

Information-Theoretic Capacity of Clustered Random Networks

Michele Garetto^{*}, Alessandro Nordio[†], Carla-Fabiana Chiasserini[‡], Emilio Leonardi[‡]

^{*} Università degli Studi di Torino, Corso Svizzera 185, 10149 - Torino, Italy

Email: michele.garetto@unito.it

[†] IEIIT-CNR, Corso Duca degli Abruzzi 24, 10129 - Torino, Italy

Email: alessandro.nordio@polito.it

[‡] Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 - Torino, Italy

Email: {chiasserini,leonardi}@polito.it

Abstract

We analyze the capacity scaling laws of clustered ad hoc networks comprising significant inhomogeneities in the node spatial distribution over the area. In particular, we consider the class of networks in which nodes are distributed according to a doubly stochastic shot-noise Cox process, which allows to model a wide variety of inhomogeneous topologies. For this class of networks, we derive information theoretic upper-bounds to the capacity, identifying six operational regions. We provide also constructive lower bounds, by devising, for each region, an optimal communication strategy to achieve the maximum network throughput. The performance of our communication schemes match, in terms of scaling exponent, the theoretical upper-bounds.

I. INTRODUCTION AND RELATED WORK

The capacity of ad hoc wireless networks has been traditionally studied considering single-user communication schemes over point-to-point links [1]. Only recently [2], [3], [4], [5], information-theoretic scaling laws of ad hoc networks have been investigated, showing that multi-user cooperative schemes can achieve much better performance than traditional single-user schemes, especially in the low power attenuation regime.

This work has been supported by the European Commission through STAMINA (Statistical Mechanics Inspired Methods for Green Autonomous Networking) FET-Open Project, ICT FP7-ICT-2007-8.0 G.A. n. 265496

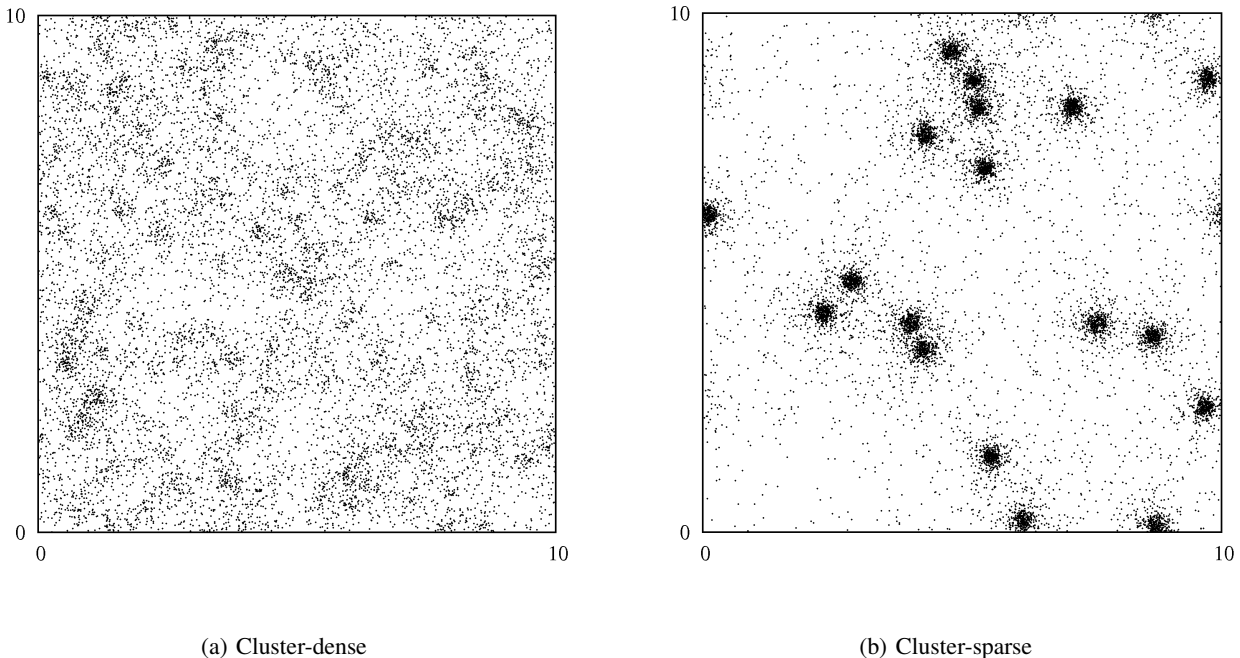


Fig. 1. Examples of SNCP node distribution: (a) *cluster-dense* and (b) *cluster-sparse* network topologies.

In this paper, we analyze the information-theoretic capacity of clustered random networks containing significant inhomogeneities in the node spatial distribution. The interest in such network topologies is driven by the fact that they well represent several systems found in the real world. In particular, we consider nodes distributed according to a doubly stochastic Shot-Noise Cox Process (SNCP) over a square region whose edge size can scale with the number of nodes. The SNCP model allows us to analyze a very broad class of network topologies by varying the model parameters, including the uniform distribution as a special case (hereinafter referred to as homogeneous network).

As the SNCP parameter settings vary, two different regimes can be identified: *cluster-dense* and *cluster-sparse*. In the former, clusters largely overlap and the node density does not exhibit significant inhomogeneities (in order sense) across the network area. In the latter, clusters are well distinct from each other and the node density tends to become strongly inhomogeneous as the number of nodes grows. Examples of the two network regimes are depicted in Fig. 1.

To analyze the network capacity in these two regimes, we follow the stream of work [2], [3], [4], [5], which relies on the fundamental assumption that all channel gains among the nodes exhibit i.i.d. random phase shifts with uniform distribution in $[0, 2\pi]$. This assumption allows to fully exploit the additional

capacity provided by cooperative communication schemes (such as MIMO or interference alignment).

We emphasize that the i.i.d. uniform phase model, although widely accepted in the field of wireless communication, may not hold in some operating conditions. Indeed, as shown in [11], [12], the network capacity is ultimately constrained by the total degrees of freedom in the network, which can be derived from physical principles by studying basic properties of the radiated field.

In the case of traditional geometries (e.g., planar networks), the study in [13] has shown that, when the ratio of the diameter of the network area to the carrier wavelength is smaller than \sqrt{n} , the network capacity is already achieved by traditional multi-hop in the case of homogeneous topologies. This means that, unless the network area scales more than linearly with the number of nodes (leading to lower and lower node density), cooperative schemes do not help. We emphasize that in [13] authors consider free-space propagation, homogeneous networks, and assume that nodes can scale up the transmission power to compensate for the increasing inter-node distances occurring at low node density. They also argue that in many realistic scenarios there is a huge range of values of the number of nodes for which the degrees of freedom limitation is not an issue.

In this paper, we rely on the above practical argument, and we assume the cut-set integral, as defined in [12], to be large compared to the number of nodes in the network (i.e. the i.i.d. uniform phase model). In our work, we are interested in studying the limitations to the capacity imposed by the distribution of nodes in the area, and we think that isolating the impact of the topology is a necessary first step before considering other factors. However, we do consider the possibility that the network capacity results constrained by the amount of power available in the nodes, since, in contrast to [13], we do not allow the transmission power to scale up as the network area expands (more precisely, we assume that the average power budget of each node is a fixed constant).

The goal of our work is not to show the potential benefit of cooperative communication schemes with respect to traditional point-to-point schemes, but to study the impact on the capacity of the nodes topology, at least for a significant class of clustered networks. In particular, for both the *cluster-dense* and *cluster-sparse* regimes, we provide both information-theoretic upper-bounds to the achievable capacity and constructive lower-bounds, which are asymptotically tight to within a poly-log factor (in the number of nodes).

With respect to previous work based on the i.i.d. phase model, our findings are as follows. Recall that in the case of homogeneous networks [2], [3], three communication strategies can be identified, namely, hierarchical cooperation, multihop, and multihop MIMO hierarchical cooperation, which allow to achieve the network capacity as the system parameters vary. In the case of *cluster-dense* topologies, we find that

the node density is ‘sufficiently uniform’ to ensure that capacity can be achieved by a simple modification of the above three strategies.

In the case of *cluster-sparse* topologies, the picture is more complex. For small path-loss exponents α (i.e., $\alpha \in (2, 3)$), similarly to [4], we find that the network capacity does not depend on how nodes are distributed over the area. For large path-loss exponents ($\alpha > 3$), instead, the network capacity is significantly affected by how nodes are distributed over the area, as already observed in [4]. However, in [4] the characterization of the capacity achievable for large path-loss exponents is limited to the case of adversarial node placement under a deterministic (given) degree of network regularity. Moreover, the authors of [4] impose a minimum separation constraint between the nodes, which does not allow the study of highly dense clusters over the area, and they focus on extended networks only (i.e., networks whose area grows linearly with the number of nodes¹). In our work, instead, we consider generic clustered network topologies, and we do not make any assumption on the minimum separation between nodes. In the *cluster-sparse* regime, we devise different strategies to achieve the network capacity, by properly combining previous communication schemes. Such combinations are obtained by applying multi-hop and hierarchical cooperation at different spatial scales, e.g., sub-cluster or super-cluster, requiring us to develop totally different scheduling/routing strategies with respect to previous work.

In [5], authors show that the classical multi-hop communication strategy is throughput-optimal in arbitrary networks whose area grows linearly with the number of nodes, under the assumption that the signal power decays exponentially with the distance. In contrast to [5], we allow arbitrary scaling of the network area, and we consider the usual propagation model in which that power decays with the distance d as $d^{-\alpha}$, with $\alpha \geq 2$.

In summary, our main contributions are as follows. Under the i.i.d. uniform phase model, we give a complete characterization of the network capacity achievable w.h.p.² for SNCP node placement, which extends the capacity analysis to a much broader class of network topologies than those studied in previous work. Our constructive lower bounds employ novel scheduling/routing strategies in combination to existing cooperative communication schemes. Such strategies represent an important contribution in themselves, as they could be adopted to cope with the nodes spatial inhomogeneity in more general topologies that cannot be described by the SNCP model.

¹In [3] authors recognized the importance of letting the network area scale with the number of nodes in a general way, as this gives rise to a richer set of operational regimes.

²In this paper we adopt the expression ‘with high probability’ (w.h.p.) to indicate events/properties that occur with a probability $p = 1 - O(\frac{1}{n})$.

Finally, we mention that this work extends [6], [7], where we have analyzed the capacity of networks in which nodes are distributed according to an SNCP model, but considering single-user communication schemes only (i.e., traditional point-to-point links). An early version of this paper can be found in [8].

The rest of the paper is organized as follows. In Section II, we introduce our system assumptions and notation, describing the network topology, the communication model and the traffic model. In Section III, we provide a summary of our results. In Sections IV and V, we detail the derivation of the lower and upper bounds, respectively. Some concluding remarks are reported in Section VI.

II. SYSTEM ASSUMPTIONS AND NOTATION

A. Network Topology

We consider a network composed of a random number N of nodes (being $\mathbb{E}[N] = n$) distributed over a square region \mathcal{O} of edge length L , where L takes units of distance. The network physical extension L scales with the average number of nodes, since this is expected to occur in many growing systems. Throughout this work we will assume that $L = n^\gamma$, with $\gamma \geq 0$. To avoid border effects, we consider wrap-around conditions at the network edges (i.e., the network area is assumed to be the surface of a bi-dimensional Torus).

The clustering behavior typically encountered in realistic, large-scale topologies is taken into account assuming that nodes are placed according to a shot-noise Cox process (SNCP). An SNCP [9] over an area \mathcal{O} can be conveniently described by the following construction. We first specify a homogeneous Poisson point process of cluster centres, whose positions are denoted by $\mathcal{C} = \{c_j\}_{j=1}^M$, where M is a random number with average $\mathbb{E}[M] = m$. In the literature the centre points c_j are also called parent or mother points. Each centre point c_j in turn independently generates a point process of nodes whose intensity at ξ is given by $q s(\|\xi - c_j\|)$, where $q \in (0, \infty)$ and $s(\|\xi - c_j\|)$ is a non-negative, non-increasing, bounded and continuous function, also called kernel (or shot) whose integral $\int_{\mathcal{O}} s(\|\xi - c_j\|) d\xi$ over the entire network area is equal to 1.

In practice, we consider

$$s(\|\xi - c_j\|) \propto \min(1, \|\xi - c_j\|^{-\delta}) \quad \text{for } \delta > 2, \quad (1)$$

although our results apply to more general shapes as well.

The overall node process is given by the superposition of the individual processes generated by the cluster centres. The local intensity at $\xi \in \mathcal{O}$ of the resulting SNCP is

$$\Phi(\xi) = \sum_{j=1}^M q s(\|\xi - c_j\|).$$

Under the above assumptions on the kernel shape, quantity q equals the average number of nodes generated by each cluster centre (all cluster centres generate on average the same number of nodes). In our work, we let q scale with n as well (clusters are expected to grow in size as the number of nodes increases). This is achieved assuming that the average number of cluster centres scales as $m = n^\nu$, with $\nu \in (0, 1)$. Consequently, the number of nodes per cluster scales as $q = n^{1-\nu}$.

Notice that $\Phi(\xi)$ is a random field, in the sense that, conditionally over the locations of cluster centres \mathcal{C} , the node process is an (inhomogeneous) Poisson point process with intensity function Φ . We denote by $\mathcal{X} = \{X_i\}_{i=1}^N$ the collection of all nodes generated in a given realization of the SNCP.

Moreover, let

$$d_c = L/\sqrt{m} = n^{\gamma-\nu/2} \quad (2)$$

be the typical distance between cluster centres. More precisely, d_c is the edge of the square where the expected number of cluster centres falling in it equals 1. We call³

- *cluster-dense* condition the case $\gamma < \nu/2$, in which d_c tends to zero as n increases;
- *cluster-sparse* condition the case $\gamma > \nu/2$, in which d_c tends to infinity as n increases.

Under both conditions, we define the following two quantities: $\bar{\Phi} = \sup_{\xi \in \mathcal{O}} \Phi(\xi)$ and $\underline{\Phi} = \inf_{\xi \in \mathcal{O}} \Phi(\xi)$, representing (loosely speaking) the maximum and the minimum node density in the network. Notice that the above two quantities are random variables depending on the positions \mathcal{C} of the cluster centres.

The asymptotic characterization of $\bar{\Phi}$ and $\underline{\Phi}$ for the class of topologies considered in this work has been already done in previous work [6] (see Lemma 4 in Appendix D). The basic results are: in the *cluster-dense* condition, we have $\underline{\Phi} = \Theta(\bar{\Phi}) = \Theta(n/L^2)$, which means that the network is uniformly dense in order sense, since $\bar{\Phi}$ and $\underline{\Phi}$ differ at most by a constant factor; instead, in the *cluster-sparse* condition, we have $\underline{\Phi} = o(\bar{\Phi})$, hence the network contains significant inhomogeneities in the node density.

B. Communication Model

We use the same channel model as in [2], [3], [4]. Consider the generic time t , and let $\tau(t)$ be the set of nodes transmitting at time t . The signal received at time t by a node k is

$$y_k(t) = \sum_{i \in \tau(t) \setminus \{k\}} h_{ik}(t)x_i(t) + z_k(t)$$

³In this work we do not explicitly consider $\gamma = \nu/2$ since it requires a separate analysis. However, it can be shown that results obtained for the cluster dense regime hold also in this case.

TABLE I
SYSTEM PARAMETERS (n.a. = NOT APPLICABLE)

Parameter	Definition	Scaling exponent
L	edge length of network area	$\gamma \geq 0$
m	average number of clusters	$0 < \nu < 1$
P	per-node power budget	0
α	path-loss exponent	n.a.
δ	decay exponent of the kernel	n.a.
d_c	typical distance between cluster centres	$\gamma - \nu/2$
q	average number of nodes per cluster	$1 - \nu$

where $x_i(t)$ is the signal emitted by node i , and $\{z_k(t)\}_{k,t}$ are white circularly symmetric Gaussian noise, independently and identically distributed (i.i.d.) with distribution $\mathcal{N}_{\mathbb{C}}(0, N_0)$ (with zero mean and variance N_0). The complex baseband-equivalent channel gain $h_{ik}[t]$ between i and k at time t is

$$h_{ik}(t) = d_{ik}^{-\alpha/2} e^{j\theta_{ik}(t)}$$

where $\alpha \geq 2$ is the path-loss exponent, and $\{\theta_{ik}(t)\}_{i,k}$ are i.i.d. random phases with uniform distribution in $[0, 2\pi)$, which are assumed to vary in a stationary ergodic manner over time (fast fading). Moreover, $\{\theta_{ik}(t)\}_{i,k}$ and $\{d_{ik}\}_{i,k}$ are also assumed to be independent, $\forall i, k$. The validity of the assumption on independence of phases is still an open issue, as discussed in [11], [12], [13].

At last, we impose an average power constraint of P on the transmissions performed by each node, where P is a constant.

C. Traffic Model

We assume that each node is source and destination of a single flow, and that the resulting N flows (with $\mathbb{E}[N] = n$) are established at random without any consideration of node locations. Given a communication strategy, we denote by $\lambda(n)$ the average rate between source and destination over the traffic flow duration (hereinafter referred to as per-node throughput).

Table I summarizes the parameters of our model. The scaling exponents reported in the third column of the table are the base- n logarithms of the corresponding parameters. Note that d_c and q are not native parameters, since they are derived from others, however we have included them in the table for convenience. To simplify the notation, in the following we will omit the dependence of the variables on n unless strictly necessary.

TABLE II
CAPACITY SCALING EXPONENTS AND OPERATIONAL REGIONS, $\beta = 1 - \nu - \delta(\gamma - \nu/2)$

e_C	regions	conditions	strategy
1	I	$\alpha\gamma \leq 1$	GL
$2 - \alpha\gamma$	II	$\alpha\gamma > 1 \wedge \alpha \leq 3$	GL
$\frac{\alpha-1-\alpha\gamma}{\alpha-2}$	III	$\alpha\gamma > 1 \wedge \alpha > 3 \wedge \gamma < \frac{1}{2} \wedge \frac{1-2\gamma}{\alpha-2} \geq \gamma - \frac{\nu}{2}$	SC
$\gamma + \beta \frac{\alpha-1}{\alpha-2}$	V	$\alpha\gamma > 1 \wedge \alpha > 3 \wedge \gamma \geq \frac{\nu}{2} \wedge \beta > 0 \wedge \gamma + \beta \frac{\alpha-1}{\alpha-2} > 2 - \alpha\gamma + (\alpha-3)\frac{\nu}{2}$	sC
$\gamma + \beta \frac{\alpha+1}{2}$	VI	$\alpha\gamma > 1 \wedge \alpha > 3 \wedge \beta \leq 0 \wedge \gamma + \beta \frac{\alpha+1}{2} > 2 - \alpha\gamma + (\alpha-3)\frac{\nu}{2}$	MH
$2 - \alpha\gamma + (\alpha-3)\frac{\nu}{2}$	IV	elsewhere	IC

III. MAIN RESULTS

Similarly to previous work [2], [3], we express our results in terms of the scaling exponent e_C of the network capacity $C(n) = n\lambda(n)$, defined as,

$$e_C = \lim_{n \rightarrow \infty} \log_n C(n)$$

The scaling exponent⁴ allows to ignore factors scaling with $o(n^\epsilon), \forall \epsilon > 0$. Since the lower and upper bounds on the system capacity that we derive in Sections IV and V are within $o(n^\epsilon), \forall \epsilon > 0$, the corresponding scaling exponents match, and we can claim that our characterization of the network capacity in terms of the scaling exponent is exact.

In the case of *cluster-dense* network topologies, we obtain for e_C the same results derived in [3] for homogeneous networks. Indeed, in this case the node density can be considered, in order sense, as uniform over the network area.

In the *cluster-sparse* regime, we obtain an upper bound to the network capacity by extending the results in [3], [4] and taking into account the variability of the node density over the network area. The basic idea underlying the computation of the upper bound is to evaluate the overall power transfer through a cut dividing the network area in two halves, and to relate such power transfer to the maximum possible information flow among the nodes.

In particular, under the SNCP model, we identify six operational regions, denoted by Latin numbers (I, II, ..., VI), as in Table II. In each of these regions, we find that the dominant contribution to the network capacity is due to transmissions between nodes separated by distance:

- $\Theta(L)$, in regions I and II;
- jointly $\omega(d_c \sqrt{\log n})$ and $o(L)$, in region III;

⁴In general, we denote the scaling exponent of an arbitrary function $f(n)$ as $e_f = \lim_{n \rightarrow \infty} \log_n f(n)$.

- jointly $O(d_c\sqrt{\log n})$ and $\Omega(d_c)$, in region IV;
- jointly $o(d_c)$ and $\omega(1/\sqrt{\Phi})$, in region V;
- $\Theta(1/\sqrt{\Phi})$, in region VI.

The above observation guided us also in the derivation of lower bounds to the network capacity. For each of the above regions, we have found a proper communication scheme that allows to achieve a network capacity characterized by the same scaling exponent of the corresponding upper bound. More specifically, we identify two subsets of nodes: the main infrastructure and the set of access/delivery nodes. The former set is used to transfer data across the network area, thus giving the main contribution to the network throughput; the latter set is responsible for sending/receiving data to/from the main infrastructure, without acting as a bottleneck for the network throughput.

Depending on the operational region, we select two different main infrastructures:

- **CC** infrastructure, which is composed by nodes located at finite distance from the center of their cluster;
- **HI** infrastructure, which forms an HPP process of intensity equal to the minimum (asymptotic) node density over the area.

With the first infrastructure (**CC**), the communication schemes that maximize the throughput are similar to the ones proposed in [3], [4]; we refer to them as Global Cooperative (**GL**), cooperative Super-Cluster hopping (**SC**) and cooperative Inter-Cluster hopping (**IC**), depending on the typical distance between transmitters and receivers used in the largest-scale MIMO communications.

With the second infrastructure (**HI**), the schemes maximizing the network throughput are the traditional Multi-Hop (**MH**) and the cooperative sub-Cluster hopping (**sC**). In the latter scheme we still employ MIMO communications, among nodes located at a distance smaller (in order sense) than the centers of neighboring clusters but larger than the typical distance between two neighboring nodes on the **HI** infrastructure.

A graphical representation of our results on e_C is given in Figure 2, for $\nu = 0.3$ and $\delta = 2.5$, while varying α and γ . Figure 3 depicts the behavior of e_C for $\alpha = 5$ and $\gamma = 0.4$, while varying ν and δ . It can be noticed that e_C varies with continuity in the four-dimensional space of the parameters $\{\alpha, \gamma, \delta, \nu\}$, and that, in any region, e_C is a non-increasing function of parameters $\{\alpha, \gamma, \delta\}$ and a non-decreasing function of ν .

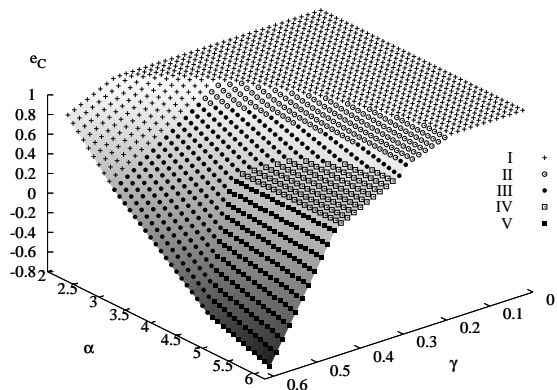


Fig. 2. Scaling exponent of network capacity as function of α and γ , for $\nu = 0.3$ and $\delta = 2.5$. Different marks are associated to the six possible operating regions.

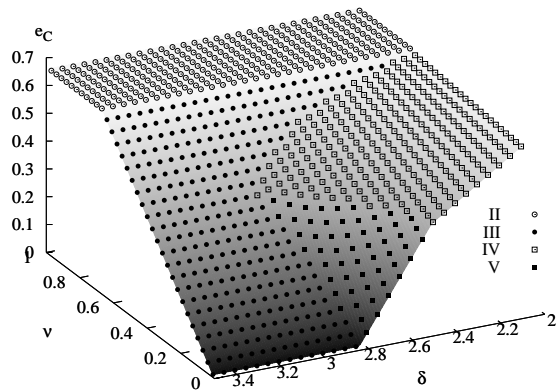


Fig. 3. Scaling exponent of network capacity as function of ν and δ , for $\alpha = 5$ and $\gamma = 0.4$. Different marks are associated to the six possible operating regions.

IV. LOWER BOUNDS

In this section, we present the family of communication schemes that allows to achieve the capacity lower bound of the clustered networks considered in our work. Existing lower bounds that have been obtained in previous work for the case of homogeneous networks can be found in Appendix C.

Depending on the operational region, different schemes must be employed to achieve optimal throughput performance. Before going into details, we provide a brief classification of the proposed schemes according to their main properties.

All of our proposed schemes work as follows: first, a subset of nodes is identified, which forms the main infrastructure through which data are transferred across the network area. A finite fraction of time is assigned to the rest of the nodes to exchange traffic with the nodes belonging to the main infrastructure (if needed). More precisely, time is divided into regular frames, each one comprising three phases of equal duration: i) an *access* phase, in which sources not belonging to the main infrastructure send data to the infrastructure; ii) a *transport* phase, in which data are transferred over the infrastructure; iii) a *delivery* phase, in which data are sent from the infrastructure to destinations not belonging to it.

We emphasize that in all cases the *delivery* phase is analogous to the *access* phase (by exchanging the role of transmitters and receivers). Thus, we classify the schemes on the only basis of their *transport* phase and *access* phase.

1) *Transport phase*: for what concerns the *transport* phase, our schemes follow the same principle adopted in the case of homogeneous networks (see Appendix C), according to which the network area is

partitioned into cells of edge size c : MIMO communications are established between the nodes belonging to neighboring cells, and global multi-hopping at the cell level is employed to transfer data through the network. In particular, the value of the cell edge size c allows us to classify our schemes into the following five strategies:

- 1) **GL: cooperative global**, in which $c = \Theta(L)$ and nodes employ a MIMO communication scheme at global network scale, without the need of cell multi-hopping;
- 2) **SC: cooperative super-cluster hopping**, in which nodes employ a cooperative multi-hop scheme, where $c = \omega(d_c \sqrt{\log n})$;
- 3) **IC: cooperative inter-cluster hopping**, in which $c = \Theta(d_c \sqrt{\log n})$, i.e., the cell edge size is closely related to the typical distance d_c between cluster centres;
- 4) **sC: cooperative sub-cluster hopping**, in which $c = o(d_c)$ and $c = \omega(1/\underline{\Phi})$, i.e., the cell edge size is smaller (in order sense) than the typical distance between cluster centres, yet the cell is large enough that the number of nodes falling in it tends to infinity, allowing cooperation among an increasingly number of nodes;
- 5) **MH: traditional multi-hop**, in which $c = \Theta(1/\sqrt{\underline{\Phi}})$, and nodes resort to the traditional point-to-point multi-hop scheme, since there is no advantage (in order sense) in employing cooperative techniques.

The above five strategies for the *transport* phase are applied to different main infrastructures, depending on the combination of system parameters. We identify the following two main infrastructures:

- **CC: clusters-core infrastructure**, which is used only in the the *cluster-sparse* condition and it is formed by a set \mathcal{Z} of nodes falling within a finite distance from their cluster centre. The cardinality of this set is still $\Theta(n)$, as shown later.
- **HI: homogeneous infrastructure**, which is obtained extracting a subset \mathcal{Z} of nodes forming an Homogeneous Poisson Process of nodes with intensity equal to the minimum node density in the network, denoted by $\underline{\Phi}$. In particular, under the *cluster-dense* condition ($\gamma < \nu/2$) the cardinality of \mathcal{Z} is $\Theta(n)$, whereas under the *cluster-sparse* condition the cardinality of \mathcal{Z} is $o(n)$;

As we will see, the network throughput is ultimately determined by the transport phase, i.e., by the throughput of the selected strategy (one of the five strategies above) over the main infrastructure (one of the two infrastructures above). However, not all combinations are meaningful: in the *cluster-dense* condition, we apply strategies **GL** or **SC** or **MH** over the **HI** main infrastructure. In the *cluster-sparse* condition, we apply strategies **GL** or **SC** or **IC** over the **CC** main-infrastructure, and strategies **sC** or

MH over the **HI** main infrastructure. We emphasize that, although the above combinations (either for the *cluster-dense* or the *cluster-sparse* condition) can always be applied, only one of them provides the maximum network throughput for a given set of system parameters.

2) *Access phase*: recall that the *access* phase is used by sources to inject their traffic over the main infrastructure. Since the system throughput is ultimately determined by the main infrastructure, the goal is to design a communication scheme that does not shift the system bottleneck to the *access* phase, while at the same time balancing the traffic uniformly over the nodes of the main infrastructure. These design principles led us to consider the following three *access* scheme:

- **SISO**. This scheme is used to access the **HI** main infrastructure under the *cluster-dense* condition. In this case, it is sufficient to employ a single-hop point-to-point transmission (SISO) between each source and one of the closest nodes belonging to \mathcal{Z} , thanks to the fact that the network is almost uniformly dense;
- **Hierarchical multi-hop**. This scheme is used to access the **HI** main infrastructure under the *cluster-sparse* condition. For this case, we could not just adapt existing schemes proposed in the literature. Therefore we have developed a novel scheduling-routing strategy specifically tailored to this case.
- **SIMO**. This scheme is used to access the **CC** main-infrastructure under the *cluster-sparse* condition, employing a SIMO technique similar to the relaying scheme proposed in [4]⁵.

For the detailed presentation of our schemes, we will follow the latter classification based on the *access* phase: we start in Section IV-A considering the simpler class of schemes devised for the *cluster-dense* condition. As already said, in this case we combine a **SISO** access scheme with one of the three transport strategies (**GL** or **SC** or **MH**) applied over the **HI** main infrastructure.

Then, in Section IV-B we present the schemes devised to operate under the *cluster-sparse* condition over the **HI** main infrastructure. Such schemes combine either the **sC** or the **MH** transport strategy with the **Hierarchical multi-hop** access scheme.

At last, in Section IV-C we present those schemes exploiting the **CC** main-infrastructure under the *cluster-sparse* condition. Such schemes combine one of the three transport strategies **GL**, **SC**, **MH** with the **SIMO** access scheme.

⁵The corresponding scheme for the *delivery* phase is a MISO scheme analogous to the broadcast phase proposed in [4].

A. Communication schemes under the cluster-dense condition

Recall that under the *cluster-dense* condition the node density is in order sense uniform over the network area, since $\underline{\Phi} = \Theta(\overline{\Phi})$ (see Lemma 4 in Appendix D). Hence, we can reasonably expect the network throughput to scale in the same way as in the homogeneous case treated in [3], whose main results have been summarized in Appendix C. The following simple arguments show that this is indeed the case.

Using a standard thinning technique (see Lemma 7 in Appendix D), we can extract from \mathcal{X} (recall that $\mathcal{X} = \{X_i\}_{i=1}^N$ is the collection of all nodes generated in a given realization of the SNCP) a set of nodes \mathcal{Z} forming the main infrastructure and distributed according to an HPP process of intensity $\underline{\Phi}$, and consider the throughput of the sub-system comprising only the nodes in \mathcal{Z} . The latter can be characterized by (19) using $\beta = 1 - 2\gamma$, since $\underline{\Phi} = \Theta(n^{1-2\gamma})$. Nodes in \mathcal{Z} form the **HI** main infrastructure through which data are transported across the network adopting the schemes in [3], and whose aggregate throughput is exploited by all of the nodes.

It remains to show that the **SISO** access scheme used by nodes in $\mathcal{Y} = \mathcal{X} \setminus \mathcal{Z}$ to send/receive data from nodes in \mathcal{Z} does not throttle the throughput on the main infrastructure. This is guaranteed by the following lemma:

Lemma 1: . Under the *cluster-dense* condition, the rate achievable by each node of \mathcal{Y} during the *access* (or *delivery*) phase is $\Omega(1/\log^{1+\alpha/2} n)$.

Proof: The proof can be found in Appendix A. ■

Since nodes in \mathcal{Z} form an HPP process, the rate achievable by nodes in \mathcal{Z} during the *transport* phase can be immediately obtained applying existing results for homogeneous system. Furthermore, our construction guarantees that every node in \mathcal{Z} injects/collect in/from the main infrastructure the traffic associated to at most $\lceil 4\overline{\Phi}/\underline{\Phi} \rceil + 1 = \Theta(1)$ end-to-end traffic flows. As a result of previous discussion,

Proposition 1: Under the *cluster-dense* condition, i.e., for any set of system parameters in which $\gamma < \nu/2$, the network throughput is $\Theta(T_u(n, L, \underline{\Phi}, \alpha))$, where $T_u(\cdot)$ is given by (19) (in Appendix C). Therefore, using the expressions (20) of the scaling exponents, we obtain,

$$e_T(\alpha, \gamma, \nu, \delta) = \begin{cases} 1 & \alpha\gamma \leq 1, \delta < \nu/2 \\ 2 - \alpha\gamma & \alpha\gamma > 1, \alpha < 3, \delta < \nu/2 \\ \frac{\alpha-1-\alpha\gamma}{\alpha-2} & \alpha\gamma > 1, \alpha \geq 3, \gamma < \frac{1}{2}, \delta < \nu/2 \\ \frac{\alpha+1}{2} - \alpha\gamma & \alpha \geq 3, \gamma \geq \frac{1}{2}, \delta < \nu/2 \end{cases} \quad (3)$$

*B. Communication schemes employing the **HI** main infrastructure under the cluster-sparse condition*

Recall that under the *cluster-sparse* condition the node density is characterized by significant inhomogeneity (in order sense), being $\underline{\Phi} = o(\overline{\Phi})$. Nevertheless, using the same thinning technique adopted under the *cluster-dense* condition, it is still possible to extract from \mathcal{X} a set of nodes \mathcal{Z} distributed according to an HPP process of intensity $\underline{\Phi}$, and rely on the transport throughput of the sub-system comprising only the nodes in \mathcal{Z} .

However, the traffic exchange between the nodes in \mathcal{Z} and the rest of the nodes, i.e., nodes $\mathcal{Y} = \mathcal{X} \setminus \mathcal{Z}$, becomes problematic when the node density is highly inhomogeneous. Indeed, under the *cluster-sparse* condition, traffic is not uniformly balanced in the network when it leaves the sources, or when it is in proximity of the destinations, because the majority of sources/destinations are concentrated in proximity of the cluster centres. As consequence, a simple point-to-point single-hop scheme to send/receive data to/from the nodes of the main infrastructure (as done under the *cluster-dense* condition) can lead to the formation of system bottlenecks around the cluster centres, which would throttle the throughput over the main infrastructure. To overcome this problem, a significantly more complex communication strategy is needed during the *access* (or *delivery*) phase, as described in the following section.

1) **Hierarchical multi-hop access scheme:** During the *access* (*delivery*) phase, the formation of system bottlenecks around highly dense regions of the network area can be avoided adding a traffic spreading (traffic densification) phase before (after) data are transferred across the network area over the main transport infrastructure. A finite fraction of time can be devoted to each of these phases without modifying (in order sense) the throughput provided by nodes in \mathcal{Z} , with the additional benefit of uniformly balancing the traffic over the nodes of the main infrastructure. In this section we provide an intuitive understanding of how the traffic spreading (or densification) technique works. More details on the hierarchical multi-hop access scheme, which is essentially based on this technique, are relegated to Appendix E.

First of all, we can restrict ourselves to the traffic spreading phase, since the traffic densification phase is identical if we look at the data transmission process backward in time. The idea is to spread the traffic out of highly dense regions gradually, through a sequence of intermediate, local transport infrastructure nested one within the other (see Figure 4).

Intermediate transport infrastructures are obtained by thinning the overall point process of nodes within certain domains (specified later) surrounding the clusters centres. More specifically, in each domain we extract a set of points distributed according to an HPP of intensity equal to the minimum node density within the domain. Such set forms the local transport infrastructure of the corresponding domain.

The sequence $k = 0, 1, \dots, K_{\max}$ of nested domains is carefully chosen in such a way that: i) the

first domain in the sequence coincides with the network area, hence the corresponding infrastructure is the main transport infrastructure of the network, of density $\underline{\Phi}$, which is shared by all data flows; ii) the infrastructure extracted in each domain $k > 0$ can pass to the infrastructure of domain $k - 1$ all traffic generated by nodes contained in it; iii) the total number of domains grows at most like $\log n$, i.e., $k_{\max} = O(\log n)$.

Conditions i) and ii) guarantee that the system throughput is throttled by the lowest infrastructure (the main transport infrastructure) and no bottleneck arises within any higher infrastructure. Condition iii) guarantees that, even if we devote to each layer- k infrastructure a finite fraction of time, the total overhead due to the *access* phase causes at most a $\log n$ loss to the overall system throughput.

We now specify one possible way to jointly achieve the three conditions above. We build a sequence of nested domains \mathcal{O}_k , $k = 0, 1, 2, \dots, K_{\max}$, as follows. The first domain, denoted by \mathcal{O}_0 , coincides with the entire network area, i.e., $\mathcal{O}_0 = \mathcal{O}$, satisfying condition i).

For the generic point $\xi \in \mathcal{O}$, let $d_{\min}(\xi) = \min_j \|\xi - c_j\|$ be the distance between ξ and the closest cluster centre. We define domains \mathcal{O}_k , for $k \geq 1$, as follows: $\mathcal{O}_k = \{\xi \in \mathcal{O} : d_{\min}(\xi) \leq d_k\}$, where $\{d_k\}_k$ are a set of decreasing distances, i.e., $d_1 > d_2 > \dots > d_{K_{\max}}$. Domain \mathcal{O}_k is, in general, composed of a random number J_k of disjoint regions ($J_k \leq M$), corresponding to the connected components of the standard Gilbert's model of continuum percolation [18] with ball radius d_k . Figure 4 shows examples of domains \mathcal{O}_k having different values of d_k . Let $\{\mathcal{I}_k^j\}_j$ be the set of disjoint regions ($1 \leq j \leq J_k$) forming domain \mathcal{O}_k .

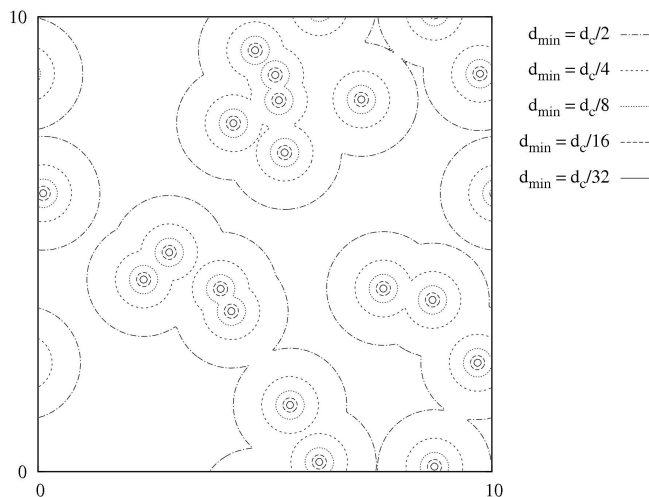


Fig. 4. Example of construction of nested domains \mathcal{O}_k for the topology depicted in Figure 1(b). Domain \mathcal{O}_1 is characterized by $d_1 = 0.5d_c$.

We set the largest d_k , namely d_1 , equal to $d_1 = \mu d_c$, where μ is a small constant. Choosing μ sufficiently small, in such a way that the associated Gilbert's model is below the percolation threshold (we need $\mu < \mu^*$, where $\mu^* \approx 0.6$, [17]), we have the property that the maximum number of clusters centres belonging to the same region \mathcal{I}_1^j is w.h.p. $O(\log n)$, $\forall j$ [18]. Since by construction $\mathcal{O}_{k+1} \subset \mathcal{O}_k$, the same property holds for all $k > 1$. It follows that, in terms of physical extension, the area $|\mathcal{I}_k^j|$ of region \mathcal{I}_k^j lies w.h.p. in the interval $\pi d_k^2 \leq |\mathcal{I}_k^j| \leq \pi d_k^2 \log n$.

We further observe that the density of nodes at any point within \mathcal{O}_k ($k \geq 1$) can be lower bounded by $\lambda_k = q d_k^{-\delta}$, by considering the contribution of the closest cluster centre only. Hence, it is possible to extract from \mathcal{O}_k ($k \geq 1$) a set of points \mathcal{Z}_k forming an HPP with intensity λ_k . Note that in the domain \mathcal{O}_0 we have $\lambda_0 = \Phi$. Distances d_k , for $k \geq 2$, are then assigned in such a way that $\lambda_k = 2^{k-1} \lambda_1$, i.e., the intensities of the nested transport infrastructures form a geometric progression. This requires to set $d_k = d_1 2^{-\frac{k-1}{\delta}}$. Since the maximum node density in the network is $\bar{\Phi} < G_2 q \log m$ (the constant G_2 is specified in Lemma 4), we have $K_{\max} = 1 + \lceil \log_2(q \log m / \lambda_1) \rceil = O(\log n)$, hence the total number of domains satisfies condition iii).

It remains to show that each domain $k < K_{\max}$ can sustain the traffic generated by domain $k+1$. To this purpose, we need to show that each region \mathcal{I}_k^j can handle the traffic produced by all components of domain $k+1$ nested in it. Let \mathcal{H}_{k+1}^j be the set of indexes h of regions \mathcal{I}_{k+1}^h falling in \mathcal{I}_k^j . Moreover, let M_k^j be the number of cluster centres falling within \mathcal{I}_k^j .

The area of \mathcal{I}_k^j can be expressed as $|\mathcal{I}_k^j| = M_k^j \pi d_k^2 \zeta_k(\mathcal{I}_k^j)$, where $\zeta_k(\mathcal{I}_k^j) < 1$ is a reduction factor that accounts for the overlapping among the discs of radius d_k forming region \mathcal{I}_k^j . The sum of the areas of all nested regions \mathcal{I}_{k+1}^h is instead given by $\sum_{h \in \mathcal{H}_k^j} |\mathcal{I}_{k+1}^h| = M_k^j \pi d_{k+1}^2 \zeta_{k+1}(\mathcal{I}_k^j)$, where, $\zeta_{k+1}(\mathcal{I}_k^j)$ is the degree of overlapping among the discs of radius d_{k+1} falling within \mathcal{I}_k^j .

Observe that $\zeta_{k+1}(\mathcal{I}_k^j) > \zeta_k(\mathcal{I}_k^j)$ because the degree of overlapping among the discs reduces for decreasing values of d_k . Since $(d_k/d_{k+1})^2 = 2^{2/\delta}$, we conclude that the ratio between $|\mathcal{I}_k^j|$ and $\sum_{h \in \mathcal{H}_k^j} |\mathcal{I}_{k+1}^h|$ is bounded. This is important, because it allows to exploit the full capacity of the infrastructure extracted in \mathcal{I}_k^j to evenly spread out the traffic coming from nested regions \mathcal{I}_{k+1}^h over the larger region \mathcal{I}_k^j .

Moreover, using the results on homogeneous networks reported in (19), it can be shown that the aggregate throughput of regions \mathcal{I}_{k+1}^h , with $h \in \mathcal{H}_{k+1}^j$, is larger than the throughput of region \mathcal{I}_k^j . This is true for all k, j , and it allows to conclude that domain 0 (i.e., the main infrastructure) acts as the system bottleneck.

To show this fact, we need to use the expressions (19) for the capacity of a generic sub-system in

which nodes are distributed according to an HPP process. We can discard the second case in (19), since the hierarchical access scheme is not needed when $\alpha < 3$.

Now, let us suppose that the throughput of region \mathcal{I}_k^j (and that of regions \mathcal{I}_{k+1}^h nested in \mathcal{I}_k^j) is given by the first case of (19), in which $T_u = \omega(\bar{N}^{1-\epsilon})$. The number of points in \mathcal{I}_k^j is

$$M_k^j \pi \lambda_k d_k^2 \zeta_k(\mathcal{I}_k^j) = M_k^j \pi \lambda_1 d_1 2^{(k-1)(1-\frac{2}{\delta})} \zeta_k(\mathcal{I}_k^j)$$

recalling that $\zeta_{k+1}(\mathcal{I}_k^j) > \zeta_k(\mathcal{I}_k^j)$ the total number of points in regions \mathcal{I}_{k+1}^h , with $h \in \mathcal{H}_{k+1}^j$, can be lower bounded by

$$M_k^j \pi \lambda_{k+1} d_{k+1}^2 \zeta_k(\mathcal{I}_k^j) = M_k^j \pi \lambda_1 d_1 2^{k(1-\frac{2}{\delta})} \zeta_k(\mathcal{I}_k^j)$$

Since $\delta > 2$, the total number of points in regions \mathcal{I}_{k+1}^h , with $h \in \mathcal{H}_{k+1}^j$, is larger than the number of points in \mathcal{I}_k^j . This guarantees that, in the first case of (19), the aggregate throughput of the infrastructures nested in \mathcal{I}_k^j is higher than the throughput of region \mathcal{I}_k^j , $\forall k, j$.

In the third case of (19), the throughput (either of region \mathcal{I}_{k+1}^h or the aggregate throughput of regions \mathcal{I}_{k+1}^h nested in \mathcal{I}_k^j) would be proportional to $2^{k[\frac{\alpha-1}{\alpha-2}-\frac{1}{\delta}-\epsilon(1-\frac{2}{\delta})]}$. Since $\frac{\alpha-1}{\alpha-2} > 1 > \frac{1}{\delta}$, and ϵ is small, the throughput increases with k . At last, in the fourth case of (19) the throughput would be proportional to $2^{k[\frac{\alpha+1}{2}-\frac{1}{\delta}-\epsilon(1-\frac{2}{\delta})]}$ which again increases with k .

We conclude that the chosen sequence of nested local infrastructures satisfies the conditions that allow to balance the traffic towards the nodes of the main infrastructure at most with a $\log n$ penalty factor to the system throughput.

The precise details about how the scheduling/routing scheme works during the *access* (or *delivery*) phase are reported in Appendix E.

2) *Performance of the schemes employing the **HI** main infrastructure under the cluster-sparse condition:* The hierarchical multi-hop access scheme introduced in Section IV-B1 guarantees that we can always achieve *at least* the throughput of a system in which nodes are distributed according to an HPP of intensity $\underline{\Phi}$, under the only condition that the number of nodes \mathcal{Z} belonging to the main infrastructure tends to infinity (See Proposition 3). Hence we can write,

$$T(n, \alpha, \gamma, \nu, \delta) = \Omega(T_u(n, L, \underline{\Phi}, \alpha))$$

under the condition $|\mathcal{Z}| = L^2 \underline{\Phi} = \omega(1)$.

In particular, two schemes that arise in the homogeneous scenario (see Appendix C), when applied to the **HI** infrastructure of density $\underline{\Phi}$, allow to achieve the maximum network throughput, for some range of system parameters:

- **MH: traditional multi-hop**, in the case of $\alpha \geq 3$ and $\underline{\Phi} = O(1)$. This scheme provides a throughput $\omega(|\mathcal{Z}|^{-\epsilon} L \underline{\Phi}^{\frac{\alpha+1}{2}})$ (see Table II).
- **sC: cooperative sub-cluster hopping**, in the case $\alpha \geq 3$, $|\mathcal{Z}| < L^\alpha$, and $\underline{\Phi} = \omega(1)$. This scheme provides a throughput $\omega(|\mathcal{Z}|^{-\epsilon} L \underline{\Phi}^{\frac{\alpha-1}{\alpha-2}})$ (see Table II).

C. Communication schemes employing the CC main infrastructure under the cluster-sparse condition

Recall that the CC main infrastructure is formed by a subset \mathcal{V} of nodes, of cardinality $|\mathcal{V}| = \Theta(n)$, lying within a finite distance from their cluster centres. The subset \mathcal{V} is constructed as follows: let $\mathcal{O}_C = \{\xi \in \mathcal{O} : d_{\min}(\xi) \leq C\}$ be the set of points whose distance from the closest cluster centre is smaller than a finite constant C . Since C is finite, while the typical distance between cluster centres tends to infinity as $d_C = n^{\gamma-\nu/2}$, domain \mathcal{O}_C is formed w.h.p. by M disjoint discs of radius C . Within each disc, the node density can be lower bounded by $\lambda_C = q \min(1, C^{-\delta})$, hence from \mathcal{O}_C we can extract, w.h.p., a set of points \mathcal{V} distributed according to an HPP of intensity λ_C within M disjoint discs of constant area. It follows that the cardinality of set \mathcal{V} is $|\mathcal{V}| = \Theta(Mq) = \Theta(n)$. Notice that the number of points \mathcal{V} belonging to a given cluster is a finite fraction of all nodes belonging to the same cluster.

Set \mathcal{V} can be used as the main transport infrastructure of the system by employing a cooperative multi-hop strategy similar to the one adopted in the case of homogeneous networks. More specifically, we partition the network area into square cells of edge size c , and perform MIMO communications between the nodes in \mathcal{V} belonging to neighboring cells. Global multi-hopping at the cell level is employed to transfer data throughout the network. This strategy has the advantage of employing a rich set of nodes, but it is constrained to using a cell size $c = \Omega(d_c \sqrt{\log m})$, i.e., to operate at a spatial resolution higher than the typical distance d_c between cluster centres. Indeed, if $c = \Omega(d_c \sqrt{\log m})$ we can apply Lemma 2 to the set of cluster centres and conclude that each cell contains in order sense the same number of nodes belonging to set \mathcal{V} .

The following proposition guarantees that the throughput of the above cooperative multi-hop strategy, applied over the CC main infrastructure, is the same as the throughput achievable by an analogous scheme (i.e., a cooperative multi-hop strategy with the same cell size) applied over a homogeneous network of n nodes.

Proposition 2: Adopting a cooperative multi-hop strategy with cell size $c = \Omega(d_c \sqrt{\log m})$, the CC main infrastructure can sustain the same throughput achievable by a homogeneous network of n nodes, adopting the same scheme.

Proof: The proof can be found in Appendix B. ■

Before computing the resulting throughput, we need to premise how the rest of nodes $\mathcal{Y} = \mathcal{X} \setminus \mathcal{V}$ can exchange traffic with nodes in \mathcal{V} without shifting the system bottleneck to the *access* (or *delivery*) phase.

1) **SIMO access scheme:** during the *access* phase, data generated by nodes \mathcal{Y} are transferred to nodes \mathcal{V} residing in the same cell employing a single-hop SIMO communication scheme similar to the one proposed in [4]. In [4], authors present a technique that allows nodes located in low-density areas to relay their data over densely populated areas, by exploiting the diversity gain intrinsically available in high-density regions thanks to the presence of many nodes acting as an array of receiving antennas. Indeed, the additional overhead due to cooperatively decode the received signals in dense areas can be easily sustained thanks to the large capacity intrinsically available within these dense regions.

More in detail, nodes \mathcal{Y} belonging to the same cell simultaneously transmit (without prior coordination) their data (appropriately encoded) to nodes \mathcal{V} , which later on exchange their (quantized) observed signals realizing the multi-user diversity gain. Interference coming from nearby cells can be made negligible employing again the standard TDMA technique according to which only a subset of weakly interfering cells are active at any time, in such a way that the performance of the scheme can be analyzed (in order sense) considering each cell in isolation.

The scheme in [4] requires an equal number of transmitting and receiving nodes within each cell. In our setting, in each cell the number of nodes belonging to \mathcal{Y} and the number of nodes belonging to \mathcal{V} are both, w.h.p., $\Theta(n(c/L)^2)$. Although the cardinalities of transmitters and receivers are not exactly the same, nodes in \mathcal{Y} can equally distribute their traffic to nodes in \mathcal{V} in such a way that each node in \mathcal{V} has to sustain the traffic of a bounded number of nodes \mathcal{Y} . By applying the scheme in [4] sequentially a bounded number of times, all nodes in \mathcal{Y} get an opportunity to transmit, achieving a constant fraction of the data-rate sustainable by nodes in \mathcal{V} .

Furthermore, in [4] it is assumed that transmitters \mathcal{Y} can adapt their power so that signals are received roughly at the same strength by relays \mathcal{V} . This can be achieved also in our case adopting the usual trick of partitioning each cell into a finite number of sub-cells, and assigning as receivers of a given transmitter y only the nodes \mathcal{V} residing in a suitable sub-cell, such that the minimum distance between a transmitter and its closest receiver is Kc , where K is some constant.

In [4] it is shown that the per-node rate achievable by the above SIMO access phase is

$$R_{\text{SIMO}} \geq KVPc^{-\alpha} \quad (4)$$

where V is the number of receiving nodes (in the above expression it is assumed that $VPc^{-\alpha} \leq 1$).

2) **MISO delivery scheme:** in the *delivery* phase, nodes \mathcal{V} deliver data to final destinations belonging to \mathcal{Y} , in a way similar to the *access phase*, except that in this case a single-hop MISO communication

scheme is employed, according to which the relays in \mathcal{V} perform transmit beam-forming (by exchanging information among them) and then concurrently send the (quantized and encoded) messages to all nodes of the cell. The per-node rate achievable by the above MISO broadcast phase is (see [4]),

$$R_{\text{MISO}} \geq KVPc^{-\alpha} \quad (5)$$

where V is the number of transmitting nodes (again, in the above expression it is assumed that $VPC^{-\alpha} \leq 1$).

3) *Performance of the schemes employing the CC main infrastructure under the cluster-sparse condition:* by comparing (4) (or (5)) with (21), we can see that the per-node rate achievable in the *access* or *delivery* phase is in order sense the same as the per-node rate achievable during a single-hop MIMO communication between two adjacent cells. As consequence, the performance is limited by the *transport* phase, in which traffic is sent across the network area employing cell multi-hopping.

For this phase, it remains to optimize the cell edge size c in a way analogous to the one described in Appendix C, but with the additional constraint that $c = \Omega(d_c\sqrt{\log m})$, which allows us to see the network as being uniformly populated by nodes \mathcal{V} .

This leads to the following three schemes,

- **IC: cooperative inter-cluster hopping**, when the optimal choice is to set $c = d_c\sqrt{\log m}$. This occurs when $\alpha > 3$, and either $\gamma \geq 1/2$ or jointly $1/\alpha < \gamma < 1/2$ and $2(1 - 2\gamma) < (2\gamma - \nu)(\alpha - 2)$. This scheme provides a throughput $\omega(N^{2-\epsilon}d_c^{-\alpha}/m^{3/2})$ (see Table II).
- **SC: cooperative super-cluster hopping**, when the optimal c takes an intermediate value in between d_c and L . This occurs when $\alpha > 3$, $1/\alpha < \gamma < 1/2$ and $2(1 - 2\gamma) \geq (2\gamma - \nu)(\alpha - 2)$. Then the optimal choice is to set $c = n^{\frac{1-2\gamma}{\alpha-2}}$, which allows to achieve a throughput $\omega(N^{\frac{\alpha-1}{\alpha-2}-\epsilon}L^{-\frac{\alpha}{\alpha-2}})$ (see Table II).
- **GL: cooperative global**, when the optimal choice is to set $c = L$, i.e., to perform single hop MIMO communication at global scale. This occurs when $\alpha\gamma \leq 1$, providing a throughput $\omega(N^{1-\epsilon})$, or when $\alpha\gamma > 1$ and $\alpha < 3$, leading to a throughput $\omega(N^{2-\epsilon}L^{-\alpha})$ (see Table II).

Note that the above three schemes can always be applied (under the *cluster-sparse* condition), however they outperform those based on the **HI** main infrastructure (see Section IV-B2), thus providing an order-optimal scheme, only in a specific range of system parameters, as specified in Table II.

V. UPPER BOUNDS

We first describe how to upper-bound the system capacity under the *cluster-dense* condition, which can be done in a very simple way exploiting existing results for homogeneous networks [3].

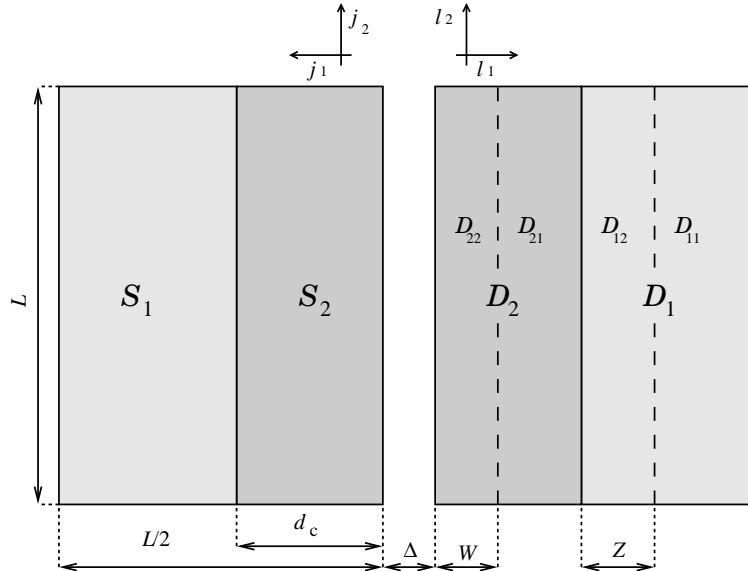


Fig. 5. Network area

Under the *cluster-dense* (i.e., for $\gamma < \nu/2$) condition, we can complete (w.h.p.) \mathcal{X} to a superset of nodes \mathcal{W} distributed according to an HPP process of intensity $\bar{\Phi} = \bar{\Phi}_0 = G_1 \frac{n}{L^2}$, where G_1 is the constant defined in Lemma 7. We can then upper-bound the capacity of the clustered network formed by nodes \mathcal{X} with the capacity of the homogeneous network formed by nodes \mathcal{W} . Indeed, the presence of additional nodes that can be (possibly) used as intermediate relays can only improve the resulting network capacity. Notice that since the cardinality of \mathcal{W} is $\Theta(n)$, the resulting upper bounds to the scaling exponent of the system capacity (derived in [3]) are tight, i.e., the proposed simple technique to upper bound the system capacity does not introduce any gap in order sense.

Under the *cluster-sparse* condition, the derivation of upper-bounds to the system capacity is more involved. We extend the approach in [2], [3], [4], which is based on the computation of an upper bound to the information flow passing through a cut that divides the network in two parts.

First, by leveraging percolative arguments (see [6]), we find a strip of width Δ , with $\Delta = \Theta((qs(d_c) \log n)^{-1/2})$ which divides the network area in two parts, and satisfies the following properties: i) the considered strip does not include any node; ii) every cluster centre lies at a distance greater than gd_c from the strip, for a sufficiently small constant g . Then, we bound the information flow $\mathcal{C}(\mathcal{S}, \mathcal{D})$ from sources \mathcal{S} on the left of the strip to destinations \mathcal{D} on the right of the strip with the power transfer $\mathcal{P}_{\mathcal{S}, \mathcal{D}}$ through the strip, as in [2], [3], [4].

To do so, we proceed as follows. We isolate the contribution of the nodes that are located close to the

strip, more specifically the nodes located in the low density region of width $\Theta(d_c)$, from the rest of the network (see Figure 5). We denote by \mathcal{S}_1 and \mathcal{S}_2 the high and low density regions, respectively, which are located on the left side of the strip, and with \mathcal{D}_1 and \mathcal{D}_2 the corresponding regions on the right side. Then, we first focus on the case $\alpha \leq 3$. By exploiting the methodology developed in [2], [3], [4], we can write

$$\begin{aligned}
C(\mathcal{S}, \mathcal{D}) &= \max_{\substack{\mathbf{Q}(\mathbf{H}) \geq 0 \\ \mathbb{E}[\mathbf{Q}_{kk}(\mathbf{H})] \leq P, \forall k \in \mathcal{S}}} \mathbb{E} \left[\log \det \left(\mathbf{I} + \mathbf{H}\mathbf{Q}(\mathbf{H})\mathbf{H}^H \right) \right] \\
&\leq \max_{\substack{\mathbf{Q}(\mathbf{H}) \geq 0 \\ \mathbb{E}[\mathbf{Q}_{kk}(\mathbf{H})] \leq P, \forall k \in \mathcal{S}}} \mathbb{E} \left[\text{tr} \{ \mathbf{H}\mathbf{Q}(\mathbf{H})\mathbf{H}^H \} \right] \\
&\leq bP_{\mathcal{S}_1, \mathcal{D}_1} + bP_{\mathcal{S}_1, \mathcal{D}_2} + bP_{\mathcal{S}_2, \mathcal{D}_1} + bP_{\mathcal{S}_2, \mathcal{D}_2}
\end{aligned} \tag{6}$$

where

$$b > 4 \max \left(1, \max_{k \in \mathcal{D}} \sum_{i \in \mathcal{S}} \frac{|h_{ik}|^2}{\sum_{h \in \mathcal{D}} d_{ih}^{-\alpha}} \right).$$

By applying similar arguments to those presented in [4], we can prove⁶ that $\max_{k \in \mathcal{D}} \sum_{i \in \mathcal{S}} \frac{|h_{ik}|^2}{\sum_{h \in \mathcal{D}} d_{ih}^{-\alpha}} \leq \log^3 n$ for $\alpha = 2$ and $\max_{k \in \mathcal{D}} \sum_{i \in \mathcal{S}} \frac{|h_{ik}|^2}{\sum_{h \in \mathcal{D}} d_{ih}^{-\alpha}} \leq \log^{1+\frac{\alpha}{2}} n$ for $\alpha > 2$.

In order to compute the term $P_{\mathcal{S}_1, \mathcal{D}_1}$, we divide the regions \mathcal{S}_1 and \mathcal{D}_1 in squarelets of side d_c . Indeed, the minimal distance between a source in \mathcal{S}_1 and a destination in \mathcal{D}_1 is $\Theta(d_c)$. The j -th squarelet in \mathcal{S}_1 is denoted by $\mathcal{S}_{1,j}$ and has coordinates $\mathbf{j} = [j_1, j_2]$. Likewise the ℓ -th squarelet in \mathcal{D}_1 is denoted by $\mathcal{D}_{1,\ell}$ and has coordinates $\boldsymbol{\ell} = [\ell_1, \ell_2]$. Thus, we have

$$\begin{aligned}
P_{\mathcal{S}_1, \mathcal{D}_1} &\leq \sum_{\mathbf{j}} \sum_{\boldsymbol{\ell}} \underline{d}_{\mathbf{j}\boldsymbol{\ell}}^{-\alpha} U(d_c)^2 \\
&= \sum_{\mathbf{j}} \sum_{\boldsymbol{\ell}} \left((j_1 d_c + \ell_1 d_c + K d_c)^2 + (j_2 d_c - \ell_2 d_c)^2 \right)^{-\alpha/2} \left(\frac{n}{m} \right)^2 \log^2 n \\
&= d_c^{-\alpha} \left(\frac{n}{m} \right)^2 \log^2 n \sum_{\mathbf{j}} \sum_{\boldsymbol{\ell}} \left((j_1 + \ell_1 + K)^2 + (j_2 - \ell_2)^2 \right)^{-\alpha/2}
\end{aligned} \tag{7}$$

for $0 \leq j_1, \ell_1 \leq L/d_c$ and $-L/2d_c \leq j_2, \ell_2 \leq L/2d_c$. We define $\underline{d}_{\mathbf{j}\boldsymbol{\ell}}$ as the minimum distance between squarelets $\mathcal{S}_{1,j}$ and $\mathcal{D}_{1,\ell}$, which is given by

$$\underline{d}_{\mathbf{j}\boldsymbol{\ell}} = d_c \sqrt{(j_1 + \ell_1 + K)^2 + (j_2 - \ell_2)^2}$$

⁶We consider the source nodes located in the regions \mathcal{S}_1 and \mathcal{S}_2 , separately, and compute the term at the denominator in the two cases by dividing the region \mathcal{D} into squarelets of side $d_c \log n$ and $\Delta \log n$, respectively. By doing so, the average node density in each type of squarelet results to be $\Theta(\frac{n}{L^2})$ and $\Theta(\frac{1}{\Delta^2})$, respectively. We then divide \mathcal{S}_1 and \mathcal{S}_2 into squarelets of side d_c and Δ , respectively, and upper bound the number of nodes in each type of squarelet by using the Corollaries 1 and 2 in Appendix F.

where Kd_c is a lower bound for the width of regions \mathcal{S}_2 and \mathcal{D}_2 , as well as of the strip. By using the results in Appendix F-A with $N = 2M$ and $M = L/2d_c$, we obtain

$$P_{\mathcal{S}_1, \mathcal{D}_1} \leq \begin{cases} Kd_c^{-2} \left(\frac{L}{d_c}\right)^2 \left(\frac{n}{m}\right)^2 \log^2 n \log\left(\frac{L}{d_c}\right) & \alpha = 2 \\ Kd_c^{-\alpha} \left(\frac{L}{d_c}\right)^{4-\alpha} \left(\frac{n}{m}\right)^2 \log^2 n & 2 < \alpha < 3 \\ Kd_c^{-3} \frac{L}{d_c} \left(\frac{n}{m}\right)^2 \log^2 n \log\left(\frac{L}{d_c}\right) & \alpha = 3. \end{cases} \quad (8)$$

Likewise, by dividing the regions \mathcal{S}_2 and \mathcal{D}_1 in squarelets of side d_c we can upper bound $P_{\mathcal{S}_2, \mathcal{D}_1}$ as follows (see Appendix F):

$$P_{\mathcal{S}_2, \mathcal{D}_1} \leq \begin{cases} Kd_c^{-3} \left(\frac{n}{m}\right)^2 L \log^3 n & \alpha = 2 \\ Kd_c^{-1-\alpha} \left(\frac{n}{m}\right)^2 L \log^2 n & 2 < \alpha \leq 3. \end{cases} \quad (9)$$

An equivalent result can be obtained in order sense for the term $P_{\mathcal{S}_1, \mathcal{D}_2}$.

Next, we evaluate the contribution of the term $P_{\mathcal{S}_2, \mathcal{D}_2}$. To do so, we partition the regions \mathcal{S}_2 and \mathcal{D}_2 into squarelets of side Δ . Similarly to what has been done before, we obtain (see Appendix F):

$$P_{\mathcal{S}_2, \mathcal{D}_2} \leq \begin{cases} K\Delta^{-4} L d_c \log^3 n & \alpha = 2 \\ K\Delta^{-4} L d_c^{3-\alpha} \log^2 n & 2 < \alpha < 3 \\ K\Delta^{-4} L \log^3 n & \alpha = 3. \end{cases} \quad (10)$$

Observe that, for $\alpha \leq 3$, the terms in (9) and (10) are negligible with respect to $P_{\mathcal{S}_1, \mathcal{D}_1}$. From (8), it follows that the scaling exponent e_C of the network capacity is given by

$$2 - \alpha\gamma.$$

Next, we focus on $\alpha > 3$. In order to obtain tight bounds, we further divide \mathcal{D}_1 into \mathcal{D}_{11} and \mathcal{D}_{12} , and \mathcal{D}_2 into \mathcal{D}_{21} and \mathcal{D}_{22} (see Figure 5). We have that:

- \mathcal{D}_{22} is the sub-region of \mathcal{D}_2 of width $W = n^w$ and close to the cut; we constrain W to be smaller than or equal to the width of \mathcal{D}_2 , i.e., $w \leq \gamma - \nu/2$;
- \mathcal{D}_{12} is the sub-region of \mathcal{D}_1 of width $Z = n^z$, with $\gamma - \frac{\nu}{2} \leq z < \gamma$.

Let us consider the channel matrix \mathbf{X} connecting the subset of source nodes \mathcal{X}_s to the destination nodes \mathcal{X}_d . We define $f(\mathbf{X}) = \max_{\substack{\mathbf{Q}(\mathbf{X}) \geq 0 \\ \mathbb{E}[\mathbf{Q}_{kk}(\mathbf{X})] \leq P, \forall k \in \mathcal{X}_s}} \mathbb{E}[\log \det(\mathbf{I} + \mathbf{X}\mathbf{Q}(\mathbf{X})\mathbf{X}^H)]$.

We upper bound the network capacity as

$$C(\mathcal{S}, \mathcal{D}) \leq f(\mathbf{H}_1) + f(\mathbf{H}_2) + f(\mathbf{H}_3) + f(\mathbf{H}_4) + f(\mathbf{H}_5)$$

where \mathbf{H}_1 connects the sources in \mathcal{S} with the destinations in \mathcal{D}_{11} , \mathbf{H}_2 connects the sources in \mathcal{S}_2 with the destinations in \mathcal{D}_{21} , \mathbf{H}_3 connects the nodes in \mathcal{S}_1 with those in \mathcal{D}_{21} , while \mathbf{H}_4 and \mathbf{H}_5 connect the

sources in \mathcal{S} with the destinations in \mathcal{D}_{12} and \mathcal{D}_{22} , respectively. As done in (6), for the first two terms we can write

$$f(\mathbf{H}_1) + f(\mathbf{H}_2) \leq bP_{\mathcal{S}_1, \mathcal{D}_{11}} + bP_{\mathcal{S}_2, \mathcal{D}_{11}} + bP_{\mathcal{S}_2, \mathcal{D}_{21}}. \quad (11)$$

We compute terms $P_{\mathcal{S}_1, \mathcal{D}_{11}}$ and $P_{\mathcal{S}_2, \mathcal{D}_{21}}$ following the same procedure described above but dividing the regions in squarelets of side Z and W respectively. As for $P_{\mathcal{S}_2, \mathcal{D}_{11}}$ we divide \mathcal{S}_2 into rectangles of area $d_c \times Z$ and \mathcal{D}_{11} into squarelets of side Z . Then we obtain

$$P_{\mathcal{S}_1, \mathcal{D}_{11}} \leq K d_c^{-4} Z^{3-\alpha} L \left(\frac{n}{m}\right)^2 \log^2 n = K n^{2-3\gamma-z(\alpha-3)} \log^2 n \quad (12)$$

$$P_{\mathcal{S}_2, \mathcal{D}_{11}} \leq K d_c^{-3} Z^{2-\alpha} L \left(\frac{n}{m}\right)^2 \log^2 n = K n^{2-2\gamma-z(\alpha-2)-\nu/2} \log^2 n \quad (13)$$

$$P_{\mathcal{S}_2, \mathcal{D}_{21}} \leq K \Delta^{-4} W^{3-\alpha} L \log^2 n = K n^{2\beta+\gamma-w(\alpha-3)} \log^2 n \quad (14)$$

For the remaining terms, we repeatedly apply the Hadamard inequality and use the following result:

$$\begin{aligned} f(\mathbf{X}) &\leq \sum_{i \in \mathcal{X}_d} \max_{\substack{\mathbf{Q}(\mathbf{x}_i) \geq 0 \\ \mathbb{E}[\mathbf{Q}_{kk}(\mathbf{x}_i)] \leq P, \forall k \in \mathcal{X}_s}} \mathbb{E} \left[\log \left(1 + \mathbf{x}_i \mathbf{Q}(\mathbf{x}_i) \mathbf{x}_i^H \right) \right] \\ &\leq \sum_{i \in \mathcal{X}_s} \log(1 + K P n \underline{d}_x^{-\alpha}) \\ &\leq K |\mathcal{X}_s| \log n. \end{aligned} \quad (15)$$

where \mathbf{x}_i is the i -th row of \mathbf{X} and \underline{d}_x is a lower-bound to the distance between source and destination nodes. It follows that

$$f(\mathbf{H}_3) \leq K |\mathcal{D}_{21}| \log n \leq K \frac{L}{d_c} \frac{n}{m} \log^2 n = K n^{1-\nu/2} \log^2 n \quad (16)$$

$$f(\mathbf{H}_4) \leq K |\mathcal{D}_{12}| \log n \leq K \frac{LZ}{d_c^2} \frac{n}{m} \log^2 n = K n^{1-\gamma+z} \log^2 n \quad (17)$$

$$f(\mathbf{H}_5) \leq K |\mathcal{D}_{22}| \log n \leq K \frac{LW}{\Delta^2} \log^2 n = K n^{\gamma+\beta+w} \log^3 n. \quad (18)$$

In order to find the scaling exponent of the network capacity, we have to identify the dominant term among the above expressions. We first observe that (12) always dominates (13), and that (17) always dominates (16) since by definition $z > \gamma - \nu/2$.

We then consider equations (12) and (17), which depend on the parameter z , and note that the former decreases while the latter increases as z grows. To obtain a tight bound we need to determine the value of z^* such that $2-3\gamma-z^*(\alpha-3) = 1-\gamma+z^*$, i.e., $z^* = (1-2\gamma)/(\alpha-2)$. Note that since $z > \gamma - \nu/2 > 0$ (cluster sparse regime), then we must have $\gamma < 1/2$.

For $z^* > \gamma - \nu/2$ the scaling exponent becomes $e_1 = 1 + \gamma - z^* = (\alpha - 1 - \alpha\gamma)/(\alpha - 2)$. Otherwise, the region \mathcal{D}_{12} can be neglected ($f(\mathbf{H}_4) = |\mathcal{D}_{12}| = 0$) and the bound is obtained by replacing $z = \gamma - \nu/2$

in (12). In conclusion, we have

$$e_1 = \begin{cases} \frac{\alpha-1-\alpha\gamma}{\alpha-2} & \text{if } z^* > \gamma - \nu/2 \wedge \gamma < \frac{1}{2} \\ 2 - \alpha\gamma + (\alpha-3)\nu/2 & \text{otherwise.} \end{cases}$$

Next we consider equations (14) and (18) which depend on w and note that the former decreases while the latter increases as w grows. We determine the value of w^* such that the bound is as tight as possible. To do so, we solve the equality $2\beta + \gamma - w^*(\alpha-3) = \gamma + \beta + w^*$ in w^* and we obtain $w^* = \beta/(\alpha-2)$.

As for the value of β , we consider the following two cases:

- If $\beta \leq 0$, then we have $w^* < 0$. This is not acceptable because by hypothesis $w \geq -\beta/2$. In this case a tighter bound is then obtained by considering the region \mathcal{D}_{22} as negligible ($f(\mathbf{H}_5) = |\mathcal{D}_{22}| = 0$) and substituting $w = -\beta/2$ in (14) thus obtaining $e_2 = \gamma + \beta(\alpha+1)/2$;
- If $\beta \geq 0$, the scaling exponent is given by $e_2 = \gamma + \beta + w^* = \gamma + \beta\frac{\alpha-1}{\alpha-2}$.

In conclusion,

$$e_2 = \begin{cases} \gamma + \frac{\beta(\alpha+1)}{2} & \beta < 0 \\ \gamma + \beta\frac{\alpha-1}{\alpha-2} & \beta \geq 0. \end{cases}$$

By comparing e_1 and e_2 , it can be easily verified that for $z^* > \gamma - \nu/2$ and $\gamma < 1/2$ the dominant scaling exponent is $e_C = (\alpha-1-\alpha\gamma)/(\alpha-2)$; otherwise, the network capacity scaling exponent is given by

$$e_c = \begin{cases} \max \left\{ 2 - \alpha\gamma + (\alpha-3)\nu/2, \gamma + \frac{\beta(\alpha+1)}{2} \right\} & \beta < 0 \\ \max \left\{ 2 - \alpha\gamma + (\alpha-3)\nu/2, \gamma + \beta\frac{\alpha-1}{\alpha-2} \right\} & \beta \geq 0 \end{cases}$$

With reference to the above expressions of the scaling exponent and to the different operational regions (see Table II), we can therefore conclude that the dominant contribution to the power transfer from the sources to the destinations regions is due to transmissions between nodes at distance:

- $\Theta(L)$ from the cut, in regions I and II;
- jointly $\omega(d_c\sqrt{\log n})$ and $o(L)$, in region III;
- jointly $O(d_c\sqrt{\log n})$ and $\Omega(d_c)$, in region IV;
- jointly $o(d_c)$ and $\omega(1/\sqrt{\Phi})$, in region V;
- $\Theta(1/\sqrt{\Phi})$, in region VI.

VI. CONCLUSIONS

We have characterized the asymptotic capacity of networks whose nodes are distributed according to a doubly stochastic shot-noise Cox process. This point process provides an interesting, analytically tractable model of clustered random networks containing large inhomogeneities in the node density. Our study has

revealed the existence of additional operational regions with respect to those identified in previous work, and the need of novel scheduling and routing strategies, specifically tailored to clustered networks, to approach the system capacity.

REFERENCES

- [1] P. Gupta, P.R. Kumar, “The capacity of wireless networks”, *IEEE Trans. on Inf. Theory*, vol. 46(2), pp. 388–404, March 2000.
- [2] A. Ozgur, O. Leveque, D. Tse, “Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks”, *IEEE Trans. on Inf. Theory*, vol. 53(10), pp. 3549–3572, Oct. 2007.
- [3] A. Ozgur, R. Johari, D. Tse, O. Leveque, “Information theoretic operating regimes of large wireless networks”, *IEEE Transactions on Information Theory*, Vol. 56 n 1, pp. 427 - 437, Jan. 2010.
- [4] U. Niesen, P. Gupta, D. Shah, “On capacity scaling in arbitrary wireless networks”, *IEEE Trans. on Inf. Theory*, vol. 55(9), Sept. 2009.
- [5] U. Niesen, P. Gupta, D. Tse, “On the optimality of multi-hop communication in large wireless networks”, *IEEE ISIT 2010*, Austin, TX, June 2010.
- [6] G. Alfano, M. Garetto, E. Leonardi, “Capacity scaling of wireless networks with inhomogeneous node density: Upper bounds”, *IEEE JSAC*, 27(7), pp. 1147–1157, Sept. 2009.
- [7] G. Alfano, M. Garetto, E. Leonardi, “Capacity scaling of wireless networks with inhomogeneous node density: Lower bounds”, *IEEE Infocom 2009*, Rio de Janeiro, Brazil, Apr. 2009.
- [8] M. Garetto, A. Nardio, C.-F. Chiasserini, E. Leonardi, “Information-theoretic capacity of clustered random networks”, *IEEE ISIT 2010*, Austin, TX, June 2010.
- [9] J. Møller, “Shot noise Cox processes”, *Adv. Appl. Prob.*, vol. 35, pp. 614–640, 2003.
- [10] M. Franceschetti, O. Dousse, D.N.C. Tse, P. Thiran, “Closing the gap in the capacity of random wireless networks via percolation theory”, *IEEE Trans. on Inf. Theory*, vol. 53(3), pp. 1009–1018, Mar. 2007.
- [11] M. Franceschetti, M.D. Migliore, P. Minero, “The capacity of wireless networks: information-theoretic and physical limits”, *IEEE Trans. on Inf. Theory*, 55(8), pp. 3413–3424, Aug. 2009.
- [12] M. Franceschetti, M. D. Migliore, P. Minero, F. Schettino, “The degrees of freedom of wireless networks via cut-set integrals”, *IEEE Trans. on Inf. Theory*, to appear, 2011.
- [13] A. Ozgur, O. Leveque, D. Tse, “Linear capacity scaling in wireless networks: Beyond physical limits?”, *5th Workshop on Information Theory and Applications*, San Diego, CA, Feb. 2010.
- [14] T. Weller, B. Hajek, “Scheduling nonuniform traffic in a packet-switching system with small propagation delay”, *IEEE/ACM Trans. on Networking*, vol. 5(6), pp. 813–823, 1997.
- [15] D. Stoyan, W. Kendall, J. Mecke, “Stochastic geometry and its applications,” Wiley, 1995.
- [16] R. Motwani, P. Raghavan, *Randomized algorithms*, Cambridge University Press, 1995.
- [17] J. Quintanilla, S. Torquato, R.M. Ziff, “Efficient measurement of the percolation threshold for fully penetrable discs”, *J. Phys. A: Math. General*, vol. 33, pp. 399–407, 2000.
- [18] R. Meester, R. Roy, *Continuum percolation*, Cambridge University Press, 1996.

APPENDIX A
PROOF OF LEMMA 1

Proof: First, we provide a construction that allows each node in \mathcal{Y} to select one of the closest node belonging to \mathcal{Z} as its first-hop relay. We partition \mathcal{O} into squarelets of area $16 \log n / \underline{\Phi} \leq A \leq 32 \log n / \underline{\Phi}$ (the exact dimension of the squarelets is chosen so as to exactly cover \mathcal{O} with an integer number of squarelets, see Lemma 2 for details). According to Lemma 2 (in Appendix D), there are w.h.p. at least $\lfloor \underline{\Phi} A / 2 \rfloor$ nodes belonging to \mathcal{Z} , while the number of points of \mathcal{X} in each squarelet is upper-bounded w.h.p. by $\lceil 2A \overline{\Phi} \rceil$. Indeed \mathcal{X} can be completed to \mathcal{W} , an HPP with intensity $\overline{\Phi}$, and Lemma 2 can be applied to bound the number of points of \mathcal{W} in every squarelet. Moreover, observe that since by construction $\mathcal{Y} \subseteq \mathcal{X} \subset \mathcal{W}$, any upper bound on the number of points \mathcal{W} falling in a squarelet can be regarded as an upper bound on the number of points \mathcal{X} (or \mathcal{Y}) falling in the same squarelet.

Then, we uniformly partition the nodes of \mathcal{Y} falling in each squarelet into $\lfloor \underline{\Phi} A / 2 \rfloor$ groups, assigning each group to a different node of \mathcal{Z} belonging to the same squarelet. By construction, each group contains at most $\lceil 4 \overline{\Phi} / \underline{\Phi} \rceil = \Theta(1)$ nodes. Lying both transmitter and receiver in the same squarelet, their relative distance d_T is upper bounded by the diagonal of the squarelet: $d_T \leq \sqrt{2A}$, being $\sqrt{A} = \Theta(L \sqrt{\log n / n}) \leq \sqrt{\log n}$ (notice that under the *cluster-dense* condition $\gamma < \nu/2 < 1/2$).

To mitigate interference, we employ a standard technique (see [10]) according to which transmissions occurring within the same squarelet, as well as transmissions occurring in squarelets containing points closer than $3\sqrt{2A}$, are orthogonalized over time. To this aim, the squarelets are partitioned into a finite number of subsets, each subset comprising regularly spaced, weakly-interfering squarelets. At any time only one subset of squarelets is activated, and at most one transmission is enabled in each activated squarelet. As a result, each transmission achieves a rate $\Omega(\log(1 + d_T^{-\alpha}))$. Recall that we can devote to the *access* phase a finite fraction of time, without affecting (in order sense) the throughput achievable during the *transport* phase. Considering the fact that in each squarelet there are $O(\log n)$ competing nodes belonging to \mathcal{Y} , the fraction of time in which each of them can transmit scales as $\Omega(1/\log n)$, the achievable rate by each node of \mathcal{Y} is then $\Omega(1/\log^{1+\alpha/2} n)$.

During the *delivery* phase, data are sent from nodes in \mathcal{Z} to their final destinations belonging to \mathcal{Y} , again using a point-to-point single-hop transmission. As already mentioned, the analysis of this phase is identical to that of the *access* phase, exchanging the role of transmitters and receivers. ■

APPENDIX B
PROOF OF PROPOSITION 2

Proof: First notice that the performance of a MIMO communication established between neighboring cells does not depend on how nodes are distributed within a cell. The difficulty then lies in showing that, within a cell, the overhead required to set up cooperative transmission and cooperative decoding does not constitute a bottleneck with respect to the throughput achievable by MIMO communication between cells. This can be done in an easy way by devising a hierarchical scheme in which the size (in terms of number of nodes) of the clusters to be used at the basic layer of the hierarchy exactly matches the size of the discs forming the CC main infrastructure. Indeed, we can already start our hierarchy construction by exploiting the capacity available in a disc, which can be regarded as a sub-system of $q = n^{1-\nu}$ nodes uniformly distributed. Since the disc radius C is finite, in each disc we can achieve almost an aggregate throughput $\omega(q^{1-\epsilon'})$, for any $\epsilon' > 0$.

For the next layer, we take squarelets of edge length $c' = d_c \sqrt{\log m}$, containing $\Theta(\log m)$ discs each. Performing the basic step of the hierarchical construction described in [2], we obtain that each squarelet now forms a sub-system of size $\Theta(q \log m)$ with aggregate throughput $\omega(q^{1-\epsilon'} \log m \log(1 + qPc'^{-\alpha}))$, which is non smaller than the throughput achievable by a sub-system of the same size in a homogeneous network employing a similar hierarchical scheme. The above sub-system can then be used to construct bigger systems (until we reach squarelets of size c) with the guarantee that every intermediate throughput is at least equal to those achievable in a homogeneous network, following exactly the same approach as in [2]. By construction, the throughput of every level is always non smaller than the throughput of the corresponding level (i.e. the level employing the same cluster size) in a homogeneous network. As a conclusion, all phases needed to set up transmit and decoding cooperation do not throttle the throughput provided by the highest layer of the hierarchy, in which we perform MIMO transmissions between cells of size c . ■

APPENDIX C
COMMUNICATION SCHEMES HOMOGENEOUS NETWORKS

Here, we present a slight generalization of the results established in [2], [3], in a convenient form that will allow us to reuse them in the derivation of our lower bounds.

Consider a sub-area of the region \mathcal{O} shaped as a square (or disc) of edge (radius) $E = \Theta(n^\ell)$, where nodes are distributed according to a homogeneous Poisson process of intensity $\psi = \Theta(n^\kappa)$. This process generates on average $\bar{N} = \psi E^2 = \Theta(n^{\kappa+2\ell})$ nodes within the considered sub-area. We do not impose

here any specific restriction on the real numbers κ and ℓ , except that we require the average number of nodes \bar{N} in the system to go to infinity, which implies $\kappa + 2\ell > 0$. In [3] authors consider the special case $\kappa = 1 - 2\ell$, since they derive results for the entire network area containing $\Theta(n)$ nodes⁷. The power loss exponent is $\alpha > 2$.

Given a communication strategy, let $T_u(n, E, \psi, \alpha)$ be the aggregate throughput of the above system, i.e., the maximum amount of data that can be transported network-wide across the area. The following proposition characterizes the behavior of $T_u(n, E, \psi, \alpha)$. To simplify the expressions, we write them as functions also of $\bar{N} = \psi E^2$.

Proposition 3: For any given combination of system parameters, such that $\kappa + 2\ell > 0$, there exists a communication strategy that allows to achieve the aggregate throughput:

$$T_u(n, E, \psi, \alpha) = \begin{cases} \omega(\bar{N}^{1-\epsilon}) & \bar{N} \geq E^\alpha \\ \omega(\bar{N}^{2-\epsilon} E^{-\alpha}) & \bar{N} < E^\alpha, \alpha < 3 \\ \omega(\bar{N}^{-\epsilon} E \psi^{\frac{\alpha-1}{\alpha-2}}) & \bar{N} < E^\alpha, \alpha \geq 3, \psi = \omega(1) \\ \omega(\bar{N}^{-\epsilon} E \psi^{\frac{\alpha+1}{2}}) & \alpha \geq 3, \psi = O(1) \end{cases} \quad (19)$$

w.h.p. for any $\epsilon > 0$. Moreover, the associated scaling exponents

$$e_{T_u}(\alpha, \kappa, \ell) = \begin{cases} 2\ell + \kappa & \kappa + 2\ell \geq \alpha\ell \\ \ell(4 - \alpha) + 2\kappa & \kappa + 2\ell < \alpha\ell, \alpha < 3 \\ \ell + \kappa \frac{\alpha-1}{\alpha-2} & \kappa + 2\ell < \alpha\ell, \alpha \geq 3, \kappa > 0 \\ \ell + \kappa \frac{\alpha+1}{2} & \alpha \geq 3, \kappa \leq 0 \end{cases} \quad (20)$$

exactly match the scaling exponent of the upper bounds to the network capacity, hence they precisely characterize how the aggregate capacity scales with n if we neglect poly-log terms.

Proposition 3 is a straightforward consequence of the main theorem in [3], so we do not provide a detailed proof here. Instead, to help the reader grasp the basic facts leading to the expressions in (19) and (20), we provide an intuitive explanation of the scheduling-routing schemes that allow to achieve (in order sense) the maximum capacity in the various regimes that arise while varying the system parameters. Such intuitive description will be especially useful to understand the construction of the novel schemes proposed for our network topologies.

A general class of scheduling-routing strategies, which includes as special cases the schemes leading to (19), is obtained partitioning the network area into cells of side $c = \Theta(n^\theta)$, and applying a cooperative

⁷We remark that here we consider a more general model with the additional parameter κ because in our analysis we need to consider many different sub-systems having variable area and number of nodes, and it turns out that it is not possible to just re-scale the system parameters and obtain general results which are function of only two parameters.

multi-hop strategy, in which multiple input multiple output (MIMO) communications are established among the nodes belonging to neighboring cells and global multi-hopping at the cell level is employed to transfer data throughout the network.

When we set c equal to the edge of the network area, we end up performing MIMO communications at global scale, without the need of cell multi-hopping. On the other extreme, we can set c so small that each cell contains on average only $\Theta(1)$ nodes (by setting $c = \Theta(n^{-\kappa/2})$). This case occurs when there is no gain in employing MIMO communications at any scale, hence the scaling exponent of the system capacity is the same as in the traditional multi-hop scheme employing point-to-point communications [1], [10].

For assigned system parameters, the adoption of the above strategy, coupled with the optimization of θ in the range $[-\kappa/2, \gamma]$ essentially provides an order optimal scheme achieving the results in (20).

The performance analysis can be intuitively understood as follows. First of all, local MIMO transmissions between neighboring cells can be parallelized with a TDMA scheme in such a way that any pair of cells is active for a finite fraction of time, while at the same time the interference due to other cell pairs which are active concurrently can be neglected. This comes from standard arguments, which are by now well understood and widely adopted in previous work (see e.g., [10]). Hence, we can restrict ourselves to computing the rate achievable by MIMO communications between two neighboring cells as considered in isolation, multiply it by the total number of cells in the network, and consider the multi-hop penalty due to (re)transmitting the same data over each hop toward the destination.

From [3], [2], the rate R_{HC} achievable by MIMO communication between two neighboring cells comprising $\Theta(z)$ nodes each, can be lower bounded as,

$$R_{\text{HC}} \geq K z^{1-\epsilon} \log \left(1 + z \frac{P}{N_0 B c^\alpha} \right) \quad (21)$$

w.h.p. for any $\epsilon > 0$ and a constant $K > 0$, where B is the transmission bandwidth and $z^{-\epsilon}$ is the loss in performance due to cooperation overhead. Notice that the term $z \frac{P}{c^\alpha}$ can be regarded as the total power arriving at a node from a set of z nodes, all located at distance c from the receiving node, and transmitting independently at full rate. When this term is $o(1)$, we are in the power-limited (or low-SNR) regime. When this term is $\Omega(1)$, we are in the bandwidth-limited (or high-SNR) regime.

In practice, individual links established between two nodes belonging to neighboring cells have variable length (arbitrarily small). However, by dividing each cell into two halves and enabling communication only between non-adjacent halves (see [3]), we can assume that the power transfer on the links of each MIMO sub-system is roughly the same, i.e., $\Theta(c^{-\alpha})$ (see [3] for the details).

From the above discussion, it follows that there are essentially three factors that can affect the scaling order achievable by the considered class of strategies:

- 1) The **power transfer** $\Theta(c^{-\alpha}) = \Theta(n^{-\theta\alpha})$ on each link established between nodes belonging to neighboring (sub)-cells. The power transfer decreases when the cell size is set larger and larger, resulting into a performance penalty for increasing values of θ .
- 2) The **diversity gain** z due to cooperative MIMO communications between neighboring cells, which grows linearly with the number of transmitters (or receivers). This gain factor is thus equal to the number of nodes in a cell, which is $z = \Theta(n^{2\theta+\kappa})$, growing with θ .
- 3) The **multi-hop penalty** factor, due to the fact that data need to be (re)transmitted on each hop towards the destination. Since any flow has to go through a number of hops $O(E/c)$, the resulting penalty factor is at most $O(n^{\ell-\theta})$, which decreases with θ .

The first two factors above determine the scaling order of the inter-cell rate R_{HC} achievable by MIMO communication (21). In the power-limited regime, the scaling order of R_{HC} is $4\theta + 2\kappa - \theta\alpha$. In the bandwidth-limited regime, the scaling order of R_{HC} is $2\theta + \kappa$, since the per-node rate can grow at most⁸ like $\log n$.

We can concisely express this result stating that the rate achievable by inter-cell MIMO communication has scaling exponent $2\theta + \kappa + \min(0, 2\theta + \kappa - \theta\alpha)$. As already said, the overall system throughput is obtained multiplying the inter-cell rate by the total number of cells in the network, and dividing it by the multi-hop penalty factor, obtaining

$$e_{T_u}(\alpha, \kappa, \ell) = \min(0, 2\theta + \kappa - \theta\alpha) + \ell + \kappa + \theta.$$

Maximizing the above quantity over $\theta \in [-\kappa/2, \gamma]$ essentially provides the scaling orders expressed in (20). In particular, θ is set equal to ℓ (which means global MIMO communication) whenever $\kappa + 2\ell \geq \alpha\ell$, or $\alpha < 3$. When $\alpha \geq 3$ and $\kappa \leq 0$, instead, θ is minimized and the optimal scheme is the traditional multi-hop scheme of Gupta-Kumar. A non-degenerate cooperative multi-hop scheme is only employed when $\kappa + 2\ell < \alpha\ell$, $\alpha \geq 3$, and $\kappa > 0$, for which the optimal θ takes an intermediate value, equal to $\kappa/(\alpha - 2)$, for which the diversity gain exactly compensates the power loss over inter-cell MIMO communications.

Notice that from (20) we can immediately derive the scaling exponent $e_\lambda(\alpha, \kappa, \ell)$ of the per-node throughput $\lambda(n, \alpha, \kappa, \ell)$. Indeed, since the per-node throughput is equal to the maximum system

⁸This limit is established in [2] by upper-bounding the rate of a node by the single-input multiple-output (SIMO) channel between the node and the rest of the network.

throughput $T_u(n, \alpha, \kappa, \ell)$ divided by the total number of nodes, $e_\lambda(\alpha, \kappa, \ell)$ is simply obtained by subtracting $2\ell + \kappa$ from $e_{T_u}(\alpha, \kappa, \ell)$.

APPENDIX D

Lemma 2: (Concentration result for HPP). Consider a set of points \mathcal{X} distributed over a bi-dimensional domain \mathcal{O} of area $|\mathcal{O}|$ according to an HPP of rate $\Phi = n/|\mathcal{O}|$. Let \mathcal{T}_k be a regular tessellation of \mathcal{O} (or any sub-region of \mathcal{O}), whose tiles T_k have a surface $|T_k|$ non smaller than $16\frac{\log n}{\Phi}$. Let $U(T_k)$ be the number of points of \mathcal{X} falling within T_k . Then, uniformly over the tessellation, $U(T_k)$ is comprised w.h.p. between $\frac{\Phi|T_k|}{2}$ and $2\Phi|T_k|$, i.e., $\frac{\Phi|T_k|}{2} < \inf_k U(T_k) \leq \sup_k U(T_k) < 2\Phi|T_k|$.

The proof of this statement follows directly from the Chernoff bound and can be found in [10], [16].

The previous result can be easily generalized to the case of Inhomogeneous Poisson Point (IPP) processes:

Lemma 3: (Concentration result for IPP). Consider a set of points \mathcal{X} distributed over \mathcal{O} according to an IPP process having local intensity $\Phi(\xi)$, such that $\int_{\mathcal{O}} \Phi(\xi) d\xi = n$. Let \mathcal{T}_k be any tessellation of \mathcal{O} (or any sub-region of \mathcal{O}), whose tiles T_k satisfy: $\int_{T_k} \Phi(\xi) d\xi \geq 16 \log n$. Then uniformly over the tessellation $U(T_k) = \Theta(\int_{T_k} \Phi(\xi) d\xi)$.

Lemma 4: (Asymptotic analysis of the node density). Consider a set of nodes distributed according to the clustered point process introduced in Section II-A. Recall that there are on average m clusters whose centres are distributed according to an HPP on area \mathcal{O} . Let $\eta(m) = d_c \sqrt{\log m}$, where d_c is the typical distance between cluster centres as defined in (2). If $\eta(m) = o(1)$ (or, equivalently, $d_c = o(1/\sqrt{\log m})$), then it is possible to find two positive constants g_1, G_1 with $g_1 < G_1$ such that $\forall \xi \in \mathcal{O}$

$$g_1 \frac{n}{L^2} < \Phi(\xi) < G_1 \frac{n}{L^2} \quad \text{w.h.p.} \quad (22)$$

which means that $\underline{\Phi} = \Theta(\overline{\Phi})$. More in general, when $\eta(m) = \Omega(1)$ it is possible to find two positive constants g_2, G_2 , such that, w.h.p., $\underline{\Phi} > g_2 q \log m s(d_c \sqrt{\log m})$ and $\overline{\Phi} < G_2 q \log m$.

The proof can be found in [6].

We next introduce a couple of standard properties of Poisson point processes (see [15]) which are necessary for the construction of our scheduling-routing schemes.

Lemma 5: (Thinning of IPP process). Consider a set $\mathcal{X} = \{X\}_1^N$ of points distributed over a compact domain \mathcal{O} according to an IPP process of intensity $\Phi(\xi)$. Let $\underline{\Phi}$ be $\inf_{\xi \in \mathcal{O}} \Phi(\xi)$. Then for any $\Phi_0 \leq \underline{\Phi}$ it is possible to extract from \mathcal{X} a subset of points $\mathcal{Z} \subseteq \mathcal{X}$ distributed over \mathcal{O} according to an HPP process of rate Φ_0 .

Lemma 6: (Completion of IPP process). Consider a set $\mathcal{X} = \{X\}_1^N$ of points distributed over a compact domain \mathcal{O} according to an inhomogeneous Poisson process of intensity $\Phi(\xi)$. Let $\bar{\Phi}$ be $\sup_{\xi \in \mathcal{O}} \Phi(\xi)$. Then it is possible to complete \mathcal{X} (adding some extra points) to a superset $\mathcal{W} \supseteq \mathcal{X}$ of points distributed according to an HPP process of rate $\bar{\Phi}$.

The above two results can be extended to the doubly stochastic point process considered in this work, with the only difference that they hold w.h.p., instead of with probability one. This is stated more precisely in the following:

Lemma 7: (Thinning and completion of the clustered point process). Consider nodes $\mathcal{X} = \{X\}_1^N$, with $\mathbb{E}[N] = n$, placed according to the doubly stochastic clustered point process introduced in Section II-A. Then a subset of nodes $\mathcal{Z} \subseteq \mathcal{X}$ can be found w.h.p. such that \mathcal{Z} forms an HPP with intensity $\underline{\Phi}_0$, where $\underline{\Phi}_0 = g_1 \frac{n}{L^2}$ under the *cluster-dense* condition and $\underline{\Phi}_0 = g_2 q \log ms(d_c \sqrt{\log m})$ under the *cluster-sparse* condition. Here g_1 and g_2 are the constants defined in Lemma 4. Moreover, \mathcal{X} can be completed w.h.p. to a superset $\mathcal{W} \supseteq \mathcal{X}$ that forms an HPP with intensity $\bar{\Phi}_0$, where $\bar{\Phi}_0 = G_1 \frac{n}{L^2}$ under the *cluster-dense* condition and $\bar{\Phi}_0 = G_2 q \log m$ under the *cluster-sparse* condition. Again, G_1 and G_2 are the constants defined in Lemma 4.

The following is a classical result on doubly stochastic matrices, known as the Birkhoff-von-Neumann (BvN) Theorem:

Lemma 8: (BvN Theorem). Any doubly stochastic matrix can be decomposed into a convex combination of permutation matrices.

From the BvN Theorem, it descends that every non-negative, integer-valued matrix $A = [a_{ij}]$ can be decomposed into the sum of at most H sub-permutation matrices (i.e., binary-valued doubly sub-stochastic matrices), being $H = \max(\sum_i a_{ij}, \sum_j a_{ij})$ the maximum column/row sum [14].

APPENDIX E

DETAILS ON THE HIERARCHICAL MULTI-HOP SCHEME

In this appendix we provide additional details on the hierarchical multi-hop scheme introduced in Section IV-B1, describing the complete scheduling-routing scheme combining the hierarchical multi-hop scheme (for both the *access* and the *delivery* phase) with the optimal communication scheme (for the *transport* phase) available over the **HI** main-infrastructure of density $\underline{\Phi}$. The associated analysis will confirm that the system throughput is indeed given by the aggregate throughput of the main-infrastructure, and no bottleneck arises during the *access* and *delivery* phases.

TABLE III
SUMMARY OF NOTATION

Symb.	Definition
\mathcal{O}_k	Layer- k infrastructure domain
\mathcal{I}_k^j	j -th layer- k component of domain \mathcal{O}_k
$ \mathcal{I}_k^j $	Area of component \mathcal{I}_k^j
\mathcal{X}_k	Nodes within \mathcal{O}_k
\mathcal{Z}_k	Nodes within \mathcal{O}_k forming the layer- k infrastructure
\mathcal{Y}_k	Nodes within $\mathcal{O}_k \setminus \mathcal{O}_{k+1}$ not belonging to \mathcal{Z}_k
λ_k	Density of nodes \mathcal{Z}_k within \mathcal{O}_k

Recall that the construction introduced in Section IV-B1 (see Figure 4) allows to build a sequence of nested domains \mathcal{O}_k , $k = 0, 1, 2, \dots, K_{\max}$, where $K_{\max} = O(\log n)$. Domain k is composed by J_k connected components $\{\mathcal{I}_k^j\}_j$, referred to in the following as layer- k components.

We define by \mathcal{X}_k the restriction of \mathcal{X} on \mathcal{O}_k ; i.e., \mathcal{X}_k comprises all points of \mathcal{X} lying in \mathcal{O}_k . Applying the standard thinning procedure of Lemma 7 to each domain \mathcal{O}_k , a set of nodes $\mathcal{Z}_k \subseteq \mathcal{X}_k$ can be found such that: i) \mathcal{Z}_k is an HPP process of intensity λ_k on \mathcal{O}_k ; ii) any node belonging to \mathcal{Z}_{k-1} and to \mathcal{X}_k , also belongs to \mathcal{Z}_k , for $k \geq 1$. Let $\mathcal{Y}_k = \mathcal{X}_k \setminus (\mathcal{X}_{k+1} \cup \mathcal{Z}_k)$, i.e., \mathcal{Y}_k comprises those points of \mathcal{X}_k lying in $\mathcal{O}_k - \mathcal{O}_{k+1}$ which do not belong to \mathcal{Z}_k . Table III summarizes the notation.

Consider the following scheduling/routing scheme according to which time is partitioned into frames, each frame comprising $K_{\max} + 1$ descending phases $K_{\max}, K_{\max} - 1, \dots, 1, 0$ followed by $K_{\max} + 1$ ascending phases $0, 1, \dots, K_{\max}$. Each phase is in turn partitioned into two periods.

Within descending phase k (here index k runs from K_{\max} down to 0), during the first period all nodes in \mathcal{Y}_k are allowed to transmit their data to a close node in \mathcal{Z}_k within the same layer- k component, balancing the traffic among the candidate receivers. During the second period of the descending phase, nodes in \mathcal{Z}_k transmit to: i) randomly selected nodes lying within the same layer- $(k-1)$ component, and belonging to $\mathcal{Z}_k \cap \mathcal{Z}_{k-1}$, when $k > 0$; ii) randomly selected nodes belonging to \mathcal{Z}_0 , when $k = 0$.

Nodes \mathcal{Z}_k transmit data gathered in the previous phase (if $k < K_{\max}$), data gathered during the first period of the current phase, and their own data, again balancing the traffic among the feasible receivers. The data transport is achieved exploiting the intrinsic capacity of the sub-system formed by nodes belonging to \mathcal{Z}_k , referred to as layer- k infrastructure.

In ascending phase k , within the first period data directed to destinations in $\mathcal{X}_k \setminus \mathcal{X}_{k+1}$, for $k < K_{\max}$ (i.e., destinations lying in \mathcal{O}_k but not in \mathcal{O}_{k+1}), or to destinations in $\mathcal{X}_{K_{\max}}$, for $k = K_{\max}$, are transmitted

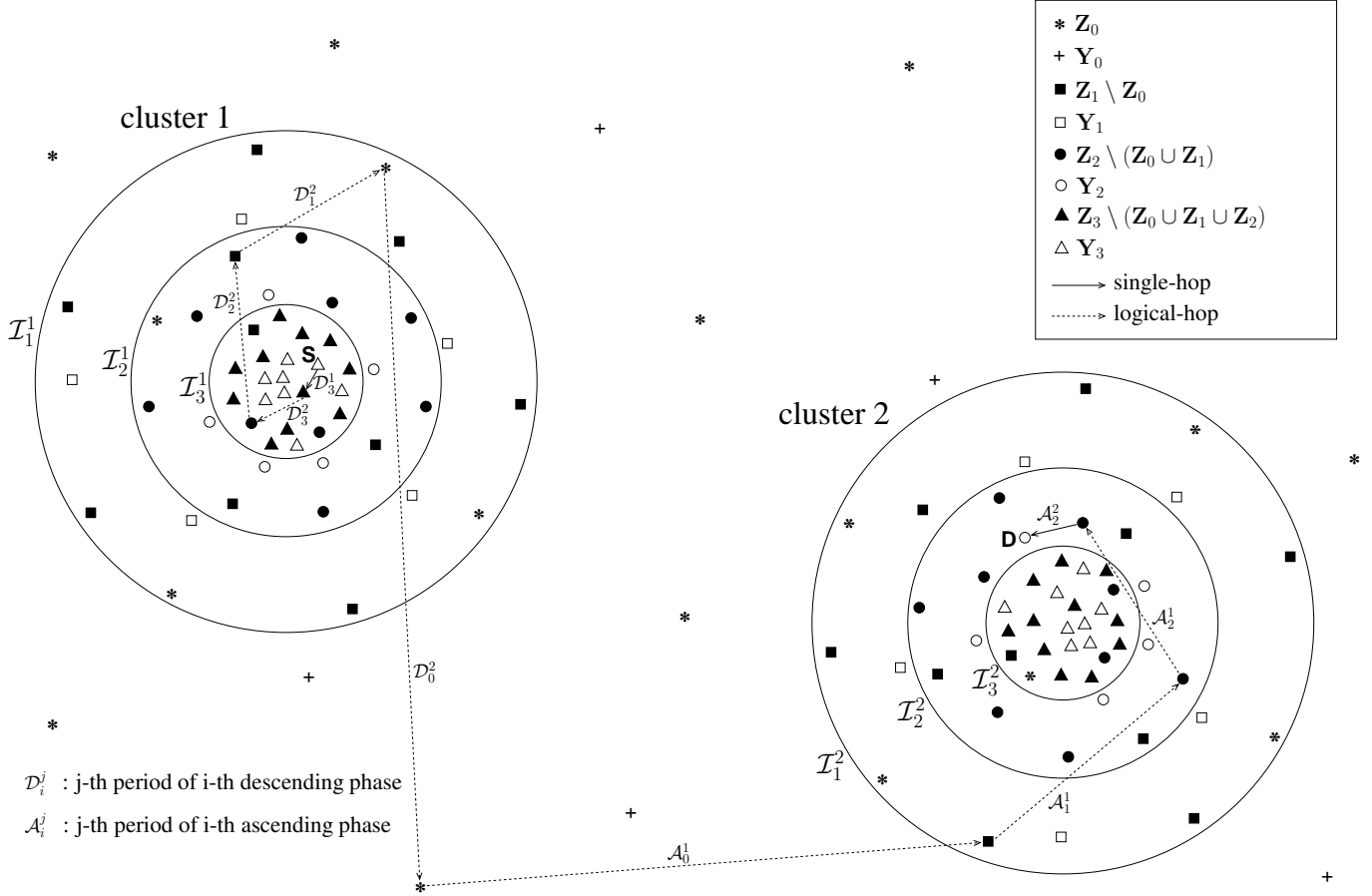


Fig. 6. Example of scheduling/routing of a flow established between node S belonging to cluster 1 and node D belonging to cluster 2. Nodes filled in white belong to sets \mathcal{Y}_i , whereas nodes filled in black belong to one or more of sets \mathcal{Z}_i . Different marks have been used to denote the nodes belonging to \mathcal{Z}_i , so as to illustrate one important property of our construction, namely the fact that, if a node belongs to both \mathcal{Z}_{k-1} and \mathcal{Z}_k , it also belongs to \mathcal{Z}_k , for $k \geq 1$.

exploiting layer- k infrastructure, either directly to their destination (whenever the destination belongs to \mathcal{Z}_k) or to a close node in \mathcal{Z}_k , while at the same time, data directed to nodes lying within \mathcal{O}_{k+1} (only for $k < K_{\max}$) are routed to nodes of $\mathcal{Z}_k \cap \mathcal{Z}_{k+1}$ lying within the same layer- $(k+1)$ component of the destination. During the second period of ascending phases $k \leq K_{\max}$, data directed to nodes in \mathcal{Y}_k are delivered to their final destination through single-hop transmissions.

Figure 6 illustrates an example of scheduling and routing of a flow established between two nodes belonging to different clusters (let these clusters be cluster 1 and cluster 2). For simplicity, we have assumed that cluster 1 and cluster 2 belong to different layer-1 components, and they are the only cluster centres falling in their respective layer-1 component. As consequence, components of any layer $k > 0$

around either cluster have a circular shape. In the specific example in Figure 6, source node S belongs to set \mathcal{Y}_3 of cluster 1, whereas destination node D belongs to set \mathcal{Y}_2 of cluster 2. The figure shows one possible logical route for flow S-D, represented as a path connecting the source to the destination through a sequence of significant relay nodes. Dashed edges represent logical hop, in the sense that data are sent from one vertex to the other using the optimal communication scheme of the underlying infrastructure, which might employ a rather complex cooperative multi-hop strategy to forward the data.

Edges are labeled with the phase and period in which the corresponding communication can be scheduled. The following notation has been used: \mathcal{D}_i^j stands for the j -th period of descending phase i ; \mathcal{A}_i^j stands for the j -th period of ascending phase i . We observe that the chosen route is significantly more tortuous than other multi-hop routes between S and D that one could follow. This is due to the randomness introduced in our scheme in the selection of the next-hop relay belonging to a given component. Such randomness does not penalize performance in order sense, while having the benefit of uniformly balancing the traffic among the candidate receivers.

Consider descending phase k , with $0 \leq k \leq K_{\max}$. We fix the duration of this phase to $\beta^k = 1/(2K_{\max})$, and suppose that the two periods within the same phase are of equal duration.

We start looking at the first period of this phase, in which nodes in \mathcal{Y}_k are allowed to transmit to nodes in \mathcal{Z}_k lying in the same component. We apply Lemma 3 to the network region $\mathcal{O}_k \setminus \mathcal{O}_{k+1}$ (if $k < K_{\max}$, otherwise we take domain $\mathcal{O}_{K_{\max}}$), partitioning it into tiles T_k having surface⁹ $|T_k| = \Theta(\log n / \lambda_k)$. Uniformly over the tiles, the number of points of \mathcal{Y}_k falling in each tile is $O(\log n)$, whereas the number of points of \mathcal{Z}_k in each tile is $\Omega(\log n)$ (actually, there are $\Theta(\log n)$ nodes of either sets in each tile).

Hence we can assign nodes \mathcal{Z}_k to nodes \mathcal{Y}_k residing within the same tile in such a way that the number of nodes \mathcal{Y}_k associated to the same node \mathcal{Z}_k is uniformly bounded by a constant.

Highly interfering transmissions are orthogonalized over time, adopting the usual technique of partitioning tiles into a finite number of subsets, each comprising mutually weakly interfering tiles, and enabling one subset at a time. Since the number of conflicting transmitters per tile is $O(\log n)$, the fraction of time devoted to each transmitter is $\Omega(\beta^k / (2 \log n))$ and consequently the achievable rate by every node in \mathcal{Y}_k is $\Omega(\beta^k (d_k^T)^{-\alpha} / (2 \log n))$, being $d_k^T = \sqrt{|T_k|}$.

In the second period, for $k > 0$, data are transported within the same component \mathcal{I}_k^j by nodes \mathcal{Z}_k , adopting the best communication scheme available in the underlying layer- k infrastructure. We observe that, by construction, in each component \mathcal{I}_k^j the amount of data to be transferred is that generated by

⁹We shape the tiles in such a way that the maximum distance between two points in the same tile T_k is $\Theta(\sqrt{|T_k|})$ for any k .

nodes lying within the same component. The number of nodes belonging to the infrastructure covering \mathcal{I}_k^j is $z_k^j = \Theta(\lambda_k |\mathcal{I}_k^j|)$ (we are assuming that $\lambda_k |\mathcal{I}_k^j| = \Omega(\log n)$, as can be easily verified). The total number of nodes n_k^j lying within \mathcal{I}_k^j can be upper bounded by $n_k^j = O(\log nq)$, since each component contains $O(\log n)$ cluster centres (see Section IV-B1).

Since the throughput of the infrastructure covering \mathcal{I}_k^j is $T_u(n, d_k, \lambda_k, \alpha)$, and considering that such infrastructure is used only for a fraction of time equal to $\beta^k/2$, the aggregate data rate that can be sustained by the layer- k infrastructure covering \mathcal{I}_k^j is $\beta^k/2 T_u(n, d_k, \lambda_k, \alpha)$. Thus, every node of \mathcal{Z}_k is allowed to exchange with layer- k infrastructure a data rate which is $\beta^k/(2z_k^j) T_u(n, d_k, \lambda_k, \alpha)$. Considering that, by construction, every node of \mathcal{Z}_k is pushing/pulling into layer- k infrastructure the aggregate data of $O(n_k^j/z_k^j)$ end-to-end flows, the achievable per-flow throughput by layer- k infrastructure is $\omega(\beta^k/(2n_k^j T_u(n, d_k, \lambda_k, \alpha)))$. Indeed note that thanks to the BvN Theorem (see Lemma 8), we can decompose our traffic pattern into $O(n_k^j/z_k^j)$ permutation traffic patterns, and devote to each of them an equal fraction of the system bandwidth.

In the case of descending phase $k = 0$, using similar arguments, it turns out that the achievable per-flow throughput by the ground infrastructure (layer-0) is $\omega(\beta^0/n_0^j T_u(n, d_0, \lambda_0, \alpha))$.

Turning our attention to ascending phase k , we fix the duration of this phase to $\beta^k = 1/2K_{\max}$ and assume that the two periods within the same phase are of equal duration. Then ascending phase k can be mapped to the corresponding descending phase k by reversing the time; i.e., by observing the data transmission process backward. As a consequence the maximum throughput sustainable in ascending phase k equals the throughput sustainable in descending phase k .

We can conclude that the maximum per-flow throughput sustainable by the whole system is given by the minimum among the per-flow throughput sustainable in every period of the frame. By construction, the system bottleneck is due to the throughput of the ground infrastructure, thus the per-node throughput that can be achieved is $\omega(\beta^0/n_0^j T_u(n, d_0, \lambda_0, \alpha))$.

APPENDIX F

Let us first introduce two useful results that immediately descend from Lemma 3.

Corollary 1: Given an IPP of intensity $\Phi(\xi)$ the number of nodes falling in a squarelet of side Δ located in the region of minimal node density is upper-bounded by

$$U(\Delta) = \log n$$

Corollary 2: The number of nodes falling in a squarelet of side d_c is upper-bounded by:

$$U(d_c) = \frac{n}{m} \log n$$

To compute an upper bound for $P_{\mathcal{S}_2, \mathcal{D}_1}$, we write:

$$\begin{aligned} P_{\mathcal{S}_2, \mathcal{D}_1} &= \sum_{i \in \mathcal{S}_2} \sum_{k \in \mathcal{D}_1} r_{ik}^{-\alpha} \\ &= \sum_{\mathcal{S}_{2,j}} \sum_{\mathcal{D}_{1,\ell}} \sum_{i \in \mathcal{S}_{2,j}} \sum_{k \in \mathcal{D}_{1,\ell}} d_{ik}^{-\alpha} \\ &\leq \sum_{\mathcal{S}_{2,j}} \sum_{\mathcal{D}_{1,\ell}} (\underline{d}_{j\ell})^{-\alpha} U(d_c) U(\Delta) \\ &= d_c^{-\alpha} \left(\frac{n}{m}\right) \left(\frac{d_c}{\Delta}\right)^2 \log^2 n \sum_{j_1, j_2} \sum_{\ell_2} ((j_1 + K)^2 + (j_2 - \ell_2)^2)^{-\alpha/2} \end{aligned} \quad (23)$$

where Kd_c takes into account the width of region \mathcal{D}_2 . Using the results in Appendix F-B and considering $N = L/d_c$ and $M = L/2d_c$, we obtain

$$P_{\mathcal{S}_2, \mathcal{D}_1} \leq \begin{cases} K d_c^{-2} \left(\frac{n}{m}\right) \left(\frac{L}{d_c}\right) \left(\frac{d_c}{\Delta}\right)^2 \log^2 n \log\left(\frac{L}{d_c}\right) & \alpha = 2 \\ k_2 d_c^{-\alpha} \left(\frac{n}{m}\right) \left(\frac{L}{d_c}\right) \left(\frac{d_c}{\Delta}\right)^2 \log^2 n & \alpha > 2 \end{cases} \quad (24)$$

Similarly, to upper-bound $P_{\mathcal{S}_2, \bar{\mathcal{D}}_2}$, we divide the regions \mathcal{S}_2 and $\bar{\mathcal{D}}_2$ in squarelets of side W since the minimal distance between a source in \mathcal{S}_2 and a destination in $\bar{\mathcal{D}}_2$ is $\Theta(W)$. Also, recall that each of such squarelets contains at most $\Theta(W^2 \Delta^{-2} \log n)$ nodes. The j -th squarelet in \mathcal{S}_2 is denoted by $\mathcal{S}_{2,j}$ and has coordinates $\mathbf{j} = [j_1, j_2]$; the ℓ -th squarelet in $\bar{\mathcal{D}}_2$ is denoted by $\bar{\mathcal{D}}_{2,\ell}$ and has coordinates $\ell = [\ell_1, \ell_2]$.

Then, we write

$$\begin{aligned} P_{\mathcal{S}_2, \bar{\mathcal{D}}_2} &= \sum_{i \in \mathcal{S}_2} \sum_{k \in \bar{\mathcal{D}}_2} d_{ik}^{-\alpha} = \sum_{\mathcal{S}_{2,j}} \sum_{\bar{\mathcal{D}}_{2,\ell}} \sum_{i \in \mathcal{S}_{2,j}} \sum_{k \in \bar{\mathcal{D}}_{2,\ell}} d_{ik}^{-\alpha} \\ &\leq \sum_{\mathcal{S}_{2,j}} \sum_{\bar{\mathcal{D}}_{2,\ell}} \underline{d}_{j\ell}^{-\alpha} \frac{W^4}{\Delta^4} U(\Delta)^2 \\ &= W^{4-\alpha} \Delta^{-4} \log^2 n \sum_{\mathbf{j}} \sum_{\ell} ((j_1 + \ell_1 + 1)^2 + (j_2 - \ell_2)^2)^{-\alpha/2} \end{aligned} \quad (25)$$

for $0 \leq j_1, \ell_1 \leq d_c/\Delta$. $-L/2\Delta \leq j_2, \ell_2 \leq L/2\Delta$. Here we define $\underline{d}_{j\ell}$ as the minimum distance between squarelets $\mathcal{S}_{2,j}$ and $\bar{\mathcal{D}}_{2,\ell}$, which is given by $\underline{d}_{j\ell} = W ((j_1 + \ell_1 + 1)^2 + (j_2 - \ell_2)^2)^{1/2}$.

By using the results in Appendix F-A with $N = d_c/\Delta$ and $M = L/2\Delta$, we obtain (10).

A. Useful Series Bounds

We upper-bound the following term

$$\sum_{j_1, j_2} \sum_{\ell_1, \ell_2} ((j_1 + \ell_1 + K)^2 + (j_2 - \ell_2)^2)^{-\alpha/2} \quad (26)$$

where $0 \leq j_1, \ell_1 \leq N$, $-M \leq j_2, \ell_2 \leq M$.

We first notice that for any positive even function $f(x)$ with support in \mathbb{R}

$$\sum_{j=-M}^M \sum_{\ell=-M}^M f(j - \ell) = \sum_{a=-2M}^{2M} (2M - |a|)f(a) \leq 4M \sum_{a=0}^{2M} f(a)$$

Moreover, if $f(x)$ has a maximum in $x = 0$ and is decreasing in $[0, +\infty)$,

$$\sum_{h=h_1}^{h_2} f(h) \leq f(h_1) + \int_{h_1}^{h_2} f(x) dx \quad (27)$$

for any h_1, h_2 such that $0 \leq h_1 < h_2$. If the integral in (27) converges, then we can write

$$\sum_{h=h_1}^{h_2} f(h) \leq f(h_1) + \int_{h_1}^{\infty} f(x) dx \quad (28)$$

Let us consider $f(x) = ((j_1 + \ell_1 + K)^2 + x^2)^{-\alpha/2}$. Then, we can upper-bound (26) as

$$\sum_{j_1, j_2, \ell_1, \ell_2} ((j_1 + \ell_1 + K)^2 + (j_2 - \ell_2)^2)^{-\alpha/2} \leq 4M \sum_{j_1, \ell_1} \sum_{a=0}^{2M} ((j_1 + \ell_1 + K)^2 + a^2)^{-\alpha/2} \quad (29)$$

For $\alpha > 1$, we can use (28) with $h_1 = 0$ and write

$$\begin{aligned} & \sum_{j_1, j_2, \ell_1, \ell_2} ((j_1 + \ell_1 + K)^2 + (j_2 - \ell_2)^2)^{-\alpha/2} \\ & \leq 4M \sum_{j_1, \ell_1} \left[(j_1 + \ell_1 + K)^{-\alpha} + \int_0^{\infty} ((j_1 + \ell_1 + K)^2 + x^2)^{-\alpha/2} dx \right] \\ & \leq 4M \sum_{j_1=0}^N \sum_{\ell_1=0}^N [(j_1 + \ell_1 + K)^{-\alpha} + Y(j_1 + \ell_1 + K)^{1-\alpha}] \end{aligned} \quad (30)$$

for a positive constant Y independent of M .

For $\alpha = 2$, we apply (28) and (27) to the first term of (30) and we obtain

$$\begin{aligned} \sum_{j_1=0}^N \sum_{\ell_1=0}^N (j_1 + \ell_1 + K)^{-2} & \leq \sum_{\ell_1=0}^N \left[(\ell_1 + K)^{-2} + \int_0^{\infty} (x + \ell_1 + K)^{-2} dx \right] \\ & = \sum_{\ell_1=0}^N [(\ell_1 + K)^{-2} + (\ell_1 + K)^{-1}] \\ & \leq 4M \int_0^{\infty} (x + K)^{-2} dx + \int_0^N (x + K)^{-1} dx \\ & \leq \log \left(\frac{N + K}{K} \right) \\ & \leq K \log N \end{aligned} \quad (31)$$

Similarly, for the second term of (30), we obtain

$$\begin{aligned} \sum_{j_1=0}^N \sum_{\ell_1=0}^N (j_1 + \ell_1 + K)^{-1} &\leq \sum_{\ell_1=0}^N \left[(\ell_1 + K)^{-1} + \log \left(\frac{N + \ell_1 + K}{\ell_1 + K} \right) \right] \\ &\leq KN \log N \end{aligned} \quad (32)$$

Following a similar approach, for $\alpha > 2$ we have:

$$\sum_{j_1, \ell_1=0}^N \sum_{j_2, \ell_2=-M}^M ((j_1 + \ell_1 + K)^2 + (j_2 - \ell_2)^2)^{-\alpha/2} \leq \begin{cases} KMN \log N & \alpha = 2 \\ KMN^{3-\alpha} & 2 < \alpha < 3 \\ KM \log N & \alpha = 3 \\ KM & \alpha > 3 \end{cases} \quad (33)$$

B. Further Series Bounds

By using the results of Appendix F-A, we have

$$\sum_{j_1=0}^N \sum_{j_2, \ell_2=-M}^M ((j_1 + K)^2 + (j_2 - \ell_2)^2)^{-\alpha/2} \leq 4M \sum_{j_1=0}^N [(j_1 + K)^{-\alpha} + Y_1(j_1 + K)^{1-\alpha}]$$

for a positive constant Y_1 independent of M . Similarly to the derivation in Appendix F-A, we obtain

$$\sum_{j_1=0}^N \sum_{j_2, \ell_2=-M}^M ((j_1 + K)^2 + (j_2 - \ell_2)^2)^{-\alpha/2} \leq \begin{cases} KM \log N & \alpha = 2 \\ KM & \alpha > 2 \end{cases} \quad (34)$$