A Scalable Model for Energy Load Balancing in Large-scale Sensor Networks

Seung Jun Baek and Gustavo de Veciana

Abstract— In this paper we propose a stochastic geometric model to study the energy burdens seen in a large scale hirarchical sensor network. The network makes a natural use of aggregation nodes, for compression, filtering or data fusion of local sensed data. Aggregation nodes (AGN) then relay the traffic to mobile sinks. While aggregation may substantially reduce the overall traffic on the network it may have a deleterious effect of concentrating loads on paths between AGNs and the sinks–such inhomogeneities in energy burdens may in turn lead to nodes with depleted energy reserves. To remedy this problem we consider how one might achieve more balanced energy burdens across the network by spreading traffic, i.e., using a multiplicity of paths between AGNs and sinks. The proposed model reveals, how various aspects of the task at hand impact the characteristics of energy burdens on the network and in turn the likely lifetime for the system. We show that the scale of aggregation and degree of spreading might need and can be optimized. Additionally if the sensing activity involves large amounts of data flowing to sinks, then inhomogeneities in the energy burdens seen by nodes around the sinks will be hard to overcome, and indeed the network appears to scale poorly. By contrast if the sensed data is bursty in space and time, then one can reap substantial benefits from aggregation and balancing.

I. INTRODUCTION

In large-scale wireless ad-hoc/sensor networks enabled by multi-hop relaying, it is critical to provide mechanisms to conserve energy due to nodes' limited capabilities to store and replenish energy. Additional challenges may result from inhomogeneous spatial patterns in traffic loads resulting in uneven energy burdens seen by nodes resulting in a shortening the network's lifetime. In this paper we consider a widely considered *hierarchical organization* of network resources based on local aggregation of sensor data followed by forwarding to a set of information sinks. The rationale is to have nodes in close proximity elect an aggregation node, whose role might be to compress, filter, or perform data fusion on spatially correlated data prior to forwarding it to an information sink. Unfortunately, such a hierarchy still faces an intrinsic problem from an energy perspective: the nodes close to sinks will still see a disproportionate energy burden as they will see the higher loads of traffic that flow to the sinks. One solution is to increase the density of sinks, however it may end up being too costly. In this paper we will assume that only a small number of sinks are available as compared to the size of network, and thus by large-scale we mean that on average there are a large number of sensors associated with sinks.

A second possibility is to make sinks *mobile*, i.e., have sinks change their locations to balance the energy burdens incurred accross the network nodes [1]. As shown in Fig. 1, the network can operate in two stages. In Stage 1 local aggregation nodes
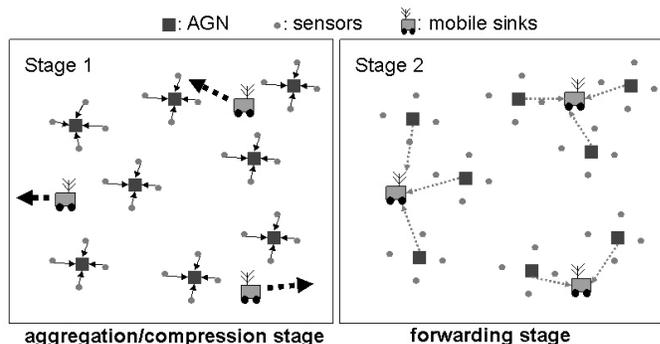

Fig. 1.   Stages of operation for sensor network with mobile sinks.

(AGN) may aggregate information from sensors in their cell while sinks move around. Then, in Stage 2, the sinks may 'probe' the network at high power, and collect information from AGNs. Not surprisingly the effectiveness of this scheme depends on a number of factors including the very nature of information being gathered. Among those we consider to be critical we have: the timescale of sink mobility, the spatio-temporal periodicity of sensed data, and delay sensitivity of data delivery.

For brevity let us consider a few extreme examples. If *every* sensor generates data periodically on relatively short timescales versus that on which sinks move, and if this data must be relayed to sinks immediately, then the network will scale very poorly for two reasons: concentration of energy burdens and throughput collapse around sinks. In this case the only reasonable solution is to put more sinks. However if data delivery is delay insensitive AGNs may forward data only when a sink is close by i.e., in an *opportunistic* sense the sink mobility increases energy efficiency and throughput capacity [2].

By contrast let us consider the case of applications where the 'events' being sensed correspond to spatio-temporal bursts of information and can tolerate delays on the order of the timescales of sink mobility. We believe a number of interesting scenarios for large-scale sensor networks fall in this category, e.g., environmental monitoring. In this case the network can significantly benefit from the load and energy balancing resulting from the sink mobility. However this has the following problem: as shown on the left of Fig. 2, the aggregated data may still be substantial even once it is successfully compressed/filtered, and may be forwarded on 'narrow' paths to the sinks. This may result in inhomogeneities

■:AGN　•:sensors　🚚:sinks　★:landmarks
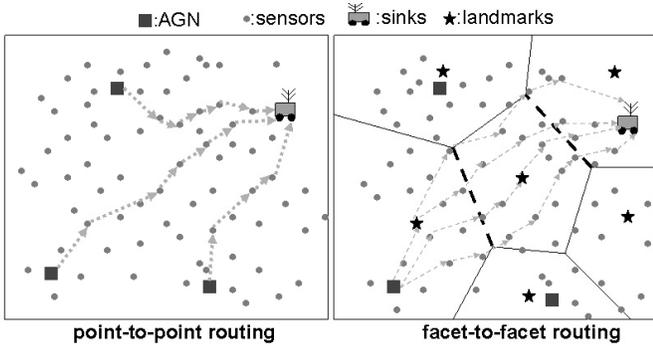
point-to-point routing　　facet-to-facet routing

Fig. 2. Comparison of the point-to-point routing (on the left) and a facet-to-facet routing (on the right). Note the figure on the right shows the facet routing for only one of the aggregation nodes for simplicity.

in the energy burdens over network which will eventually shorten the network's lifetime.

To mitigate the concentration of load associated with aggregation it is reasonable to *spatially spread* aggregated traffic in a *proactive* manner. That is we propose to automatically have nodes *spatially cooperate* to create multi-path routes over which traffic is spread. An efficient, scalable way to achieve this it is to use a *facet routing* scheme as shown on the right of Fig. 2. The idea is spread the traffic over 'facets' of Voronoi cells induced by a set of *landmarks* where such hypothetical facets are depicted as dotted lines in the figure. The scheme is decentralized and light-weight in nature and may be efficiently implemented using geographic routing techniques. Moreover one can use local information of depletion levels within landmark cells to efficiently 'detour' highly depleted local regions. In this paper we will largely focus on analyzing stochastic geometric models which allow one to evaluate several tradeoffs associated with aggregation of traffic an then the degree of spreading, e.g., facet widths, over which traffic is carried to the sinks.

We will explore the following tradeoffs associated with spatial spreading of traffic combined with aggregation/compression. First by spreading traffic over several paths, one can certainly balance energy burdens but at the expense of having traffic traverse longer distances and thus larger average energy costs. Second by aggregating information from nodes one can reduce the traffic load through compression, filtering and/or data fusion, but this may lead to undesirable inhomogeneities of load and thus energy burdens from AGNs to sinks. The fundamental design questions are:

- *How much spreading is beneficial for traffic from the AGNs to the sinks?*
- *When is the benefit of aggregation (compression) of traffic counteracted by the concentration of energy burdens in the network ?*

To capture the characteristics of the problem in this paper we devise a mathematical model for traffic aggregation and spreading. It is without a doubt a very simple caricature, based on first-order models for energy and compression at AGNs, yet

it will allow us to study how the network lifetime is affected by a number of design parameters, including the effectiveness of aggregation/compression, the density of sinks, battery capacity of sensors, etc. We will show how one can further jointly optimize the spatial scales for aggregation and spreading so as to maximize network lifetime and provide numerical study of the results.

This paper is organized as follows: in Section II we discuss related work. In Section III we briefly present our models and assumptions. In Section IV we derive mean and variance of energy burdens, and based on those we discuss how to optimize operation to maximize the network's lifetime in Section V. This is followed by a numerical study in Section VI. Finally we conclude with Section VII.

II. RELATED WORK

There has been substantial amount of research on energy conserving routing for large-scale ad hoc/sensor networks, for different network scenarios and contexts. The idea of using mobile relays or base stations in large sensor network was proposed and studied in [3], [4], [1] in which automated robots are deployed to solve scalability problems in sensor networks. Particularly in [1] the authors show how the energy concentration problem around sinks can be mitigated by showing that the fluctuation of energy burdens gets smoothed out by a single mobile base station. In this paper we tackle a more general problem by considering hierarchical aggregation with multiple mobile sinks, and see how the performance can be optimized. In particular we provide an analytical model to capture the impact that aggregation/compression, traffic spreading and moving sinks will have on network lifetime.

In this respect this paper is closely related to our previous work [5] where we model the random, unstructured traffic in homogeneous network as inducing *spatial traces* of energy burdens and show how the proactive multipath routing is connected to network lifetime. We will use part of these results to model the behavior of spatial patterns of structured, i.e., hierarchical traffic in the sequel. Finally we are inspired by studies modelling network structure via hierarchies of Voronoi tessellations and exploit such geometric structures for various purposes. Notably modelling of telecommunication networks using stochastic geometry has been proposed in [6] for analyzing the cost of a network with a hierarchy associated with proximity, where we borrow their framework and notations in part.

III. MODELS AND ASSUMPTIONS

As mentioned earlier assume that mobile sinks periodically coordinate to 'probe' the network, i.e., broadcast their locations and announce they are ready to serve as data sinks. If some AGNs have data to send they forward the data to the nearest sink as depicted in Fig. 1 - we refer to this two-stage operation as a *round*. The sinks navigate network at random and we assume the spatio-temporal load on network is relatively light and that at each round AGNs see sinks at a set of random locations which are effectively independent of
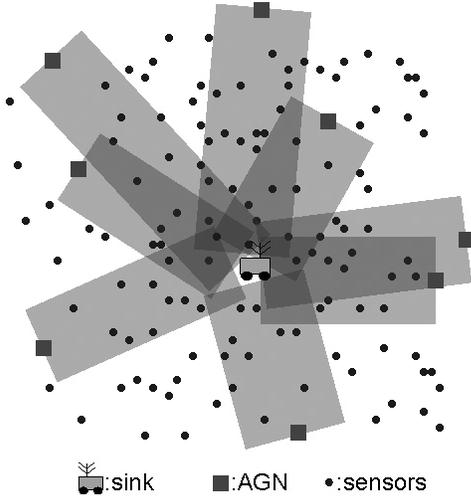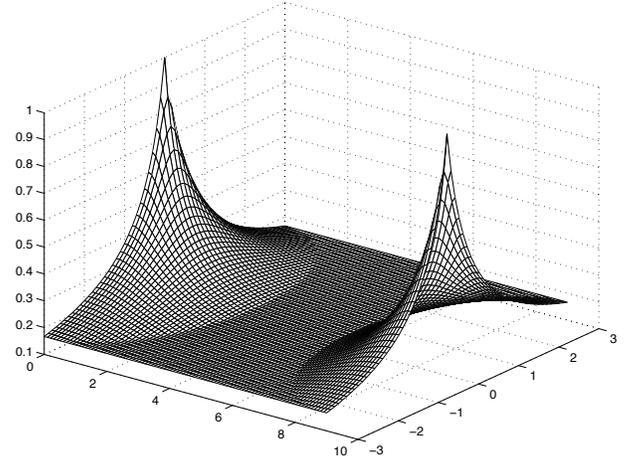
Fig. 3. Strip model.

Fig. 4. Energy burden density on a strip.

past rounds. We also assume that AGNs are selected based on proximity, i.e., once selected the sensors associate themselves with the closest AGN.

For modelling purposes we assume the locations of sensors constitute a homogeneous Poisson point process (PPP), and the locations of sinks at each round also constitute PPP with the identical density but are independent across the rounds. We assume the density of sinks is much smaller than that of sensors, i.e., the cost for placing a large number of sinks is high, thus at each round a sink is associated with hundreds or thousands of sensors on average.

We will use a simplified first order model for energy expenditures associated with data transmissions where energy burdens are proportional to the traffic incident on a node [5]. For analysis purposes we will capture the energy burdens induced by routing traffic from AGNs to sinks as two dimensional functions capturing the energy burden per unit area and whose support set captures a footprint of spatially clustered multi-path routes between the AGNs and sinks and corresponds to a closed set in $\mathbb{R}^2$. More specifically we shall assume traffic is spread over rectangular 'strips' of certain width which represents the spatial scale of spreading. Fig. 3 depicts the strip model. Originating from each AGN the strips represents the spatial footprints for energy burdens towards a sink. For example in a region where multiple strips overlap, sensors would see energy burdens proportional to the sum of energy burdens contributed by overlapping strips. Intuitively if the width of strips becomes larger one would see larger regions for overlaps, however this may not translate into higher energy burden in those regions since each unit of area in a larger strip would carry less traffic due to spreading.

In fact energy burdens are not homogenous over a strip. We incorporate the cost associated with traffic spreading in the strip model as follows. We associated with each strip a function capturing the energy burdens at different locations as shown by Fig. 4. The proposed function is motivated by

'continuum' analogy for flows in high-density networks [5], [7]. For example, the flow density corresponds to field strength within homogeneous medium in electrostatic problems. Note that the solutions of such system dynamics the field strength exhibit harmonic decay in the radial distance from the source. We will make a similar analogy such that 'well-balanced' spreading of traffic leads to a harmonic distribution of flow strengths around source and destination. Let the set $S_w(s, d)$ be a rectangular strip of width w with its ends located at s and d. The density for energy burden, i.e., burden per unit area, is given by a function $h_w : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ such that, for $x \in S_w(s, d)$,

$$h_w(x, s, d) = \begin{cases} \frac{w}{2(w-1)|z-x|+w} & \text{if } |z-x| < \frac{w}{2} \text{ for } z = s, d, \\ \frac{1}{w} & \text{otherwise.} \end{cases}$$

If $x \notin S_w(s, d)$, h is defined to be 0. An example of h for $w = 6$ and $|s - d| = 9$ is shown in Fig. 4. The function has a peak value 1 at s and d, harmonically decreases to w^{-1} between s and d and assumes constant value in the remaining area of the strip. We use this function to capture the energy burden density associated with change of traffic load when the traffic diverges and converges at endpoints, and is spread by a factor of w.

IV. DERIVATION OF COSTS UNDER STRIP MODEL

A. Mean Cost

First we introduce our notation:

- Π_0, Π_1, Π_2 denote PPP of sensors, AGNs and mobile sinks at probing instant respectively.
- λ_k is the intensity of point process Π_k.
- $\mathcal{N}_y$ is the number of points from Π_0 associated with a AGN at location y. Also we define $\mathcal{N}_y^+ := \mathcal{N}_y + 1$.
- $V_z(\Pi)$ is the Voronoi cell with the nucleus $z \in \Pi$ induced by the point process Π.
- $\mathbf{E}_k^x$ is the Palm expectation given that a point in Π_k is located at x.

- $S_w(y, z)$ is a strip of width w with its ends located at y and z.
- $B_r(x)$: a ball of radius r with its center at x.

We are interested in the overall contribution of strips originating from each AGN to its closest sink that overlaps at the typical sensor from Π_0, e.g., see Fig. 3. Since this contribution is stationary, one can write the expectation of the energy burdens burden from the perspective of a typical sensor as:

$$\mathbf{E}_0^0 \left[\sum_{y_j \in V_{z_0}(\Pi_2) \cap \Pi_1} g_w(O, y_j, z_0) \mathcal{N}_{y_j}^+ \right] \quad (1)$$

Here $g_w(x, y, z)$ represents the energy burden density experienced at a sensor at location x incurred by a strip originating from y towards the sink located at z. By z_0 we denote a random point denoting the location of the closest sink to the origin: note this is an approximation which ignores the edge effect on cell boundaries of Π_2. Also the term $N_{y_j}^+$ accounts for the number of Π_0 points aggregated at AGN plus the AGN point itself. By multiplying it by $g_w(O, y_j, z_0)$ we obtain the energy burden density experienced at the origin when the aggregated traffic $N_{y_j}^+$ is forwarded to a sink at z_0 originating from a AGN[1] at y_j.

Using Neveu exchange formula [6], one can rewrite (1) as

$$\frac{\lambda_2}{\lambda_0} \mathbf{E}_2^0 \int_{V_0(\Pi_2)} \left[\sum_{y_j \in V_0(\Pi_2) \cap \Pi_1} g_w(x, y_j, O) \mathcal{N}_{y_j}^+ \right] \Pi_0(dx) \quad (2)$$

$$= \lambda_2 \mathbf{E}_2^0 \int_{V_0(\Pi_2)} \mathbf{E}_{0,1} \left[\sum_{y_j \in V_0(\Pi_2) \cap \Pi_1} g_w(x, y_j, O) \mathcal{N}_{y_j}^+ \right] dx \quad (3)$$

where we have moved the expectation with respect to Π_0 and Π_1, denoted by $\mathbf{E}_{0,1}$, inside the integral from independence among Π_0, Π_1 and Π_2. Applying Campbell's formula [8] for the term inside the integral, we can write

$$\lambda_2 \mathbf{E}_2^0 \int_{V_0(\Pi_2)} \int_{V_0(\Pi_2)} \mathbf{E}_1^y \left[g_w(x, y, O) \mathcal{N}_y^+ \right] \Pi_1(dy) dx \quad (4)$$

$$= \lambda_2 \lambda_1 \mathbf{E}_2^0 \int_{V_0(\Pi_2)} \int_{V_0(\Pi_2)} \left[g_w(x, y, O) \mathbf{E}_1^0 \mathcal{N}_0^+ \right] dy dx \quad (5)$$

We have that $\mathbf{E}_1^0 \mathcal{N}_0 = \frac{\lambda_0}{\lambda_1}$, thus $\mathbf{E}_1^0 \mathcal{N}_0^+ = \frac{\lambda_0 + \lambda_1}{\lambda_1}$.

Note that (5) should be evaluated by considering the void probability regions of Π_2, e.g., union of balls, however it leads to an expression that cannot be reduced to closed form. Thus we approximate the Voronoi cell region $V_0(\Pi_2)$ as a ball centered at origin with radius R which is a random variable distributed according to the distance from the nucleus to a
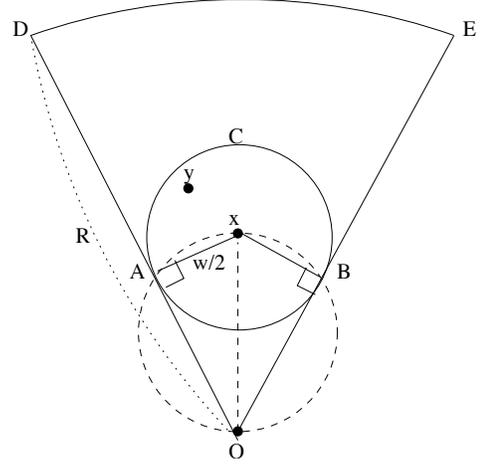

Fig. 5. Illustration for the computation involving $h_w(x, y, O)$ in the region I_1.

vertex on the boundary of the cell. Its distribution is given by [9]

$$f_R(r) = 2\pi^2 \lambda_2 r^3 e^{-\pi \lambda_2 r^2}$$

Thus (5) is approximated by

$$\lambda_2(\lambda_1 + \lambda_0)\mathbf{E} \int_{B_R(O)} \int_{B_R(O)} g_w(x, y, O) dy dx \quad (6)$$

We decompose the integration region with respect to x in (6) to $I_1 := B_R(O) \setminus B_{w/2}(O)$ and $I_2 := B_{w/2}(O)$. Approximately in the region I_1 the energy burden seen at location x is mainly an accumulation of the strips which carry the traffic spread at each aggregation nodes. In I_2, i.e., the region 'nearby' the sink, the traffic will 'converge' to the sink and the sensors in that region will experience the total traffic aggregated at the sink.

We first consider the integration (6) with x in the region I_1, i.e., when $\frac{w}{2} < |x| < R$ in which case $g_w(x, y, z)$ is given by $h_w(x, y, z)$ and in order to compute this integral, we first identify the region which is the set of y locations such that strips having end points at y and O will hit the location x. As shown in Fig. 5, such region is the intersection of the complement of $B_{\frac{|x|}{2}}(\frac{x}{2})$ and the sector of radius R formed by extension of segments OA and OB, where A and B are the tangent points of the lines originating from O and the ball $B_{\frac{w}{2}}(x)$. Let us denote that region by $K_w(x)$.

By inspecting $h_w(x, y, O)$, the integral can be divided into two parts where $h_w(x, y, O)$ is constant or varies with y. We see that, in the region $K_w(x) \cap B_{\frac{w}{2}}^c(x)$, the contribution of the strips originating within that region to x is constant by $\frac{1}{w}$, thus the inner integral of (6) reduces to $|K_w(x) \cap B_{\frac{w}{2}}^c(x)|/w$.

When the strips are originated from y in the region $K_w(x) \cap B_{w/2}(x)$, the energy density contribution changes with y. If we consider $y \in K_w(x) \cap B_{\frac{w}{2}}(x)$ satisfying $|x - y|$ is constant, $h_w(x, y, O)$ is also constant since it depends only on $|x - y|$, which facilitates the evaluation of the surface integral in (6). Also we denote the sector $xBCA$ by $C_w(x)$
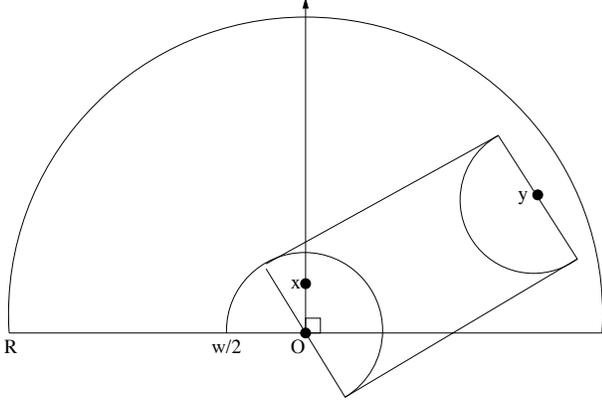
Fig. 6. Illustration for the computation in the region I_2.

and approximate the region of integration $K_w(x) \cap B_{\frac{w}{2}}(x) = C_w(x) \setminus B_{|x|/2}(x/2)$ by $C_w(x)$ which induces only a small error when $x \gg w$. Thus by defining a function $\bar{h}(\rho)$ as

$$\bar{h}(\rho) = \frac{w}{2(w-1)\rho + w},$$

we can rewrite the inner integral of (6) as

$$\int_{B_R(O)} h_w(x, y, O) dy$$
$$\approx \frac{|K_w(x) \cap B_{\frac{w}{2}}^c(x)|}{w} + \int_{C_w(x)} h_w(x, y, O) dy$$
$$= \frac{|K_w(x) \cap B_{\frac{w}{2}}^c(x)|}{w} + (\pi + 2\arcsin(\frac{w}{2|x|})) \int_0^{\frac{w}{2}} \rho \bar{h}(\rho) d\rho.$$

When evaluating the expression inside the expectation of (6), we use *large-cell* approximation, such that $\lambda_2 \ll \lambda_1$ and w varies on the order of λ_1, so that $R \gg w$ with high probability, for example we let $\sqrt{R^2 - (\frac{w}{2})^2} \approx R$, $\arcsin(\frac{w}{2R}) \approx \frac{w}{2R}$, etc. Then we have that

$$\lambda_2(\lambda_1 + \lambda_0)\mathbf{E} \int_{I_1} \int_{B_R(O)} h_w(x, y, O) dy dx \quad (7)$$
$$= 2\pi\lambda_2(\lambda_1 + \lambda_0)\mathbf{E}\left[\frac{R^3}{3} + \pi\left\{ \frac{m(w)}{2} - \frac{w}{8} \right\} R^2 \right.$$
$$\left. + \left\{ wm(w) - \frac{w^2}{8} \right\} R - \frac{\pi w^2}{4} m(w) + \frac{\pi w^3}{32} \right] \quad (8)$$

where $m(w)$ is defined as

$$m(w) = \int_0^{\frac{w}{2}} \rho \bar{h}(\rho) d\rho = \frac{w^2}{4(w-1)^2}(w - 1 - \log(w)).$$

Note $m(w)$ captures the average overall cost for spreading within a strip, i.e., it roughly increases linearly with w.

Next let us consider (6) in the region I_2 for x. As presented in Fig. 6, we see that for a sensor at location x all the strips originating from the half-circle region of radius R hit the location x. Specifically let us denote the half circle whose diameter segment is orthogonal to x and centered at O by $\bar{B}_R(O, x)$, then we assume all the y locations which are the originating points of the strips within $\bar{B}_R(O, x) \setminus B_w(O)$

contributes the burdens at x location. Note again from the large-cell approximation we ignore the strips that are originated within the region $B_w(0)$ for simplicity. Moreover such contribution of burdens depends not on y but only on $|x|$ if we consider the shape of energy burden density of a strip, e.g., see Fig. 6. In specific one can write

$$g_w(x, y, O) = \bar{h}(|x|),$$

thus we have that, from (6),

$$\lambda_2(\lambda_1 + \lambda_0)\mathbf{E} \int_{I_2} \int_{\bar{B}_R(O,x) \setminus B_w(O)} \bar{h}(|x|) dy dx$$
$$= \lambda_2(\lambda_1 + \lambda_0)\mathbf{E}\left[|\bar{B}_R(O, x) \setminus B_w(O)| \int_{I_2} \bar{h}(|x|) dx \right]$$
$$= 2\pi\lambda_2(\lambda_1 + \lambda_0) \int_0^{\frac{w}{2}} \rho \bar{h}(\rho) d\rho \mathbf{E}\left[|\bar{B}_R(O, x) \setminus B_w(O)| \right]$$
$$= 2\pi\lambda_2(\lambda_1 + \lambda_0)m(w)\mathbf{E}\left[\frac{\pi R^2}{2} - \frac{\pi w^2}{2} \right]$$

Combining this with (8) and discarding one minor term, i.e., $o(R)$ for simplicity and using

$$\mathbf{E}[R^k] = \frac{\Gamma(\frac{k}{2} + 2)}{(\pi\lambda_2)^{k/2}}$$

finally we have that

$$2\pi(\lambda_1 + \lambda_0)\mu(w, \lambda_2)$$

where we define $\mu(w, \lambda_2)$ as

$$\left\{ 2m(w) - \frac{w}{4} + \frac{5}{8\pi\sqrt{\lambda_2}} + \frac{3\sqrt{\lambda_2}}{4}(wm(w) - \frac{w^2}{8}) \right\} \quad (9)$$

B. Variance of Cost

To evaluate the variance of the energy burden for a typical sensor under the strip model, consider the following:

$$\mathbf{E}_0^0 \left[\sum_{y_j \in V_{z_0}(\Pi_2) \cap \Pi_1} h_w(O, y_j, z_0)\mathcal{N}_{y_j}^+ \right]^2 \quad (10)$$

$$= \frac{\lambda_2}{\lambda_0}\mathbf{E}_2^0 \int_{V_0(\Pi_2)} \left[\sum_{y_j \in V_0(\Pi_2) \cap \Pi_1} h_w(x, y_j, O)\mathcal{N}_{y_j}^+ \right]^2 \Pi_0(dx) \quad (11)$$

$$= \frac{\lambda_2}{\lambda_0}\mathbf{E}_2^0 \int_{V_0(\Pi_2)} \left[\sum_{y_j \in V_0(\Pi_2) \cap \Pi_1} h_w^2(x, y_j, O)(\mathcal{N}_{y_j}^+)^2 \right.$$

$$\left. + \sum_{\substack{y_j, y_k \in V_0(\Pi_2) \cap \Pi_1 \\ y_j \neq y_k}} h_w(x, y_j, O)h_w(x, y_k, O)\mathcal{N}_{y_j}^+ \mathcal{N}_{y_k}^+ \right] \Pi_0(dx) \quad (12)$$

Assuming $\lambda_2 \ll \lambda_0, \lambda_1$, the intensity measure for the second term in (12) is approximated by $\Pi_1 \times \Pi_1$ from the property of PPP, and furthermore $\mathcal{N}_{y_j}$ and $\mathcal{N}_{y_k}$ are independent conditioned on the area of cells originated by y_i, y_k since they

are the number of points in Π_0 fall on disjoint regions , i.e., $V_{y_j}(\Pi_1)$ and $V_{y_k}(\Pi_2)$. Also we assume that the area of $V_0(\Pi_2)$ is large with high probability w.r.t. density of Π_1, so that their contributions to the indicator terms in (12) are uncorrelated with high probability considering the shape of a strip, thus the $\mathcal{N}_{y_j}$ and $\mathcal{N}_{y_k}$ are assumed to be roughly independent marks associated with y_j and y_k.

Thus the term can be written

$$\mathbf{E}\left[\sum_{\substack{y_j,y_k \in V_0(\Pi_2)\cap\Pi_1 \\ y_j \neq y_k}} \{h_w(x,y_j,O)h_w(x,y_k,O)\mathcal{N}_{y_j}^+\mathcal{N}_{y_k}^+\}\right]$$

$$\approx \left[\mathbf{E}\left\{\sum_{y_j\in V_0(\Pi_2)\cap\Pi_1} h_w(x,y_j,0)\mathcal{N}_{y_j}^+\right\}\right]^2$$

Thus this term cancels out when evaluating the variance which reduces to

$$\mathbf{Var}_0^0\left[\sum_{y_j\in V_{z_0}(\Pi_2)\cap\Pi_1} h_w(O,y_j,z_0)\mathcal{N}_{y_j}^+\right] \qquad (13)$$

$$\approx \frac{\lambda_2}{\lambda_0}\mathbf{E}_2^0\int_{V_0(\Pi_2)}\left[\sum_{y_j\in V_0(\Pi_2)\cap\Pi_1} h_w^2(x,y_j,O)(\mathcal{N}_{y_j}^+)^2\right]\Pi_0(dx) \qquad (14)$$

This can be evaluated in the similar way as before, except that the term including $\mathcal{N}_0^2$.

Recall that $\mathcal{N}_0$ is the number of points in Π_0 which lie within the typical Voronoi cell of Π_1. To compute $\mathbf{E}_1^0\mathcal{N}_0^2$ we have that

$$\mathbf{E}_1^0\mathcal{N}_0^2 = \mathbf{E}[\mathbf{E}_1^0\{\Pi_0(V_0(\Pi_1))\}^2 \mid |V_0(\Pi_1)|]$$
$$= \mathbf{E}[\lambda_0|V_0(\Pi_1)| + \{\lambda_0|V_0(\Pi_1)|\}^2]$$

where $\Pi_k(V)$ denotes the number of Π_k points in a compact set $V \in \mathbb{R}^2$. since given the area $|V_0(\Pi_1)|$ the mean and variance can be obtained using the PPP property of Π_0. Using

$$\mathbf{E}|V_0(\Pi_1)| = \frac{1}{\lambda_1},\ \ \mathbf{E}|V_0(\Pi_1)|^2 \approx \frac{1.28}{\lambda_1^2}$$

we have that

$$\mathbf{E}_1^0[\mathcal{N}_0^+]^2 = 1 + 3\frac{\lambda_0}{\lambda_1} + 1.28\frac{\lambda_0^2}{\lambda_1^2}.$$

Thus the variance is given by

$$2\pi\left(\lambda_1 + 3\lambda_0 + 1.28\frac{\lambda_0^2}{\lambda_1}\right)\sigma^2(w,\lambda_2) \qquad (15)$$

where we define $\sigma^2(w,\lambda_2)$ as

$$\left\{v(w) - \frac{1}{4} + \frac{5}{8\pi w\sqrt{\lambda_2}} + \frac{3\sqrt{\lambda_2}}{4}(wv(w) - \frac{w}{8})\right\} \qquad (16)$$

and $v(w)$ is defined as

$$v(w) = \int_0^{\frac{w}{2}}\rho\bar{h}^2(\rho)d\rho = \frac{w^2}{4(w-1)^2}(\frac{1}{w} - 1 + \log(w)).$$

V. Design for Jointly Optimized Compression, Aggregation and Energy Burden Balancing

In this section we investigate how one can jointly optimize the degree of aggregation and traffic spreading subject to compression performance and density of mobile sinks, based on the strip model. Specifically we find the optimal densities for aggregation nodes and landmarks for a given compression model and sink density in order to minimize the network depletion probability. In the following we introduce compression models and specify how we estimate the depletion probability.

A. Compression and Data Models

We factor in the compression gains associated with the number of sensors per aggregation node. Assume every sensor generates a packet of size 1. Let us denote the size of packet after compression by $f(n)$ when the total number of aggregated packets is n. One would expect $f(n) \leq n$, however it is more important to point out, firstly, the asymptotic gain for the compression, and secondly, the rate of convergence to that gain as n grows. Thus we have the following assumptions for $f(n)$:

1) $\lim_{n\to\infty}\frac{f(n)}{n} = \alpha,\ 0 \leq \alpha \leq 1$,
2) $\frac{f(n)}{n}$ is a monotonically nonincreasing function of n.

If we consider the compression via entropy coding, these assumption corresponds to those on the entropy rate of a sequence of random variables. For example, suppose each sensor i generates discrete random variable X_i where it is identically distributed according to a random variable X whose entropy is 1. If we let $f(n) = H(X_1, X_2, \ldots, X_n)$ the first assumption is equivalent to stating the entropy rate

$$\lim_{n\to\infty}\frac{H(X_1,X_2,\ldots,X_n)}{n} = \alpha.$$

and it is known that $\frac{H(X_1,X_2,\ldots,X_n)}{n}$ is a nonincreasing function of n, i.e., monotonicity of entropy per element [10] for stationary stochastic process $\{X_i\}$, which corresponds to the second assumption.

Note we intend to capture a fairly general concept for compression, i.e., if there is any gain in reducing traffic via data aggregation we will refer to it as *compression*. Consider the following example of *data filtering*. Suppose a large sensor network monitoring seismic activity of a region, and when a number of sensors (say 10) have detected strength of seismic waves and reported to local aggregation node. The aggregation node may judiciously choose to send only the averaged wave strength contained in one packet to central data fusion point, achieving 10:1 compression ratio. One may make similar assumptions as provided, for example when the number of seismic wave samples increase, there are more chances for filtering and as a result there always exists reduction in the compressed size per data sample.

Next we consider the decaying rate of $\frac{f(n)}{n}$. Although we are able to reflect a wide range of nonincreasing $\frac{f(n)}{n}$ to our analysis, we consider a harmonically decaying model

for simplicity, i.e., it decreases in an order of $\frac{1}{n}$. It is also motivated by the following example:

Example 1: Suppose the aggregation node at the origin in a straight line collects data from sensors that are located at $d, 2d, 3d, \ldots$, i.e., the sensors are equally apart with some distance d. The sensors generate identically distributed Gaussian random variables having zero-mean unit-variance. The model for correlation is such that the correlation coefficient for a pair of sensors that are distance x apart by

$$\rho(x) = \exp(-cx)$$

where c is a positive constant. If we consider the differential entropy h_n of the total data aggregated from n sensor to the origin (thus there are $n+1$ samples including the data collected by the aggregation node itself), we have that

$$h_n = \frac{1}{2}\log(2\pi e)^{n+1}|K_{n+1}| \text{ nats}$$
$$= \frac{1}{2}\log(2\pi e)^{n+1}(1 - e^{-2cd})^n \text{ nats}$$

where K_n is the covariance matrix for n data samples, and it is easy to show that since K_n is symmetric Toeplitz matrix the determinant is given by the above expression. If we normalize h_n by the differential entropy for a single sample and refer to it as $f(n)$, we have that

$$f(n) = \frac{\log\{2\pi e(1 - e^{-2cd})\}}{\log(2\pi e)} n + 1 \text{ nats}.$$

Clearly we have that

$$\alpha = \lim_{n \to \infty} \frac{f(n)}{n} = \frac{\log\{2\pi e(1 - e^{-2cd})\}}{\log(2\pi e)} < 1$$

and $f(n)/n$ decreases on the order of $\frac{1}{n}$. ∎

In what follows we will set $f(n) = \alpha n + 1$, for example if the number of samples collected to a aggregation node is given by a random variable $\mathcal{N}_0$ the traffic generated after compression is given by $f(\mathcal{N}_0)$.

Finally we model the data generation at each strip as a Bernoulli random variable P with parameter p, i.e., at each round an aggregation node will forward data to sink with probability p. It is equivalent to saying each strip carries an independent 'mark' of P, which is a crude model for sampling spatio-temporal random field. Note p is an important parameter which should be known a priori as will be discussed in Section VI.

B. Depletion Probability

We assume the density of sensors are fixed and some portion of the sensors are elected to aggregation nodes so that one can view the point process of sensors as a mixture of two thinned independent point processes, e.g., sensors perform independent coin flipping to become aggregation nodes. Also we assume without loss of generality $\lambda_0 = 1$. Thus we assume Π_1 is PPP with density λ_1 where λ_1 is less that 1 and thus the density of sensor process Π_0 are $1 - \lambda_1$. Also we replace $\mathcal{N}_0^+$ with $f(\mathcal{N}_0) = \alpha\mathcal{N}_0 + 1$ in the previous derivations to account for the compression gain with varying number of sensors associated with an aggregation node.

Reflecting these modifications we have the following expressions for mean and the variance: the mean is given by $2\pi p \tilde{\mu}(w, \lambda_1, \lambda_2)$ where we define

$$\tilde{\mu}(w, \lambda_1, \lambda_2) := (\alpha + (1 - \alpha)\lambda_1)\mu(w, \lambda_2)$$

and the variance is given by $2\pi p \tilde{\sigma}^2(w, \lambda_1, \lambda_2)$ is given by

$$\tilde{\sigma}^2(w, \lambda_1, \lambda_2) := [(0.28\alpha^2 - 2\alpha + 1)\lambda_1 + \frac{1.28\alpha^2}{\lambda_1}$$
$$+ 2\alpha - 1.56\alpha^2]\sigma^2(w, \lambda_2).$$

Given these we estimate the energy depletion probability after multiple rounds of data collection at a typical sensor. At each round the energy burden at a typical sensor is incurred by data relaying/forwarding from landmarks to nearest sink locations which are independent across rounds. If these burdens are accumulated over sufficient number of rounds, we may apply central limit theorem (CLT) [11] for the energy depletion probability .

Suppose one would like to operate the given network for m rounds. The maximum energy reserve, denoted by b, is parameterized to the average number of data generation for a sensor during m rounds, i.e., mp. Specifically b is a multiple of mp, say $2\pi k$, i.e., $b = 2\pi kmp$ for convenience. If we denote the energy burdens at a typical node by Z_m at mth round, for m sufficiently large, we may apply CLT and the depletion probability is given by

$$P(Z_m > b) \approx \phi\left(z_k(w, \lambda_1, \lambda_2)\right)$$

where we define the following:

$$z_k(w, \lambda_1, \lambda_2) := \sqrt{2\pi mp}\left\{\frac{k - \tilde{\mu}(w, \lambda_1, \lambda_2)}{\tilde{\sigma}(w, \lambda_1, \lambda_2)}\right\}, \quad (17)$$

$$\phi(u) := \frac{1}{2\pi}\int_u^\infty e^{-v^2/2}dv. \quad (18)$$

The objective is to have as many nodes alive after m rounds, i.e., we would like to maximize $z_k(w, \lambda_1, \lambda_2)$ for a given density of mobile sinks per round, i.e., λ_2. Thus we formally define the problem as follows:

$$\begin{aligned}
&\text{maximize: } z_k(w, \lambda_1, \lambda_2),\\
&\text{subject to: } \lambda_2 < \lambda_1 \le 1 = \lambda_0,\\
&\text{variables: } w, \lambda_1,\\
&\text{given: } \lambda_2, \alpha, k, m, p.
\end{aligned}$$

C. Design for Optimal Landmark/Aggregation Densities

Unfortunately z_k depends in a complex way on w and λ_1 and moreover the function is not necessarily convex. However with the aid of numerical evaluation of the objective function we find that, for a wide range of values of k, α and λ_2 the function admits an unique maximizer, i.e., it is sufficient to estimate a local maximizer using the first-order necessary condition for the optimality. Thus we propose the following method to solve *analytically* for an approximate solution.

From the first order necessary condition for the optimality we need to find (w, λ_1) pair that satisfies the following equation:

$$\frac{\tilde{\mu}(w, \lambda_1, \lambda_2)}{\frac{\partial}{\partial s}\tilde{\mu}(w, \lambda_1, \lambda_2)} = \frac{2\tilde{\sigma}^2(w, \lambda_1, \lambda_2)}{\frac{\partial}{\partial s}\tilde{\sigma}^2(w, \lambda_1, \lambda_2)} \quad (19)$$

for $s = w$ and λ_1. Firstly let us fix λ_1 and consider the objective function in terms of w. By inspecting the functions $\mu(w, \lambda_2)$ and $\sigma(w, \lambda_2)$, we see that the optimal w for a given λ_1 will be not large, i.e., at most 10. The rationale is that $\sigma(w, \lambda_2)$ contains the dominating term $5/(8\pi)(w\sqrt{\lambda_2})^{-1}$ since $\lambda_2 \ll 1$, thus if we increase w at a small value will greatly reduce $\sigma(w, \lambda_2)$ and increase the objective function since $\sigma(w, \lambda_2)^{-1}$ is a *multiplicative factor*. However considering the characteristic of $\frac{1}{w}$ the gain will not be high especially considering the 'penalty' of increasing the mean energy burden, i.e., $\mu(w, \lambda_2)$ with increasing w. Moreover if we write the explicit first order necessary condition for λ_1, the optimal λ_1 is a solution to the following equation:

$$m(w, \lambda_2)\left\{\gamma(1-\alpha)\lambda_1^3 + 3.84\alpha^2(1-\alpha)\lambda_1 \right.$$
$$\left. +(2\delta(1-\alpha) - \alpha\gamma)\lambda_1^2 + 1.28\alpha^3\right\} = 1.28k\alpha^2 - k\gamma\lambda_1^2. \quad (20)$$

where $\gamma := 1 - 2\alpha + 0.28\alpha^2$, $\delta := 2\alpha - 1.56\alpha^2$. Note that the solution to the above equation is the intersection point of a quadratic curve and a cubic curve scaled by a positive number $m(w, \lambda_2)$. One can verify that (1) the equation admits an unique solution for $\lambda_1 > 0$, (2) using the conjecture on the optimal w value the sensitivity of the solution is quite small with respect to $m(w, \lambda_2)$ for a wide range of λ_2 values. Thus we would like to fix the scaling $m(w, \lambda_2)$ to a representative value given the expected range of the location of the optimal w. We use the lower limit of w value, i.e., $w = 1$ which corresponds to the baseline scheme, i.e., no traffic spreading if we consider the shape of energy density on a strip. The rationale is that we expect the optimal w to be a small number, thus we choose the smallest possible value of w. The choice of $m(1, \lambda_2)$ turns out to be a *conservative* estimate of λ_1 in that the solution to (20) decreases with increasing w thus it maximizes the maximizer λ_1 if we recall that higher density of aggregation nodes corresponds to lower number of nodes per aggregation node on average.

Thus we solve the cubic equation (20) by setting the scaling to $m(1, \lambda_2)$. Let us denote the solution by λ_1^* if the solution lies within the interval $[0, 1]$ or set $\lambda_1^* = 1$ otherwise. With λ_1^* determined we solve (19) with respect to w, and using second-order approximations for the logarithmic term (19) reduces to a quadratic equation admitting an unique positive solution w^* given by

$$w^* = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad (21)$$

where

$$\tau := \frac{5}{24\pi\sqrt{\lambda_2}}, \quad \beta := \alpha + (1-\alpha)\lambda_1^*,$$
$$a := \frac{\beta}{8}, \quad b := \beta\left(\frac{\tau}{4} - \frac{3}{4}\right) + \frac{k}{2},$$
$$c := \beta\left(\tau^2 - \frac{\tau}{4} + \frac{1}{2}\right) - k\left(\frac{1}{2} + \tau\right).$$

We summarize the procedure as follows:
1) Find $m(1, \lambda_2)$ and solve the cubic equation (20) to find a positive root.
2) Set λ_1^* as the minimum of 1 and the root.
3) Find w^* using (21). The resulting (w^*, λ_1^*) pair is an approximate maximizer.

Note this can be done in an analytical way, however due to complexity we do not specify the root of the cubic equation explicitly. Based on this result in the following section we provide numerical results on specific tradeoffs existing in traffic spreading and compression/aggregation schemes.

VI. NUMERICAL RESULTS AND DISCUSSION

In this section we study numerical results based on the analysis presented in the previous section. First we investigate the impact of the sink density λ_2 and the compression ratio α on the optimal spatial scales for traffic spreading and aggregation. In these results we have assumed the network aims to operate a total of $m = 100$ rounds with the thinning probability p for traffic as 0.1. The maximum reserve k is appropriately scaled to obtain each point in the plot such that the optimal spatial scales yield a depletion probability of 10^{-4}. Fig. 7 (Fig. 8) show the optimal spatial scale w^* (resp. λ_1^*) for traffic spreading (resp. aggregation density) with varying λ_2 and α, Note in the plot the axis for the density of sinks is translated to the average number of sensors associated with a sink: for example 1000 nodes per sink would correspond to $\lambda_2 = 10^{-3}$. The number of sensors per sink ranges from 500 to 2000, and α takes values 0.2, 0.5 and 0.8 which represent excellent, moderate and poor compression performance respectively.

By inspecting Fig. 7 we have the following observations. When we fix the compression ratio the optimal spreading width tends to increase as the number of sensors per sink increases. We see that in Fig. 8 the change in λ_1^* is negligible with the variations in λ_2 when α is fixed, i.e., the traffic generated per AGN is roughly constant. This implies that as λ_2 decreases the traffic travels longer distances, in which case the result in Fig 7 indicates that traffic should spread *more*. This is intuitive since we would see larger number of overlapping strips for smaller sink densities thus reducing the variability by spreading improves the performance.

Let us consider how w^* varies when λ_2 is fixed and we change α. Fig. 7 shows that we should spread more conservatively when the compression worsens. This may be interpreted as a tension between mean and variance of the energy burdens, i.e., the mean plays larger role per sensor in network lifetime with deteriorating compression ratio when
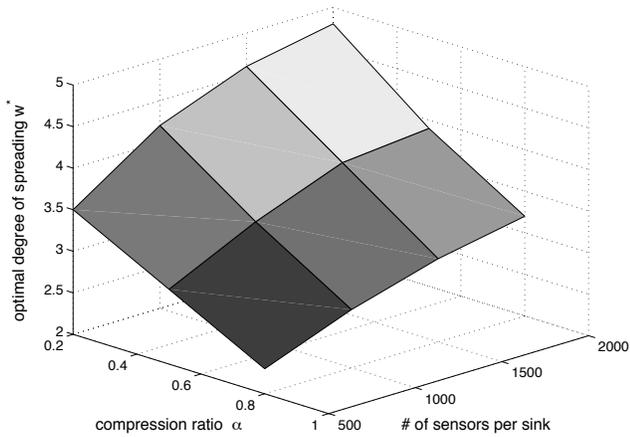
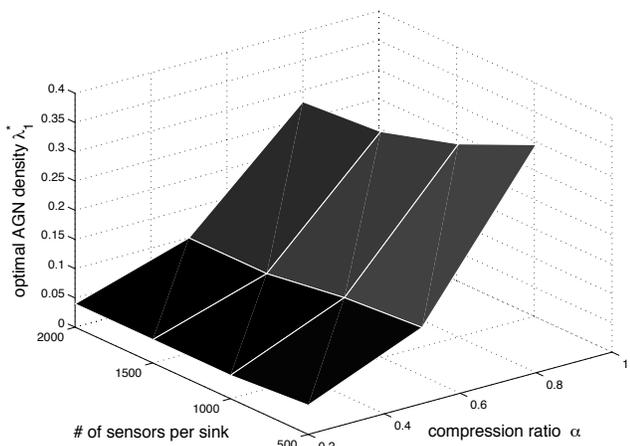Fig. 7. Optimal facet width as a function of compression ratio and sink density.

Fig. 8. Optimal aggregation density as a function of compression ratio and sink density.
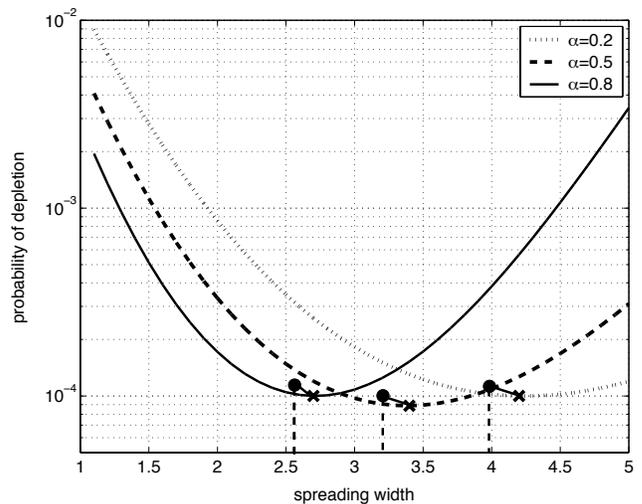

Fig. 9. The probability of depletion when $\lambda_2 = 10^{-3}$ versus spreading width.
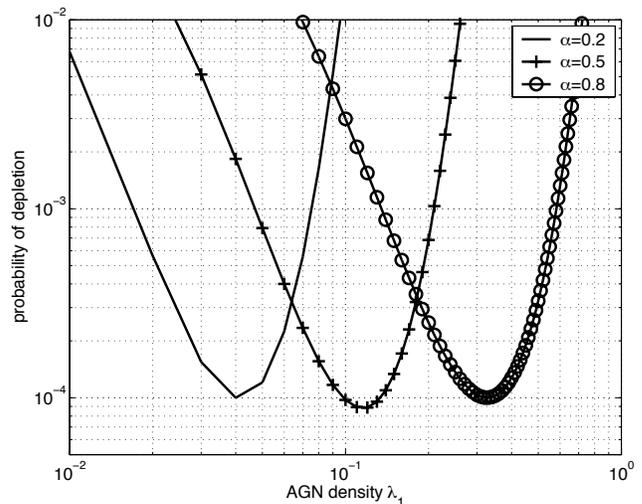
Fig. 10. The probability of depletion when $\lambda_2 = 10^{-3}$ versus AGN density.

the size of sink cells is fixed. However if we consider Fig. 8 at the same time the overall traffic per strip *increases* when the compression performance improves. For example when $\lambda_2 = 10^{-3}$ the optimal pair (w^*, λ_1^*) is given by $(2.7, 0.33)$, i.e., roughly 2 sensors per AGN, when $\alpha = 0.8$ and $(4.2, 0.04)$, i.e., roughly 25 sensors per AGN, when $\alpha = 0.2$. Thus each strip offers roughly 2.4 and 5 units of load when α is given by 0.8 and 0.2 respectively on average. Since the average distance each strip spans is identical, this also implies strips with larger loads should spread *more* given that the average length of strips is fixed. Interestingly the above observations are consistent with those in [5]. Finally in Fig. 8 we see that the degree of aggregation increases, i.e., the densities for AGN decreases with improving compression performance.

In Fig. 9 and Fig. 10 the probability of depletion versus spreading width w and AGN density λ_1, respectively with varying α where $\lambda_2 = 10^{-3}$. Similar to the previous case the value of k in each case is chosen to yield the optimal depletion probability roughly at 10^{-4}. Note Fig. 9 (Fig. 10) is obtained using $z_k(w, \lambda_1^*, \lambda_2)$ (resp. $z_k(w^*, \lambda_1, \lambda_2)$) where we

vary w (resp. λ_1). In Fig. 9 the maximizers and corresponding estimates of depletion probability obtained by using the proposed approximation approach are shown by dots for each α. When we compare these to the true minimum depletion probabilities marked by X each of which is connected by a line to each estimate, we see that the approach yields fairly accurate estimate. In the two figures we clearly observe the tradeoffs associated with spatial scales of spreading and aggregation. By choosing the optimal spatial scale a significant gain is achieved, e.g., about 10 times gain is realized in the depletion probability as compared to a scheme without traffic spreading, i.e., w near 1, see Fig. 9.

We provide some comments on these results. We find these optimal points depend on m, the number of rounds and p, the parameter for traffic intensity per round as one can see in (17). If we fix p and increase m, the optimal scale for spreading tends to decrease and for aggregation tends to

increase. Clearly in the extreme if m grows infinity the law of large numbers dictates that only the mean will dominate performance and the influence of variance will diminish. Thus the optimal points move to the directions to decrease the mean energy burdens: to spread less and more aggregation for further compression. If reducing the mean is the issue, rather than using aggregation based on proximity, one can consider schemes such as exploiting mean-optimal hierarchical organization based on Johnson-Mehl tessellation [12]. Also we note that if $p = 1$ a sensor next to a sink may experience several hundred units of energy burdens if $\lambda_2 = 10^{-3}$ in which case due to large variability the application of CLT may not work well. However this corresponds to the regime of heavily loaded networks as been addressed earlier, i.e., the network will not scale well.

By contrast for finite m and small p, there is room for optimizing network lifetime by choosing appropriate scales for aggregation and/or traffic spreading. Moreover from the results the optimal scale for spreading ranges from $2\sim5$ which is a reasonable degree of spreading in a practical point of view thus enabling efficient realization of the scheme.

VII. Conclusion and Future Work

In this paper we address a fundamental scalability problem for energy-constrained large scale sensor networks based on wireless relaying: the sensors in the vicinity of sinks will incur a much higher energy burden. To mitigate this problem we consider widely used ideas: reducing the traffic on the network through local aggregation/compression and making the sinks mobile. However in most regimes where such schemes are applicable, we argue that network operation can still be optimized to enhance the networks lifetime. The key idea is to spread out aggregated traffic when it is forwarded from AGNs to sinks in order to smooth inhomogeneity in the energy burdens the network will incur.

Interestingly the degree to which traffic should be spread is interwoven in a subtle way with spatial scales on which traffic is aggregated. The tension lies in the following key tradeoffs. Increasing the degree of traffic spreading results in smoother energy burdens but incurs additional overall energy burdens. By contrast, increasing the degree of aggregation reduces the per sensor traffic and thus overall energy burdens, but increases the spatial variability of energy burdens. Moreover these trade-offs are not 'orthogonal' to each other. In this paper we provide a stochastic geometric model to investigate this interaction. By adopting simple models for compression and energy burdens we were able to show how the mean and variance of energy burdens for a typical node relates to the scales for spreading and aggregation. We jointly optimize these scales, and find that while spreading reduces the variance induced by overlaps of 'long' routes towards sinks this only helps to extent that the increased mean energy costs do not counteract the benefit. Similarly the more aggregation the better unless this adversely affects the variability of energy burdens on the network. Using numerical results we can concretely see how these aspects counterbalance each other. For example, even when operated at

the the optimal aggregation density the probability of depletion may be decreased by a factor of 10 if traffic is correctly spread.

Our ongoing work includes study on the following conjecture: if the spreading scales are determined non-uniformly depending on load or distances to sinks, we expect additional improvement in network lifetime. Importantly note the setup in our work is fairly *benign*, i.e., the traffic loads are spatially homogeneous. If there is additional inhomogeneity in node placement, sensing events, variability in data size, etc., we conjecture that reducing variability in traffic becomes more critical issue. On top of that in the proposed facet routing scheme, the performance analysis on overheads and gains associated with routing based on local depletion levels, and devising implementation algorithms are part of our future work.

References

[1] J. Luo and J. Hubaux, "Joint mobility and routing for lifetime elongation in wireless sensor networks," in *Proceedings of IEEE INFOCOM*, 2005, pp. 1735–1746.

[2] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *Proceedings of IEEE INFOCOM*, 2001, pp. 1360–1369.

[3] R. Shah, S. Roy, S. Jain, and W. Brunette, "Data MULEs: Modeling a three-tier architecture for sparse sensor networks," in *IEEE SNPA Workshop*, May 2003.

[4] W. Wang, V. Srinivasan, and K. Chua, "Using mobile relays to prolong the lifetime of wireless sensor networks," in *Proceedings of ACM MOBICOM*, 2005, pp. 270–283.

[5] S. Baek and G. de Veciana, "Spatial energy balancing through proactive multipath routing in multi-hop wireless networks," Tech. Rep., Wireless Networking and Communications Group (WNCG), University of Texas at Austin, TX. [Online] Available: http://www.ece.utexas.edu/~gustavo, 2004.

[6] F. Baccelli and S. Zuyev, "Poisson-voronoi spanning trees with applications to the optimization of communication networks," *INRIA Research Report No.3040*, Nov. 1996.

[7] S. Toumpis and L. Tassiulas, "Packetostatics: Deployment of massively dense sensor networks as an electrostatics. problem," in *IEEE INFO-COM*, 2005.

[8] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and its Applications*, J. Wiley & Sons, Chichester, 1995.

[9] K. Brakke, "Statistics of random voronoi plane tessellations," *preprint*, 2005.

[10] T. Cover and J. Thomas, *Elements of information theory*, John Wiley & Sons, 1991.

[11] L. Heinrich and V. Schmidt, "Normal convergence of multidimensional shot noise and rates of this convergence," *Adv. Appl. Prob.*, vol. 17, pp. 709–730, 1985.

[12] S. Baek, G. de Veciana, and X. Su, "Minimizing energy consumption in large-scale sensor networks through distributed data compression and hierarchical aggregation," *IEEE JSAC Special Issue on Fundamental performance limits of wireless sensor networks*, vol. 22, no. 6, pp. 1130–1140, August 2004.