Computing and Communicating Functions Over Sensor Networks

Arvind Giridhar and P. R. Kumar
University of Illinois at Urbana-Champaign

***Abstract*— In wireless sensor networks, one is not interested in downloading all the data from all the sensors. Rather, one is only interested in collecting from a sink node a relevant function of the sensor measurements. This paper studies the maximum rate at which functions of sensor measurements can be computed and communicated to the sink node.**

It focuses on symmetric functions, where only the data from a sensor is important, not its identity. The results incude the following. (i) The maximum rate of downloading the frequency histogram in a random planar multi-hop network with n nodes is O(1/log n). (ii) A subclass of functions, called type-sensitive functions, is maximally difficult to compute. In a collocated network they can be computed at rate O(1/n), and in a random planar multi-hop network at rate O(1/log n). This class includes the mean, mode, median, etc. (iii) Another subclass of functions, called type-threshold functions, is exponentially easier to compute. In a collocated network they can be computed at rate O(1/log n), and in a random planar multi-hop network at rate O(1/log log n). This class includes the max, min, range, etc.

The results also show the architecture for processing information across sensor networks.

Keywords: Sensor networks, Data fusion, sensor network statistics collection, communication and computation rate.

I. Introduction

Wireless sensor networks are composed of nodes with sensing, wireless communication and computation capabilities. Such networks must carry out not only the task of sensing the environment, but also the task of communicating a relevant summary of the data, through a sequence of messages passed between nodes and computations at nodes, to a designated sink node.

The relevant summary of the data is in general a function of the raw sensor measurements. For example, in environmental monitoring, a relevant statistic of temperature sensor readings $x_1, x_2, \ldots, x_n$ may be the mean temperature $\frac{x_1+x_2+\ldots+x_n}{n}$, or median, or mode. In "alarm" networks, the quantity of interest might be the maximum, $\max\limits_{1\leq i\leq n} x_i$, of the n temperature readings. In general, we formulate the required task of the sensor network as one of making available at a designated sink node a specified function $f(x_1, x_2, \ldots, x_n)$.

The question examined in this paper is how the nodes in the network should cooperate to efficiently compute the desired function $f(x_1, x_2, \ldots, x_n)$, to make it available at a specified sink node. We are interested in particular in the maximum rate, or the maximum frequency, at which the function can be communicated, given the underlying constraints of the wireless medium. If such functions are required to be computed periodically, the reciprocal of this maximum rate is the minimum period at which the desired function of the environment can be sampled. Alternately, given the number of sensors and frequency of measurement generation, the reciprocal of the rate is directly proportional to the link bandwidth needed to carry out the required communication, and as such is directly related to the communication resources consumed per function computation. Finally, the dependence of this maximum rate on the number of sensors is crucial to determining the scalability of the network, a topic of much current interest.

We should emphasize that the most general version of this problem is far more complex than determining the capacity of a wireless network. This is essentially due to the possibility of combining data at intermediate nodes. In other words, this is not simply a matter of communicating bits over a wireless network. The latter problem has been the focus of a number of recent papers. In [1], scaling laws on transport capacity were derived, under a simple collision based interference model. More recently, similar results have been derived in [2] under an information-theoretically more general communication model.

It is possible to construct an information theoretic formulation of the problem of computing functions of sensor measurements. Briefly, it would be one of communicating possibly correlelated sources over a multi-terminal wireless network, of the kind described in [2], to a specified destination, at a desired fidelity with respect to a joint distortion criterion (which will depend on the particular function of interest). Such a problem combines the complexity of source coding of correlated sources with rate distortion (itself an open problem, see [3]), the manifold collaborative possibilities in wireless, the inapplicability of the separation theorem demarcating source and channel coding, and last but not least, the complications introduced by the joint distortion criterion. At the present time, this problem is open, and there has essentially been little or no work

This material is based upon work partially supported by AFOSR under Contract No. F49620-02-1-0217, USARO under Contract Nos. DAAD19-00-1-0466 and DAAD19-01010-465, AFOSR under Contract No. F49620-02-1-0325, and NSF under Contract Nos. NSF ANI 02-21357 and CCR-0325716. Original version submitted December 2, 2003. Revised version submitted July 30, 2004.

Please address all correspondence to the second author at University of Illinois, CSL, 1308 West Main St, Urbana, IL 61801. Email: prkumar@uiuc.edu

in the information theory literature addressing problems of this kind. One special case, a source coding problem involving communicating a function of two variables in a simple two node network with side information at the receiver, has been solved in [4].

A considerably simpler problem, involving computation of functions in a distributed fashion, is the subject of a rich literature on the topic of communication complexity [5]. Problems in communication complexity typically involve two agents, initially knowing numbers X and Y respectively, exchanging bits until both know $f(X,Y)$. The problems considered in this area are typically deterministic, in that no probability of error is permitted.

In this paper, we eschew a completely information theoretic approach. Rather, we adopt a packet based collision model of wireless communication. We suppose that communication between nearby nodes takes place through packets, which may collide. We adopt the *protocol* model of [1] where concurrent transmissions in the vicinity of the receiver collide. This model does not allow information-theoretic collaborative techniques such as interference cancellation, superposition coding, and coherent combining, but is fairly reflective of current technology. We also adopt a deterministic formulation for the problem of function computation, allowing zero error, and consider worst case performance as the relevant metric, just as in the communication complexity literature. However, we will allow for possible efficiencies realizable by block coding, i.e., by combining several function computations in a large epoch (recall that the function computation must be performed periodically). The result is a problem setting that is well tailored to the technology used currently in wireless communication, while at the same time capturing some of the characteristics of multi-hop communication, as well as those of the communication complexity of computing functions. At some points, the arguments in this paper are similar in flavor to those in communication complexity, although no specific results, save for some trivial bounds, are from that literature. The bulk of the proofs are essentially based on counting arguments, while some of the constructive schemes utilized are similar to those introduced in [1].

We now summarize the main results of this paper.
1) The class of symmetric functions, which are invariant to permutations of the arguments, are of interest in many applications. For example, many statistical functions such as the mean, mode, median, frequency histogram, max, and range, are symmetric functions. They embody the data centric paradigm that it is the value of a sensor reading that is important, not the identity of the sensor. We prove a $\Omega(\frac{1}{\log n})$ lower bound on the achievable rate for any symmetric function in multi-hop random planar networks[1]. This is done in Section II.
2) We prove sharp bounds on the order of the achievable rates for two subclasses of symmetric functions, namely type-sensitive functions and type-threshold functions, in collocated networks as well as random planar networks. The mean, median and mode are examples of type-sensitive functions, while the max and range are examples of type-threshold functions. Most other statistical functions also fall within these subclasses. The maximum rate for type-sensitive functions is $\Theta(\frac{1}{n})$ in collocated networks, and $\Theta(\frac{1}{\log n})$ in multi-hop random planar networks. The corresponding quantities for type-threshold functions are $\Theta(\frac{1}{\log n})$ and $\Theta(\frac{1}{\log\log n})$, respectively. These results show that multi-hop communication in a random planar network affords an exponential improvement over the single hop collocated network. They also show that type-sensitive functions are maximally difficult to compute within the class of symmetric functions, while type-threshold functions are exponentially easier. This is the focus of Section III.
3) For a certain class of functions called divisible functions, we characterize the maximum rate in terms of the cardinality of the function's range, for a certain class of graphs. The constructive scheme described in the proof of this result is the basis for the achievability part of the proofs of all the results on random planar networks. This is described in Section II.

II. Model and Problem Statement

Consider a network of n wireless-antenna equipped sensor nodes $1,2,\ldots,n$, along with a sink node s, located on a plane. Let ρ_{ij} denote the distance between two nodes i and j. We assume the *protocol* model [1] for wireless communication, with all nodes sharing a common transmission range r. Node i can successfully transmit a packet to node j if $\rho_{ij} \leq r$, and for every other simultaneously transmitting node k, $\rho_{kj} \geq (1+\Delta)r$. All such successful communications will be assumed to take place at a fixed rate of at most W bits per second.

Each sensor node i periodically makes measurements of the environment, which belong to a fixed finite set $\mathcal{X} = \{1,2,\ldots,|\mathcal{X}|\}$, where $|\cdot|$ denotes cardinality. The nodes in the network must cooperate so as to make known at the sink node s a certain function of the sensor measurements. Let $f_n : \mathcal{X}^n \longrightarrow \mathcal{Y}_n$, be the function of interest, defined for each $n \geq 1$. Abusing notation for the sake of simplicity, we however use $f(\cdot)$ to represent all the functions in this family.

Since sensor measurements are repeatedly generated, the function of interest is thus required to be computed repeatedly. We therefore permit block coding, i.e., strategies that combine several of these function computations (corresponding to long blocks of measurements). Thus, suppose that each sensor has an associated block of N readings, known a-priori. The following is an (almost) exhaustive list of definitions and notation used.[2].
1) $\bar{X} \in \mathcal{X}^{Nn}$ denotes the $n \times N$ matrix of measurements. X_{ij} is the j^{th} measurement of node i. $\underline{X}_i$ is the i^{th} row of

[1]All the results for random planar networks are with probability approaching one as the network size increases to ∞, a statement that is occasionally omitted for brevity.

[2]Variable names are in uppercase, and values in lowercase. Thus, X_{ij} refers to the j^{th} measurement of node i, whereas x may simply be an element of $\mathcal{X}$. Vector variable names are underlined.

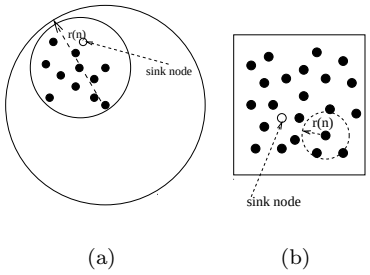

Fig. 1. (a)A collocated network (b)A random planar network

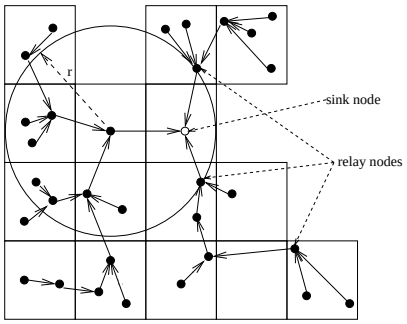

Fig. 2. Grouping nodes in cells.

$\bar{X}$, i.e., the block of N measurements of node i, which it is assumed to know a-priori. $\underline{X}^j$ is the j^{th} column of $\bar{X}$, i.e., the set of j^{th} measurements of the n nodes.
2) Given a k-vector $\underline{x} \in \mathcal{X}^k$, denote $f(\underline{x}) := f(x_1, x_2, \ldots, x_k)$.
3) $\underline{f}(\bar{X}) := [f(\underline{X}^1), f(\underline{X}^2), \ldots, f(\underline{X}^N)]$, the vector of N function values corresponding to the N measurement vectors, or equivalently the measurement matrix $\bar{X}$.
4) $\mathcal{R}(f, n)$ is the range of f, or more properly, f_n. (Recall that although f is a single name given to all the functions in the family, they could have different ranges.)
5) A scheme or strategy $\mathcal{S}_{N,n}$ determines a sequence of message passings between sensors and computations at sensors, which, given any $\bar{X} \in \mathcal{X}^{nN}$, results in $\underline{f}(\bar{X})$ becoming known to the sink.
6) $T(\mathcal{S}_{N,n})$ is the time taken by scheme $\mathcal{S}_{N,n}$, worst case over all $\bar{X} \in \mathcal{X}^{nN}$. The reciprocal $R(\mathcal{S}_{N,n}) := \frac{N}{T(\mathcal{S}_{N,n})}$ is the *rate* of the scheme $\mathcal{S}_{N,n}$.
7) $R^{(n)}_{max}$ is the supremum of rates $R(\mathcal{S}_{N,n})$ over all schemes $\mathcal{S}_{N,n}$ and block-lengths N. This is the *maximum rate*, the quantity of interest in this paper.

The spatial graph $G^{(n)}$ associated with the wireless network consists of the set of n nodes, with edges between nodes that are within a distance r of each other. The degree of the graph, i.e., the maximum of the degrees of nodes in the graph, is denoted by $d(G^{(n)})$.

There are two network topologies of interest.

Collocated networks. These are networks with $\rho_{ij} \leq r$ for all i, j, so every transmission is heard by all nodes (Figure 1(a)).

Random planar networks. The n nodes along with the sink node s are uniformly and independently distributed on a unit square (Figure 1(b)). The common range $r(n)$ of all the n nodes is so chosen that, by using multi-hop communication, the graph is connected. The following lemma shows how to so choose $r(n)$ that the graph is connected, and shows that the order of the degree that results is $O(\log n)$ with high probability.

Lemma 1: For random planar networks, if range $r(n) = \sqrt{\frac{2 \log n}{n}}$, then $\lim_{n\to\infty} Pr[G^{(n)} \text{ is connected}] = 1$, and $\lim_{n\to\infty} Pr[d(G^{(n)}) \leq c \log n] = 1$, for some $c > 0$.

Proof: The first assertion is from [6]. To prove the second, consider a square tessellation $\mathcal{T}_n$ of the unit square into $(\lceil \frac{1}{2.5r(n)} \rceil)^2$ small squares. Each square has side $> 2r(n)$. Thus, any circle of radius $r(n)$ is covered by at most 4 squares. Applying Lemma 3.1 in [7],
$\lim_{n\to\infty} Pr[\text{Some square has more than } C \log n \text{ nodes}] = 0$, for some constant $C > 0$. Thus, $\lim_{n\to\infty} Pr[d(G^{(n)}) \leq 4C \log n] = 1$. ■

III. Divisible Functions

A trivial upper bound on $R^{(n)}_{max}$ can be derived as follows. Since the sink node can receive at most W bits/sec, and since representing the value of $f(\cdot)$ requires $\log |\mathcal{R}(f, n)|$ bits for a network of n nodes (all the logs in this paper are to the base 2),

$$R^{(n)}_{max} \leq \frac{W}{\log |\mathcal{R}(f, n)|}. \tag{1}$$

In this section we will consider a special class of functions called *divisible* functions. These are functions which can be computed in a divide-and-conquer fashion, and are defined in the sequel.

Denote the set $\{1, 2, \ldots, n\}$ by $[n]$. Given $\underline{x} \in \mathcal{X}^n$, and a subset $S = \{i_1, i_2, \ldots, i_k\} \subset [n]$, where $i_1 < i_2 < \ldots < i_k$, denote by $\underline{x}_S$ the vector $[x_{i_1}, x_{i_2}, \ldots, x_{i_k}]$.

Definition: A function $f : \mathcal{X}^n \longrightarrow \mathcal{Y}_{(n)}$ is said to be a *divisible function* if:

1) $|\mathcal{R}(f, n)|$ is non-decreasing in n.
2) Given any partition $\Pi(S) = \{S_1, S_2, \ldots, S_j\}$ of $S \subset [n]$, there exists a function $g^{\Pi(S)}$, such that for any $\underline{x} \in \mathcal{X}^n$,

$$f(\underline{x}_S) = g^{\Pi(S)}(f(\underline{x}_{S_1}), f(\underline{x}_{S_2}), \ldots, f(\underline{x}_{S_k})),$$

The following result shows that on a certain class of graphs, the simple upper bound in (1) is in fact achievable for divisible functions.

Theorem 1: Let $f(\cdot)$ be a divisible function. Suppose that $G^{(n)}$ is connected, and $d(G^{(n)}) \leq k_1 \log |\mathcal{R}(f, n)|$, for some $k_1 > 0$. Then, the maximum rate $R^{(n)}_{max}$ for computing $f(\cdot)$ satisfies

$$R^{(n)}_{max} \geq \frac{c_1(k_1, \Delta, W)}{\log |\mathcal{R}(f, n)|},$$

where $c_1(k_1, \Delta, W)$ is a constant depending on k, Δ, and W, but independent of n, the particular network configuration, and $f(\cdot)$.

Proof: Consider a tessellation of the plane into square cells of side $r/\sqrt{2}$ (see Figure 2). Define the *cell graph* on the set

of non-empty cells as vertices, with two cells being adjacent if there are two nodes (belonging to $G^{(n)}$) within each cell which are adjacent in $G^{(n)}$.

Designate the cell with the sink node as the root, and consider a rooted spanning tree of the cell graph. There is a naturally defined partial order on the set of cells, corresponding to depths in this rooted tree. Let δ_{max} be the maximum depth of the tree, and $\delta(c)$ the depth of cell c. In each cell c, designate as the *relay node* a node u which is adjacent to a node v in the parent (in the rooted tree of the cell graph) of c, and designate v as the *relay parent* of u. Designate the sink as the relay node of the cell containing it. There is thus one relay node (picked out of possibly multiple choices) and possibly relay parents in each cell.

For each node $u \in G^{(n)}$, define the *descendant set* D_u as follows:
1) If u is a relay node of a cell c, D_u is the set of indices of all nodes in $G^{(n)}$ that either belong to c or to descendants of c.
2) Otherwise, if u is the relay parent of relay nodes $u_1, \ldots, u_\ell$, $D_u := \bigcup_i D_{u_i} \bigcup \{u\}$.
3) Otherwise, $D_u := u$.

Consider the following scheme: For a T_1 to be specified later, between times jT_1 and $(j+1)T_1$, the following transmissions take place for each cell c. Let $m := j - 2(\delta_{max} - \delta(c))$. If $m \leq 0$ or $m > N$, there are no transmissions scheduled for c in $[jT_1, (j+1)T_1]$. Otherwise,
1)Each non-relay node v in c transmits $f(\underline{X}^m_{D_v})$ to the relay node of c.
2) If $m > 1$ and c does not contain the sink node s, the relay node u of c transmits $f(\underline{X}^{m-1}_{S_u})$ to its relay parent.

It is clear that this scheme, if feasible, communicates $f(\underline{X}^i)$, for each $1 \leq i \leq N$, to the sink node. To prove that it is feasible, two properties must be established. First, we will need to prove that the function values to be transmitted are known to the corresponding nodes. Specifically, we will need to prove that at time $(m+2(\delta_{max}-\delta(c)))T_1$, each non-relay node v knows $f(\underline{X}^m_{D_v})$, and each relay node u knows $f(\underline{X}^{m-1}_{D_u})$. This is easily proved by induction on the epoch number j. It is trivially true for nodes that are neither relays nor relay parents, since their descendent sets are singletons. If v is a relay parent of relay nodes $u_1, \ldots, u_\ell$,

$$f(\underline{X}^m_{D_v}) = g^{\Pi(D_v)}(f(\underline{X}^m_{D_{u_1}}), \ldots, f(\underline{X}^m_{D_{u_\ell}}), f(X^m_{\{v\}})).$$

Also, the u_i's belong to cells of depth one less than the cell c containing v, so between $(j-1)T_1$ and jT_1, by the induction assumption the $j-1-2(\delta_{max}-(\delta(c)-1))-1 = m$ indexed transmission is made from each u_i to v; so the arguments of $g^{\Pi(D_v)}(\cdot)$ in the RHS are known to v at time jT_1. Similar reasoning applies to relay nodes.

Second, we will need to prove that the transmissions can be feasibly scheduled. We first note that each cell has a bounded number of children. Also, each cell has a bounded number of *interfering neighbors*, where two cells c_1 and c_2 are interfering neighbors if there exist nodes x and y in c_1 and c_2 respectively with $\rho_{xy} < (1+\Delta)r$. Both bounds can be obtained by simple geometric arguments; see [1]. Let k_2 be the bound on the number of interfering neighbors, and k_3 the bound on the number of children. Note that k_3 is a constant, while k_2 depends only on Δ; there is no dependence on r, n, or the particular placement.

Applying a simple graph coloring argument (see [1]), it follows that there exists a schedule such that each cell receives one out of every $(1+k_2)$ slots to transmit. In an interval of length T_1, each cell can thus be allotted $\frac{T_1}{1+k_2}$ slots. Each relay parent node v requires at most $\log|\mathcal{R}(f,n)|$ bits to communicate $f(\underline{X}^m_{D_v})$, and there are at most k_3 relay parent nodes per cell (since each cell contains at most one relay parent node per child cell). The same applies for a relay node. The non-relay nodes require at most $\log|\mathcal{X}|$ bits, and there are at most $k_1 \log|\mathcal{R}(f,n)|$ number of them in a single cell. This follows from the fact that any two nodes within a cell of side $r/\sqrt{2}$ are within a distance r of each other. Thus the degree of any node in a cell is greater than or equal to the number of nodes in the cell, and by assumption $d(G^{(n)}) \leq k_1 \log|\mathcal{R}(f,n)|$. The total number of bits required is upper-bounded by $(k_3 + 1+\log|\mathcal{X}|)k_1 \log|\mathcal{R}(f,n)|$. Consequently, the transmissions can be feasibly scheduled if

$$T_1 \geq \frac{k_2+1}{W}(k_3 + 1 + \log|\mathcal{X}|)k_1 \log|\mathcal{R}(f,n)|. \quad (2)$$

In particular, there is a schedule with equality in (2). We can now upper-bound the total number of time slots required. The transmissions are complete at time jT_1 for $j - 2(\delta_{max} - \delta(c)) > N$ for every cell c. Therefore,

$$\begin{aligned} R^{(n)}_{max} &\geq \lim_{N\to\infty} \frac{N}{N+2\delta_{max}} \frac{1}{T_1} \\ &= \frac{W}{(k_2+1)(k_3+1+\log|\mathcal{X}|)k_1 \log|\mathcal{R}(f,n)|}. \quad \blacksquare \end{aligned}$$

We now describe applications of Theorem 1 to several natural functions of interest.

1) The data downloading problem: Consider the identity function $f(\underline{x}) = \underline{x}$, which corresponds to the sink node downloading all the raw measurements of all the sensors. This is a divisible function, with $|\mathcal{R}(f,n)| = |\mathcal{X}|^n$. Thus $\log|\mathcal{R}(f,n)|$ is linear in n, which implies that the degree condition of Theorem 1 is true for any connected graph $G^{(n)}$. Therefore, Theorem 1 proves the existence of a scheme to communicate $f(\cdot)$ at rate $O(\frac{1}{n})$. This problem is in fact the *many-to-one* maximum throughput problem, considered in [8]. An application of Theorem 1 thus gives a more general result than proved in [8].

2) Computing the frequency histogram of the sensor measurements: Theorem 1 can also be applied to the problem of computing the frequency histogram, or *type-vector*, of the sensor readings. The type-vector $\underline{\tau}(\underline{x})$ of the vector $\underline{x} \in \mathcal{X}^n$ is defined as $\underline{\tau}(\underline{x}) = [\tau_1(\underline{x}), \tau_2(\underline{x}), \ldots, \tau_{|\mathcal{X}|}(\underline{x})]$, where $\tau_i(\underline{x}) := |\{j : x_j = i\}|$, the number of occurrences of i in $\underline{x}$.

It can be verified that $\underline{\tau}(\cdot)$ is a divisible function. The number of type-vectors of a vector of length n where

each component has alphabet size $|\mathcal{X}|$, is the number of nonnegative integer solutions to the equation $y_1 + y_2 + \ldots + y_{|\mathcal{X}|} = n$, which is given by $\binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}$ [9]. Thus $|\mathcal{R}(\underline{\tau})| = \binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}$. Simple combinatorial arguments (see [3], Chapter 12) can be used to prove that $(\frac{n}{|\mathcal{X}|})^{|\mathcal{X}|} < \binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1} < (n+1)^{|\mathcal{X}|}$. It follows that the maximum rate of computing $\underline{\tau}(\cdot)$ is $\Theta(\frac{1}{\log n})$, if $d(G^{(n)}) = O(\log n)$. Lemma 1 shows that the range $r(n)$ can be chosen so that the latter is true in a random planar network with high probability as the network size grows. Thus, we have the following result.

Theorem 2: In a random planar network, under the choice of $r(n)$ given by Lemma 1, the maximum rate $R^{(n)}_{max}$ for computing the type-vector $\underline{\tau}(\cdot)$, satisfies $\lim_{n\to\infty} Pr[R^{(n)}_{max} = \Theta(\frac{1}{\log n})] = 1$.

3) Symmetric functions: Consider the class of *symmetric functions*, or functions which are invariant with respect to permutations of their arguments, i.e.,

$$f(x_1, ...x_n) = f(\sigma(x_1, ...x_n)), \text{for all permutations } \sigma.$$

Symmetric functions are the main class of functions of interest in this paper. From an applications standpoint, many natural functions of interest, including most statistical functions, belong to this class. Furthermore, symmetric functions embody the data centric paradigm [10], where it is the data generated by a sensor that is of primary importance, rather than its identity. Note that a symmetric function $f(\underline{x})$ depends on $\underline{x}$ only through its type vector $\underline{\tau}(\underline{x})$. For any type-vector $\underline{\tau} \in \mathcal{Z}_+^{|\mathcal{X}|}$, where $\mathcal{Z}_+$ denotes the set of non-negative integers, let $f'(\underline{\tau})$ denote the value of $f(\underline{x})$ for any $\underline{x}$ with $\underline{\tau}(\underline{x}) = \underline{\tau}$. This is convenient for studying symmetric functions with different numbers of arguments, and also allows us to cleanly state scaling laws.

As a consequence of their dependence only on the type-vector, it follows that the maximum rate of computing the type-vector at the sink node is a lower bound on the maximum rate for any symmetric function. Thus, for any symmetric function, $R^{(n)}_{max} = \Omega(\frac{1}{\log n})$ with high probability in random planar networks.

IV. Communicating Symmetric Functions

An obvious strategy to compute any symmetric function is to simply communicate the entire type or frequency-histogram. Is it possible to do better? This is the focus of the rest of the paper.

It may be expected that the answer to the above question is different for different functions within the class of symmetric functions. Below, we define two disjoint subclasses of symmetric functions. It turns out the order of the maximum rate can be characterized uniformly over each subclass, in both collocated and random planar networks. This is not a complete characterization of the rates for all symmetric functions, since there are symmetric functions that belong to neither subclass. However, most statistical functions of interest do fall within these subclasses.

Type-sensitive functions. A symmetric function $f(\cdot)$ is said to be *type-sensitive* if there exists some γ with $0 < \gamma < 1$, and an integer $\bar{n}$, such that for $n \geq \bar{n}$, and any $j \leq n - \lceil \gamma n \rceil$, given any subset $\{x_1, x_2, \ldots, x_j\}$, there are two subsets of values $\{y_{j+1}, y_{j+2}, \ldots, y_n\}$ and $\{z_{j+1}, z_{j+2}, \ldots, z_n\}$, such that

$$f(x_1, \ldots, x_j, z_{j+1}, \ldots z_n) \neq f(x_1, \ldots, x_j, y_{j+1}, \ldots y_n).$$

Note that if the above is true for $j = n - \lceil \gamma n \rceil$, it is automatically true for all lower values of j.

Type-threshold functions. A symmetric function $f(\cdot)$ is said to be *type-threshold* if there exists a non-negative $|\mathcal{X}|$-vector $\underline{\theta}$, called the *threshold vector*, such that $f(\underline{x}) = f'(\underline{\tau}(\underline{x})) = f'(min(\underline{\tau}(\underline{x}), \underline{\theta}))$, for all $\underline{x} \in \mathcal{X}^n$, with *min* signifying element-wise minimum.

Intuitively, the value of a type-sensitive function cannot be determined if a large enough fraction of the arguments are unknown, whereas the value of a type-threshold function can be determined by a fixed number of known arguments. It is easy to see that the two subclasses are disjoint: Suppose $f(\cdot)$ is a type-threshold function. As $n \to \infty$ the fraction $\frac{\sum_i \theta_i}{n} \to 0$, and yet $\sum_i \theta_i$ input values can fix the value of $f(\cdot)$. Therefore, a type-threshold function cannot be a type-sensitive function.

Examples. Important examples of symmetric functions fall within these two subclasses.

1) The *mode* of $\underline{x}$ (i.e., the value which occurs the most frequently) is a type-sensitive function. If more than half the x_i's are unknown, the mode is undetermined.

2) The *mean* of $\underline{x}$ computed to within a finite precision is a type-sensitive function. So are the median and standard deviation. (The mean to exact precision is also a type-sensitive function; in fact it is a divisible function.)

3)The *max* function, i.e., the maximum among the x_i's, is a type-threshold function, with a threshold vector $[1, 1, \ldots, 1]$. The *min* function and the *range* function $(max_i\ x_i - min_i\ x_i)$ are as well.

4) The k^{th} *largest value* among the x_i's is a type-threshold function, again with a threshold vector $[1, 1, \ldots, 1]$.

5) The *mean of the k largest values* is a type-threshold function, with threshold vector $[k, k, \ldots, k]$.

6) The indicator function $I\{x_i = k, for\ some\ i\}$ is a type-threshold function with threshold vector $[0, 0, \ldots, 0, 1, 0 \ldots 0]$ (a 1 in the k^{th} position).

7) There are however symmetric functions which are neither type-sensitive nor type-threshold. Let $|\mathcal{X}| = 2$ and consider the function $f(\underline{x})$ which is equal to 1 if the number of 1's in $\underline{x}$ is no smaller than $\lceil \sqrt{n} \rceil$, and 0 otherwise. This is not a type-threshold function, since knowing a constant (i.e., independent of n) number of values cannot determine the value of the function. On the other hand, there is no $c < 1$ such that $cn > n - \lceil \sqrt{n} \rceil$ for arbitrarily large n, so $f(\cdot)$ is not type-sensitive either.

A. Collision free strategies in collocated networks

Theorem 1 is essentially useful for functions with large ranges, for which the bottleneck condition in (1) provides a good upper bound. However, when the cardinality of the range is small, which is true of many type-sensitive

as well as type-threshold symmetric functions, this simple bound is usually not good enough. More sophisticated arguments are then necessary to bound the performance of any scheme. In order to do this, it is first necessary to more precisely characterize the class of allowable strategies, which was not required for the previous results. This is a complicated issue, due to the degrees of freedom inherent in multi-hop communication, coupled with the complexity of distributed computation of a function.

We begin by focusing on collocated networks, for which characterizing the class of strategies is easier than in other multi-hop scenarios. Without loss of generality, suppose that time is slotted, and that one bit is transmitted in each slot.

Collision free strategies: A *collision free strategy* (CFS) $\mathcal{S}_{N,n}$ to communicate $\underline{f}(\bar{X}), \bar{X} \in \mathcal{X}^{nN}$ (recalling the notation from Section II) with block-length N and time $T(\mathcal{S}_{N,n})$, consists of the following:

1) A set of functions $\phi_m : \{0,1\}^{m-1} \longrightarrow \{1,2,\ldots,n\}$, $2 \leq m \leq T(\mathcal{S}_{N,n})$, and $\phi_1 \in \{1,2,\ldots,n\}$. The function $\phi_m(\cdot)$ picks the node to transmit at time m.

2) A set of functions $\psi_m : \mathcal{X}^N \times \{0,1\}^{m-1} \longrightarrow \{0,1\}$, $1 \leq m \leq T(\mathcal{S}_{N,n})$. The m^{th} transmission Z_m is (recursively) defined as follows: $Z_1 := \psi_1(\underline{X}_{\phi_1})$, $Z_m := \psi_m(\underline{X}_i, Z_{m-1}, \ldots, Z_1)$, for $1 < m \leq T(\mathcal{S}_{N,n})$, where $i = \phi_m(Z_{m-1}, Z_{m-2}, \ldots, Z_1)$.

3) A decoding function $\xi : \{0,1\}^{T(\mathcal{S}_{N,n})} \longrightarrow \mathcal{Y}^N$, such that $\underline{f}(\bar{X}) = \xi(Z_1, Z_2, \ldots, Z_T(\mathcal{S}_{N,n}))$.

The functions $\phi_m(\cdot)$ and $\psi_m(\cdot)$ (which can be thought of as the analog of a codebook [3]) are known to all nodes a-priori. The maximum rate is given by

$$R^{(n)}_{max} = \limsup_{N\to\infty} \sup_{\mathcal{S}_{N,n}\in CFS} \frac{N}{T(\mathcal{S}_{N,n})}.$$

We point out some characteristics of the class defined above. The node designated to transmit at time m is fixed by the value of $\phi_m(Z_{m-1}, Z_{m-2}, \ldots, Z_1)$, which can be computed by all the nodes. The medium access problem is thus resolved in a distributed but collision-free fashion. However, the transmission itself can depend only on what the sensor itself "knows," which knowledge is comprised by its own data vector as well as all the previous transmissions. Another consequence is that the identity of the transmitting node is automatically known to all. Therefore, addressing is not an issue.

Is this notion of strategy general enough? The strategies described above are required to explicitly avoid collisions. However, allowing a more general class of strategies where more than one node can decide to transmit in a particular slot may create packet collisions. Then, one would have to consider two possibilities in modeling: Either a collision is allowed to convey information, or it is not. By the latter is meant that the individual node and sink node decisions (i.e., the functions $h(\cdot)$ and $g(\cdot)$) can only depend on successful transmissions. In that case, a collision would simply be seen as pure noise and would provide no information. The only way to prevent collisions for all input vectors is by fixing a unique node to transmit (via the function f). We confine ourselves to this mode of operation for the following reason. Given that the basic model is that of a "packet capture" model of communication, it makes sense to allow only successfully decoded packets to convey information, since in reality packet losses could be caused by factors other than collisions. One could regard collided packets as noise corrupted packets heard by all receivers, thus providing no information on whether a transmission, or multiple transmissions, or no transmission, took place.

B. Type-sensitive functions in collocated networks

The following result shows that the maximum rate for computing a type-sensitive function in a collocated network is of the same order as communicating the entire data. In other words, type-sensitive functions are maximally difficult to compute, up to order, in the class of all functions.

Theorem 3: The maximum rate for computing a type-sensitive function in a collocated network, using any CFS, is $\Theta(\frac{1}{n})$.

Proof. For brevity, we provide the proof for $|\mathcal{X}| = 2$. It can be extended to the non-binary case as well, but the proof is somewhat cumbersome.

The strategy of all nodes communicating all their readings in round-robin fashion to the sink node has rate $\frac{1}{n}$. Thus, we only need to establish the upper bound $O(\frac{1}{n})$.

Given a strategy $\mathcal{S}_{N,n}$, we specify for each node i an "uncertainty" subset $U_i \subset \mathcal{X}^N$, such that any measurement matrix $\bar{X}$ with each $\underline{X}_i \in U_i$ produces the same $T(\mathcal{S}_{N,n})$ transmissions. This defines a subset of all measurement matrices $\mathcal{U} = \{\bar{x} \in \mathcal{X}^{nN} : \underline{x}_i \in U_i, 1 \leq i \leq n\}$. Now, if two such measurement matrices in $\mathcal{U}$ have different values under $f(\cdot)$, clearly the strategy $\mathcal{S}_{N,n}$ cannot work.

We specify each U_i by defining a sequence of subsets U_i^k, each one containing the next, with U_i being the final set of the sequence.

Let $i_1 = \phi_1$ be the first node picked to transmit. Pick as $U^1_{i_1}$ the larger of the two subsets of $\mathcal{X}^N$, mapped to 1 and 0 respectively by $\psi_1(\cdot)$. Let z_1 be the corresponding (1 or 0) transmission. Then let $i_2 = \phi_2(z_1)$, and repeat as before. At the m^{th} step , $i_m = \phi_m(z_{m-1}, z_{m-2}, \ldots, z_1)$. Let $k_1, k_2, \ldots, k_n$ be the number of transmissions made by nodes $1, 2, \ldots, n$ respectively ($\sum_i k_i = m-1$). Then, $U^{k_{i_m}+1}_{i_m}$ is chosen as the larger of the two subsets of $U^{k_{i_m}}_{i_m}$ mapping to 0 and 1, via the function $\psi_m(\cdot, z_{m-1}, \ldots, z_1)$, and z_m is the corresponding transmission. After $T(\mathcal{S}_{N,n})$ transmissions, let $U_1, U_2, \ldots, U_n$ be the "uncertainty" subsets corresponding to nodes $1, 2, \ldots, n$ respectively. The implication is that for any set of $\underline{X}_i$'s with $\underline{X}_i \in U_i$ for each i, $[Z_1, Z_2, \ldots, Z_{T(\mathcal{S}_{N,n})}] = [z_1, z_2, \ldots, z_{T(\mathcal{S}_{N,n})}]$.

Therefore, all $\bar{X} \in U_1 \times U_2 \times \ldots \times U_n$ produce the same sequence of transmissions under $\mathcal{S}_{N,n}$. Furthermore, since each $|U^k_i| \geq \frac{|U^{k-1}_i|}{2}$, and $|U^0_i| = 2^N$, $|\mathcal{U}| = \prod_i |U_i| \geq 2^{Nn - T(\mathcal{S}_{N,n})}$.

We now lower-bound the number of elements in $\mathcal{U}$. Call an entry index (i,j) *undetermined* if there are two vectors in U_i with j^{th} entries equal to 0 and 1, respectively.

Conversely, call index (i,j) *determined* if all vectors in U_i have the same j^{th} entry. Now, the definition of a type-sensitive function restricts the number of undetermined entry indices per column to be no more than $\lfloor \gamma n \rfloor$, where $0 < \gamma < 1$ is the fraction corresponding to f (see the above definition of type-sensitive functions). To see this, suppose that it is not true, and there are at least $\lceil \gamma n \rceil$ undetermined entry indices in the j^{th} column. Then, the number k of determined indices is no greater than $n - \lceil \gamma n \rceil$. Let the corresponding measurements (consisting of the value of the fixed entries in the corresponding k uncertainty subsets) be $x_1, x_2, \ldots, x_k$. Then, by definition of type-sensitive functions there are sequences $\{y_k, y_{k+1}, \ldots, y_n\}$ and $\{w_k, w_{k+1}, \ldots w_n\}$ such that

$$f(x_1, \ldots, x_k, w_{k+1}, \ldots w_n) \neq f(x_1, \ldots, x_k, y_{k+1}, \ldots y_n).$$

Furthermore, by the definition of undetermined entry indices, there exist vectors $\underline{y}_1, \underline{w}_1 \in U_1, \underline{y}_2, \underline{w}_2 \in U_2, \ldots$, and $\underline{y}_n, \underline{w}_n \in U_n$, such that the set of j^{th} indices of $\{\underline{y}_i : 1 \leq i \leq n\}$ is the same as the set $\{x_1, \ldots, x_k, y_{k+1}, \ldots, y_n\}$, and that of $\{\underline{w}_i : 1 \leq i \leq n\}$ is the same as the set $\{x_1, \ldots, x_k, w_{k+1}, \ldots, w_n\}$. The measurement matrices composed of each of these sets of vectors as rows will produce the same sequence of transmissions, which will lead to an error, since the j^{th} function value is not the same in the two matrices.

The above argument shows that for a correct strategy, there cannot be more than $\lfloor \gamma n \rfloor$ undetermined entry indices in any column. Thus, there cannot be more than $N\lfloor \gamma n \rfloor$ undetermined entry indices over all the columns. Consequently, the maximum possible number of matrices in $\mathcal{U}$ is $2^{N\lfloor \gamma n \rfloor}$. Combining the above two arguments,

$$Nn - T(\mathcal{S}_{N,n}) \leq N\lfloor \gamma n \rfloor).$$

Thus, $T(\mathcal{S}_{N,n}) \geq (1-\gamma)Nn$, and so, $R_{max}^{(n)} = \limsup_{N\to\infty} \sup_{\mathcal{S}_{N,n}} \frac{N}{T(\mathcal{S}_{N,n}} \leq \frac{1}{\gamma n}$. ■

C. Type-threshold functions in collocated networks

Now we turn to the class of type-threshold functions. We say a symmetric function $f(\cdot)$ is *non-constant*, if for any n, there exist $\underline{x}, \underline{y} \in \mathcal{X}^n$ such that $f(\underline{x}) \neq f(\underline{y})$.

Lemma 2: Let $f(.)$ be a non-constant type-threshold function with threshold vector $\underline{\theta}$. Then, there exists a non-negative type-vector $\underline{\eta} = [\eta_1, \eta_2, \ldots, \eta_{|\mathcal{X}|}]$ such that
1) $\eta_i \leq \theta_i$ for all i, with equality at some i^*.
2) For some $j^* \neq i^*$,

$$f'(\eta_1, \ldots, \eta_{j^*}, \ldots, \eta_{|\mathcal{X}|}) \neq f'(\eta_1, \ldots, \eta_{j^*}+1, \ldots, \eta_{|\mathcal{X}|}).$$

3) For all $k \geq \eta_{j^*} + 1$,

$$f'(\eta_1, \ldots, k, \ldots, \eta_{|\mathcal{X}|}) = f'(\eta_1, \ldots, \eta_{j^*}+1, \ldots, \eta_{|\mathcal{X}|}).$$

Proof. Take $n > \sum_{i=1}^{|\mathcal{X}|} \theta_i$. Note that for any $\underline{x} \in \mathcal{X}^n$, there is at least one entry i at which $\tau_i(\underline{x}) > \theta_i$. Since $f(\cdot)$ is non-constant, there exist $\underline{x}, \underline{y} \in \mathcal{X}^n$ such that $f'(min(\underline{\tau}(\underline{x}), \underline{\theta})) \equiv f(\underline{x}) \neq f(\underline{y}) \equiv f'(min(\underline{\tau}(\underline{y}), \underline{\theta}))$. Construct a sequence of type-vectors starting from $\underline{\tau}^{(1)} = \underline{\tau}(\underline{x})$ and ending at $\underline{\tau}^{(j)}$ with $min(\underline{\theta}, \underline{\tau}^{(j)}) = min(\underline{\theta}, \underline{\tau}(\underline{y}))$, as follows. Pick an entry of $\underline{\tau}(\underline{x})$, say i, at which $\tau_i(\underline{x}) \neq \tau_i(\underline{y})$, and either $\tau_i(\underline{x}) < \theta_i$ or $\tau_i(\underline{y}) < \theta_i$. If no such entry exists then $f(\underline{x}) = f(\underline{y})$ by definition of a type-threshold function, which contradicts our assumption. Without loss of generality assume $\tau_i(\underline{x}) < \tau_i(\underline{y})$. Pick another entry j at which $\tau_j(\underline{x}) > \theta_j$. Let $\underline{\tau}^{(2)}$ be the type-vector obtained from $\underline{\tau}^{(1)}$ by increasing $\tau_i^{(1)}$ by one and decreasing $\tau_j^{(1)}$ by one. The entries of $\underline{\tau}^{(2)}$ add up to n, and the Hamming distance between $min(\underline{\theta}, \underline{\tau}^{(2)})$ and $min(\underline{\theta}, \underline{\tau}(\underline{y}))$ is one less than that between $min(\underline{\theta}, \underline{\tau}(\underline{x}))$ and $min(\underline{\theta}, \underline{\tau}(\underline{y}))$. Repeat this step until the Hamming distance reaches zero.

Since $f'(\underline{\tau}^{(j)}) = f'(\underline{\tau}(\underline{y}))$, $f'(\cdot)$ must change value at some point in the sequence, say i. Let $\underline{\nu} = min(\underline{\theta}, \underline{\tau}^{(i)})$. Clearly, for some i^*, $\tau_{i^*}^{(i)} > \theta_{i^*}$, and so $\nu_{i^*} = \theta_{i^*}$. Also, for some $j^* \neq i^*$,

$$f'(\nu_1, \ldots, \nu_{j^*}, \ldots, \nu_{|\mathcal{X}|}) \neq f'(\nu_1, \ldots, \nu_{j^*}+1, \ldots, \nu_{|\mathcal{X}|}).$$

Let k be the largest nonnegative number such that $f'(\underline{\nu} + ke_{j^*}) \neq f'(\underline{\nu} + (k+1)e_{j^*})$, where e_{j^*} is the vector with j^*th entry 1 and all other entries 0. Clearly, $0 \leq k \leq \theta_{j^*} - \nu_{j^*} - 1$. Thus, $\underline{\eta} = \underline{\nu} + ke_{j^*}$ is the vector we seek. ■

The following result sharply characterizes the maximum rate for computing type-threshold functions.

Theorem 4: The maximum rate for computing a non-constant type-threshold function in a collocated network, using any CFS, is $\Theta(\frac{1}{\log n})$.

Proof: We first prove the result for the case $|\mathcal{X}| = 2$ and for the max function $f(x_1, x_2, \ldots, x_n) = max\{x_i : 1 \leq i \leq n\}$.

We first prove the achievability, by providing a sequence of CFS's $\mathcal{S}_{l(n+1),n}$, for $l = 1, 2, 3 \ldots$, asymptotically achieving the rate $\Omega(\frac{1}{\log n})$.

Take block-length $N = l(n+1)$. Let the number of $1's$ in vector $\underline{X}_i$ be N_i. Further, define $S_i := \{1 \leq j \leq N : X_i(j) = 1, X_k(j) = 0 \; for \; all \; k < i\}$ and $\bar{N}_i := |S_i|$, for each $1 \leq i \leq n$. Observe that $\bar{N}_i \leq N_i$. The $\bar{N}_i$'s count the number of 1's in the ith vector, in positions that are all 0's in the previous $k < i$ vectors. In a sense, these are the only "new" 1's as far as the max function is concerned. The S_i's are disjoint, and $\bigcup_i S_i = \{j : f(X_1(j), X_2(j), \ldots, X_n(j) = 1\}$. Also, $\sum_i \bar{N}_i \leq N$. Therefore, communicating the sets $S_1, S_2, \ldots, S_n$ to the sink node suffices to reconstruct the function.

The strategy $\mathcal{S}_{N,n}$ consists of n stages, within each of which a single node transmits. In the i^{th} stage, it will be ensured that S_j for $j < i$ is made known to all nodes. The i^{th} node can therefore compute $\bar{N}_i$, and communicate its value in $\log N$ slots. Now S_i is one of $\binom{N-\sum_{j<i}\bar{N}_j}{\bar{N}_i}$ possibilities. Also, at this stage, $\bar{N}_i$ as well as the previous S_j's are known to all nodes. Therefore, node i can communicate the identity of the set S_i in $\log \binom{N-\sum_{j<i}\bar{N}_j}{\bar{N}_i}$ slots. Decoding by the sink node, and knowing when to begin transmitting by node $i+1$, are taken care of by knowledge of $\bar{N}_i$. The total number of slots required by this scheme is

$$T(\mathcal{S}_{N,n}) = n \log N + \sum_i \log \binom{N - \sum_{j<i} \bar{N}_j}{\bar{N}_i}. \quad (3)$$

We bound the RHS of (3) as follows. First observe that $\prod_i(\binom{N-\sum_{j<i}\bar{N}_j}{\bar{N}_i})$ is the multinomial coefficient $\binom{N}{\bar{N}_1,\bar{N}_2,\ldots,\bar{N}_n,\bar{N}'}$, where $\bar{N}' = N-\sum_i^n \bar{N}_i$. The multinomial coefficient is maximum when all the $\bar{N}_i$'s and $\bar{N}'$ are equal to $N/(n+1) = l$ (see [9]) . Thus,

$$\binom{N}{\bar{N}_1,\bar{N}_2,\ldots,\bar{N}_n,\bar{N}'} \leq \binom{N}{\frac{N}{n+1},\frac{N}{n+1},\ldots,\frac{N}{n+1}} = \prod_{0\leq i\leq n}\binom{(n-i+1)l}{l} < \binom{l(n+1)}{l}^{n+1}.$$

Using the fact that $\binom{n}{k} < (\frac{ne}{k})^k$,, (see [9] ,4.1.2), we have

$$(\binom{l(n+1)}{l})^{n+1} < (\frac{l(n+1)e}{l})^{l(n+1)} = ((n+1)e)^{l(n+1)}.$$

Thus, $T(S^{(ln,n)}) < n\log l(n+1) + l(n+1)\log(n+1)e$. For n large enough so that $n^2 > e(n+1)$,

$$R_{max}^{(n)} = \limsup_{N\to\infty}\sup_{\mathcal{S}_{N,n}\in CFS}\frac{N}{T(\mathcal{S}_{N,n})} \geq \limsup_{l\to\infty}\frac{l(n+1)}{T(S_{(l(n+1),n)})} \geq \frac{1}{2\log n}.$$

Converse: We will show that within a certain large enough set of measurement matrices, every matrix maps maps uniquely to a set of transmissions. Take block length $N > 2n$. Let Ξ be the set of measurement matrices $\bar{x} \in \mathcal{X}^{Nn}$ which satisfy the following properties:
1) Each row $\underline{x}_i$ has exactly $k = \lfloor N/(2n)\rfloor$ 1's.
2) For any two sequences $\underline{x}_i$ and $\underline{x}_j$, the sets of entries with value 1 are disjoint.

A simple counting argument, omitted here due to lack of space, on the number of choices of matrices satisfying (1) and (2) above, shows that $|\Xi| = \prod_{1\leq i\leq n}\binom{N-(i-1)\lfloor\frac{N}{2n}\rfloor}{\lfloor\frac{N}{2n}\rfloor}$

Fix a strategy $\mathcal{S}_{N,n}$ of duration $T(\mathcal{S}_{N,n})$. We claim that for any sequence $z_1, z_2, \ldots, z_{T(\mathcal{S}_{N,n})}$ of transmissions produced by applying the strategy $\mathcal{S}_{N,n}$, there is a unique $\bar{x} \in \Xi$ that produces it.

Suppose not. Then let $\bar{x} = [\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n]^t$ and $\bar{y} = [\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_n]^t$ be two distinct measurement matrices producing this sequence of transmissions. There is some i such that $\underline{x}_i \neq \underline{y}_i$. We will show that the measurement matrix $\bar{w} = [\underline{y}_1, \ldots, \underline{y}_{i-1}, \underline{x}_i, \underline{y}_{i+1}, \ldots, \underline{y}_n]^t$ will produce the same sequence of transmissions $z_1, z_2, \ldots, z_{T(\mathcal{S}_{N,n})}$. It will then follow that there must be an error in the strategy. To see this, note that $\underline{y}_i \neq \underline{x}_i$, and at the same time both have the same number of 1's. So $\underline{y}_i$ has a 1 in some entry, say the j^{th}, in which $\underline{x}_i$ is 0. Furthermore, since $y \in \Xi$, none of the other $\underline{y}_j$'s have a 1 in the same entry. Thus $f(\underline{y}^j) = 1$ and $f(\underline{w}^j = 1$ but yet $\mathcal{S}_{N,n}$ produces the same sequence of transmissions for both, an error.

To show this, consider $\mathcal{S}_{N,n}$ applied to $\bar{w}$. First observe that not only do $\bar{x}$ and $\bar{y}$ produce the same sequence of transmissions, but also each transmission is by the same node in each case. This follows from the definition of a CFS; the identity of the node to transmit in a given slot is determined only by the transmissions in the previous slots.

Suppose the sequence produced by $\mathcal{S}_{N,n}$ applied to $\bar{w}$ is $z'_1, z'_2, \ldots, z'_{T(\mathcal{S}_{N,n})}$. Let z_k be the first transmission by node i, when $\mathcal{S}_{N,n}$ is applied to $\bar{y}$. Then, $z'_j = z_j$ for $j < k$. This is because all the rows of $\bar{w}$ and $\bar{y}$ except for the i^{th} are identical, and so the transmissions produced by the corresponding nodes will also be the same. Hence, i (in the application of $\mathcal{S}_{N,n}$ to $\bar{w}$) will be scheduled to transmit for the first time in slot k. Now, given that the first $k-1$ transmissions are z_1 to z_{k-1}, $\mathcal{S}_{N,n}$ dictates that node i will be scheduled to transmit in slot k. Moreover, observe that the same sequence of transmissions is also produced for $\bar{x}$, in which the vector corresponding to node i is $\underline{x}_i$. Thus what node i transmits in the k^{th} slot is the same whether its vector is $\underline{x}_i$ or $\underline{y}_i$, conditioned on $z_1, z_2, \ldots, z_{k-1}$ being the previous $(k-1)$ transmissions. Thus, $z'_k = z_k$. Repeating this argument, we can conclude that $z'_j = z_j$ for all $j \leq T(\mathcal{S}_{N,n})$.

Thus, each $\bar{x} \in \Xi$ produces a distinct sequence of transmissions $z_1, z_2, \ldots, z_{T(\mathcal{S}_{N,n})}$. We then have

$$2^{T(\mathcal{S}_{N,n})} \geq |\Xi| = \prod_{1\leq i\leq n}\binom{N-(i-1)\lfloor\frac{N}{2n}\rfloor}{\lfloor\frac{N}{2n}\rfloor}. \quad (4)$$

Using the inequalities $\binom{n}{k} > \binom{m}{k}$ and $\binom{n}{k} > (\frac{n-k}{k})^k$, for $n > m$,

$$\text{RHS of (4)} > \binom{\lfloor N/2\rfloor}{\lfloor\frac{N}{2n}\rfloor}^n > \left(\frac{(\lfloor\frac{N}{2}\rfloor - \lfloor\frac{N}{2n}\rfloor)}{\lfloor\frac{N}{2n}\rfloor}\right)^{\lfloor\frac{N}{2n}\rfloor n}.$$

Taking logs,

$$\begin{aligned} T(\mathcal{S}_{N,n}) &> \lfloor\frac{N}{2n}\rfloor n\log\left(\frac{(\lfloor\frac{N}{2}\rfloor - \lfloor\frac{N}{2n}\rfloor)}{\lfloor\frac{N}{2n}\rfloor}\right) \\ &> (\frac{N}{2n}-1)n\log(n-\frac{2n}{N}-1). \end{aligned}$$

Consequently for n large enough so that $n-1 > \sqrt{n}$,

$$R_{max}^{(n)} = \limsup_{N\to\infty}\sup_{\mathcal{S}_{N,n}}\frac{N}{T(\mathcal{S}_{N,n})} \leq \frac{4}{\log n}.$$

Now, we extend the above proof to the general case where $|\mathcal{X}| \geq 2$, and $f(\cdot)$ is an arbitrary type-threshold function. Let $\underline{\theta}$ be the threshold vector of $f(\cdot)$.
Achievability: We construct a CFS composed of $|\mathcal{X}|$ phases. The kth phase consists of a sub-strategy similar to the one constructed for the binary case, with the objective of communicating the function $\underline{f_k}(\bar{X})$ to the sink node, where $f_k(\underline{x}) := min(\tau_k(\underline{x}), \theta_k)$ ($f_k(\cdot)$ is the number of occurrences of k in the vector $\underline{x}$ upto the threshold θ_k). At the end of $|\mathcal{X}|$ phases, the sink node will know $\min(\underline{\tau}(\underline{X}^i), \underline{\theta})$ for $1 \leq i \leq N$, and by definition of a type-threshold function, it will know $f(\underline{X}^i)$, for each $1 \leq i \leq N$.

Define a set of $|\mathcal{X}|$ binary valued $n \times N$ matrices $\bar{Y}^1, \bar{Y}^2, \ldots, \bar{Y}^{|\mathcal{X}|}$, where $\bar{Y}^k$ is given by $Y_{ij}^k = I_{\{X_{ij}=k\}}$. Each node i knows $\underline{Y}_i^j$ for each $1 \leq j \leq |\mathcal{X}|$, since these vectors depend only on $\underline{X}_i$.

The k^{th} phase is as follows. As in the binary case, define for each $1 \leq i \leq n$,

$$S_i := \{1 \leq j \leq N : Y_{ij}^k = 1, |\{l < i : Y_{lj}^k = 1\}| \leq \theta_l\},$$

and $\bar{N}_i := |S_i|$. It follows from the above that knowledge of all the S_i's determines the value of $\underline{f_k}(\bar{X})$, and also that $\sum_i \bar{N}_i \leq \theta_l N$. This is essentially the only difference from the binary max case; there, a single 1 in a column of $\bar{X}$ fixed the value of the function, whereas here the function value is insensitive to a larger number of occurrences of the value l over θ_l. The rest of the proof parallels the proof for the binary max function. Sensor i knows S_i, because it knows all the S_j for $j < i$. Sensor i communicates S_i using $(\log N + \log \binom{N}{N_i})$ bits. The bounds on total time are the same as in the binary case, with N being replaced by $\theta_l N$. The total number of transmissions made in this scheme is obtained by summing over the $|\mathcal{X}|$ phases. Thus, for large enough n and N, $T(\mathcal{S}_{N,n}) \leq (\sum_i k_i) 3N \log n$, and the desired result follows by taking $N \to \infty$.

Converse: Recall that Lemma 2 shows the existence of a vector $\underline{\eta}$ with $\eta_i \leq \theta_i$ for each i, such that for some i^*, j^* and function values $f_1 \neq f_2$, $\eta_{i^*} = \theta_{i^*}$, where $f_1 := f'(\eta_1, \ldots, \eta_{j^*}, \ldots, \eta_{|\mathcal{X}|})$, and $f_2 := f'(\eta_1, \ldots, \eta_{j^*} + m, \ldots, \eta_{|\mathcal{X}|})$ for all $m \geq 1$. Denote $a := j^*$, and $b := i^*$.

We define a subset $\Xi \subset \mathcal{X}^{Nn}$ as follows. For each $\bar{x} \in \Xi$, the first η_1 rows are all $[1, 1, \ldots, 1]$, followed by η_2 rows all $[2, 2, 2, \ldots, 2]$, $\ldots$, $\eta_{|\mathcal{X}|}$ rows all $[|\mathcal{X}|, |\mathcal{X}|, |\mathcal{X}|, \ldots, |\mathcal{X}|]$, and all the remaining matrix elements are either a's or b's. Suppose that the measurement matrix of the sensor network is restricted to be an arbitrary element of Ξ. In other words, $\sum_i \eta_i \leq \sum_i \theta_i$ sensors have fixed constant measurement block vectors, while the remaining sensors have binary valued (a or b) measurement vectors. By the property of $\underline{\eta}$ and type-threshold functions,

$$f(\underline{x^i}) = \begin{cases} f_2 & \text{if } x_{ij} = b \text{ for some } \sum_k \eta_k < j \leq n, \\ f_1 & \text{if } x_{ij} = a \text{ for all } \sum_k \eta_k < j \leq n, \end{cases}$$

For each $1 \leq i \leq N$, and any $\bar{x} \in \Xi$. Thus, $f(\cdot)$ on the restricted set Ξ is equivalent to the max function on a network of size $n - \sum_i^{|\mathcal{X}|} \eta_i$, with binary valued measurements. Applying the bound derived in the binary case with n large enough so that $n - \sum_i^{|\mathcal{X}|} \eta_i > \sqrt{n}$,

$$R_{max}^{(n)} = \limsup_{N \to \infty} \sup_{\mathcal{S}_{N,n}} \frac{N}{T(\mathcal{S}_{N,n})} \leq \frac{8}{\log n}. \quad \blacksquare$$

It can be proven that strategies with block-length 1 however have a maximum achievable rate of only $\Theta(\frac{1}{n})$, showing the significant benefit realizable by block computation. We omit this result due to lack of space. The key idea is simply that prior transmissions give no information about the measurements of nodes that have not transmitted. In the max function, for instance, if the first $n-1$ transmissions are by nodes carrying zeros, the value of the function is still undetermined. Thus, coding over long blocks in this case does give an exponential improvement over coding with block-length 1.

D. Type-sensitive and type-threshold functions over random planar networks

The results for collocated networks can be extended to random planar networks under the protocol model. However, the class of allowable strategies needs to be suitably modified. CFS's are not feasible since node transmissions are not heard by all. As mentioned earlier, defining a class of allowable strategies in such a network is difficult. Rather than delving further into the construction of such a class, we upper-bound the achievable rate using the following argument: Assume that each transmission made by a node to a neighbor is instantly relayed by a "genie" to all nodes in the network. Such a system can perform at least as well as a network without a genie. The presence of the genie allows the performance in the random planar network to be bounded through the identification of an embedded collocated network of size $\log n$.

Theorem 5: Consider a random planar network, with common range $r(n)$ chosen to be large enough so that the network is connected. Let $f(\cdot)$ and $g(\cdot)$ be type-sensitive and type-threshold functions, respectively.

(i) There exist constants $k_2 > k_1 > 0$, such that

$$\lim_{n \to \infty} Pr[R_{max}^{(f,n)} \geq \frac{j}{\log n}] = \begin{cases} 1 & \text{if } j \leq k_1 \\ 0 & \text{if } j \geq k_2 \end{cases}.$$

(ii) There exist constants $k_4 > k_3 > 0$, such that

$$\lim_{n \to \infty} Pr[R_{max}^{(g,n)} \geq \frac{j}{\log \log n}] = \begin{cases} 1 & \text{if } j \leq k_3 \\ 0 & \text{if } j \geq k_4 \end{cases}.$$

Proof. (i) Achievability directly follows from Theorem 2, since symmetric functions depend only on type.

Upper bound: In the proof of Theorem 3, it is shown that the number of transmissions required to communicate $\underline{f}(\bar{X})$, where every transmission is heard by all, which is the case here due to the presence of the genie, is at least kNn. The same statement is true of the random planar network with a genie. In a collocated network, the total number of transmissions is the same as the total number of time slots required, which however is not the case here.

Consider a tessellation of the square into smaller squares of area $A(n) = \frac{(\Delta r)^2}{2}$. Then, the number of such squares is $\frac{1}{A(n)}$. The total number of transmissions required is at least kNn. By the pigeonhole principle, at least one square must make $kNnA(n)$ transmissions. As pointed out in [1] Section 5.2, no two nodes within a distance of $\Delta r(n)$ of each other can simultaneously transmit and be heard by any other node. Therefore, no two nodes in a single square can transmit simultaneously. Also, to guarantee connectivity (see [6]), $r(n)$ must be asymptotically larger than $\sqrt{\frac{\log n}{n}}$. Thus, as $n \to \infty$, the total time required is lower bounded by $kNnA(n) \geq (k/2)\Delta^2 N \log n$. The result follows by observing that the maximum rate without a genie cannot be more than the maximum rate with a genie.

(ii) We prove the result only for the max function with binary valued inputs. The extension to the general case parallels the proof in Theorem 4. Tessellate the square as described above in (i). Pick a square with at least

	Collocated networks	Random planar networks
All data	$\Theta(\frac{1}{n})$	$\Theta(\frac{1}{n})$
Histogram/type	$\Theta(\frac{1}{n})$	$\Theta(\frac{1}{\log n})$
Type-sensitive	$\Theta(\frac{1}{n})$	$\Theta(\frac{1}{\log n})$
Type-threshold	$\Theta(\frac{1}{\log n})$	$\Theta(\frac{1}{\log \log n})$

Fig. 3. Summary of results: Rates for different classes and networks.

$n' = nA(n)$ nodes. Now, for all other nodes in the network, set the corresponding vectors to consist of all 0's. Thus the max function is determined by the data vectors of nodes in the square. By Theorem 4, the number of transmissions required is greater than $\frac{N}{4} \log n'$. Only one node can transmit at a time. Thus, for large n the achievable rate is upper bounded by $\frac{1}{k \log n'} \geq \frac{1}{k' \log \log n}$, for an appropriately defined k'.

The achievability of this rate can be proved by combining the ideas of the achievable schemes in Theorems 1 and 4. We provide only a brief description. Tesselate the unit square into square cells of area $\frac{1}{\log n}$. Lemma 1 guarantees that all cells have $O(\log n)$ nodes. Also, if $r(n) = \sqrt{\frac{2 \log n}{n}}$, the network is connected with high probability as $n \to \infty$, and each cell is in effect a collocated subnetwork, with each node within range of every other node within the cell. Designate one relay per cell, and construct a schedule in which each cell has a constant fraction of all time slots in which to transmit (the existence of such a schedule is guaranteed by Theorem 1). Each cell can then be treated as a collocated network of size $O(\log n)$, with the relay functioning as the sink node. The max function of the entire network is the maximum of the individual max functions in all the cells. The latter can be communicated to the relays in $O(N \log \log n)$ time. The maximum of the cellular max functions can then be pipelined along the cells, just as in the scheme described in Theorem 1, taking $O(N)$ time per block. Note that the pipelining here would involve sending the max of the previous N-length block, while cellular max functions of the current N-length block are being computed in the first phase described above. The result follows. ■

V. Conclusions and Future Work

We have studied the problem of determining the maximum rate of computing and communicating functions of measurements taken by nodes in a sensor network to a designated sink node. We have focused on symmetric functions, since they are a natural class of functions of interest in homogeneous sensor networks. A summary of the results is given in Table 3.

There are a number of directions for future work. First, the results proved in the last section are for networks in which the number of simultaneous transmissions is the limiting constraint. There are other spatial configurations, such as grid and line, in which a constant throughput to the nearest neighbors is possible, but the constraining factor is that computation of the function required may still require certain data to be relayed.

Second, we have not considered non-symmetric functions, and neither do we obtain lower bounds on achievable rate for all possible symmetric functions. Another natural extension is to introduce joint distributions on the sensor readings, and determine bounds on the average rate of computation of functions. We believe that this is a fairly challenging problem. Finally, an information theoretic approach to the problem is wide open.

References

[1] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Trans. Inform. Th.*, vol. 46, no. 2, pp. 388–404, 2000.

[2] L.-L. Xie and P. Kumar, "A network information theory for wireless communication: Scaling laws and optimal operation," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 748–767, May 2004.

[3] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 1991.

[4] A. Orlitsky and J. R. Roche, "Coding for computing," in *IEEE Symposium on Foundations of Computer Science*, 1995, pp. 502–511.

[5] E. Kushilevitz and N. Nisan, *Communication Complexity*. Cambridge University Press, 1997.

[6] P. Gupta and P. Kumar, "Critical power for asymptotic connectivity in wireless networks," in *Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W.H. Fleming*, W. McEneaney, G. Yin, and Q. Zhang, Eds., 1998.

[7] F. Xue and P. Kumar, "The number of neighbors needed for connectivity of wireless networks," *Wireless Networks*, vol. 10, no. 2, pp. 169–181, March 2004.

[8] D. Marco, E. J. Duarte-Melo, M. Liu, and D. L. Neuhoff, "On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data," in *Proc. IPSN*, April 2003.

[9] D. West, "Combinatorial mathematics," Preliminary version.

[10] S. Shenker, S. Ratnasamy, B. Karp, R. Govindan, and D. Estrin, "Data-centric storage in sensornets," 2002. [Online]. Available: citeseer.ist.psu.edu/shenker02datacentric.html