European Journal of Operational Research 246 (2015) 744–749

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Discrete Optimization

Scheduling for data gathering networks with data compression

Joanna Berlińska*

Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Umultowska 87, Poznań 61-614, Poland

ARTICLE INFO

Article history:
Received 22 September 2014
Accepted 9 May 2015
Available online 15 May 2015

Keywords:
Scheduling
Data gathering networks
Data compression

ABSTRACT

This paper analyzes scheduling in a data gathering network with data compression. The nodes of the network collect some data and pass them to a single base station. Each node can, at some cost, preprocess the data before sending it, in order to decrease its size. Our goal is to transfer all data to the base station in given time, at the minimum possible cost. We prove that the decision version of this scheduling problem is NP-complete. Polynomial-time heuristic algorithms for solving the problem are proposed and tested in a series of computational experiments.

1. Introduction

Data gathering wireless sensor networks (WSNs) find a broad variety of applications in monitoring the environment, surveillance and other areas (Akyildiz, Su, Sankarasubramaniam, & Cayirci, 2002). A network consists of a set of measuring sensors and a base station to which the collected data should be transferred. As wireless sensors have only limited power and computing capability, these resources have to be carefully managed to ensure network efficiency. Lower energy usage or network delay can be achieved by applying specific protocols of communication (Ergen & Varaiya, 2010; Kumar & Chauhan, 2011; Shi & Fapojuwo, 2010; Wu, Li, Liu, & Lou, 2010). Another approach is constructing scheduling algorithms for the whole data gathering application. On the basis of the divisible load theory, algorithms were created for minimizing data gathering time (Choi & Robertazzi, 2008; Moges & Robertazzi, 2006) and maximizing network lifetime (Berlińska, 2014). Column generation algorithms for maximizing network lifetime were also proposed (Alfieri, Bianco, Brandimarte, & Chiasserini, 2007; Rossi, Singh, & Sevaux, 2013). One more method for increasing network performance is compressing the gathered data before sending it to the base station in order to decrease the communication time. Since general compression algorithms are often not applicable for sensor networks because of scarce resources, several compression algorithms were specifically designed for WSNs (Kimura & Latifi, 2005). Compressive sensing theory was applied to data gathering wireless sensor networks, taking advantage of the correlations among the sensor data, in paper (Luo, Wu, Sun, & Chen, 2009). This research direction became very popular, resulting in more advanced data gathering approaches based on compressive sensing (Cheng, Ye, Jiang, Wang, & Wang, 2013; Wang, Tang, Yin, & Li, 2012; Xiang, Luo, & Rosenberg, 2013; Xu, Wang, & Wang, 2011).

In this paper we do not aim at constructing one more method of compressing the data gathered by a wireless sensor network. Instead, our goal is to provide and analyze a simple theoretical model of data gathering networks with data compression. In our model each node of the network is able to preprocess the gathered data in order to compress it before sending to the base station, at monetary or energy cost proportional to its size. Let us note that this model can be applied not only to wireless sensor networks but also to wired networks whose nodes collect some data or obtain them as a result of computations. We analyze the problem of sending the whole data gathered by the network to the base station in given time, at the minimum possible cost of data compression. We prove that the decision version of this problem is NP-complete. Three polynomial-time heuristics are proposed. Their performance and its dependence on system parameters are tested in a series of computational experiments.

The rest of this paper is organized as follows. In Section 2 we propose a model of data gathering networks with data compression and formulate the scheduling problem. Its computational complexity is studied in Section 3. The following section describes the algorithms we propose for solving the problem. Section 5 comprises the results of computational experiments. The last section is dedicated to conclusions.

2. Problem formulation

We analyze a data gathering network consisting of a set of m identical nodes $\mathcal{P} = \{P_1, \dots, P_m\}$ and a single base station to which the collected information should be passed. The words node, sensor and processor will be used interchangeably. The size of data gathered on node P_i is α_i. Each node can preprocess its data, using application-

* Corresponding author. Tel.: +48 618295333.
E-mail address: joanna.berlinska@amu.edu.pl

http://dx.doi.org/10.1016/j.ejor.2015.05.026
0377-2217/

specific knowledge, to decrease its size to $\gamma\alpha_i$ in time $A\alpha_i$, where $0 < \gamma < 1$. However, this comes at cost $f\alpha_i$. A node that compresses its data can start sending it only after the compression process is finished. Although all sensors are identical, their locations are different, and hence, the communication capabilities are not the same for all of them. Sensor P_i transmits data of size x in time xC_i. Only one sensor can communicate with the base station at a time. Constructing a schedule for data gathering in the described network consists in choosing which nodes should compress their data and what should be the order of communications with the base station. Our goal is to create a schedule for transmitting all data in given time, at the minimum possible cost. This problem can be formulated as follows.

Problem 1. (DG-Compr-OptF). Given m nodes, their parameters $A, f, \gamma, (C_i)_{1\le i\le m}, (\alpha_i)_{1\le i\le m}$ and a rational number T, what is the minimum cost necessary to send all data to the base station within time T? What schedule allows for achieving this cost?

In the next section we will analyze the complexity of the following decision version of problem DG-Compr-OptF.

Problem 2. (DG-Compr). Given m nodes, their parameters A, f, γ, $(C_i)_{1\le i\le m}$, $(\alpha_i)_{1\le i\le m}$ and two rational numbers T, F, is it possible to send all data to the base station within time T at cost not greater than F?

As all nodes of the network execute the same data gathering application, it is often the case that the amounts α_i of collected data are the same for all sensors. Therefore, we will also study the complexity of the following subproblem representing this situation.

Problem 3. (DG-Compr{ $\alpha_i = \alpha$}). Given m nodes, their parameters $A, f, \gamma, (C_i)_{1\le i\le m}, \alpha_i = \alpha$ for $1 \le i \le m$ and two rational numbers T, F, is it possible to send all data to the base station within time T at cost not greater than F?

3. Computational complexity

In this section we study the computational complexity of the formulated scheduling problems. Let us start with the observation that DG-Compr is a selection problem, i.e. if a subset $\mathcal{P}' \subset \mathcal{P}$ of processors that compress their data is given, it is easy to check if it allows for constructing a feasible solution. Indeed, the cost of compression for given $\mathcal{P}'$ can be computed straightforwardly and compared with F. It remains to check if the shortest possible schedule for $\mathcal{P}'$ is not longer than T. Node P_j is ready to start communication with the base station at time r_j, where $r_j = 0$ if $P_j \notin \mathcal{P}'$ and $r_j = A\alpha_j$ if $P_j \in \mathcal{P}'$. As all sensors send their data sequentially, constructing the shortest schedule is equivalent to solving problem $1|r_j|C_{\max}$, i.e. scheduling on one processor with job release times and makespan criterion. This can be done in $O(m\log m)$ time by sorting the processors according to non-decreasing r_j (Błażewicz, Ecker, Pesch, Schmidt, & Węglarz, 2007).

We will now prove that problem DG-Compr{$\alpha_i = \alpha$}, and hence the more general problem DG-Compr, is NP-complete. Afterwards, it will be shown that problem DG-Compr{$\alpha_i = \alpha$} can be solved in polynomial time if all nodes have the same communication rate ($C_i = C$ for all $1 \le i \le m$). Finally, we will prove that problem DG-Compr remains NP-complete even if all nodes have the same communication rate and data compression takes no time.

Proposition 1. *Problem* DG-Compr{$\alpha_i = \alpha$} *is NP-complete.*

Proof. It is obvious that DG-Compr{$\alpha_i = \alpha$} belongs to NP. Indeed, as explained above, given a subset of processors that compress their data, we can create a schedule and verify its feasibility in $O(m\log m)$ time, which is polynomial.

We prove that DG-Compr{$\alpha_i = \alpha$} is NP-complete via a reduction from the NP-complete Partition problem (Garey & Johnson, 1991), which is defined as follows.

Problem 4. (Partition). Given a finite set of m integers $\mathcal{B} = \{b_1, \dots, b_m\}$, is there a subset $\mathcal{B}' \subset \{1, \dots, m\}$ such that $\sum_{i\in\mathcal{B}'} b_i = \sum_{i\notin\mathcal{B}'} b_i = \sum_{i=1}^m b_i/2$?

Given an instance of Partition, we construct an instance of DG-Compr{$\alpha_i = \alpha$} in the following way. Let $\alpha = f = 1$, $C_i = b_i$ for $i = 1, \dots, m$, $A = \sum_{i=1}^m b_i/2$, $F = m$. Parameter γ can be an arbitrary positive number smaller than one. Let $T = (1+\gamma)\sum_{i=1}^m b_i/2$. Note that F is so big that in order to solve the problem it is sufficient to find a schedule of length not exceeding T. Let $\mathcal{P}'$ be the subset of nodes which compress their data in some solution of the analyzed instance of DG-Compr{$\alpha_i = \alpha$}. Let $\mathcal{B}'$ denote the set of indices of nodes contained in $\mathcal{P}'$. The processors from subset $\mathcal{P} \setminus \mathcal{P}'$ can start communications at time 0, whereas the processors contained in $\mathcal{P}'$ can start sending their data only after time $A\alpha$. Hence, the schedule length is

$$\begin{aligned} t(\mathcal{B}') &= \max\left\{\sum_{i\notin\mathcal{B}'} C_i\alpha, A\alpha\right\} + \sum_{i\in\mathcal{B}'} C_i\gamma\alpha \\ &= \max\left\{\sum_{i\notin\mathcal{B}'} b_i, \sum_{i=1}^m b_i/2\right\} + \gamma\sum_{i\in\mathcal{B}'} b_i. \end{aligned} \tag{1}$$

If $\sum_{i\notin\mathcal{B}'} b_i < \sum_{i=1}^m b_i/2$, then

$$t(\mathcal{B}') = \sum_{i=1}^m b_i/2 + \gamma\sum_{i\in\mathcal{B}'} b_i > (1+\gamma)\sum_{i=1}^m b_i/2 = T. \tag{2}$$

For $\sum_{i\notin\mathcal{B}'} b_i > \sum_{i=1}^m b_i/2$ we obtain

$$\begin{aligned} t(\mathcal{B}') &= \sum_{i\notin\mathcal{B}'} b_i + \gamma\sum_{i\in\mathcal{B}'} b_i = \gamma\sum_{i=1}^m b_i + (1-\gamma)\sum_{i\notin\mathcal{B}'} b_i \\ &> \gamma\sum_{i=1}^m b_i + (1-\gamma)\sum_{i=1}^m b_i/2 = (1+\gamma)\sum_{i=1}^m b_i/2 = T. \end{aligned} \tag{3}$$

Finally, if $\sum_{i\notin\mathcal{B}'} b_i = \sum_{i=1}^m b_i/2$, then

$$t(\mathcal{B}') = \sum_{i=1}^m b_i/2 + \gamma\sum_{i\in\mathcal{B}'} b_i = (1+\gamma)\sum_{i=1}^m b_i/2 = T. \tag{4}$$

Hence, the required schedule of length not exceeding T exists if and only if there exists a subset $\mathcal{B}' \subset \mathcal{B}$ such that $\sum_{i\in\mathcal{B}'} b_i = \sum_{i\notin\mathcal{B}'} b_i = \sum_{i=1}^m b_i/2$, which ends the proof. □

Corollary 1. *Problem* DG-Compr *is NP-complete.*

Proposition 2. *Problem* DG-Compr{$\alpha_i = \alpha$} *can be solved in* $O(1)$ *time if* $C_i = C$ *for* $1 \le i \le m$.

Proof. Let us note that in a system with equal data sizes and equal communication rates the schedule length and cost depend only on the number n of nodes that compress their data. In order to achieve cost not exceeding F, it must hold that $n \le n_f$, where

$$n_f = \lfloor F/(f\alpha)\rfloor. \tag{5}$$

The n nodes compressing their data can start communications after time $A\alpha$, whereas the remaining $m - n$ sensors can start sending data at once. Hence, the schedule length for given n is equal to

$$t(n) = \max\{(m-n)C\alpha, A\alpha\} + nC\gamma\alpha. \tag{6}$$

We will now distinguish two cases to compute the minimum length of a schedule with cost not exceeding F.

1. Let us first assume that $n \le m - A/C$. This means that $(m-n)C\alpha \ge A\alpha$ and hence,

   $$t(n) = (m-n)C\alpha + nC\gamma\alpha = (mC - (1-\gamma)nC)\alpha. \tag{7}$$

   As $1 - \gamma > 0$, $t(n)$ decreases with increasing n and has the smallest value when $n_1 = \min\{\lfloor m - A/C\rfloor, n_f\}$ nodes compress data.