On Fundamental Trade-offs of Device-to-Device Communications in Large Wireless Networks

Andrés Altieri, *Member, IEEE*, Pablo Piantanida, *Member, IEEE*,
Leonardo Rey Vega, *Member, IEEE*, and Cecilia G. Galarza

***Abstract*—This paper studies the gains, in terms of served requests, attainable through out-of-band device-to-device (D2D) video exchanges in large cellular networks. A stochastic framework, in which users are clustered to exchange videos, is introduced, considering several aspects of this problem, i.e., the video caching policy, user matching for exchanges, and aspects regarding scheduling and transmissions. A family of *admissible protocols* is introduced: in each protocol the users are clustered by means of a hard-core point process and, within the clusters, video exchanges take place. Two metrics, quantifying the "local" and "global" fractions of video requests served through D2D are defined, and relevant trade-off regions involving these metrics, as well as quality-of-service constraints, are identified. A simple communication strategy is proposed and analyzed to obtain inner bounds to the trade-off regions and to draw conclusions on the performance attainable through D2D. To this end, the analysis of the time-varying interference that the nodes experience and the tight approximations of its Laplace transform are derived.**

***Index Terms*—Cooperative communications, Device-to-device, outage probability, stochastic geometry.**

I. Introduction

A. Motivation and Related Work

CELLULAR device-to-device (D2D) communications, in which two or more mobile users establish a direct link without going through the Base Station (BS), have emerged as a viable alternative to partially cope with the increasing requirements that cellular networks will face in the future. Generally speaking, D2D communications are opportunistic, one-hop, short range transmissions in which the BS can be used for coordination and acts as a last case fall-back alternative [1]. This allows, for a higher spatial frequency reuse, energy efficiency, coverage extension, and a reduced backhaul load. The scope of D2D communications is very wide, from machine-to-machine, gaming and relaying, to content distribution, and public safety networks [1]. Among these, an important application is video content distribution. This is because, in the next few years, traffic from wireless and mobile devices will exceed traffic from wired devices, and this will be largely related to an increase in *video on demand* (VoD) and Internet-to-TV downloads [2]. The asynchronous nature of VoD requests implies that, in many cases, multicasting strategies for video transmissions cannot be employed, even though a small library of videos may be accessed by many users [3]. A recent approach to mitigate this consists in including small distributed caching stations with a limited backhaul that can locally serve video demands [4], [5]. Another approach [3] is to take advantage of the unused storage space available in many wireless devices to store and exchange videos locally. For example, a user may keep its watched videos to satisfy nearby requests, or certain videos could be cached during moments of low network load. In this paper we focus on the second approach, which does not require dedicated storage units. Our main goal is to study the potential benefits achievable through a distributed user-caching strategy, by considering the fraction of mean video requests that could be served through D2D, without requiring the BS to transmit them. This may yield some insight on the impact of D2D in terms of video availability and backhaul load, which may have implications both economically and in terms of quality of service. To this end, we introduce a simple framework for analysis, based on a stochastic geometry model [6], [7]. In this framework, users are assumed to be grouped into clusters where D2D video exchanges take place. We attempt to consider the problem of establishing matches between requesting and caching users, and the problem of scheduling and transmitting, involving slow fading, path loss, and interference between nodes. We focus on *out-of-band* D2D, which uses bands outside the cellular ones, increasing the frequency reuse and mitigating interference in the cellular band. We also study the trade-offs between the fraction of requests served "locally" (per cluster) and "globally" (in an arbitrary region) in the network.

Some related works which focus on D2D through local distributed caching include [3], [5], [8]–[11]. These works consider finite area networks with a fixed number of users, distributed uniformly or on a regular grid. The model for transmission failure is generally the distance-based *protocol model* introduced

Manuscript received May 20, 2014; revised October 30, 2014 and April 2, 2015; accepted April 20, 2015. Date of publication May 6, 2015; date of current version September 7, 2015. This work was supported in part by the Eiffel Scholarship of France, by Project UBACyT under Grant 2002013100751BA, and by the FP7 Network of Excellence in Wireless Communications NEWCOM#. The associate editor coordinating the review of this paper and approving it for publication was T. Melodia.

A. Altieri was with the Laboratoire des Signaux et Systèmes (L2S, UMR8506), CentraleSupélec-CNRS-Université Paris-Sud, F-91190 Gif-sur-Yvette, France. He is now with the Department of Electronics, School of Engineering, Universidad de Buenos Aires (UBA), Buenos Aires, Argentina, C1063ACV (e-mail: aaltieri@fi.uba.ar).

P. Piantanida is with the Laboratoire des Signaux et Systèmes (L2S, UMR8506), CentraleSupélec-CNRS-Université Paris-Sud, F-91190 Gif-sur-Yvette, France (e-mail: pablo.piantanida@centralesupelec.fr).

L. R. Vega and C. G. Galarza are with the Department of Electronics, School of Engineering, Universidad de Buenos Aires (UBA), Buenos Aires, Argentina, C1063ACV, and also with the National Council for Scientific and Technological Research at CSC - CONICET, Buenos Aires, Argentina, C1425FQD (e-mail: lrey@fi.uba.ar, cgalar@fi.uba.ar).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2015.2430341

in [12]. They also consider *out-of-band* D2D but focus on the optimal asymptotic scaling laws of the networks. For example, in [3], [9] a one-hop network in which users are clustered and cache videos is studied with this model, and a throughput-outage trade-off is characterized for various regimes, in terms of scaling laws as the number of nodes and the library size grow to infinity. In [8] they find an optimal collaboration distance to balance interference, and analyze the scaling behavior of the benefits of D2D. On the other hand, our approach considers an infinite-area constant-density model in which transmissions are impaired by path loss and fading.

Other works on D2D through stochastic geometry models are [13]–[16]. In general, these works are not focused on video distribution, which we attempt to analyze, but on general traffic and general aspects of D2D communications. Hence, they do not consider the problems of user matching, user request statistics and caching policies, which become central in the video distribution problem. In [13], the authors study the optimal downlink spectrum partition between D2D and BS transmissions. In [14], the authors analyze a D2D *in-band overlaid* cellular network model and find expressions for several performance metrics. Finally, [16] considers the problem of video distribution through distributed storage BSs.

B. Main Contributions

The main goal of this paper is to study the number of requests that could be served by D2D instead of asking the BS for a transmission. To this end, we propose a stochastic geometry framework, with the following characteristics:

- Requesting users (destinations) and cooperative users (with cached videos) are distributed in space as a Poisson point process (PP). Transmissions are affected by path loss, slow fading and interference.
- Users are grouped in disjoint clusters where D2D exchanges take place. A family of *admissible protocols* is introduced: each protocol is composed of a *clustering strategy* induced by any hard-core PP [6], and any suitable *in-cluster communication strategy*, which dictates the communication schemes of users inside the clusters.

In this setting, we define two metrics of interest, which characterize the performance of D2D in terms of served requests:

- A *global* metric, that measures the ratio between the spatial density of served requests and the total density of requests. This gives the global fraction of the video requests which could be served through D2D exchanges without using the downlink of the cellular network.
- A *local* metric, that measures the ratio between the average number of served requests and total requests in a cluster. This can be interpreted as an indication of what gains could be expected in a localized region in space, in which certain level of service is required.

Although these metrics address the three aspects of the problem mentioned earlier, introducing a link-quality constraint is reasonable to model the delay constraints which may be required in video distribution. For this reason, we introduce three trade-off regions pertaining these metrics:

- The *global metric-average rate* trade-off region, which pertains the fraction of requests than can be served considering an average rate requirement over the cluster.
- The *global metric-average rate and cluster density* trade-off region, which refers to the local fraction of requests that can be served considering that an average rate and a minimal cluster density are required.
- The *global-local* trade-off regions, which pertain the balance between the global and the local density of served requests which are attainable simultaneously.

Determining these regions implies characterizing the optimal communication scheme among the family of admissible protocols mentioned before. Since optimal communication schemes remain unknown for each trade-off region, we analyze a simple in-cluster communication strategy, which can be paired with any clustering strategy to obtain a protocol. This will give an inner bound to the trade-off regions. In this strategy, users which request videos and those with cached videos are paired and a one-hop transmission takes place. Interference within clusters is avoided by precluding simultaneous transmissions through a time-division multiple access (TDMA) scheme which shares the time resource between transmitters. It is shown that the TDMA scheme implies that a user will experience a time-varying interference during a slot. An analysis which takes this into account is performed in order to determine the rates that a user can achieve. This analysis, which is not usually considered in a stochastic geometry setup, may be of interest for scenarios other than D2D. We then evaluate the global and local metrics for all the protocols obtained by pairing this in-cluster communication strategy with any clustering strategy. In this way, we obtain a set of inner bounds to the optimal trade-off regions, which give an indication of the possible gains through D2D. Finally, we numerically evaluate these inner bounds in different scenarios, considering the clustering strategy induced by a type II Matérn hard-core PP and the translated grid PP [6], under a Rayleigh fading model and a lognormal shadow fading model in which line of sight (LOS) may be present inside the clusters. In the Matérn hard-core PP with Rayleigh fading, we also develop approximations to the Laplace transform (LT) of the interference anywhere in a cluster. Through this analysis, we draw conclusions regarding the performance of D2D communications in cellular networks.

The paper is organized as follows. In Section II, we introduce the network model, the family of admissible protocols and the main metrics and trade-offs under study. In Section III, we introduce and analyze a simple communication strategy. In Section IV, we present some plots and comments. In Section V we discuss our findings, and proofs are in the Appendix.

Notation: $F_X(\cdot)$ is the cumulative distribution function (cdf) of the random variable (RV) X and $\bar{F}_X(\cdot)$ is its complementary cdf. $\mathcal{B}(x, y)$ is the ball of radius y centered at x. $\mathbb{1}_A$ is the indicator for an event A. All logarithms are to base two unless specified, $\mathcal{C}(x) \triangleq \log(1+x)$, $(x)^+ \triangleq \max\{x, 0\}$.

II. System Model and Admissible Protocols

A. Clustering Strategies and Spatial Model

We consider an infinite planar network in which:

- Users who request videos are distributed according to an homogeneous Poisson PP Φ_r, of intensity λ_r. Users who cache videos are distributed according to an homogeneous Poisson PP Φ_u, of intensity λ_u, independent of Φ_r.[1] Users attempt to exchange videos through D2D to reduce the load on the downlink of the cellular network, and do so outside the downlink band so there is no interference between cellular and D2D communications.
- Each user in Φ_r requests a video, which is selected independently according to a discrete distribution $p_V(v)$, where $1 \leq v \leq L$, and L is the library size. In numerical results, we assume that the videos are sorted according to their popularity, which implies that $p_V(v)$ is the probability of requesting the v-th most popular video. This distribution is commonly [3], [5], [8], [9] assumed to be a Zipf distribution of parameter $0 < \gamma < 1, p_V(v) = \frac{v^{-\gamma}}{\sum_{i=1}^{L} i^{-\gamma}}$.
- Each user of Φ_u has M (fixed) videos cached, which, for simplicity, are selected independently according to a discrete distribution $p_A(a)$, $1 \leq a \leq L$. For numerical results we assume that $p_A \equiv p_V$; this can be motivated assuming that users cache videos they watch.

To exchange videos, users of Φ_r and Φ_u are grouped into disjoint clusters which, for simplicity, are approximated as discs of radius R_c. The users of Φ_r which are not clustered will ask for the videos directly to the BS, while the users who are clustered can search in their cluster for a user from Φ_u who has the video, and request a transmission through D2D. Assuming that the clusters are disjoint implies there is a minimum distance of at least $2R_c$ between their centers, so we can model the spatial distribution of cluster centers as a hard-core PP [6], which guarantees this clearance. This leads to the following definition.

Definition 2.1 (Process of Clustered Users): Given a cluster radius $R_c > 0$, a PP of clustered users Φ_c is constructed from Φ_u and Φ_r as follows:

$$\Phi_c = \bigcup_{x \in \Phi_p} \mathcal{B}(x, R_c) \cap (\Phi_r \cup \Phi_u), \tag{1}$$

where $\Phi_p = \{x_i\}$ is a stationary *parent* hard-core PP of intensity $\lambda_p > 0$ and clearance $\delta \geq 2R_c$.

Any hard-core stationary PP with clearance $\delta \geq 2R_c$ will generate a cluster PP. In Fig. 1 we can see a representation of the network.

Remark 2.1: The same model is obtained if we first deploy the discs of radius R_c with centers in Φ_p, then create independent Poisson PPs of the same intensities as Φ_u and Φ_r inside and outside the discs.

We consider an attenuation model with slow fading and path loss, with possibly a different attenuation model for transmissions inside a cluster and one between clusters. This is for scenarios in which transmitters inside a cluster are collocated or that LOS is present. For this we consider:

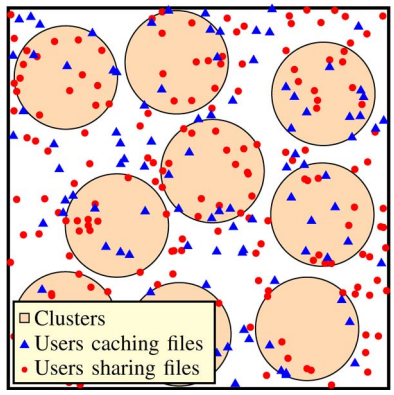


Fig. 1. Representation of the network with clusters. The users in the clusters, form the cluster PP Φ_c according to (1), in which D2D takes place.

- *Inter-cluster attenuation model:* a transmission of power P from x to y in different clusters is received with power:

$$P|h_{x,y}|^2 l(x, y), \tag{2}$$

with $|h_{xy}|^2$ a fading coefficient, independent of everything, and $l(x, y) \equiv l(||x - y||)$ is a path loss function.
- *Intra-cluster attenuation model:* a transmission of power P from x to y in the same cluster is received with power $P|g_{xy}|^2$, where $|g_{xy}|^2$ is a random power attenuation coefficient which contains fading and path loss, and whose distribution depends on $||x - y||$.

We focus on the interference generated between the nodes using D2D, and on the signal-to-interference ratio (SIR). Independent background interference or noise could be added directly. Each cluster will have a family of associated users which cache or request videos, which will be the points of the original PPs Φ_u and Φ_r which fall in the discs. Each of these users will be represented by some information, mainly, their positions, the video(s) they cache or request, and fading coefficients towards other users. This information is represented as a vector of RVs, associated to each cluster center, and, according to the assumptions stated earlier, independent among clusters. So we can represent the network of clusters as a stationary independently marked PP [6], [7] $\tilde{\Phi} = \{(x_i, \mathbf{m}_{x_i})\}$, where $\Phi_p = \{x_i\}$ is the PP of cluster centers from Def. 2.1, and $\mathbf{m}_{x_i}$ is a mark vector, containing all the RVs characterizing the users of the cluster at x_i, which are:

- $N_{x,u}$: the number of users which cache videos in the cluster centered at x. They are Poisson RVs of mean $\lambda_u \pi R_c^2$. $\mathbf{S}_x = (S_{x,1}, \ldots, S_{x,N_{x,u}})$ is the vector positions of these users relative to x, which, conditioned on $N_{x,u}$, are i.i.d. uniform RVs on the cluster (Remark 2.1).
- $N_{x,r}$: the number of users requesting videos within the cluster centered at x. They are Poisson RVs of mean $\lambda_r \pi R_c^2$. $\mathbf{D}_x = (D_{x,1}, \ldots, D_{x,N_{x,r}})$ is the vector of positions of these users relative to x, which, conditioned on $N_{x,r}$, are i.i.d. uniform RVs on the cluster (Rem. 2.1).
- $\tilde{\mathbf{A}}_x = \{\mathbf{A}_{x,1}, \ldots, \mathbf{A}_{x,N_{x,u}}\}$ are the videos which are stored in the users, such that $S_{x,i}$ stores $\mathbf{A}_{x,i} = (A_{x,i,1}, \ldots, A_{x,i,M})$.

[1] We can consider these two PPs as originating from a single Poisson PP of intensity $\lambda_u + \lambda_r$, where then a user decides to become a requesting user with probability $\lambda_r/(\lambda_u + \lambda_r)$, independent of everything else, and the rest are caching users. This separation is done to simplify the model.

They are selected independently according to p_A as indicated before. $\mathbf{V}_x = \{V_{x,1}, \ldots, V_{x,N_{x,r}}\}$ are the requested videos such that $D_{x,i}$ requests $V_{x,i}$. They are selected independently according to p_V, as mentioned earlier.

- $\mathbf{H}_x$ is a family of independent power fading coefficients between the users inside the cluster and towards users in other clusters.

When clear from context the subscript x in $\mathbf{m}_x$ is omitted. The dependence between the variables in the mark vector is characterized by their joint distribution $F_{\mathbf{m}_x}$, which, from previous assumptions, factors as:

$$\begin{aligned} F_{\mathbf{m}_x} &= F_{\mathbf{H}_x|\mathbf{S}_x,\mathbf{D}_x,N_{x,u},N_{x,r}} F_{\mathbf{S}_x|N_{x,u}} F_{\tilde{\mathbf{A}}_x|N_{x,u}} F_{\mathbf{D}_x|N_{x,u}} \\ &\quad \times F_{\mathbf{V}_x|N_{x,r}} F_{N_{x,u}} F_{N_{x,r}}, \qquad (3) \\ &= N_{x,u} F_{N_{x,r}} \left(\prod_{i=1}^{n_r} F_{V_{x,i}|N_{x,r}=n_r} F_{D_{x,i}|N_{x,r}=n_r} \right) \\ &\quad \times \left[\prod_{i=1}^{n_u} F_{S_{x,i}|N_{x,u}=n_u} \left(\prod_{j=1}^{M} F_{A_{x,i,j}|N_{x,u}=n_u} \right) \right] \\ &\quad \times F_{\mathbf{H}|\mathbf{S}_x,\mathbf{D}_x,N_{x,u},N_{x,r}}. \qquad (4) \end{aligned}$$

For shortness, unless mandatory like in the last step, we have not included the point where the distributions are evaluated. For example, $F_{\mathbf{D}_x|N_{r,x}} \equiv F_{\mathbf{D}_x|N_{r,x}=n_{r,x}}(\mathbf{d}_x)$.

There are several hard-core PPs which can be used as a parent PP Φ_p. Here we consider two possible examples. One is the type II Matérn hard-core model [6], which is obtained through a position-dependent thinning of a Poisson PP Φ of intensity λ; a uniform RV is drawn for each point of Φ and, for each pair of points which are separated by less than δ, only the one with the smallest RV is kept. This leads to a cluster density $\frac{(1-e^{-\lambda\pi\delta^2})}{\pi\delta^2}$. We also consider the translated-grid PP [6], which gives a regular square grid. It is obtained by using two independent uniform RVs in $[0, \delta)$, U_1 and U_2, and by considering the grid formed by the pairs $(m\delta + U_1, n\delta + U_2)$, for all integers m, n. This gives a density of clusters of δ^{-2}.

B. In-Cluster Communication and Admissible Protocols

Given a realization of $\tilde{\Phi}$, in each cluster, a memoryless network is defined, where users who cache videos are sources, which have a subset of all possible messages, and requesting users are receivers, requesting one of the possible messages each. To conserve the network symmetry, keep a simple structure and not require long-range coordination, we assume:

- Transmitters may have different degrees of CSI pertaining only to their own cluster, that is, some knowledge about the cluster, which is contained in its mark vector, $\mathbf{m}_x$.
- Clusters are uncoordinated, interference between them is treated as noise and there is no interaction between them to reduce interference.
- Transmissions in the network take place at a rate R in a block, in which all the clusters attempt to serve some or all of the requests inside at the same time, and an average-power constraint of P is imposed on each user.

With the above assumptions, we can focus on a single cluster to describe the behavior of any cluster in the network.

Definition 2.2 (In-Cluster Communication Strategy): An in-cluster communication strategy is given by any coding scheme that guarantees an achievable rate region $\mathcal{R}(\mathbf{m}_x, \tilde{\Phi}) \subset \mathbb{R}_+^{n_r}$ for the involved cluster with n_u transmit nodes to the n_r receiver nodes, where a symmetric rate R is attempted to all users.

Definition 2.3 (Served Request): A request from the i-th user at the cluster centered at x is said to be served whenever:

- The video is available in the cluster, that is, there is a match for this user, an event which writes as: $\mathcal{M}_{x,i} = \bigcup_{j=1}^{N_{x,u}} \{V_{x,i} \in \mathbf{A}_{x,j}\}$.
- The transmission is scheduled during the block, that is, the user with the match is scheduled to receive a transmission from one or more users with the video.
- The i-th transmission rate R belongs to the rate-region $\mathcal{R}(\mathbf{m}_x, \tilde{\Phi})$ induced by the strategy.

The probability of having a match is the same for all the users in the network, i.e., $\mathbb{P}(\mathcal{M}_{x,i}) \triangleq p_{\mathcal{M}}$. However, these events are correlated between users because they use the same cache.

Definition 2.4 (Admissible Protocol): An admissible protocol is any pair $(\Phi_p, \mathcal{F})$, where Φ_p is an admissible parent PPs, which defines a clustered network $\tilde{\Phi}$ and $\mathcal{F}$ is an in-cluster communication strategy as in Def. 2.2.

C. Performance Metrics and Trade-Offs

For every admissible protocol $(\Phi_p, \mathcal{F})$ and given a compact set $K \subset \mathbb{R}^2$ we define $N_s(K, \Phi_p, \mathcal{F}, R)$ as the number of served requests in K during a transmission block:

$$N_s(K, \Phi_p, \mathcal{F}, R) \triangleq \sum_{x \in \Phi_p} \sum_{i=1}^{N_{x,r}} \mathbb{1}_{\mathcal{K}_{x,i}} \mathbb{1}_{\mathcal{S}_{x,i}}, \qquad (5)$$

where $\mathcal{S}_{x,i} = \{\text{Req. of user } i \text{ in } (x, \mathbf{m}_x) \text{ is served}\}$ and $\mathcal{K}_{x,i} = \{x + D_{x,i} \in K\}$. Also, we define $N_{sc}(\Phi_p, \mathcal{F}, R, x)$ as the number of served requests in a cluster centered at x:

$$N_{sc}(\Phi_p, \mathcal{F}, R, x) \triangleq \sum_{i=1}^{N_{x,r}} \mathbb{1}_{\mathcal{S}_{x,i}}. \qquad (6)$$

Lemma 2.1: Given a compact set $K \subset \mathbb{R}^2$, the average number of served requests in K is:

$$\mathbb{E}\left[N_s(K, \Phi_p, \mathcal{F}, R)\right] = \lambda_p |K| \mathbb{E}^0\left[N_{sc}(\Phi_p, \mathcal{F}, R, 0)\right], \qquad (7)$$

where $|K|$ is the area of K. $\mathbb{E}^0$ is the expectation with respect to the Palm distribution of the PP with a cluster at the origin, a conditional distribution of the realizations of the PP with a cluster at the origin. The term $\mathbb{E}^0[N_{sc}(\Phi_p, \mathcal{F}, R, 0)]$ is the average number of users served in any cluster in the network.

Proof: Please see Appendix A. □

We next define the main metrics under study.

Definition 2.5 (Local Metric): The ratio of mean served requests per cluster is:

$$T_L(\mathcal{F}, R) = \frac{\mathbb{E}^0\left[N_{sc}(\Phi_p, \mathcal{F}, R, 0)\right]}{\mathbb{E}^0[N_{0,r}]}, \qquad (8)$$

where $\mathbb{E}^0[N_{0,r}] = \lambda_r \pi R_c^2$ is the average number of requests within any cluster of the network. This ratio indicates how many requests are served on average in any cluster of the network, relative to the average number of requests per cluster.

Linking this metric with (7) we define a global metric:

Definition 2.6 (Global Metric): Chosen a compact set K, the ratio of mean served requests is:

$$T_G(\mathcal{F}, R) = \frac{\mathbb{E}\left[N_s(K, \Phi_p, \mathcal{F}, R)\right]}{\mathbb{E}[N_r(K)]} = \lambda_p |\mathcal{B}(0, R_c)| T_L(\mathcal{F}, R), \quad (9)$$

where $\mathbb{E}[N_r(K)] = \lambda_r |K|$ is the average number of requests in the set K. Due to the stationarity of the PP, this does not depend on K and it can be interpreted as the spatial density of served requests, normalized by the density of requests λ_r. These metrics are determined by many factors, such as the caching policy and video request statistics, the attempted rate R, the cluster radius, and the in-cluster communication strategy. The local metric always benefits from a reduction in the cluster density because as $\lambda_p \to 0$ the interference decreases on average. If the cluster radius R_c is fixed and λ_p is diminished, the local metric will benefit, but if the density becomes too small, the global metric will eventually have to decrease. This means there is a trade-off between the metrics.

Definition 2.7 (Average Rate): Given an admissible protocol $(\Phi_p, \mathcal{F})$, a cluster at the origin has an average rate:

$$\bar{R}(\mathcal{F}, R) = R\, \mathbb{E}^0 \left[\sum_{i=1}^{N_{0,r}} \frac{\mathbb{1}_{\mathcal{S}_{0,i}}}{\sum_{j=1}^{N_{0,r}} \mathbb{1}_{\mathcal{M}_{0,j} \cap \mathcal{P}_{0,j}}} \right], \quad (10)$$

where $\mathcal{P}_{0,j}$ is the event that a transmission to user j is scheduled in the transmission block.

Considering the metrics in Defs. 2.5 and 2.6 with an average-rate constraint, which models requirements in terms of delay and link-reliability, we define the following trade-off regions:

Definition 2.8 (Trade-Off Regions):

- *Global-metric trade-off region:* a pair (r, t) of average rate and global metric is said to be achievable if there exists an admissible protocol $(\breve{\Phi}_p, \breve{\mathcal{F}})$ with rate R and density λ_p satisfying:

$$\begin{cases} T_G(\breve{\mathcal{F}}, R) \geq t, \\ \bar{R}(\breve{\mathcal{F}}, R) \geq r. \end{cases} \quad (11)$$

  The set of all achievable pairs (r, t) is the global-metric trade-off region.
- *Local-metric trade-off region:* a tuple (r, t, λ_l) of average rate, local metric and parent density is said to be achievable if there exists an admissible protocol $(\breve{\Phi}_p, \breve{\mathcal{F}})$ with rate R and density λ_p satisfying:

$$\begin{cases} T_L(\breve{\Phi}_p, \breve{\mathcal{F}}, R) \geq t, \\ \bar{R}(\breve{\Phi}_p, \breve{\mathcal{F}}, R) \geq r, \\ \lambda_p(R_c, \delta) \geq \lambda_l. \end{cases} \quad (12)$$

  The set of all achievable tuples (r, t, λ_l) is the local-metric trade-off region.
- *Local-global trade-off region:* given an attempted rate R, a pair (t_g, t_c) of global and local metrics is said to be achievable if there exists an admissible protocol $(\breve{\Phi}_p, \breve{\mathcal{F}})$ with rate R and density λ_p satisfying:

$$\begin{cases} T_G(\breve{\Phi}_p, \breve{\mathcal{F}}, R) \geq t_g, \\ T_L(\breve{\Phi}_p, \breve{\mathcal{F}}, R) \geq t_c. \end{cases} \quad (13)$$

  The set of all achievable pairs (t_g, t_c) is the local-global trade-off region.

The first region refers to the maximum fraction of users which receive videos successfully globally, subject to an average rate constraint. This metric is limited by the fraction of the users of the network that are clustered, because unclustered users cannot exchange videos. A PP $\breve{\Phi}_p$ should be chosen such that the network is almost fully covered by clusters, and a strategy $\breve{\mathcal{F}}$ such that all the requests in a cluster can be served ($T_L \approx 1$), while fulfilling the rate constraint. The second region refers to the fraction of the users inside a cluster that receive videos successfully. In this case, both a rate constraint and a certain density of clusters are required. Otherwise, we could set $\lambda_p \approx 0$ and a achieve a large level of service at the typical cluster, but there would be almost no other cluster in the network. Since this region is defined by what happens in a cluster, we could have $T_L \approx 1$, for any clustering PP. The third region refers to maximizing one the metrics, with a constraint on the other one, balancing the global and local benefits of D2D.

We cannot find the optimal protocol in terms of each trade-off region. However, analyzing a specific protocol will yield inner regions and thus yield insights on the performance attainable through D2D.

The following Lemma provides a straightforward bound for the local and global metrics (proof found in Appendix B).

Lemma 2.2: Given a protocol $(\Phi_p, \mathcal{F})$, we have $T_L(\mathcal{F}, R) \leq p_{\mathcal{M}}$, with equality as $R \to 0$. This implies for the global metric that $T_G(\mathcal{F}, R) \leq \lambda_p p_{\mathcal{M}} |\mathcal{B}(0, R_c)|$. In addition, under the caching scheme described, the probability of a match is: $p_{\mathcal{M}} = 1 - \mathbb{E}_V \left[e^{-\lambda_u \pi R_c^2 [1 - (1 - p_A(V))^M]} \right]$.

This Lemma states that if no constraint is imposed on the transmission quality, the local metric T_L is limited by the probability of finding a match in a cluster $p_{\mathcal{M}}$, and the global metric T_G is limited by this probability and the fraction of clustered users, which is $\lambda_p |\mathcal{B}(0, R_c)|$.

III. Analysis of An In-Cluster Communication Strategy

A. Strategy Definition

In this section we define and analyze an in-cluster communication strategy which, paired with any parent PP, forms an admissible D2D protocol via Def. 2.4. We focus on the cluster at the origin $(0, \mathbf{m}_0)$, omitting the subscript 0 in all its RVs ($N_{0,u} \equiv N_u$, etc.).

To consider a simple strategy we assume that:

- To reduce the interference inside the clusters, a TDMA scheme is employed to divide the transmission block into equal sized slots, in which only one request is served, through a point-to-point transmission of power P.
- At most $n_{m,\max}$ slots are defined in each cluster, regardless of the number of matches. This does not have any

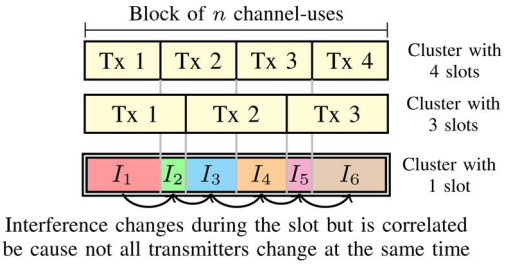


Fig. 2. Dividing the transmission block in a number of slots equal to the number of matches in a cluster, makes the interference time-correlated and non-stationary, because transmissions among clusters are unsynchronized.

practical implications, since we can choose this such that $F_{N_m}(n_{m,\max}) \approx 1$, i.e., fraction of clusters with dropped requests is negligible.[2]

To fully occupy the block with this TDMA scheme, each cluster would split the block into as many matches as it has, and clusters with the same number of matches would have the same number of slots. This assignment is reasonable in terms of theoretical performance, but leads to an interference model which is not tractable, mainly because the transmissions between clusters with a different number of matches is unsynchronized. Fig. 2, which focuses on a cluster with one slot, is provided to help understand this. When a slot is over in another cluster, its transmitter is replaced by a new one, which changes a part of the interference. Meanwhile, the other clusters will generate the same interference as before. This results in a time-correlated interference, whose statistics are very involved to model, specially considering the large range of number of slots in a very large network. Also, the distribution of the interference is not the same in all slots of the cluster. This is inherent to any wireless system using a similar TDMA scheme. To overcome this, some degree of regularity is required. Splitting the block in $n_{m,\max}$ slots would solve the problem but lead to an inefficient use of resources, since the block would be mostly unoccupied in all clusters. For this, we propose a strategy in which the block always has a power of two number of slots. This allows a different number of slots between clusters, improves the use of resources, and considers interference changes during a slot.

Definition 3.1 (Strategy): In a cluster with N_m matches:

- The block is split into $W(N_m, \varepsilon)$ slots:

$$W(N_m, \varepsilon) = \begin{cases} W_L(N_m) & \text{if } \frac{N_m - W_L(N_m)}{W_H(N_m) - W_L(N_m)} < \varepsilon, \\ W_H(N_m) & \text{otherwise,} \end{cases} \quad (14)$$

where $0 < \varepsilon \leq 1$ is a design parameter, and:

$$W_H(N_m) = 2^{\lceil \log(N_m) \rceil}, \qquad W_L(N_m) = 2^{\lfloor \log(N_m) \rfloor},$$

the powers of two closest to N_m from above and below, respectively. The RVs $W(N_m, \varepsilon) \equiv W_x(N_{m,x}, \varepsilon)$, defined for each cluster, are independent like the $\{N_{m,x}\}$.

[2]This requirement is considered for mathematical reasons. In any cluster, the number of slots/matches is always finite, but, since there is an infinite number of clusters, the maximum number of slots over all clusters is unbounded, which is in conflict with the finite (yet long) length of the block.

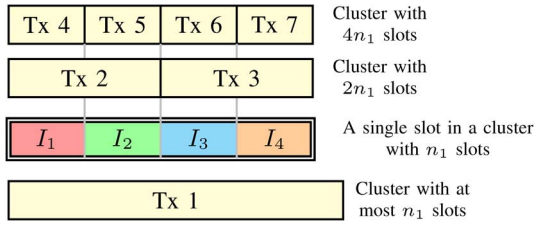


Fig. 3. We focus on a single slot in a cluster with n_1 slots (double line). Clusters with more than n_1 slots cause time variations in the interference, because different transmitters are active in each slot. Clusters with at most n_1 slots cause a constant interference, because only one transmitter is active.

- For each request with a match, a caching user is selected at random from the set of users who have the video: $\mathcal{A}(V_i, \tilde{\mathbf{A}}) = \{j : V_i \in \mathbf{A}_j, 1 \leq j \leq N_u\}$, as a candidate to serve the request.
- If there are $W = W_H(N_m)$ slots, transmissions are scheduled by selecting N_m out of the $W_H(N_m)$, and generating a random permutation of the transmissions in these slots. Otherwise, $N_m - W_L(N_m)$ requests are dropped at random, and the rest are assigned by generating a random permutation of the slots.

The parameter ε balances the fraction of time in which the channel is occupied per cluster, with the fraction of the requests that are served. When $\varepsilon = 0$, all the requests are served, maybe leaving some slots unused. When $\varepsilon = 1$ some requests are dropped, but all the slots are occupied. Also, by dropping requests, a higher success probability per user may be achieved because each transmission gets more channel uses.[3] Since only one transmitter is chosen at random from $\mathcal{A}(V_i, \tilde{\mathbf{A}})$ to serve a request, for each user $1 \leq i \leq N_r$ in a cluster, we define an RV C_i indicating which user transmits:

$$C_i|_{V_i, \tilde{\mathbf{A}}} \sim U\left(\mathcal{A}(V_i, \tilde{\mathbf{A}})\right) \quad \text{if } \mathcal{A}(V_i, \tilde{\mathbf{A}}) \neq \emptyset, \quad (15)$$

where $U(\cdot)$ denotes a uniform RV over the set, and $C_i = 0$ if $\mathcal{A}(V_i, \tilde{\mathbf{A}}) = \emptyset$.

B. Interference Characterization and Achievable Rates

We now analyze the interference generated by our scheme, to define the achievable rates of the users. We assume that is $\tilde{\Phi}$ fixed, and also, that transmissions have been scheduled in each cluster, according to Def. 3.1. We consider a fixed point in the cluster at the origin, during one of the slots, which we call the *slot under study*. Notice that the interference seen by a user depends on the total number of slots in its own cluster.

Assume that the cluster at the origin has n_1 slots, n_1 being a power of two, as indicated in Def. 3.1. Fig. 3 shows how the interference behaves during the slot under study:

- Clusters with at most n_1 slots generate a constant interference power because only one transmitter is active.
- Clusters with more than n_1 slots, say $2^k n_1$, will generate 2^k interference powers during the slot.

[3]We haven't considered the possibility that $N_m > n_{m,\max}$. We could do this, but it would further complicate the exposition without adding any substantial modifications, because both the probability of this event in the typical cluster, and the fraction of the clusters in this condition are negligible.

If the maximum number of slots in a cluster is $\Delta \triangleq W_H(n_{m,\max})$, there will be Δ/n_1 different interference powers during the slot under study. If $\{\tilde{I}_1(y), \ldots, \tilde{I}_{\Delta/n_1}(y)\}$ are these powers, and we consider a long block-length, a transmission from x to y could achieve a rate R:

$$R < \frac{1}{\Delta} \sum_{i=1}^{\Delta/n_1} \mathcal{C}(\mathrm{SIR}_i), \tag{16}$$

where $\mathrm{SIR}_i = |g_{xy}|^2/\tilde{I}_i(y)$. Notice that SIR changes are caused by the interference, while the source-destination channel is the same. The complexity of (16) precludes finding the probability of a failed transmission, i.e. the probability that (16) does not happen, so we develop a lower bound on the achievable rate. To do this, we consider that all the clusters with at most n_1 slots will generate a constant interference, which we can add to define $I_b(n_1, y)$:

$$I_b(n_1, y) = \sum_{j=0}^{\log n_1} I_j(y), \tag{17}$$

where $I_j(y)$ is the aggregate interference at y from the clusters with 2^j slots during the slot under study. Clusters with more than n_1 slots, say $2^k n_1$, generate 2^k interference values; we index these 2^k values using a binary expansion, as $I_k(u_1, \ldots, u_k, y)$, with $(u_1, \ldots, u_k) \in \{1, 2\}^k$, for any $1 \leq k \leq \log(\Delta/n_1)$. This indexing is only used because it is convenient for the proof of Lemma 3.1. We do not need to specify and order in which these interferences appear during the slot under study; we are only interested in the time average of these values:

$$\bar{I}_k(n_1, y) = \frac{1}{2^k} \sum_{(u_1,\ldots,u_k)\in\{0,1\}^k} I_k(u_1, \ldots, u_k, y). \tag{18}$$

That is, $\bar{I}_k(n_1, y)$ is the time-average of the interference coming from clusters with $2^k n_1$ slots, seen at y during the slot under study. With these definitions we introduce the following:

Lemma 3.1 (Achievable Rate): The rate (16) of any slot in a cluster with a total of n_1 slots can be lower bounded by:

$$R_a(n_1, x, y) = \frac{1}{n_1} \mathcal{C}\left(\frac{P|g_{xy}|^2}{I_b(n_1, y) + \sum_{i=1}^{\log(\Delta/n1)} \bar{I}_k(n_1, y)} \right), \tag{19}$$

where I_b comes from (17) and $\bar{I}_k(n_1, y)$ is given by (18).

Proof: Details can be found in Appendix C. □

This means that any rate $R < R_a$ is achievable, where R_a is a time-sharing, point-to-point rate with a constant interference, made up of the original interferences added together, with components that change over the slot being time-averaged. We now write the expression of the interference in (19) as seen from the slot under study. If we focus on an interfering cluster with $2^k n_1$ slots, we see that the same transmitter can be active in more than one slot, or some slots may be empty, so we cannot assume that there will be 2^k different or non-null interference powers. Since we have time-averaged the interference, we are only interested in how many times each user transmits, and not in the transmission order. By assuming that all the slots are occupied, which gives a worst case, the interference in (19) at a point d of the cluster at the origin with n_1 slots, during a slot, is:

$$I(d, n_1) = \sum_{x\in\Phi\setminus\{0\}} P \left\{ \mathbb{1}_{\{W_x \leq n_1\}} |h_{xd}|^2 l(x + S_x, d) + \sum_{i=\log(n_1)+1}^{\log\Delta} \mathbb{1}_{\{W_x = 2^i\}} \frac{n_1}{2^i} \sum_{j=1}^{2^i/n_1} B_{x,j,i} |h_{x+S_{x,j},d}|^2 l(x + S_{x,j}, d) \right\}. \tag{20}$$

The terms with $\mathbb{1}_{\{W_x \leq n_1\}}$ represent the interference from users in clusters with at most n_1 slots. The terms with $\mathbb{1}_{\{W_x = 2^i\}}$ represent the interference from a cluster with 2^i slots. The RVs $B_{x,j,i} \geq 0$ indicate how many times each transmitter is active during slot. Since we take that all the slots are occupied, we have:

$$\sum_{j=1}^{2^i/n_1} B_{x,j,i} = 2^i/n_1. \tag{21}$$

If in a cluster $B_{x,j,i} = 1$, for all j, then every slot that took place during the slot under study is used by a different transmitter; if only one satisfies $B_{x,j,i} = 2^i/n_1$ and the rest are zero, all the slots are used by the same transmitter.

Remark 3.1: Actually, there is an abuse of notation in (20). The terms $|h_{x+S_{x,j},d}|^2 l(x + S_{x,j}, d)$ do not use the same indexing over j as in the definition of $\tilde{\Phi}$. In fact, for example, for a cluster with $2^i > n_1$ slots, we cannot guarantee that there will be $2^i/n_1$ different users with videos stored, i.e. that the sum over j is well defined. We have obtained an upper bound to the interference seen at d during the slot under study. To do this, we take each cluster, see which transmitters will be active during the slot under study, and count how many times each will transmit. Then, we may add more fictitious transmitters to have $2^i/n_1$ transmitters, and assign them slots such that (21) is met. We do not need to consider which of the users actually transmits in each cluster, because we are focused on a single slot of the cluster at the origin, and because, conditioned on the cluster centers, the fading and user positions are independent.

C. Performance Metric Analysis

We now evaluate the metrics from Def. 2.8 for the strategy given by Def 3.1, which for simplicity we denote by $\mathcal{F}^*$. For a compact set K and a protocol $(\Phi_p, \mathcal{F}^*)$, the average number of served requests is given by (5). The event of a served requests $\mathcal{S}_{x,i}$ can be written as: $\mathcal{S}_{x,i} = \mathcal{M}_{x,i} \cap \mathcal{T}_{x,i}$, where $\mathcal{M}_{x,i}$ indicates a match, and $\mathcal{T}_{x,i}$ means a transmission was scheduled and succeeded ($R < R_a$, R_a from (19)).

Theorem 3.1 (Local and Global Metrics): Given a parent PP Φ_p, as in Def. 2.1, the ratio of mean served requests per cluster, or local metric T_L (8), for $\mathcal{F}^*$ is given by (22), shown at bottom of the next page. The spatial density of served requests or global metric (9) for the strategy $\mathcal{F}^*$ is $T_G(\mathcal{F}^*, R) = \lambda_p |\mathcal{B}(0, R_c)| T_L(\mathcal{F}^*, R)$. The average rate achieved is:

$$\bar{R}(\mathcal{F}^*, R) = R\, \mathbb{E}^0\left[\bar{F}_{|g_{S,D}|^2|S,D} \left(I\left(D, W(N_m)\right) \left(2^{W(N_m)R} - 1 \right) \right) \right].$$

Proof: See Appendix D. □

Achieving a high level of service within a cluster may imply that only a few clusters need be created, and hence, globally the effects of D2D may not be significant. Locally, larger clusters imply an increased likelihood of matches, but also a larger average transmission distance, and a reduced slot duration, which increases the chances of failed transmissions.

As we mentioned before, the intra-cluster attenuation model could have different expressions. One case of interest is when it is the same as the inter-cluster model (2), that is, when we have $|g_{xy}|^2 = |h_{xy}|^2 l(x, y)$. Also, it is interesting to consider the specific case in which the fading is Rayleigh, that is, $|h_{xy}|^2$ follows a unit mean exponential distribution. In this case we have:

$$\mathbb{E}^0\left[\bar{F}_{|g_{x,y}|^2}\left(\eta I(d,n)\right)\right] = \mathcal{L}^0_{I(d,n)}\left(\frac{\eta}{l(x,y)}\right), \tag{23}$$

where $\mathcal{L}^0_{I(d,n)}(\eta) = \mathbb{E}^0[e^{-\eta I(d,n)}]$ is the LT of $I(d, n)$ with respect to $\mathbb{P}^0$.

The main issue in this case is the distribution of the variables $B_{x,j,i}$. Fortunately, the following Lemma, whose proof is in Appendix E, helps us avoid this:

Lemma 3.2: The LT $\mathcal{L}^0_{I(z,n)}$ of (20) for Rayleigh fading can be lower bounded by setting $B_{x,j,i} = 1$ for all x, j, k.

Notice that these results are valid for any PP of cluster centers Φ_p. We now consider that Φ_p is a type II Matérn hard-core PP and approximate the LT (23).

D. Approximations and Bounds for Matérn Type II Processes

The main issue to evaluate the metrics for a type II PP is finding the reduced LT of the interference (23). This PP has been used mostly to model networks using carrier-sense multiple access schemes [7], [17], and even the most simple transforms are not known in closed form. For this reason, we approximate the PP by a more tractable one, following the approach in [18]. In our setting this equates to considering the cluster centers, as seen from the cluster at the origin, to be distributed as a non-homogeneous Poisson PP of intensity: $\lambda_p \mathbb{1}_{\{||x||>\delta\}}$, where λ_p is the intensity of the original hard-core PP (Section II-A). If we use Lemma 3.2 to bound the true interference (20), and approximate the Matérn PP in this way, we have the following approximate interference:

$$I(d,n) \approx \hat{I}(d,n) = P \sum_{x\in\Phi} \tilde{\psi}(x, \mathbf{m}_x, d, n) \mathbb{1}_{\{||x||>\delta\}}, \tag{24}$$

where $\tilde{\psi}(x, \mathbf{m}_x, d, n)$ is the function in the sum (20) with all the $B_{x,i,j} = 1$. The sum in (24) is over an homogeneous Poisson PP of intensity λ_p and the non-homogeneity is given by $\mathbb{1}_{\{||x||>\delta\}}$. The LT of a Poisson PP is [7]:

$$\mathcal{L}_{\hat{I}(d,n)}(\eta) = \exp\left\{-\lambda_p \int_{||x||>\delta} \mathbb{E}_{\mathbf{m}_x}\left[1 - e^{-\eta\tilde{\psi}(x,\mathbf{m}_x,d,n)}\right] dx\right\}. \tag{25}$$

Taking the expectation over W_x and over the independent unit-mean exponential fading RVs we have:

$$\mathbb{E}\left[e^{-\eta\tilde{\psi}(x,\mathbf{m}_x,d,n)}\right] = \mathbb{E}\left[\frac{\mathbb{P}(W_x \le n)}{1 + l(x + S_x, d)\eta}\right] + \sum_{i=\log n+1}^{\log\Delta} \prod_{j=1}^{2^i/n} \mathbb{E}\left[\frac{\mathbb{P}\left(W_x = 2^i\right)}{1 + l(x + S_{x,j}, d)\eta n/2^i}\right], \tag{26}$$

where the remaining expectations, which cannot be computed in closed form, are over $\{S_x\}$, $\{S_{x,j}\}$. However, it is reasonable to introduce the far field approximation that the interfering users are seen from the typical cluster as if located at the center of their cluster. This is because the favorable and unfavorable positions of the interferers will be approximately canceled, because they are uniformly distributed around the center. With this, (26) is approximated as:

$$\mathbb{E}\left[e^{-\eta\tilde{\psi}(x,\mathbf{m}_x,d,n)}\right] \approx \sum_{i=0}^{\Delta} \frac{\mathbb{P}\left(W_x = 2^i\right)}{\left(1 + \frac{l(x,d)\eta}{\lceil 2^i/n\rceil}\right)^{\lceil \frac{2^i}{n}\rceil}}. \tag{27}$$

Replacing this equation in (25), we have:

$$\mathcal{L}_{\hat{I}(d,n)}(\eta) \approx \exp\left\{-\lambda_p \sum_{i=0}^{\Delta} \mathbb{P}(W = 2^i) \times \int_{||x||>\delta} \frac{\left(1 + \frac{l(x,d)\eta}{\lceil 2^i/n\rceil}\right)^{\lceil \frac{2^i}{n}\rceil} - 1}{\left(1 + \frac{l(x,d)\eta}{\lceil 2^i/n\rceil}\right)^{\lceil \frac{2^i}{n}\rceil}} dx\right\}. \tag{28}$$

The summation in (28) has few terms and hence is straightforward to implement numerically.

IV. Relevant Plots and Comments

In this section we study the trade-off regions for strategy $\mathcal{F}^*$ (Theorem 3.1), in order to study the performance attainable through D2D. We consider two PPs for the clusters centers: the type II Matérn PP, denoted as Φ_{HC}, and the translated grid model, denoted as Φ_{TG}. The inter-cluster attenuation model is given by (2), with $l(x, y) = \tilde{C}||x - y||^{-\alpha}$, $\tilde{C}$ is a constant and $\alpha = 4$. For the intra-cluster attenuation model given by $|g_{x,y}|^2$ we consider two scenarios: one in which it is the same as the inter-cluster model, and another one with lognormal shadow fading in which there may be LOS inside a cluster.

First, we consider the intra-cluster attenuation model is the same as the inter-cluster model, that is, $|g_{x,y}|^2 = |h_{x,y}|^2 l(x, y)$. We focus on the clustering PP Φ_{HC} and the fading is Rayleigh, that is, $|h_{xy}|^2$ are unit mean exponentials. We take $P = \tilde{C} = 1$, since these constants will cancel out when computing the SIR.

$$T_L(\mathcal{F}^*, R) = \frac{\mathbb{E}^0\left[\left(N_m \mathbb{1}_{\{W(N_m)=W_H(N_m)\}} + W_L(N_m)\mathbb{1}_{\{W(N_m)=W_L(N_m)\}}\right) \bar{F}_{|g_{S,D}|^2|S,D}\left(I\left(D, W(N_m)\right)\left(2^{W(N_m)R} - 1\right)\right)\right]}{\mathbb{E}^0[N_r]}. \tag{22}$$

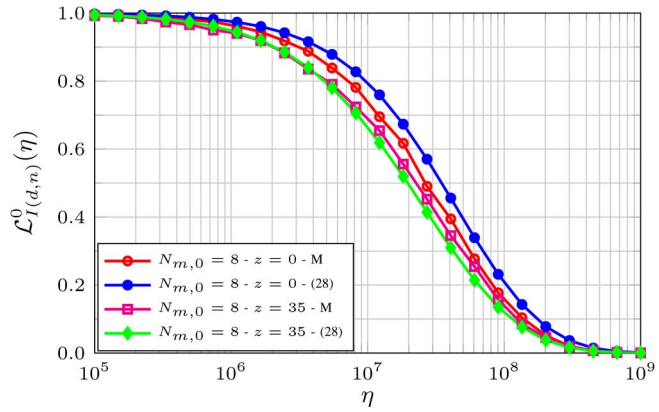


Fig. 4. Evaluation of (28) as an approximation of the LT of the interference for a type II Matérn PP. M: Monte Carlo simulation (2000 runs each point) of the LT of (20) at z by simulating the PP. $\delta/2 = R_c = 50$, $\alpha = 4$, $\lambda_u = 4\lambda_r = 0.072$, $M = 6$, $\varepsilon = 0.5$.

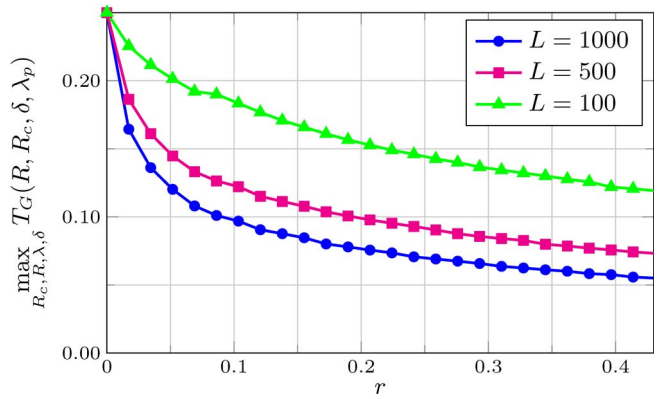


Fig. 5. Inner bound (29) given by $(\Phi_{HC}, \mathcal{F}^*)$ to the Global trade-off region (11) for a Rayleigh fading model. L is the library size. $\lambda_u = 0.012 = 4\lambda_r$, $M = 6$, $\gamma = 0.6$, $\alpha = 4$. $\varepsilon = 0.05$.

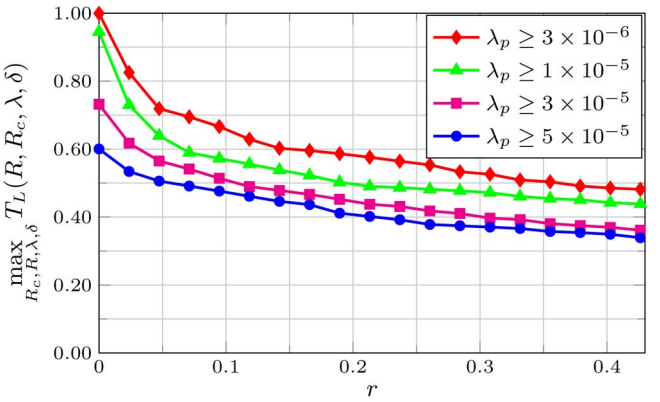


Fig. 6. Inner bound (31) given by $(\Phi_{HC}, \mathcal{F}^*)$ to the local trade-off region (12) for a Rayleigh fading model. $\lambda_u = 0.012 = 4\lambda_r$, $L = 500$, $M = 6$, $\gamma = 0.6$, $\alpha = 4$. $\varepsilon = 0.05$.

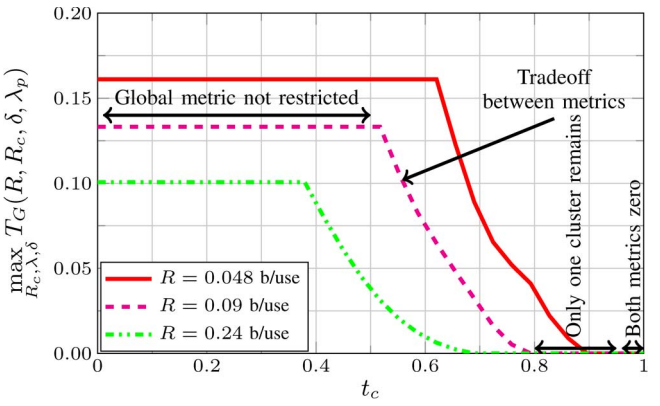


Fig. 7. Inner bound (32) given by $(\Phi_{HC}, \mathcal{F}^*)$ to the Local-Global trade-off region (13) for a Rayleigh fading model. R is the attempted rate. $\lambda_r = 0.003 = \lambda_u/4$, $L = 500$, $M = 6$, $\gamma = 0.6$, $\alpha = 4$. $\varepsilon = 0.05$.

The user densities are $\lambda_u = 4\lambda_r = 0.012$ users/area, each caching user stores $M = 6$ videos. In Fig. 4, we plot the LT approximation (28), when compared to the Monte Carlo simulation of the interference (20) using Lemma 3.2. We evaluate this at different places in the cluster and taking a number of slots $N_{m,0} = 8$, which is the most likely for the distribution of matches (the rest of the parameters are indicated in the caption). We see that the approximation of Palm distribution of the hard core PP by a non-homogeneous Poisson PP as proposed in [17] is very accurate. Although we do not show it here, considering the interfering users as located at the centers of the cluster introduces a larger error than the approximation on the PP.

In Figs. 5, 6, and 7 we study the inner bounds to the trade-off regions under the same setup as the previous figure. We use (28) to approximate the LT of the interference and Monte Carlo simulations to average over S and D and to estimate the distribution of N_m. In Fig. 5 we plot the inner bound of the global metric trade-off region (11) achievable by $(\Phi_{HC}, \mathcal{F}^*)$, changing the library size L, that is:

$$\max_{R_c,R,\lambda,\delta} T_G(R, R_c, \delta, \lambda_p) \text{ subject to } \bar{R}(R, R_c, \lambda, \delta) \geq r. \quad (29)$$

Optimizing over all parameters R_c, R, λ, δ implies this is the best global fraction of served requests achievable by $(\Phi_{HC}, \mathcal{F}^*)$. Locally, increasing L requires, on average, larger clusters to find matches, increasing the chance of failed transmissions through path loss. This can be mitigated with a bigger cluster separation $\delta \geq 2R_c$, which reduces the density of clusters, and hence the average level of interference, but may negatively impact the density of served requests, because the clusters cover a smaller fraction of the network. A pair (R_c, δ) should be chosen to balance these effects; although not plotted, simulations show that in this setup, the optimal value was always $\delta = 2R_c$. In addition, by using Lemma 2.2 and the density of the hard-core PP (Section II-A), we can find an upper bound for the global metric of this protocol:

$$T_G(R, R_c, \delta) \leq (1 - e^{-\lambda\delta^2}) p_{\mathcal{M}} \frac{R_c^2}{\delta} \leq \frac{1}{4}, \quad (30)$$

for any R and hence any average rate constraint. In the last inequality we used that $\delta \geq 2R_c$. The plot shows that when the rate is unconstrained ($r = 0$ in (29)), then with $\delta = 2R_c$, large λ, and $R \to 0$, the bound is achieved. If the global metric indicates the fraction of users not requiring a downlink transmission, the plot shows that, even through this simple strategy, D2D could serve a reasonable number of requests. The global trade-off does not guarantee a certain percentage of service within a cluster, which may be important in some scenarios: there may

be many clusters with a small percentage of served requests (small T_L) which results in large overall benefits (large T_G). In Fig. 6 we plot the inner bound of the local metric trade-off region (12) achievable by $(\Phi_{HC}, \mathcal{F}^*)$, which maximizes the mean ratio of served requests:

$$\max_{R_c,R,\lambda_p,\delta} T_L(R, R_c, \delta, \lambda_p) \text{ subject to}$$
$$\bar{R}(R, R_c, \lambda, \delta) \geq r \text{ and } \lambda_p(\delta, \lambda) \geq \lambda_t. \quad (31)$$

The fraction of per cluster served requests, which is not bounded by the coverage percentage of cluster PP, can be much larger according to setup parameters and the value of the restrictions.

In Fig. 7 we plot the local-global metrics trade-off region (13) inner bound given by $(\Phi_{HC}, \mathcal{F}^*)$:

$$\max_{R_c,\lambda,\delta} T_G(R, R_c, \delta, \lambda_p) \text{ subject to } T_L(R, R_c, \delta, \lambda_p) \geq t_c. \quad (32)$$

We see that if t_c is small enough, then the global metric can be maximized without restriction, and in this regime, the local metric can be set as needed. A smaller local metric implies there is a large number of clusters with a low fraction of served requests, while a larger local metric implies fewer, larger clusters with a higher level of service. If t_c is larger, both constraints cannot be satisfied at the same time, and there is a trade-off between the metrics, which depends on the setup. If t_c is even larger, then the maximal density of served requests becomes negligible, which implies that only the cluster at the origin remains. Finally, if the local constraint is too large it cannot be satisfied for any parameters.

After exploring the inner bounds to the trade-off regions for Rayleigh fading, we now introduce a model with shadow fading in which LOS may be present between the users in the clusters. For the intra-cluster model, we assume an indoor office model like the A1 Winner II model [19], in which the power attenuation coefficient for a transmission from x to y (in dB) is given by:

$$|g_{x,y}|^2[\text{dB}] = C_1 \log_{10}(||x-y||) + C_2 + C_3 \log_{10}(f_c[\text{GHz}]/5) + 5N_w(x,y) + \chi_{x,y}, \quad (33)$$

where f_c is the carrier frequency, $\chi_{x,y}$ is a lognormal shadow fading coefficient of zero mean, and N_w is the number of walls between x and y. The constants C_1, C_2, C_3, the value of N_w and the standard deviation of $\chi_{x,y}$ depend on whether there is LOS between the transmitter and the destination or not. The event of LOS between x and y is determined according to the LOS probabilities of the Winner II A1 model [19], which is:

$$\mathbb{P}(\text{LOS}(x,y)) = 1 - 0.9\left(1 - \left(1.24 - 0.61\log_{10}||x-y||\right)^3\right)^{\frac{1}{3}},$$

when $||x-y|| > 5$ and one, otherwise. The value of the constants under LOS and NLOS are:

$$C_1 = 18.7, \quad C_2 = 46.8, \quad C_3 = 20 \quad \text{LOS} \quad (34)$$
$$C_1 = 36.8, \quad C_2 = 43.8, \quad C_3 = 23 \quad \text{NLOS}. \quad (35)$$

The standard deviation of χ_{xy} are $\sigma_{\text{LOS}} = 3$ (for the LOS case) and $\sigma_{\text{NLOS}} = 6$. Since we do not define a deterministic wall distribution in the clusters, we introduce a simple model for

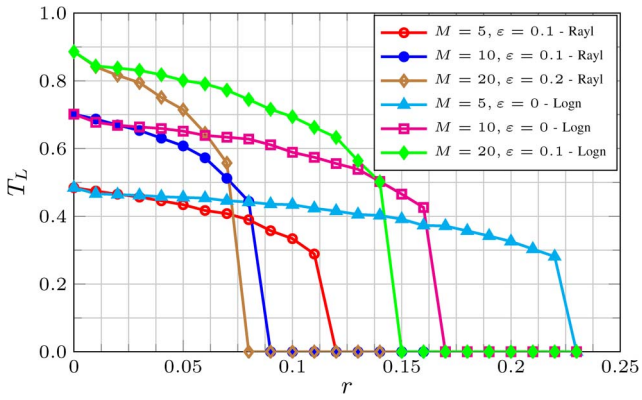


Fig. 8. Fraction of mean served requests for an average rate constraint given by $(\Phi_{HC}, \mathcal{F}^*)$ (inner bound to the local trade-off region (12)) for the lognormal shadowing and LOS model, and the Rayleigh fading model. $R_c = 20$, $\delta = 2R_c$, $\lambda = 2 \times 10^{-4}$, $\lambda_r = \lambda_u = 0.0278$, $L = 300$. $\gamma = 0.4$.

N_w, which determines the number of walls as a function of the distance between the nodes:

$$N_w(x,y) = \begin{cases} 0 & \text{LOS} \\ 1 + \lfloor (||x-y||/5 - 1)^+ \rfloor & \text{NLOS}. \end{cases} \quad (36)$$

We assume that transmissions take place in the WiFi band, with $f_c = 2.45$ GHz. Users transmit with power $P = 10^{G_t G_r P_{tx}/10}$, where $G_t = 12$ dB is the transmit antenna gain, $G_r = 0$ dB is the receive antenna gain, and $P_{tx} = 20$ dBm. For the inter-cluster attenuation model, we use a variation of the B4 Winner model [19, (3.23), part II]. We assume that no LOS is possible, and a transmission from x to y in different clusters to be attenuated (in dB) as:

$$|h_{x,y}|^2 l(x,y)[\text{dB}] = 40\log_{10}(||x-y||) + 41 + 22.7\log_{10}(f_c[\text{GHz}]/5) + 28N_b(x,y) + \chi_{x,y}, \quad (37)$$

where $\chi_{x,y}$ is a lognormal RV with zero mean and standard deviation $\sigma = 7$. $N_b = 1, 2, \ldots$ takes into account the attenuation due to the clusters which are on the line between the source and the destination. This is a simple model considering a 14 dB attenuation for each time the transmission penetrates or leaves a cluster between the source and the destination. We now evaluate the performance with this shadow fading model with possible LOS and the Rayleigh fading model without LOS (used in Figs. 5, 6, and 7) in different scenarios. We always consider $\lambda_r = \lambda_u = 0.0278$, $L = 300$ and change the number of cached videos M.

In Figs. 8 and 9 we consider both attenuation models and evaluate the local trade-off region (11). For the Rayleigh model we approximate the LT of the interference using (28), while for the shadowing model we use a Monte Carlo simulation of the PP. The value of ε is chosen to maximize the local metric T_L for an attempted rate $R = 0.05$ b/use. The number of videos cached M and the fraction of dropped requests ε leverage the spectral efficiency and the probability of finding videos in the cluster. A smaller value of M results in a smaller fraction of served requests but allows the system to achieve a higher average rate. In Fig. 8 the cluster PP is Matérn type II; since this PP is not regular, to simplify the simulations we take $N_b = 1$. We

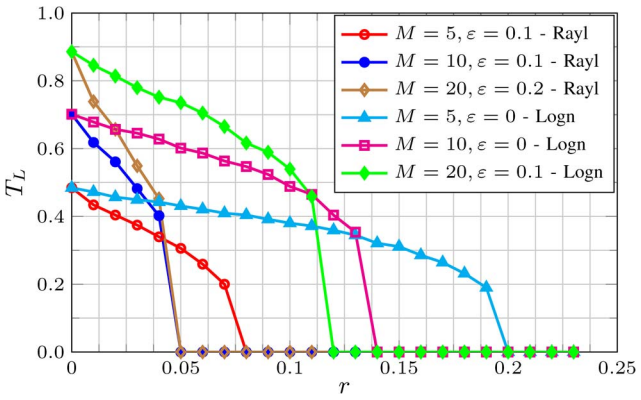


Fig. 9. Fraction of mean served requests for an average rate constraint given by $(\Phi_{TG}, \mathcal{F}^*)$ (inner bound to the local trade-off region (12)) for the lognormal shadowing and LOS model, and the Rayleigh fading model. $R_c = 20$, $\delta = 50$, $\lambda_r = \lambda_u = 0.0278$, $L = 300$, $\gamma = 0.4$.

consider a cluster radius $R_c = 20$, with minimal clearance $\delta = 2R_c$ (the rest of the parameters are in the caption). We can see both models provide similar results when the required average rate is small, while the model with LOS has a better performance when a higher average throughput is required. This is caused by considering the possibility of LOS and also the penetration loss to the cluster. Finally, in Fig. 9 the cluster PP is the translated grid with clearance $\delta = 50$ m and cluster radius $R_c = 20$ m. Each cluster could represent a block of buildings in an urban scenario. Since the process is regular, the coefficient N_b between two clusters centered at x_1 and x_2 is $||x_1 - x_2||_\infty / R_c$, where $||\cdot||_\infty$ is the standard infinite norm in $\mathbb{R}^2$. All the other parameters are the same as in Fig. 8. In this case, the lognormal model predicts more important gains than the Rayleigh fading one. This is mainly because the Rayleigh model does not consider attenuation of the interference when penetrating the cluster, while the other one does. We see that the LOS model, which is more realistic in an urban scenario, predicts that a large number of requests could be served through D2D communications.

V. Summary and Discussion

We described a framework which can be used to evaluate the theoretical performance attainable through D2D. In this framework we considered a simple communication strategy, which gives an inner bound to the performance trade-off regions introduced. We have studied the performance of this strategy under several attenuation models. The analysis shows that, even through our simple communication protocol, a substantial number of requests could be served through a distributed caching policy without a dedicated caching infrastructure, reducing the downlink traffic on the cellular band and a reduced load on the backhaul. Through the analysis in which we have considered a whole communications block, we have observed that the design of D2D architectures should balance the density of videos available through local caching with a proper use of time/frequency resources, so as to maximize the number of served requests. Increasing the number of cached videos will generally improve performance but choosing a transmitter randomly may imply that the average source-transmitter distance is not decreased. Therefore, we believe that the results presented in the simulation sections could be improved by considering a strategy which pairs users according to their distance. In this way, the attenuation incurred by randomly choosing a far away transmitter or by a poor choice of the cluster size could be mitigated without tuning any parameters. However, this would also required further information and coordination. Other medium access strategies which require less coordination, such as random time/frequency hopping, could also be studied to better understand how performance could be degraded if less coordination was required.

Appendix

A. Proof of Lemma 2.1

For shortness we write $N_s(K)$, and using the Campbell-Mecke formula for an independently marked point PP [6], [7] we have:

$$\mathbb{E}\left[N_s(K)\right] = \lambda_p \int_{\mathbb{R}^2} \mathbb{E}^x \left[\sum_{i=1}^{N_{x,r}} \mathbb{1}_{\{x+D_{x,i} \in K\}} \mathbb{1}_{\mathcal{S}_{x,i}} \right] dx,$$

where $\mathbb{E}^x$ is the Palm distribution of the PP with a cluster centered at x. With all this we can write:

$$\mathbb{E}\left[N_s(K)\right] = \lambda_p \int_{\mathbb{R}^2} \mathbb{E}^0 \left[\sum_{i=1}^{N_{0,r}} \mathbb{1}_{\{D_{0,i} \in K_{-x}\}} \mathbb{1}_{\mathcal{S}_{0,i}} \right] dx,$$

where K_{-x} is the set obtained by shift every point in K by $(-x)$. Moving the integral inside we conclude by noting that: $\int_{\mathbb{R}^2} \mathbb{1}_{\{D_{0,i} \in K_{-x}\}} dx = |K|$.

B. Proof of Lemma 2.2

Suppose a protocol $(\Phi_p, \mathcal{F})$ is chosen. We can write $\mathcal{S}_{0,i} \equiv \mathcal{S}_{0,i}(R)$ to indicate the dependence of the service event with the required rate R. For a given realization of $\tilde{\Phi}$, decreasing R cannot decrease the number of served requests, that is:

$$\sum_{i=0}^{N_{r,0}} \mathbb{1}_{\mathcal{S}_{0,i}(R_1)} \leq \sum_{i=0}^{N_{r,0}} \mathbb{1}_{\mathcal{S}_{0,i}(R_2)}, \tag{38}$$

when $R_1 > R_2$. Taking expectation $\mathbb{E}^0[\cdot]$ on both sides, we can upper bound the right by taking the limit as $R_2 \to 0$. But when $R = 0$ we have that: $\mathcal{S}_{0,i} = \mathcal{M}_{0,i}$, that is, when the required rate is 0, a request is served whenever the video is available in the cluster, and hence, applying the monotone convergence theorem [20] (with equality as $R \to 0$):

$$\mathbb{E}^0 \left[\sum_{i=0}^{N_{0,r}} \mathbb{1}_{\mathcal{S}_{0,i}(R)} \right] \leq \mathbb{E}^0 \left[\sum_{i=0}^{N_{0,r}} \mathbb{1}_{\mathcal{M}_{0,i}} \right] = \mathbb{E}^0[N_{0,r}] p_{\mathcal{M}}.$$

C. Proof of Lemma 3.1

It is straightforward to show that, given two constants $A, B > 0$, the function $\phi(u) = \mathcal{C}(A/(B+u))$ is convex for $u \geq 0$. Now,

we rewrite the rate (19) in a way suitable to use the convexity of $\phi(u)$. Using (17) and (18), (16) is written as:

$$\frac{1}{n_1}\frac{1}{2}\sum_{u_1=1}^{2}\cdots\frac{1}{2}\sum_{u_{\hat{k}}=1}^{2}\mathcal{C}\left(\frac{|g_{xy}|^2}{I_b+I_1(u_1)+\ldots+I_{u_{\hat{k}}}(u_1,\ldots u_{\hat{k}})}\right),$$

where for shortness we defined $\hat{k}=\log(\Delta/n_1)$. This expression consists of $\hat{k}$ nested convex combinations with two terms each, giving the Δ/n_1 terms seen in (16). Notice that the k-th term, $I_k(u_1,\ldots,u_k)$, depends only on the summations over $(u_1,\ldots,u_k)$ and is constant for indexes $(u_{k+1},\ldots,u_{\hat{k}})$. We can therefore recursively use the convexity of $\phi(u)$ to transfer the summations inside $\mathcal{C}(\cdot)$, innermost to outermost. Defining $B(k,u_1,\ldots,u_k)=I_b+\sum_{i=1}^{k}I_i(u_1,\ldots,u_i)$. In the first step, for any set of indexes $u_1,\ldots,u_{\hat{k}-1}$ we can write:

$$\begin{aligned}&\frac{1}{2}\sum_{u_{\hat{k}}=1}^{2}\mathcal{C}\left(\frac{|g_{xy}|^2}{I_b+I_1(u_1)+\ldots+I_{u_{\hat{k}}}(u_1,\ldots u_{\hat{k}})}\right)\\&=\frac{1}{2}\sum_{u_{\hat{k}}=1}^{2}\mathcal{C}\left(\frac{|g_{xy}|^2}{B(\hat{k}-1,u_1,\ldots,u_{\hat{k}-1})+I_{u_{\hat{k}}}(u_1,\ldots u_{\hat{k}})}\right)\\&\geq\mathcal{C}\left(\frac{|g_{xy}|^2}{B(\hat{k}-1,u_1,\ldots,u_{\hat{k}-1})+\frac{1}{2}\sum_{u_{\hat{k}}=1}^{2}I_{u_{\hat{k}}}(u_1,\ldots u_{\hat{k}})}\right).\end{aligned}$$

Continuing this procedure $\log(\Delta/n_1)$ times, we get the result.

D. Proof of Theorem 3.1

We focus on the cluster at the origin, removing all subindexes 0 (that is, $N_{0,r}\equiv N_r$, $\mathcal{T}_{0,i}\equiv\mathcal{T}_i$, etc.). For the local metric, we need to find:

$$\mathbb{E}^0\left[\sum_{i=1}^{N_r}\mathbb{1}_{\mathcal{M}_i}\mathbb{1}_{\mathcal{T}_i}\right]=\mathbb{E}^0\left[\sum_{i=1}^{N_r}\mathbb{E}^0[\mathbb{1}_{\mathcal{M}_i}\mathbb{1}_{\mathcal{T}_i}|N_r,N_m]\right].\quad(39)$$

In what follows, unless needed for clarity, we do not write the specific values taken by the RVs in the expectations and probabilities; for example, $\mathbb{P}^0(\mathcal{M}_i|N_r=n_r,N_m=n_m)\equiv\mathbb{P}^0(\mathcal{M}_i|N_r,N_m)$. We have:

$$\begin{aligned}\mathbb{E}^0[\mathbb{1}_{\mathcal{M}_i}\mathbb{1}_{\mathcal{T}_i}|N_r,N_m]&=\mathbb{P}^0(\mathcal{T}_i|N_r,N_m,\mathcal{M}_i)\\&\quad\times\mathbb{P}^0(\mathcal{M}_i|N_r,N_m).\quad(40)\end{aligned}$$

To find $\mathbb{P}^0(\mathcal{M}_i|N_r,N_m)$ we condition on $(N_u,\tilde{\mathbf{A}})$, so:

$$\mathbb{P}^0\left(V_i\in\tilde{\mathbf{A}}|N_r,N_m,N_u,\tilde{\mathbf{A}}\right)=\frac{\binom{N_r-1}{N_m-1}}{\binom{N_r}{N_m}}=\frac{N_m}{N_r}.\quad(41)$$

To prove (41) we consider that once $\tilde{\mathbf{A}}$ is fixed, the event $V_i\in\tilde{\mathbf{A}}$ can be interpreted as a success in a Bernoulli trial. Then, we are asking for the probability of a success on the i-th trial out of N_r Bernoulli trials given that there were a total of N_m successes. Using (41), we have:

$$\mathbb{P}^0(\mathcal{M}_i|N_r,N_m)=\mathbb{E}^0\left[\frac{N_m}{N_r}\middle|N_r,N_m\right]=\frac{N_m}{N_r}.$$

We now only need to find $\mathbb{P}^0(\mathcal{T}_i|N_r,N_m,\mathcal{M}_i)$ in (40). The event $\mathcal{T}_i$ can be written as $\mathcal{T}_i=\mathcal{O}_i(W(N_m))\cap\mathcal{P}_i$, where $\mathcal{P}_i$ indicates the user was scheduled for a transmission and $\mathcal{O}_i(W(N_m))=\{\text{Tx successful for user } i\}$. In what follows we do not write the dependence of W, W_L and W_H with N_m (that is, we write $W(N_m)\equiv W$, etc.). When N_m is such that $W(N_m)=W_H(N_m)$, the user is always scheduled, while when $W(N_m)=W_L(N_m)$ the user may not be scheduled. So we may write:

$$\begin{aligned}\mathbb{P}(\mathcal{T}_i|N_r,N_m,\mathcal{M}_i)&=\mathbb{P}^0\left(\mathcal{O}_i(W_H)|N_r,N_m,\mathcal{M}_i\right)\\&\quad\times\mathbb{1}_{\{W=W_H\}}+\mathbb{P}^0(\mathcal{O}_i(W_L)|\mathcal{P}_i,N_r,N_m,\mathcal{M}_i)\\&\quad\times\mathbb{P}^0(\mathcal{P}_i|N_r,N_m,\mathcal{M}_i)\mathbb{1}_{\{W=W_L\}}.\quad(42)\end{aligned}$$

By a similar argument to the one used in (41), when N_m is such that $W=W_L$ we have:

$$\mathbb{P}^0(\mathcal{P}_i|N_r,N_m,\mathcal{M}_i)=\frac{W_L(N_m)}{N_m}.\quad(43)$$

Now we need to find the probability of $\mathcal{O}_i(W)$ appearing in both terms in (42). For shortness, let us define the vector of RVs $Z\triangleq(N_u,N_r,N_m,\tilde{\mathbf{A}})$. We have:

$$\begin{aligned}&\mathbb{P}^0\left(\mathcal{O}_i(W_L)|\mathcal{P}_i,N_r,N_m,\mathcal{M}_i\right)\\&\quad=\mathbb{E}^0\left[\sum_{v\in\tilde{\mathbf{A}}}\mathbb{P}^0\left(\mathcal{O}_i(W_L),V_i=v|\mathcal{P}_i,\mathcal{M}_i,Z\right)|N_r,N_m\right].\quad(44)\end{aligned}$$

Notice that, we only need to add over the $v\in\tilde{\mathbf{A}}$ because we are conditioning on $\mathcal{M}_i$. This guarantees that in the following step we will not condition with respect to an event of zero probability. Now, since there is a match and only one transmitter will serve the request we have:

$$\begin{aligned}&\mathbb{P}^0\left(\mathcal{O}_i(W_L)|V_i,\mathcal{P}_i,\mathcal{M}_i,Z\right)\\&\quad=\sum_{c\in\mathcal{A}(\tilde{\mathbf{A}},V_i)}\frac{\mathbb{P}^0\left(\mathcal{O}_i(W_L)|C_i=c,V_i,\mathcal{P}_i,\mathcal{M}_i,Z\right)}{\#\mathcal{A}(\tilde{\mathbf{A}},V_i)}\\&\quad=\mathbb{P}^0\left(R_a(W_L,S,D_i)>R|N_r,N_m,N_u\right),\quad(45)\end{aligned}$$

where S and D follow the distribution of any user in the cluster (uniform) and R_a is given by (19). In the last step we use that the event of failed transmission to a user depends only on the number of slots in the block (that is, on N_m) and on the number of requests and caching users (otherwise the transmission may not be well defined). In this last step, the distribution of S and D_i are the same for any c,i. Thus, noticing that (45) does not depend on V_i and replacing it in (44) we can sum over V_i. This sum yields one, because we are conditioning on a match and on $\tilde{\mathbf{A}}$. Thus we obtain:

$$\begin{aligned}&\mathbb{P}^0\left(\mathcal{O}_i(W_L)|N_r,N_m,\mathcal{M}_i,\mathcal{P}_i\right)&(46)\\&\quad=\mathbb{P}^0\left(R_a(W_L,S,D_i)>R|N_m,N_r\right)&(47)\\&\quad=\mathbb{E}^0\left[\bar{F}_{|g_{S,D_i}|^2|S,D_i}\left(I(D,N_m)W_L(N_m)R\right)|N_m,N_r\right].&(48)\end{aligned}$$

$I(D,N_m)$ is the interference at D (20). With this we can find $\mathbb{P}^0(\mathcal{O}_i(W_L)|\mathcal{P}_i,N_r,N_m,\mathcal{M}_i)$ in (42). Following the same

procedure we find $\mathbb{P}^0(\mathcal{O}_i(W_H)|N_r, N_m, \mathcal{M}_i)$ in (42), which concludes the proof. Similarly we find the average rate:

$$\bar{R}(\mathcal{F}, R) = R\,\mathbb{E}^0\left[\sum_{i=1}^{N_r} \frac{\mathbb{1}_{\mathcal{S}_i}}{\sum_{j=1}^{N_r} \mathbb{1}_{\mathcal{M}_j \cap \mathcal{P}_j}}\right] = R\,\mathbb{E}^0\left[\sum_{i=1}^{N_r} \frac{\mathbb{1}_{\mathcal{M}_i \cap \mathcal{P}_i} \mathbb{E}^0[\mathbb{1}_{\mathcal{O}_i}|\mathcal{M}_i, \mathcal{P}_i, N_r, N_m]}{\sum_{j=1}^{N_r} \mathbb{1}_{\mathcal{M}_j \cap \mathcal{P}_j}}\right]. \quad (49)$$

In the last step the conditional expectation does not depend on i and is the same as (48) with $W(N_m)$ instead of $W_L(N_m)$.

E. Proof of Lemma 3.2

If X, Y are real RVs, X second-order stochastically dominates [21] Y, written $X \geq_2 Y$ if, for any monotone increasing concave function $\phi(u)$: $\mathbb{E}[\phi(X)] \geq \mathbb{E}[\phi(Y)]$. We have [22]:

Theorem A.1: If $\{X_1, \ldots X_n\}$ and $\{Y_1, \ldots, Y_n\}$ are sets of independent real RVs and $\{\beta_i\}_{i=1}^n$ are non-negative real numbers:

$$X_i \geq_2 Y_i \;\forall\, i \Longleftrightarrow \sum_{i=1}^{n} \beta_i X_i \geq_2 \sum_{i=1}^{n} \beta_i Y_i, \quad (50)$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq_2 \sum_{i=1}^{n} \beta_i X_i, \text{ when } \sum_{i=1}^{n} \beta_i = 1. \quad (51)$$

We now consider the interference (20) generated only by the clusters with centers inside $\mathcal{B}(0, \rho)$:

$$I_\rho(d, n_1) = \sum_{x \in (\Phi \cap \mathcal{B}(0,\rho)) \setminus \{0\}} \psi(x, d, \mathbf{m}_x, n_1), \quad (52)$$

where $\psi(x, d, \mathbf{m}_x, n_1)$ is the function in the sum in (20). This sum has a finite number of terms almost surely. Now, we condition on the cluster centers Φ, on the number of slots $\{W_x\}$ and on the variables $\{B_{x,i,j}\}$ of all the clusters. Then the only randomness in the sum (52) comes from the fading coefficients and the positions of the users around the cluster centers, which are all independent RVs. With this, for a cluster with $W_x = 2^i > n_1$ and using (51) we have:

$$\psi(x, d, \mathbf{m}_x, n_1) = \frac{n_1}{2^i} \sum_{j=1}^{2^i/n_1} B_{x,j,i} |h_{x,j,d}|^2 l(x + r_j, d) \geq_2 \frac{n_1}{2^i} \sum_{j=1}^{2^i/n_1} |h_{x,j,d}|^2 l(x + r_j, d) = \tilde{\psi}(x, d, \mathbf{m}_x, n_1)$$

where $\tilde{\psi}(x, d, \mathbf{m}_x, n_1)$ is the same $\psi(x, d, \mathbf{m}_x, n_1)$ but has $B_{x,i,j} = 1$. Keeping the conditioning, and summing over all the clusters which have more that n_1 slots and using (50):

$$\sum_{x \in (\Phi \cap \mathcal{B}(0,\rho)) \setminus \{0\}} \tilde{\psi}(x, d, \mathbf{m}_x, n_1) \mathbb{1}_{\{W_x > n_1\}} \geq_2 \sum_{x \in (\Phi \cap \mathcal{B}(0,\rho)) \setminus \{0\}} \psi(x, d, \mathbf{m}_x, n_1) \mathbb{1}_{\{W_x > n_1\}}. \quad (53)$$

With the conditioning, the interference from clusters such that $W_x \leq n_1$ is independent of the ones such that $W_x > n_1$, so using (50) once more get: $\sum_{x \in (\Phi \cap \mathcal{B}(\rho)) \setminus \{0\}} \tilde{\psi}(x, d, \mathbf{m}_x, n_1) \geq_2 I_\rho(d, n_1)$. Using the definition of stochastic dominance with $\phi(u) = -e^{-su}$, $s > 0$, which is concave and increasing in u. Averaging over the conditioned RVs we get the desired result but for $I_\rho(d, n_1)$ instead of the full interference. We conclude by letting $\rho \to \infty$ and using monotonicity arguments.

ACKNOWLEDGMENT

The authors are grateful to the Associate Editor and the anonymous reviewers for their constructive comments, which helped to improve the manuscript.

REFERENCES

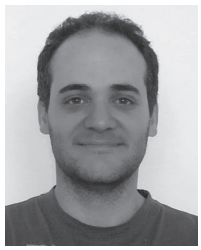
[1] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP D2D proximity services," *IEEE Commun. Mag.*, Dec. 2013.
[2] C. Systems, "The zettabyte era—Trends and analysis," Cisco System, San Jose, CA, USA, White paper, 2013.
[3] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage trade-off of wireless one-hop caching networks," *preprint arXiv:1312.2637*, 2013.
[4] N. Golrezaei, A. F. Molisch, A. Dimakis, and G. Caire, "Femtocaching and D2D collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.
[5] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.
[6] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd ed. Hoboken, NJ, USA: Wiley, 2009.
[7] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry and Wireless Networks*. Boston, MA, USA: Now Publishers, 2009.
[8] N. Golrezaei, A. Dimakis, and A. F. Molisch, "Wireless device-to-device communications with distributed caching," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 2781–2785.
[9] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 1461–1465.
[10] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *preprint arXiv: 1305.5216*, 2013.
[11] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *preprint arXiv:1405.5336*, 2014.
[12] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.
[13] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Resource optimization in device-to-device cellular systems using time-frequency hopping," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5467–5480, Oct. 2013.
[14] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Device-to-device modeling and analysis with a modified Matérn hardcore BS location model," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 1825–1830.
[15] 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, Dec. 2013.
[16] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and trade-offs," *EURASIP J. Wireless Commun. Netw.—Special Issue Tech. Adv. Design Deployment Future Heterogeneous Netw.*, vol. 2015, no. 1, p. 41, Feb. 2015.
[17] H. Q. Nguyen, F. Baccelli, and D. Kofman, "A stochastic geometry analysis of dense IEEE 802.11 networks," in *Proc. IEEE INFOCOM*, 2007, pp. 1199–1207.
[18] M. Haenggi, "Mean interference in hard-core wireless networks," *IEEE Commun. Lett.*, vol. 15, no. 8, pp. 792–794, Aug. 2011.
[19] D1. 1.2, Winner II Channel Models IST-4-027756 WINNER II, 2007. [Online]. Available: http://www.ist-winner.org/deliverables.html
[20] R. B. Ash and C. Doléans-Dade, *Probability and Measure Theory*. San Diego, CA, USA: Academic, 2000.
[21] E. Wolfstetter, *Topics in Microeconomics*. Cambridge, U.K.: Cambridge Univ. Press, 1999.
[22] C.-K. Li and W.-K. Wong, "Extension of stochastic dominance theory to random variables," *RAIRO-Oper. Res.*, vol. 33, no. 4, pp. 509–524, Oct. 1999.

Andrés Altieri (S'12–M'15) received the B.Sc. and M.Sc. degrees (with honors) in electrical engineering from the University of Buenos Aires, Buenos Aires, Argentina, in 2009 and the joint Ph.D. degree from the University of Buenos Aires and École Supérieure d'Électricité (SUPELEC), Gif-sur-Yvette, France, in 2014.

He is currently a Postdoctoral Researcher with the Department of Electronics, School of Engineering, Universidad de Buenos Aires (UBA), Buenos Aires, Argentina. He was with the Laboratoire des Signaux et Systèmes, CentraleSupélec-CNRS-Université Paris-Sud, Gif-sur-Yvette, France. His research interests include cooperative wireless networks, information theory, and stochastic geometry.

Pablo Piantanida (S'04–M'08) received the B.Sc. and M.Sc. degrees (with honors) in electrical engineering from the University of Buenos Aires, Buenos Aires, Argentina, in 2003 and the Ph.D. degree from the Paris-Sud University, Orsay, France, in 2007.

In 2006, he was with the Department of Communications and Radio-Frequency Engineering, Vienna University of Technology, Vienna, Austria. Since October 2007, he has been with Laboratoire des Signaux et Systèmes, CentraleSupélec-CNRS-Université Paris-Sud, Gif-sur-Yvette, France, as an Assistant Professor in network information theory. His research interests include multiterminal information theory, cooperative communications, physical-layer security, and distributed source coding.



Leonardo Rey Vega (S'07–M'11) received the M.Sc. (with honors) and Ph.D. degrees in electrical engineering from the University of Buenos Aires, Buenos Aires, Argentina, in 2004 and 2010, respectively.

In 2007 and 2008, he was invited by Énergie Matériaux Télécommunications, Institut National de la Recherche Scientifique, Montreal, QC, Canada, and in the first semester of 2012, he was a Visitor with the Department of Telecommunications, CentraleSupélec, Gif-sur-Yvette, France. He is currently an Assistant Professor with the Department of Electronics, School of Engineering, Universidad de Buenos Aires (UBA). His research interests include information theory, cooperative communications, and statistical signal processing.

Cecilia G. Galarza received the Ingeniera Electrónica degree from the University of Buenos Aires, Buenos Aires, Argentina, in 1990 and the M.S. and Ph.D degrees in electrical engineering and computer science from the University of Michigan, Ann Arbor, MI, USA, in 1995 and 1998, respectively.

She is currently an Associate Professor with the Department of Electronics, School of Engineering, Universidad de Buenos Aires (FIUBA). Before joining FIUBA, she worked with Voyan Technologies, Santa Clara, CA, USA. She is also a Researcher with the National Scientific and Technological Research at CSC-CONICET, Buenos Aires, Argentina. Her research interests include wireless networks, detection on sensor networks, and sensor management.