moving at speed v, is fulfilled through a D2D transmission by a device encountered during the content timeout following the request, is given by the probability that the request has *not* been fulfilled immediately, $\left(1 - P^{(d_{\max})}(\text{off.imm} \mid z)\right)$, times the probability (13) of encountering, within the content timeout following the request, a device with the desired content z in its cache, i.e.

$$Pr^{(d_{\max})}(\text{off.del} \mid z, v) = \left(1 - Pr^{(d_{\max})}(\text{off.imm} \mid z)\right) Pr(\text{enc} \mid z, v). \tag{15}$$

This quantity depends on $d_{\max}$ only through $Pr^{(d_{\max})}(\text{off.imm} \mid z)$.

Removing the dependence on the speed v at which the requesting device is moving, we can compute the probability, indicated with $Pr^{(d_{\max})}(\text{off.del} \mid z)$, that a non-repeated request for a specific content z, issued by a vehicle moving at *any* speed, is fulfilled through a D2D transmission by a device encountered during the content timeout following the request. By the law of total probability, we get

$$\begin{aligned} Pr^{(d_{\max})}(\text{off.del} \mid z) &= \int_{-\infty}^{\infty} Pr^{(d_{\max})}(\text{off.del} \mid z, v)\, p_V(v)\, dv \\ &= \left(1 - Pr^{(d_{\max})}(\text{off. imm}|z)\right) \int_{-\infty}^{\infty} Pr(\text{enc} \mid z, v)\, p_V(v)\, dv. \end{aligned} \tag{16}$$

This expression is general with respect to the speed PDF $p_V(v)$. In a practical scenario, to quantitatively evaluate the probability of offloading, a specific model for $p_V(v)$ should be provided in input to (16). For instance, taking $p_V(v)$ as in (2), it is straightforward to show that (16) becomes

$$\begin{aligned} Pr^{(d_{\max})}(\text{off.del} \mid z) &= \left(1 - Pr^{(d_{\max})}(\text{off.imm} \mid z)\right) \\ &\quad \cdot \left(1 - \frac{e^{-\frac{\lambda_t v_b Pr(z \in \mathcal{C}) \tau_c}{(v_b - v_a)}}}{(v_b - v_a)} \int_{v_a}^{v_b} e^{-\frac{\lambda_t Pr(z \in \mathcal{C}) \tau_c}{(v_b - v_a)} v (\ln v - \ln v_a - 1)} dv \right). \end{aligned} \tag{17}$$

The total probability that a non-repeated request for content z is fulfilled through offloading is obviously given by

$$Pr^{(d_{\max})}(\text{off}|z) = Pr^{(d_{\max})}(\text{off.imm}|z) + Pr^{(d_{\max})}(\text{off.del}|z)$$

Finally, the probability that a non-repeated request for content z is fulfilled through an I2D transmission, i.e., it is not offloaded, is given by

$$Pr^{(d_{\max})}(\text{non-off} \mid z) = 1 - Pr^{(d_{\max})}(\text{off.imm} \mid z) - Pr^{(d_{\max})}(\text{off.del} \mid z). \tag{18}$$

We now proceed removing the dependence on the requested content z. First, we prove the following

Lemma 3. *Under the assumptions in Subsection III-A, the probability that a content request is not repeated, is given by*

$$Pr\left(\text{NR}\right)=\sum_{z}Pr\left(Z=z\right)Pr\left(\mathcal{C}\not\ni z\right), \tag{19}$$

and the probability that the content Z requested in a content request, conditioned to the fact that the request is not repeated, is given by

$$p_Z\left(z\mid \text{NR}\right)=\frac{Pr\left(Z=z\right)Pr\left(\mathcal{C}\not\ni z\right)}{\sum_{z\in\mathcal{L}}Pr\left(Z=z\right)Pr\left(\mathcal{C}\not\ni z\right)}. \tag{20}$$

Proof: See the Appendix A. ■

Finally, we obtain the following

Theorem 4. *The probability of offloading for a non-repeated request, irrespective of the requested content, is given by*

$$\begin{aligned}Pr^{(d_{max})}\left(\text{off}\mid \text{NR}\right)&=\sum_{z\in\mathcal{L}}p_Z\left(z\mid \text{NR}\right)Pr^{(d_{\max})}\left(\text{off}\mid z\right)\\&=\sum_{z\in\mathcal{L}}\frac{p_Z\left(z\right)Pr\left(\mathcal{C}\not\ni z\right)}{\sum_{z\in\mathcal{L}}p_Z\left(z\right)Pr\left(\mathcal{C}\not\ni z\right)}Pr^{(d_{\max})}\left(\text{off}\mid z\right).\end{aligned} \tag{21}$$

Proof: This comes straightforward from applying the law of total probability, and replacing the probability of requesting z (conditioned to the event that the request is not repeated) with Eq. (20). Eq. (21) expresses the law of total probability applied to the event of offloading a non-repeated request. i.e., it simply states that the probability is the sum, over all the possible realizations z of the requested content Z, of the probability of the offloading event conditioned to each specific realization z, or $Pr^{(d_{\max})}\left(\text{off}\mid z\right)$, weighted by the probability that the requested content is z (conditioned to the fact that the request is not repeated), or $p_Z\left(z\mid \text{NR}\right)$. The specific expression of $p_Z\left(z\mid \text{NR}\right)$ is given by (20) in Lemma 3. ■

C. Energy consumption minimization

We consider, without loss of generality an LTE-like multi-carrier communication system. A set of Physical Resource Blocks (PRBs), corresponding to the elements of a time-frequency grid, is allocated to each communication, on the basis of the size of the content that needs to be transmitted, see below. In the following, we assume a fixed content size of D bits. As stated in Subsection III-B the CDMS is aware of the nominal (scalar) channel gain between any D2D or I2D pair. Therefore, power is allocated uniformly over the subcarriers. Let $\overline{e}$ be a nominal target normalized (i.e., measured in bps/Hz) information rate that a link is required to be able support (in this work, this is consider a fixed system parameter). Let $\mathcal{P}^{(w_c)}$ be the transmit power allocated *on each subcarrier*, $g(d)$ be a generic monotonically decreasing propagation loss formula which relates distance d to the nominal channel gain g in a deterministic way, w_c the subcarrier spacing, F_{rc} the noise figure at the receiver, N_0 the thermal noise power spectral density, and $\sigma_c^2 = w_c F_{rc} N_0$ the noise power on each subcarrier. Let $\mathcal{P}_{tx}^{(w_c)}$ be the transmit power allocated on each subcarrier, and let the nominal channel gain be $g(d)$. We define the normalized nominal information rate (measured in bps/Hz) e as the Shannon capacity on that subcarrier divided by the subcarrier width, or

$$e = \frac{1}{w_c} w_c \log_2 \left(1 + \frac{\mathcal{P}_{tx}^{(w_c)} g(d)}{\sigma_c^2}\right) = \log_2 \left(1 + \frac{\mathcal{P}_{tx}^{(w_c)} g(d)}{\sigma_c^2}\right). \tag{22}$$

We assume that, to transmit to a receiver located d meters away, the transmitter sets the the transmit power over each subcarrier to

$$\mathcal{P}_{tx,w_c}^{(\bar{e})}(d) = \frac{1}{g(d)} \sigma_c^2 \left(2^{\bar{e}} - 1\right). \tag{23}$$

This is obtained by inverting (22) with respect to $\mathcal{P}_{tx,w_c}^{(\bar{e})}$, with the objective to match the target nominal normalized information rate $\bar{e}$.

If a non-repeated request for content z if fulfilled immediately through offloading, the distance at which the transmitter is located is the distance of the nearest neighbor (with content z in its cache), *conditioned to the fact that the nearest neighbor is within a range* $d_{\max}$ *off the requesting device.* Let us indicate the nearest neighbor distance with the random variable D. The required transmit power to fulfill a non-repeated request is the random variable resulting from the transformation of the random variable D to the random variable $Y_{\text{NR,off,im}}$ defined as

$$Y_{\text{NR,off,im}} \triangleq \mathcal{P}_{tx}^{(\bar{e})}(D) = \frac{1}{g(D)} \sigma_c^2 \left(2^{\bar{e}} - 1\right). \tag{24}$$

With relatively straightforward integral calculus steps, it can be showed that the CDF $F_{Y,\text{NR,off,im}}(y)$, the PDF $p_{Y,\text{NR,off,im}}(y)$, and average value $\overline{Y}_{\text{NR,off,im}}(d_{\max})$ of $Y_{\text{NR,off,im}}$, computed as a function of the system parameter $d_{\max}$, are given by[15] (see Appendix B)

$$F_{Y,\text{NR,off,im}}(y) = \frac{1}{1 - e^{-\rho_z 2 d_{\max}}} \left(1 - e^{-\rho_z 2 g_{\text{dB}}^{-1}\left(10 \log_{10}\left(\frac{1}{y}\sigma_c^2 (2^{\bar{e}}-1)\right)\right)}\right) u_{[0, \sigma_c^2(2^{\bar{e}}-1)/g(d_{\max})]}(y). \tag{25}$$

$$p_{Y,\text{NR,off,im}}(y) = -\frac{1}{y^2} \frac{2\rho_z \sigma_c^2 \left(2^{\bar{e}} - 1\right)}{\left(1 - e^{-\rho_z 2 d_{\max}}\right)} \frac{e^{-\rho_z 2 g^{-1}\left(\frac{1}{y}\sigma_c^2 (2^{\bar{e}}-1)\right)}}{g'\left(g^{-1}\left(\frac{1}{y}\sigma_c^2 \left(2^{\bar{e}} - 1\right)\right)\right)} u_{[0, \sigma(2^{\bar{e}}-1)/g(d_{\max})]}(y). \tag{26}$$

$$\overline{Y}_{\text{NR,off,im}}(d_{\max}) = \frac{\sigma_c^2 (2^{\bar{e}} - 1)}{g(d_{\max})} \left(\frac{1}{1 - e^{\rho_z 2 d_{\max}}}\right) + \frac{\sigma_c^2 \left(2^{\bar{e}} - 1\right)}{1 - e^{-\rho_z 2 d_{\max}}} \frac{\ln 10}{10} \int_{g_{\text{dB}}(d_{\max})}^{+\infty} 10^{-y'/10} e^{-\rho_z 2 g_{\text{dB}}^{-1}(y')} dy'. \tag{27}$$

[15]Expression (27) is provided in terms of the function $g_{\text{dB}}(d) = 10 \cdot \log_{10}(g(d))$, as it it is more suitable to a numeric integration of the last term.

In the case the request is not fulfilled immediately, but it is fulfilled within the content timeout, through an encounter with another device, the transmission distance is always equal to $d_{\max}$, since we are assuming that as soon as the two devices get within each other's range, the content is transmitted. Therefore, the transmit power is, in this case, see (23)

$$Y_{\text{NR,off,del}}\left(d_{\max}\right) = \sigma_c^2\left(2^{\bar{e}}-1\right)/g(d_{\max}). \tag{28}$$

In the case that the request is fulfilled using an I2D transmission, the transmit power is a function of the distance $d^{(\text{I2D})}$ between the eNodeB and the receiving device as in

$$Y_{\text{NR,non-off}}\left(d^{(\text{I2D})}\right) = \sigma_c^2\left(2^{\bar{e}}-1\right)/g(d^{(\text{I2D})})$$

The distance $d^{(\text{I2D})}$ is distributed uniformly in the range $[0, d_{\max}^{(\text{I2D})}]$, where $d_{\max}^{(\text{I2D})}$ is the cell radius[16].

With some integral calculus, see Appendix B, it can be showed that the CDF, PDF, and average value of $Y_{\text{NR,non-off}}$ are given by

$$F_{Y,\text{NR,non-off}}\left(y\right) = \frac{1}{d_{\max}^{(\text{I2D})}} g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) u_{\left[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{(\text{I2D})}\right)\right]}\left(y\right), \tag{29}$$

$$p_{Y,\text{NR,non-off}}\left(y\right) = -\frac{1}{y}\frac{1}{d_{\max}^{(\text{I2D})}}\frac{1}{g'_{\text{dB}}\left(g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)\right)}\frac{10}{\ln 10} u_{\left[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{(\text{I2D})}\right)\right]}\left(y\right), \tag{30}$$

$$\overline{Y}_{\text{non-off}}^{d_{\max}^{(\text{I2D})}} = \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max}^{(\text{I2D})})} - \frac{\sigma_c^2\left(2^{\bar{e}}-1\right)}{d_{\max}^{(\text{I2D})}}\frac{\ln 10}{10}\int_{g_{\text{dB}}\left(d_{\max}^{(\text{I2D})}\right)}^{+\infty} 10^{-y'/10} g_{\text{dB}}^{-1}\left(y'\right)dy'. \tag{31}$$

The overall average transmit power used for fulfilling a non-repeated request for content z through D2D offloading, expressed as a function of $d_{\max}$ (whereas the cell radius $d_{\max}^{(\text{I2D})}$ is considered as a parameter), and of the specific requested content z, is given by

$$\begin{aligned}\overline{\mathcal{P}}_{tx,\text{off}}\left(d_{\max},z\right) = & Pr^{(d_{\max})}\left(\text{off.imm} \mid z\right)\overline{Y}_{\text{off,imm}}\left(d_{\max}\right)\\ & + Pr^{(d_{\max})}\left(\text{off.del} \mid z\right)Y_{\text{off,del}}\left(d_{\max}\right)\\ & + \left(1 - Pr^{(d_{\max})}\left(\text{off.imm} \mid z\right) - Pr^{(d_{\max})}\left(\text{off.del} \mid z\right)\right)\overline{Y}_{\text{non-off}}^{d_{\max}^{(\text{I2D})}}.\end{aligned} \tag{32}$$

Finally, our concluding result is the following

[16] We assume that the ROI is fully covered by a set of eNodeBs, each with coverage $d_{\max}^{(\text{I2D})}$.

Theorem 5. *Under the assumptions in Subsection III-A, the average power for fulfilling a non-repeated request is given by*

$$\overline{\mathcal{P}}_{tx}\left(d_{\max}\right)=\sum_{z=1}^{N_Z} p_Z\left(z \mid \mathrm{NR}\right)\overline{\mathcal{P}}_{tx}\left(d_{\max}, z\right), \tag{33}$$

where $p_Z\left(z \mid NR\right)$ *is given by* (20) *and* $\overline{\mathcal{P}}_{tx}\left(d_{\max}, z\right)$ *is given by* (32)*.*

Proof: The average transmit power of a non-repeated request, irrespective of which content is requested, can be obtained by averaging (32) over all the possible events $\left(Z = z \mid \mathrm{NR}\right)$ that the requested content is z, conditioned on the fact that the request is not repeated. The average is obtained assigning a weight $Pr\left(Z = z \mid \mathrm{NR}\right) = p_Z\left(z \mid \mathrm{NR}\right)$, see (20), to the value of the average transmit power required to transmit content z, and summing the product over all the possible realizations of Z. In this way, the desired result (33) is obtained. ■

The expression of the average transmit power as a function of $d_{\max}$ allows to select this system parameter to minimize the overall system energy consumption or, formally,

$$d_{\max}^{(\mathrm{opt})}=\underset{d_{\max}\in\mathbb{R}^{+}}{\arg\min}\overline{\mathcal{P}}_{tx}\left(d_{\max}\right). \tag{34}$$

Despite the the analytical expressions involved in (33) (specifically, (27), (28), (31)) make it hard to compute the optimal $d_{\max}$ in closed form, (33) can be computed through numerical integration in a relatively straightforward way, under the assumption of specific models for $p_Z\left(z\right)$ and $p_V(v)$. Our results (see Section V) show that, as intuition suggests (it can also be proved formally, but we do not do it here for space reasons), $\overline{\mathcal{P}}_{tx}\left(d_{\max}\right)$ is convex with respect to $d_{\max}$, making it straightforward to find the $d_{\max}$ which minimizes the average transmit power, and hence the energy consumption (see footnote 9).

V. Performance evaluation

To evaluate our results, we used a custom simulator written in Matlab. We used the following settings. The ROI consists of a street chunk of length 1,8 Km and width 20 m. The distance between the centers of the two lanes (one lane per marching direction), is 10 m. There is an eNodeBs every 600 m, located at 0, 600, 1200, and 1800 m, respectively, from the left edge of the ROI[17]. Vehicles enter the street with an arrival rate of $\lambda_t = 1/3$ (one vehicle every 3 seconds). The vehicles speed is distributed uniformly in the range $[v_a, v_b]$, with $v_a = 6\,$m/s and $v_b = 16\,$m/s. Each user issues content requests at a rate of 10 requests per minute (including repeated requests). The content library $\mathcal{L}$ has size $N_Z = 10^4$ contents, and the PMF representing the content popularity is a Zipf distribution, i.e $p_Z(z) \sim \frac{1}{\zeta(\alpha)} z^{-\alpha}$, truncated at the value of the library size and with $\alpha = 1.1$. We have assumed contents of equal size of 500 KBytes.

[17]To avoid border effects, we only consider content requests fulfilled when the receiving device is located under the coverage of the two central eNodeBs.

Our simulator reproduces a MAC/physical layer which allows concurrent D2D and/or I2D transmissions to be possibly allocated the same portions of spectrum, provided that the corresponding transmitter-receiver pairs are sufficiently far apart. More specifically, in our implementation, we have used a slightly modified version of the presented in [9]. For space reasons, we do not provide full details of our implementation. Suffice it to say that, differently from [9], we do allow for different transmit power selection in different links, as we assume that transmit power is a function of the transmitter-receiver distance through Eq. (23).

Time is organized in frames, and in each frame both D2D and I2D communications can be scheduled, as a result of the decisions of the CDMS described in Subsection III-B. We have assumed contents of equal size equal to 500 KBytes. The nominal channel model $g(d)$ is the one provided by Equations 5-4 and 5.5 of [16][18]. The system bandwidth is 10 MHz. The central carrier frequency is $2.3\,\mathrm{GHz}$. Content deliveries are scheduled by the CDMS every second, and the radio resource allocation scheduler (see Subsection IV-C) allocates PRBs of width 200 KHz and duration $1\,\mathrm{ms}$. In each scheduling period, the overall number of PRBs that can be assigned is 50000. In each 200 KHz block there are 12 equally spaced subcarriers. The target normalized nominal information rate $\bar{e}$ is set to 5 bps/Hz. Accordingly, a PRB carries 1 Kbit, and each content transmission requires 400 PRBs. The transmit power is selected according to (23), with $\sigma_c^2 = -164\,\mathrm{dBm/Hz}$, plus a link margin of 15 dB[19].

We considered 29 equally spaced values for the maximum D2D transmission distance $d_{\max}$ in the range [20,300] m. For each value of $d_{\max}$, we run 10 independent i.i.d simulations, each lasting 15 minutes, reinitializing the random number generator seed with the same state at the beginning of each batch of 10 simulations. Each simulation is initialized with a random number of vehicles, position and speed of each vehicle, according to the assumptions and the preliminary results in Subsections III-A and IV-A, respectively. The content cache of each node is initialized according to (8). The content timeout is set to $\tau_c = 20$ s, and the sharing timeout to $\tau_s = 600$ s.

Figures 2 and 3 represent the average performance obtained in the simulations (with 95% confidence intervals) and the performance predicted by our model, in terms of offloading efficiency and average required transmit power per subcarrier. It can be seen that the theoretical results provide a perfect match of the performance obtained through simulation, despite our theoretical model overlooks some details, for instance through the unidimensional representation in the model of the ROI (in simulations, we did reproduce two different lanes corresponding to the opposite marching directions). As intuition suggests, the offloading efficiency increases with the maximum transmission range of the devices, since the probability that within a distance $d_{\max}$ there is a neighbor with the desired content, increases. However, keeping increasing the distance indefinitely, does not help in terms of power consumption, as the transmit power to reach "far" neighbors increases exponentially. Moreover, keeping increasing $d_{\max}$, the marginal

[18]The models in [16] are based on large scale measurements campaigns conducted in collaboration with large Telecom companies like Nokia and Docomo. As such, it represents a state of the art reference channel models for a variety of scenarios.

[19]The value selected for σ_c^2 stems from adding a typical noise figure of $F_{rc} = 10\,\mathrm{dB}$ to the typical thermal noise spectral density $N_{0,[\mathrm{dBm}]/\mathrm{Hz}} = -174\,\mathrm{dBm/Hz}$.

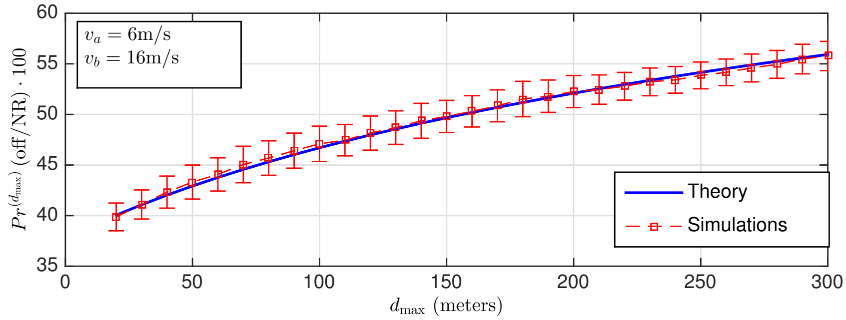

Fig. 2. Offloading efficiency

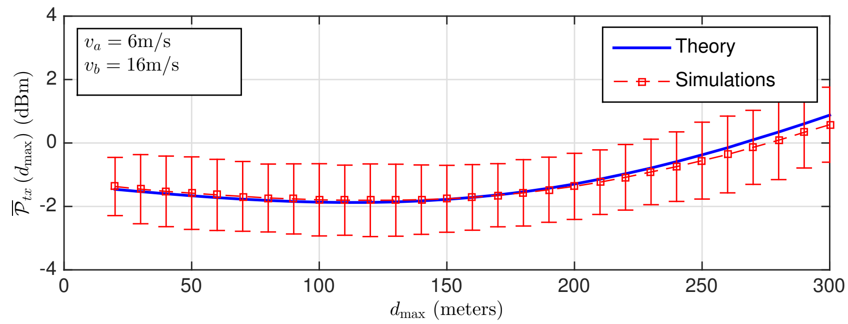

Fig. 3. System-wise average transmit power per subcarrier

gain in terms of offloading efficiency (represented by the slope of the offloading efficiency curve), progressively diminishes. On the other side, decreasing $d_{\max}$ reduces the probability to find neighbors, and hence, for low values of $d_{\max}$, the transmit power is dominated by the term related to I2D transmissions, which may require transmissions at a distance larger than $d_{\max}$ (up to 300 m in our example). As a result, there is an optimal value for the maximum distance, whose selection guarantees the minimization of the average transmit power, and therefore of the overall system energy consumption, which is related to the average transmit power through a constant term.

VI. Conclusion

We have considered a D2D data offloading content delivery system for a mobile environment, and specifically a vehicular scenario. We have derived an analytical model based on results of the theory of point processes (in this case unidimensional) to compute the average system offloading probability (or equivalently, offloading efficiency) and average transmit power (or equivalently, the system wise energy consumption). The derived expressions are function of different parameters related to various domains, namely, to the content request process (content interest distribution and requests arrival rate), the vehicles mobility model and density, to the path loss model, and to the range within which devices should be considered as neighbors, or, equivalently, the maximum power at which devices should be allowed to transmit. We have checked that the proposed model, although obtained by overlooking several details of the system geometry, is able to predict in an accurate way the optimal value for the devices transmission range or, equivalently, the maximum transmit power that should be allowed to the devices.

ACKNOWLEDGMENT

This work was partially funded by the EC under the H2020 REPLICATE (691735), SoBigData (654024) and AUTOWARE (723909) projects.

REFERENCES

[1] J. Whitbeck, Y. Lopez, J. Leguay, V. Conan, and M. D. de Amorim, “Push-and-track: Saving infrastructure bandwidth through opportunistic forwarding,” in *Pervasive and Mobile Computing*, vol. 8, no. 5, 2012, pp. 682–697.

[2] R. Bruno, A. Masaracchia, and A. Passarella, “Offloading through opportunistic networks with dynamic content requests,” in *Proc. IEEE MASS ’14*, Oct. 2014.

[3] M. V. Barbera, A. C. Viana, M. D. De Amorim, and J. Stefa, “Data offloading in social mobile networks through VIP delegation,” *Ad Hoc Networks*, vol. 19, pp. 92–110, 2014.

[4] F. Rebecchi, M. Dias de Amorim, and V. Conan, “Circumventing plateaux in cellular data offloading using adaptive content reinjection,” *Computer Networks*, vol. 106, pp. 49–63, 2016.

[5] F. Rebecchi, L. Valerio, R. Bruno, V. Conan, M. D. De Amorim, and A. Passarella, “A joint multicast/D2D learning-based approach to LTE traffic offloading,” *Computer Communications*, vol. 72, pp. 26–37, 2015.

[6] F. Rebecchi et al., “Data Offloading Techniques in Cellular Networks: A Survey,” *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 580–603, 2015.

[7] A. Asadi and V. Mancuso, “Network-Assisted Outband D2D-Clustering in 5G Cellular Networks: Theory and Practice,” *IEEE Trans. Mob. Comput.*, vol. 16, no. 8, pp. 2246–2259, 2017.

[8] X. Lin, J. G. Andrews, and A. Ghosh, “Spectrum Sharing for Device-to-Device Communication in Cellular Networks,” *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6727–6740, 2014.

[9] Y. Yang, T. Liu, X. Ma, H. Jiang, and J. Liu, “FRESH: Push the Limit of D2D Communication Underlaying Cellular Networks,” *IEEE Tran. Mob. Comput.*, vol. 16, no. 6, pp. 1630–1643, jun 2017.

[10] 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. 1, pp. 176–189, Jan. 2016.

[11] ——, “Fundamental Limits of Caching in Wireless D2D Networks,” *IEEE Trans. Information Theory*, vol. 62, no. 2, pp. 849–869, feb 2016.

[12] L. Pescosolido, M. Conti, and A. Passarella, “Performance evaluation of an energy efficient traffic offloading protocol for vehicular networks,” in *Proc. 1st International Balkan Conference on Communications and Networking (BalkanCom ’17)*, Tirana, AL, May30 - Jun. 2 2017.

[13] M. Afshang, H. S. Dhillon, and P. H. Joo Chong, “Modeling and performance analysis of clustered device-to-device networks,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4957–4972, Jul. 2016.

[14] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes. Vol. II*, 2008, vol. I.

[15] D. Moltchanov, “Distance distributions in random networks,” *Ad Hoc Networks*, vol. 10, no. 6, pp. 1146–1166, 2012.

[16] ICT METIS Project Deliverable 1.4, “METIS Channel Models,” Tech. Rep., 2015.

APPENDIX A
PROOF OF LEMMAS 1-3

Proof of Lemma 1: Under the assumption that the speeds of the vehicles are statistically independent, the fraction of vehicles with speed in the range $[v, v + \Delta v]$, for a finite Δv, is equal to the probability that the speed of a vehicle entering the ROI is in the range $[v, v + \Delta v]$, i.e., $Pr\left(V \in [v, v + \Delta v]\right) = \int_v^{v+\Delta v} p_V(v')dv'$. The process of arrival of vehicles with speed in the same range corresponds to the overall homogeneous TPPP of vehicles arrivals, thinned with the probability $Pr\left(V \in [v, v + \Delta v]\right)$. Accordingly, the arrival rate of the thinned TPPP is $\lambda_{t,[v,v+\Delta v]} = \lambda_t \int_v^{v+\Delta v} p_V(v')dv'$. This expression allows us to write the differential of the (overall) arrival rate λ_t as a function of the differential of the speed dv. Specifically[20]

$$d\lambda_t = \lambda_t p_V(v)dv. \tag{35}$$

This can be considered as the arrival rate of the elementary process of arrival of vehicles with speed equal to an exact value v.

In one second, a vehicle traverses a street span of $|v| \cdot 1$ meters. In the meantime, on average, $d\lambda_t$ vehicles have entered the same street span of length $|v| \cdot 1$ meters. Transforming the time elapsed from entering the ROI into the distance spanned, since the speed is constant and the arrival instants are independent, it can be easily showed that the *spatial* distribution of such vehicles is a homogeneous SPPP with density, obtained using (35),

$$d\rho = \frac{1}{|v|}d\lambda_t = \frac{1}{|v|}\lambda_t p_V(v)dv. \tag{36}$$

The overall spatial point process of vehicles present on the street, at any given instant, is the superposition of an infinite number of elementary homogeneous SPPPs of the kind above (one for each value of v). Therefore, it is again a homogeneous SPPP with linear density given by[21] $\rho = \int_0^\rho d\rho' = \int_{-\infty}^{+\infty} \frac{1}{|v|}\lambda_t p_V(v)dv$, which coincides with (3).

Using (2) in (3) we obtain

$$\begin{aligned}\rho &= \frac{\lambda_t}{2(v_b - v_a)}\int_{-\infty}^{+\infty}\frac{1}{|v|}\left(u_{[-v_b,-v_a]}(v) + u_{[v_a,v_b]}(v)\right)dv\\ &= \frac{\lambda_t}{2(v_b - v_a)}2\int_{v_a}^{v_b}\frac{1}{v}dv = \frac{\lambda_t(\ln v_b - \ln v_a)}{(v_b - v_a)},\end{aligned}$$

i.e, (4).

Let $x(t)$ be the trajectory of a vehicle moving at speed $v*$, and $x'(t) = x(t) + d_{\max}$ the trajectory of a point displaced at a distance $d_{\max}$ from it. The time process of the instants at which $x'(t)$ coincides with the position of other vehicles, *that have a specific speed* v, is determined by the relative speed $(v - v^*)$. Consider a reference system which moves across space and has origin, at

[20] We use the identity $x^* = \int_0^{x^*} dx$, that holds for any Real quantity x^*.

[21] We use the transformation of the integration variable from ρ to v using the relation (36) between their differentials.

each instant, in $x'(t)$. In this reference system, the speed of the considered vehicles is $(v - v^*)$. The vehicles "encountered" by the initially considered vehicle moving at speed v^*, during an interval of 1 second starting at a given instant t_0, are those that, at instant t_0, are positioned (in the new reference system) along the segment originating at the position $x_a = 0$ (the position of the displaced point in the new reference system) and the position $x_b = (v - v^*)$. Since the positions of the vehicles at t_0 is a homogeneous SPPP (with density $d\lambda_s$ given by (36)) and $(v - v^*)$ is constant, transforming travelled distances into time intervals we can claim that the set of the instants at which the devices are encountered is a homogeneous TPPP. We call the rate of this process "elementary encountering rate", and indicate it with $d\lambda_e^*$. We can compute $d\lambda_e^*$ using (36)[22], by replacing $d\lambda_t$ with $d\lambda_e^*$, obtaining[23]

$$d\lambda_e^* = |v^* - v|\, d\rho = \frac{|v^* - v|}{|v|} \lambda_t p_V(v) dv, \tag{37}$$

Following the same line of reasoning used above, the temporal process of vehicles moving at *any* speed encountered by a vehicle moving at speed v^* is the superposition of an infinite number of elementary "encountering processes" of the kind above. Accordingly, the rate at which the a point moving at speed v^* encounters other vehicles at *any* speed is

$$\lambda_e^* = \int_{-\infty}^{\infty} \lambda_t p_V(v) \frac{|v^* - v|}{|v|} dv. \tag{38}$$

Thus, we have obtained (5). Finally, using (2) in (38), with a few integral calculus steps, it is easy to obtain (6).

Proof of Lemma 2: Result (7) can be obtained as follows: since the content request arrival process of each device is a homogeneous TPPP, and the requests are statistically independent, the process of issuing requests for a specific content z by a given device is again a homogeneous TPPP, which results from thinning the TPPP of the overall requests issued by a device, with the probability that the requested content is z, i.e., $p_Z(z)$. The arrival rate of the thinned process, λ_z, is the product of the arrival rate of the content request arrival process of each device, λ_Z, by the same probability, or $p_Z(z)$, or $\lambda_z = p_Z(z)\lambda_Z$, i.e., (7).

Result (8) provides upper and lower bounds to the probability $Pr(\mathcal{C} \ni z)$ that at a given instant t, the cache $\mathcal{C}$ of a generic device contains a specific content z. We observe that a device cache holds a content z at a given instant t if the device has previously requested the content, has obtained it, and has not yet removed it from its cache (due to the expiration of the sharing timeout τ_s). The probability of this event is lower bounded by the probability that the device has requested the content (one or more times) in the interval $[t - \tau_s, t - \tau_c]$ and upper bounded by the probability that the device has requested the content (one or more times) in the interval $[t - \tau_s, t]$.

[22] Essentially, (36) expresses the relation between spatial and temporal elementary intensities (i.e., a linear density and a rate) of the two elementary PPPs we are considering: the SPPP of the positions of vehicles traveling at a specific speed v, and the TPPP of the instant they cross a given point.

[23] In (37), we use the modulus to let all the possible values of speeds v^* and v result in positive encountering rates. These possibilities encompass having vehicles traveling in the same or opposite directions.

Since the content request process for content z by a generic device is a TPPP with rate given by (7), the number of content requests in a given interval is a Poisson random variable with parameter equal to the product of λ_z times the duration of the interval. Accordingly, $Pr\left(\mathcal{C} \ni z\right)$ has the upper and lower bounds in (8).

Result (9) can be obtained building on the SPPP with density ρ of the devices' positions at a given instant. Since the content requests are independent across devices, the content of their caches at a given instant are approximately independent[24]. The presence of content z in the caches of devices located in the ROI is a again a SPPP of the same type, but with density given by $\rho_z = Pr\left(\mathcal{C} \ni z\right)\rho$, i.e, (9).

Result (10) provides the rate at which a point moving at speed v^* encounters vehicles with a specific content z in their caches. In Subsection IV-A (Lemma 1), we showed that the process of encountering devices, for a point moving at speed v^*, is a homogeneous TPPP with rate $\lambda_e^{(v^*)}$. Due to the (approximate) independence of the caches contents, the process of encountering devices that have a specific content z in their caches is still a homogeneous TPPP resulting from thinning the overall encountering TPPP with the probability $Pr\left(\mathcal{C} \ni z\right)$ that content z is cached at the encountered devices. The corresponding encountering rate is given by the product of the overall encountering rate (5) times $Pr\left(\mathcal{C} \ni z\right)$, or $\lambda_{e,z}^{(v^*)} = Pr\left(\mathcal{C} \ni z\right)\lambda_e^{(v^*)}$, i.e., (10), with $\lambda_e^{(v^*)}$ given by (5).

Proof of Lemma 3: Consider a device whose cache content is $\mathcal{C}$ which issues a content request. The event that the request is non-repeated, prior to the realization of the specific requested content, is the union of the (overlapping) events $\left(\mathcal{C} \not\ni z\right), \forall z \in \mathcal{L}$. By the law of total probability, we have

$$Pr\left(\mathrm{NR}\right) = \sum_z Pr\left(Z = z\right) Pr\left(\mathcal{C} \not\ni z / Z = z\right). \tag{39}$$

Under the assumption on the content request process in Subsection III-A, the random variable Z representing the requested content, and the set $\mathcal{C}$ of the contents in the cache at the time of request, are statistically independent. Therefore, we can write $Pr\left(\mathcal{C} \not\ni z \mid Z = z\right) = Pr\left(\mathcal{C} \not\ni z\right)$, which plugged in (39), gives the desired result (19).

Now, the probability that the requested content is z, conditioned on the fact that the request is not repeated, is

$$p_Z\left(z \mid \mathrm{NR}\right) \triangleq Pr\left(Z = z \mid \mathrm{NR}\right) = \frac{Pr\left(Z = z, \mathrm{NR}\right)}{Pr\left(\mathrm{NR}\right)}. \tag{40}$$

The joint probability $Pr\left(Z = z, \mathrm{NR}\right)$ can be written, by the Bayes Theorem, as $Pr\left(Z = z, \mathrm{NR}\right) = Pr\left(\mathrm{NR} \mid Z = z\right) Pr(Z = z)$. But $Pr\left(\mathrm{NR} \mid Z = z\right)$ is anything but the

[24]There is a small correlation among the contents present in the devices' caches at a given instant t. This is related to the presence of the contents requested in the interval $[t - \tau_c, t]$, whose reception depends on the composition of the surrounding nodes' caches during that interval. Compared to the amount of contents received in the interval $[t, t - \tau_c]$, which are certainly in the devices' caches, since $\tau_s \gg \tau_c$, the former set of contents has a minimal weight in the overall statistics of the caches. The latter sets of contents are certainly independent since they have been received, irrespective of the devices trajectories and caches during the interval $[t - \tau_s, t]$.

probability that the cache of the requesting node does not contain the specific content z (once it has been determined), i.e., $Pr\left(\mathcal{C} \not\ni z\right)$. Therefore, we can replace the numerator in (40) with the product $Pr\left(Z = z\right) Pr\left(\mathcal{C} \not\ni z\right)$. Finally, replacing the denominator in (40) with (39), we obtain the desired result (19).

APPENDIX B
PROOF OF THE RESULTS IN SUBSECTION IV-C

A. Cumulative Distribution Function, Probability Density Function, and average value of the transmit power used for immediate content delivery through D2D.

In the following, we provide the steps to obtain expressions (25), (26), and (27). The CDF of $Y_{\text{NR,off,im}}$ in (25) can be computed as

$$\begin{aligned} F_{Y,\text{NR,off,im}}\left(y/D \le d_{\max}\right) &= Pr\left(Y_{\text{NR,off,im}} \le y | D \le d_{\max}\right) \\ &= Pr\left(\frac{1}{g(D)}\sigma_c^2\left(2^{\bar{e}}-1\right) \le y \middle| D \le d_{\max}\right) \\ &= Pr\left(g(D) \ge \frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right) \middle| D \le d_{\max}\right) \\ &= Pr\left(g_{\text{dB}}(D) \ge 10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right) \middle| D \le d_{\max}\right) \\ &= Pr\left(D \le g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) \middle| D \le d_{\max}\right). \end{aligned}$$

i.e, it is given by the distribution function of the distance among a point the SPPP process and its closest neighbor, conditioned to the distance being less than $d_{\max}$, and evaluated at $g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)$. The conditional distribution $F_D\left(d/D \le d_{max}\right) \triangleq Pr\left(D \le d/D \le d_{max}\right) = \frac{Pr(D \le d, D \le d_{max})}{Pr(D \le d_{max})}$, for the distances $d \le d_{\max}$ of interest[25], coincides with $Pr\left(D \le d\right)/Pr\left(D \le d_{max}\right)$. Using the expression of the CDF of the nearest neighbor distance for homogeneous unidimensional SPPPs $F_D\left(d\right) = 1 - e^{-\rho_z 2d}$, we obtain $F_D\left(d/D \le d_{max}\right) = u_{[0,d_{\max}]}(d)\left(1 - e^{-\rho_z 2d}\right)/\left(1 - e^{-\rho_z 2d_{\max}}\right)$, and ultimately,

$$F_{Y,\text{NR,off,im}}\left(y/D \le d_{\max}\right) = \frac{1}{1 - e^{-\rho_z 2d_{\max}}}\left(1 - e^{-\rho_z 2 g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)}\right) u_{[0,\sigma(2^{\bar{e}}-1)/g(d_{\max})]}\left(y\right),$$

i.e., Equation (25).

For the sake of completeness, it is worth also computing the PDF of $Y_{\text{NR,off,im}}$, even though it is not required to compute its average value, since the integration by parts in the computation

[25]By construction, $g^{-1}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)$ is certainly less than $d_{\max}$

of the average value gets rid of the PDF (see below). The PDF of $Y_{\text{NR,off,im}}$ is given by the first derivative of the CDF, i.e.:

$$\begin{aligned}
p_{Y,\text{NR,off,im}}(y) &= \frac{d}{dy} F_{Y,\text{off,im}}\left(y/D \leq d_{\max}\right) \\
&= \frac{d}{dy}\left(\frac{1-e^{-\rho_z 2 g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)\right)}}{1-e^{-2\rho_z d_{\max}}} u_{[0,\sigma(2^{\bar{e}}-1)/g(d_{\max})]}\left(y\right)\right) \\
&= \frac{1}{1-e^{-2\rho_z d_{\max}}}\frac{d}{dy}\left(1-e^{-\rho_z 2 g^{-1}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)}\right) u_{[0,\sigma(2^{\bar{e}}-1)/g(d_{\max})]}\left(y\right) \\
&= \frac{1}{1-e^{-2\rho_z d_{\max}}}\frac{d}{dy}\left(\rho_z 2 g^{-1}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) e^{-\rho_z 2 g^{-1}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)} u_{[0,\sigma(2^{\bar{e}}-1)/g(d_{\max})]}\left(y\right) \\
&= \frac{2\rho_z}{1-e^{-\rho_z 2 d_{\max}}}\frac{d}{dy}\left(g^{-1}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) e^{-\rho_z 2 g^{-1}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)} u_{[0,\sigma(2^{\bar{e}}-1)/g(d_{\max})]}\left(y\right).
\end{aligned}$$

Applying the rule for derivative of an inverse function $\frac{d}{dx}f^{-1}\left(x\right) = \frac{1}{f'\left(f^{-1}(x)\right)}$, and the chain rule for the derivative of nested functions $\frac{d}{dx}f\left(g\left(x\right)\right) = f'\left(g\left(x\right)\right)g'\left(x\right)$, we obtain

$$\begin{aligned}
p_{Y,\text{NR,off,im}}(y) =& \frac{2\rho_z}{1-e^{-\rho_z 2 d_{\max}}}\frac{1}{g'\left(g^{-1}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)}\frac{d}{dy}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)e^{-\rho_z 2 g^{-1}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)} \\
&\cdot u_{[0,\sigma(2^{\bar{e}}-1)/g(d_{\max})]} \\
=& \frac{2\rho_z}{1-e^{-\rho_z 2 d_{\max}}}\frac{1}{g'\left(g^{-1}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)}\frac{-\sigma_c^2\left(2^{\bar{e}}-1\right)}{y^2}e^{-\rho_z 2 g^{-1}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)} \\
&\cdot u_{[0,\sigma(2^{\bar{e}}-1)/g(d_{\max})]} \\
=& -\frac{1}{y^2}\frac{2\rho_z\sigma_c^2\left(2^{\bar{e}}-1\right)}{\left(1-e^{-\rho_z 2 d_{\max}}\right)}\frac{e^{-\rho_z 2 g^{-1}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)}}{g'\left(g^{-1}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)}u_{[0,\sigma(2^{\bar{e}}-1)/g(d_{\max})]}.
\end{aligned}$$

Note that this expression is always a non-negative quantity, since $g'\left(\cdot\right)$ is always negative, because the channel gain is a decreasing function of its argument.

The average value of $Y_{\text{off,im}}$ in can be computed as follows

$$\begin{aligned}
\overline{Y}_{\text{NR,off,im}}\left(d_{\max}\right) &= \int_0^{\sigma_c^2\left(2^{\bar{e}}-1\right)/g\left(d_{\max}\right)} y \cdot p_{Y,\text{off,im}}\left(y\right) dy \\
&= \left[y \cdot F_{Y,\text{off,im}}\left(y\right)\right]_0^{\frac{\sigma_c^2\left(2^{\bar{e}}-1\right)}{g\left(d_{\max}\right)}} - \int_0^{\sigma_c^2\left(2^{\bar{e}}-1\right)/g\left(d_{\max}\right)} F_{\text{NR,off,im}}\left(y\right) dy \\
&= \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max})} \frac{\left(1 - e^{-\rho_z 2 g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{g(d_{\max})}{\sigma_c^2(2^{\bar{e}}-1)}\sigma_c^2(2^{\bar{e}}-1)\right)\right)}\right)}{1 - e^{-\rho_z 2 d_{\max}}} \\
&\quad - \frac{1}{1 - e^{-\rho_z 2 d_{\max}}} \int_0^{\sigma_c^2\left(2^{\bar{e}}-1\right)/g\left(d_{\max}\right)} \left(1 - e^{-\rho_z 2 g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)\right)}\right) dy \\
&= \frac{1}{1 - e^{-\rho_z 2 d_{\max}}} \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max})} \left(1 - e^{-\rho_z 2 g_{\text{dB}}^{-1}\left(g_{\text{dB}}\left(d_{\max}\right)\right)}\right) \\
&\quad - \frac{1}{1 - e^{-\rho_z 2 d_{\max}}} \int_0^{\sigma_c^2\left(2^{\bar{e}}-1\right)/g\left(d_{\max}\right)} \left(1 - e^{-\rho_z 2 g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)\right)}\right) dy \qquad (41) \\
&= \frac{1}{1 - e^{-\rho_z 2 d_{\max}}} \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max})} \left(1 - e^{-\rho_z 2 d_{\max}}\right) - \frac{1}{1 - e^{-\rho_z 2 d_{\max}}} \int_0^{\sigma_c^2\left(2^{\bar{e}}-1\right)/g\left(d_{\max}\right)} dy \\
&\quad + \frac{1}{1 - e^{-\rho_z 2 d_{\max}}} \int_0^{\sigma_c^2\left(2^{\bar{e}}-1\right)/g\left(d_{\max}\right)} e^{-\rho_z 2 g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)\right)} dy \qquad (42)
\end{aligned}$$

We now apply a change of the integration variable y to the variable

$$y'\left(y\right) \triangleq 10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right), \qquad (43)$$

which entails the inverse relation

$$y = \frac{\sigma_c^2\left(2^{\bar{e}}-1\right)}{10^{y'/10}}.$$

The derivative of y' with respect to y is

$$\begin{aligned}
\frac{d}{dy}y'(y) &= \frac{10}{\ln 10}\frac{d}{dy}\ln\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right) \\
&= \frac{10}{\ln 10}\frac{1}{\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)}\frac{d}{dy}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right) \\
&= \frac{10}{\ln 10}\frac{y}{\sigma_c^2\left(2^{\bar{e}}-1\right)}\frac{d}{dy}\left(\frac{1}{y/\sigma_c^2\left(2^{\bar{e}}-1\right)}\right) \\
&= \frac{10}{\ln 10}\frac{y}{\sigma_c^2\left(2^{\bar{e}}-1\right)}\left(-\frac{1}{\left(y/\sigma_c^2\left(2^{\bar{e}}-1\right)\right)^2}\right)\frac{d}{dy}\frac{y}{\sigma_c^2\left(2^{\bar{e}}-1\right)} \\
&= -\frac{10}{\ln 10}\frac{y}{\sigma_c^2\left(2^{\bar{e}}-1\right)}\left(\frac{\sigma_c^2\left(2^{\bar{e}}-1\right)}{y}\right)^2\frac{1}{\sigma_c^2\left(2^{\bar{e}}-1\right)} \\
&= -\frac{10}{\ln 10}\frac{\sigma_c^2\left(2^{\bar{e}}-1\right)}{y}\frac{1}{\sigma_c^2\left(2^{\bar{e}}-1\right)} \\
&= -\frac{1}{y}\frac{10}{\ln 10}
\end{aligned}$$

The upper and lower integration extremes of the integral appearing in (42), in terms of y', are

$$y'\left(\frac{\sigma_c^2(2^{\bar{e}}-1)}{g\left(d_{\max}^{(\mathrm{I2D})}\right)}\right) = g_{\mathrm{dB}}\left(d_{\max}^{(\mathrm{I2D})}\right)$$
$$y'(0) = +\infty.$$

So, (41) becomes

$$\begin{aligned}
\overline{Y}_{\mathrm{NR,off,im}}\left(d_{\max}\right) =& \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max})}\left(1-\frac{1}{1-e^{-\rho_z 2d_{\max}}}\right) \\
&+\frac{1}{1-e^{-\rho_z 2d_{\max}}}\int_0^{\sigma_c^2(2^{\bar{e}}-1)/g(d_{\max})}\left(-y\frac{\ln 10}{10}\right)\left(-\frac{1}{y}\frac{10}{\ln 10}\right)e^{-\rho_z 2 g_{\mathrm{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2(2^{\bar{e}}-1)\right)\right)}dy \\
=& \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max})}\left(1-\frac{1}{1-e^{-\rho_z 2d_{\max}}}\right) \\
&-\frac{1}{1-e^{-\rho_z 2d_{\max}}}\frac{\ln 10}{10}\int_{+\infty}^{g_{\mathrm{dB}}(d_{\max})}\frac{\sigma_c^2\left(2^{\bar{e}}-1\right)}{10^{y'/10}}e^{-\rho_z 2 g_{\mathrm{dB}}^{-1}(y')}dy' \\
=& \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max})}\left(\frac{1}{1-e^{\rho_z 2d_{\max}}}\right)+\frac{\sigma_c^2\left(2^{\bar{e}}-1\right)}{1-e^{-\rho_z 2d_{\max}}}\frac{\ln 10}{10}\int_{g_{\mathrm{dB}}(d_{\max})}^{+\infty}10^{-y'/10}e^{-\rho_z 2 g_{\mathrm{dB}}^{-1}(y')}dy',
\end{aligned} \tag{44}$$

i.e., (27).

B. Probability Distribution Function, Probability Density Function, and average value of the transmit power used for delayed content delivery through I2D.(31)

The CDF of the random variable $Y_{\text{NR,non-off}}$ defined in (24) can be computed from the distribution of the transmission distance from the eNodeB to the device, which is uniform in the interval $[0, d_{\max}^{(\text{I2D})}]$, or $F_D(d) = \frac{1}{d_{\max}^{(\text{I2D})}} d \cdot u_{[0,d_{\max}^{(\text{I2D})}]}(d)$. Specifically, we have:

$$\begin{aligned}
F_{Y,\text{NR,non-off}}(y) &= Pr\left(Y_{\text{NR,non-off}} \leq y\right) \\
&= Pr\left(\frac{1}{g(D)}\sigma_c^2\left(2^{\bar{e}}-1\right) \leq y\right) \\
&= Pr\left(g(D) \geq \frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right) \\
&= Pr\left(g_{\text{dB}}(D) \geq 10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) \\
&= Pr\left(D \leq g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)\right)
\end{aligned}$$

i.e, and ultimately,

$$F_{Y,\text{NR,non-off}}(y) = \frac{1}{d_{\max}^{(\text{I2D})}} g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) u_{[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{(\text{I2D})}\right)]}(y),$$

i.e., Equation (29).

For the sake of completeness, it is worth also computing the PDF of $Y_{\text{NR,non-off}}$, even though it is not required to compute its average value, since the integration by parts in the computation of the average value gets rid of the PDF (see below). The PDF of $Y_{\text{NR,non-off}}$ is given by the first derivative of the CDF, i.e.:

$$\begin{aligned}
p_{Y,\text{NR,non-off}}(y) &= \frac{d}{dy} F_{Y,\text{non-off}}\left(y/D \leq d_{\max}^{(\text{I2D})}\right) \\
&= \frac{d}{dy}\left(\frac{1}{d_{\max}^{(\text{I2D})}} g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) u_{[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{(\text{I2D})}\right)]}(y)\right) \\
&= \frac{1}{d_{\max}^{(\text{I2D})}} \frac{d}{dy}\left(g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)\right) u_{[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{(\text{I2D})}\right)]}(y) \\
&= \frac{1}{d_{\max}^{(\text{I2D})}} \frac{d}{dy}\left(g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)\right) u_{[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{(\text{I2D})}\right)]}(y).
\end{aligned}$$

Applying the rule for the derivative of an inverse function $\frac{d}{dx} f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}$, and the chain rule for the derivative of nested functions $\frac{d}{dx} f\left(g\left(x\right)\right) = f'\left(g\left(x\right)\right) g'\left(x\right)$, we obtain

$$\begin{aligned}
p_{Y,\text{NR,non-off}}(y) =& \frac{1}{d_{\max}^{\text{(I2D)}}} \frac{1}{g'_{\text{dB}}\left(g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)\right)} \frac{d}{dy}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) \\
&\cdot u_{[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{\text{(I2D)}}\right)]}\left(y\right) \\
=& \frac{1}{d_{\max}^{\text{(I2D)}}} \frac{1}{g'_{\text{dB}}\left(g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)\right)} \frac{10}{\ln 10} \frac{d}{dy}\left(\ln\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) \\
&\cdot u_{[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{\text{(I2D)}}\right)]}\left(y\right) \\
=& \frac{1}{d_{\max}^{\text{(I2D)}}} \frac{1}{g'_{\text{dB}}\left(g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)\right)} \frac{10}{\ln 10} \frac{y}{\sigma_c^2\left(2^{\bar{e}}-1\right)} \frac{d}{dy}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right) \\
&\cdot u_{[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{\text{(I2D)}}\right)]}\left(y\right) \\
=& -\frac{1}{y} \frac{1}{d_{\max}^{\text{(I2D)}}} \frac{1}{g'_{\text{dB}}\left(g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right)\right)} \frac{10}{\ln 10} u_{[0,\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{\text{(I2D)}}\right)]}\left(y\right)
\end{aligned}$$

Note that this expression is always a non-negative quantity, since $g'\left(\cdot\right)$ is always negative, because the channel gain is a decreasing function of its argument.

The average value of $Y_{\text{non-off}}$ in can be computed as follows

$$\begin{aligned}
\overline{Y}_{\text{NR,non-off}}^{\left(d_{\max}^{\text{(I2D)}}\right)} &= \int_{-\infty}^{+\infty} y \cdot p_{Y,\text{non-off}}\left(y\right) dy = \int_0^{\sigma_c^2(2^{\bar{e}}-1)/g\left(d_{\max}^{\text{(I2D)}}\right)} y \cdot p_{Y,\text{non-off}}\left(y\right) dy \\
&= \left[y \cdot F_{Y,\text{non-off}}\left(y\right)\right]_0^{\frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max}^{\text{(I2D)}})}} - \int_0^{\sigma_c^2(2^{\bar{e}}-1)/g(d_{\max}^{\text{(I2D)}})} F_{\text{NR,non-off}}\left(y\right) dy \\
&= \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max}^{\text{(I2D)}})} - \int_0^{\sigma_c^2(2^{\bar{e}}-1)/g(d_{\max}^{\text{(I2D)}})} \frac{1}{d_{\max}^{\text{(I2D)}}} g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) dy \\
&= \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max}^{\text{(I2D)}})} - \frac{1}{d_{\max}^{\text{(I2D)}}} \int_0^{\sigma_c^2(2^{\bar{e}}-1)/g(d_{\max}^{\text{(I2D)}})} g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) dy. \qquad (45)
\end{aligned}$$

We now apply the same change of the integration variable (43) used to compute (44). In this way, (45) becomes

$$
\begin{aligned}
\overline{Y}^{\left(d_{\max}^{(\mathrm{I2D})}\right)}_{\text{NR,non-off}} =& \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max}^{(\mathrm{I2D})})} - \frac{1}{d_{\max}^{(\mathrm{I2D})}} \int_0^{\frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max}^{(\mathrm{I2D})})}} \left(-y\frac{\ln 10}{10}\right)\left(-\frac{1}{y}\frac{10}{\ln 10}\right) g_{\text{dB}}^{-1}\left(10\log_{10}\left(\frac{1}{y}\sigma_c^2\left(2^{\bar{e}}-1\right)\right)\right) dy \\
=& \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max}^{(\mathrm{I2D})})} + \frac{1}{d_{\max}^{(\mathrm{I2D})}} \frac{\ln 10}{10} \int_{+\infty}^{g_{\text{dB}}\left(d_{\max}^{(\mathrm{I2D})}\right)} \frac{\sigma_c^2\left(2^{\bar{e}}-1\right)}{10^{y'/10}} g_{\text{dB}}^{-1}\left(y'\right) dy' \\
=& \frac{\sigma_c^2(2^{\bar{e}}-1)}{g(d_{\max}^{(\mathrm{I2D})})} - \frac{\sigma_c^2\left(2^{\bar{e}}-1\right)}{d_{\max}^{(\mathrm{I2D})}} \frac{\ln 10}{10} \int_{g_{\text{dB}}\left(d_{\max}^{(\mathrm{I2D})}\right)}^{+\infty} 10^{-y'/10} g_{\text{dB}}^{-1}\left(y'\right) dy',
\end{aligned}
$$

i.e, (31).