Scaling Laws for Correlated Data Gathering[1]

Răzvan Cristescu[†], Baltasar Beferull-Lozano[†] and Martin Vetterli[†,‡]

[†]Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
[‡]Department of EECS, University of California at Berkeley, Berkeley CA 94720, USA
e-mail: {Razvan.Cristescu, Baltasar.Beferull, Martin.Vetterli}@epfl.ch

Abstract — **Consider a set of correlated sources located at the nodes of a network, and a sink to which the data from all the sources have to arrive. We address the minimization of a separable joint communication cost function given by the product [rate] · [edge weight]. We present two possible approaches for rate allocation, namely Slepian-Wolf coding, and coding by explicit communication, and compare asymptotically (large networks) the associated total costs by finding their corresponding scaling laws and analyzing the ratio between them. We also provide the specific conditions on the correlation structure which determine the different cases of asymptotic behaviors.**

Consider a number of data sources with a certain spatial correlation structure and which are located at the nodes of a network. The network is represented by a graph $G = (V, E)$ which connects the sources, represented by the nodes, through links, represented by the edges. The goal is to transport the data from the nodes to a particular *sink* node S, such that a total communication cost function is minimized. The cost function is related to the lossless coding rate allocation $(R_1, \ldots, R_N)$ for the sources $(X_1, \ldots, X_N)$, and the weights of the links. We restrict the optimization over the set of *data gathering trees*:

$$\left\{ \{R_i^*\}_{i=1}^N, ST^* \right\} = \arg \min_{\{R_i\}_{i=1}^N, ST} \sum_{i \in V} R_i d_{ST}(i, S) \qquad (1)$$

where ST is a spanning tree for G and $d_{ST}(i, S)$ is the cost of the path connecting node i to S on the ST tree.

A joint treatment of data aggregation and the transmission structure is found in [3], where no collaboration among nodes is considered. We consider collaboration by distributed Slepian-Wolf coding, for which communication among the nodes is not necessary [1, 4], and compare this approach with explicit communication coding [2, 3]. The choice of approach depends on the network knowledge that the nodes have.

We analyze a one-dimensional network model with N nodes equally spaced on a line (Figure 1). For this model, the optimal transmission structure is the shortest path tree (SPT) for both rate allocation approaches. Let us denote the conditional entropies by $a_i = H(X_i | X_{i-1}, \ldots, X_1)$. If nodes are assumed to know the correlation structure, then they can perform Slepian-Wolf coding. In this case, the solution of (1) is [1] $(R_1^*, R_2^*, \ldots, R_N^*) = (a_1, a_2, \ldots, a_N)$. On the contrary, in the explicit communication approach, nodes can exploit the data correlation only by receiving *explicit* side information from other nodes (that is, when other nodes use a node as relay, their data is locally available at that relaying node). In this case, due to the symmetry of the correlation structure, the rate allocation can be written as $(R_1^*, R_2^*, \ldots, R_N^*) = (a_N, a_{N-1}, \ldots, a_1)$. Consider the ratio between the total costs associated to the two coding approaches:

$$\gamma(N) = \frac{\text{cost}_{SW}(N)}{\text{cost}_{EC}(N)} = \frac{\sum_{i=1}^N i a_i}{\sum_{i=1}^N (N - i + 1) a_i}.$$

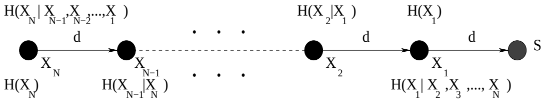
H(X_N | X_{N-1}, X_{N-2}, ..., X_1)
H(X_2|X_1)
H(X_1)
d
d
d
S
X_N
X_{N-1}
X_2
X_1
H(X_N)
H(X_{N-1}|X_N)
H(X_1| X_2, X_3, ..., X_N)

Figure 1: Rate allocation for Slepian-Wolf coding (above) and explicit communication coding (below).

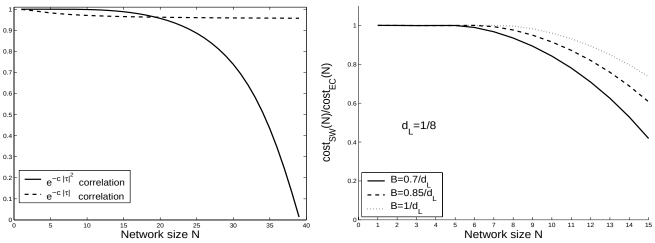

Figure 2: The ratio of total $\text{cost}_{SW}(N)/\text{cost}_{EC}(N)$ as a function of the network size for a refinement network sampling a Gaussian random field (left) and a bandlimited process (right).

Theorem 1 *If* $\lim_{i \to \infty} a_i = C > 0$*, then* $\lim_{N \to \infty} \gamma(N) = 1$ *and* $cost_{SW}(N) = \Theta(cost_{EC}(N))$*. If* $\lim_{i \to \infty} a_i = 0$*, then: (a) if* $a_i = \Theta(1/i^p), p \in (0, 1)$*, then* $\lim_{N \to \infty} \gamma(N) = 1 - p$ *and* $cost_{SW}(N) = \Theta(cost_{EC}(N))$*; (b) if* $a_i = \Theta(1/i^p), p \geq 1$*, then* $\lim_{N \to \infty} \gamma(N) = 0$ *and* $cost_{SW}(N) = o(cost_{EC}(N))$*; moreover, if* $p = 1$ *then* $\gamma(N) = \Theta(1/\log N)$*, if* $p \in (1, 2)$ *then* $\gamma(N) = \Theta(1/N^{p-1})$*, if* $p = 2$*,* $\gamma(N) = \Theta(\log N/N)$*, if* $p > 2$ *then* $\gamma(N) = \Theta(1/N)$*.*

Theorem 1 can be applied for various correlation models [1], including sampled Gaussian continuous-space WSS random processes, and bandlimited processes (see Figure 2).

References

[1] R. Cristescu, B. Beferull-Lozano, M. Vetterli, "Networked Slepian-Wolf: Theory and Algorithms," *submitted*, 2003.

[2] R. Cristescu, B. Beferull-Lozano, M. Vetterli and R. Wattenhofer, "Network Correlated Data Gathering with Explicit Communication: NP-Completeness and Algorithms," *submitted*, 2004.

[3] A. Goel, D. Estrin, "Simultaneous Optimization for Concave Costs: Single Sink Aggregation or Single Source Buy-at-Bulk," *ACM-SIAM Symposium on Discrete Algorithms*, 2003.

[4] D. Slepian, J.K. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Tr. Inf. Th. IT-19*, 1973, pp. 471-480.

[1] The work presented in this paper was supported (in part) by the National Competence Center in Research on Mobile Information and Communications Systems (NCCR-MICS, http://www.mics.org), a center supported by the Swiss National Science Foundation under grant number 5005-67322.