Customized Newspaper Broadcast: Data Broadcast with Dependencies

Sandeep Dey and Nicolas Schabanel

École normale supérieure de Lyon, UMR CNRS ENS-LYON INRIA UCBL n°5668
46 allée d'Italie, 69364 Lyon Cedex 07, France
dey.sandeep@gmail.com, http://perso.ens-lyon.fr/nicolas.schabanel

Abstract. Broadcasting has been proved to be an efficient means of disseminating data in wireless communication environments (such as Satellite, mobile phone networks; other typical broadcast networks are Videotext systems). Recent works provide strong evidence that correlation-based broadcast can significantly improve the average service time of broadcast systems. Most of the research on data broadcasting was done under the assumption that user requests are for a single item at a time and are independent of each other. However in many real world applications, such as web servers, dependencies exist among the data items, for instance: web pages on a server usually share a lot of items such as logos, style sheets, title-bar... and all these components have to be downloaded together when any individual page is requested. Such web server could take advantage of the correlations between the components of the pages, to speed up the broadcast of popular web pages. This paper presents a theoretical analysis of data dependencies and provides a polynomial time 4-approximation as well as theoretical proofs that our correlation-based approach can improve by an arbitrary factor the performances of the system. To our knowledge, our solutions are the first provably efficient algorithms to deal with dependencies involving more than two data items.

Topics: Approximation algorithms, Wireless and Push-based broadcast scheduling, Bluetooth and Satellite networks.

1 Introduction

Motivations. Broadcasting has been proved to be an efficient means of disseminating data in wireless communication environments (such as Satellite, mobile phone networks; other typical broadcast networks are Videotext systems). Recent works [15,14,10] provide strong evidence that correlation-based broadcast can significantly improve the average service time of broadcast systems. Most of the research on data broadcasting was done under the assumption that user requests are for a single item at a time and are independent of each other. However in many real world applications, such as web servers, dependencies exist among the data items, for instance: web pages on a server usually share a lot of items such as logos, style sheets, titlebar..., and all these components have to be downloaded together when any individual page is requested. Such web server could take advantage of the correlations between the components of the pages, to speed up the broadcast of popular web pages.

This paper presents a theoretical analysis of data dependencies and provides theoretical proofs that a correlation-based approach can improve arbitrarily the performance of the system. To our knowledge, our solutions are the first provably efficient algorithms to deal with dependencies involving more than two data items. Our results rely essentially on the construction of a new lower bound, based on a non-linear convex minimization program, in combination with existing heuristics.

Background. Broadcast (or push-based) server usually maintains a profile of the typical users, *i.e.*, the popularity of each item (see [1]), and schedules the broadcasts of each item accordingly, obliviously of the effective requests made by the users, so as to minimize the average service time for the users. The users connect at random instants and monitor the broadcast channel until the information they are interested in is broadcast. The user profiles can easily be obtained by analyzing the log files of the server or by asking the user to list its interests at the subscription to the service (see [1]). Such profiles provide not only informations on the popularity of each item, but also informations on their *correlations*. Several heuristics [15,14,18] have been proposed to take advantage of these dependencies.

Very little is known theoretically on the performance of these algorithms. As far as we know, the only papers that addressed this question theoretically are [8,6]. In [8], the authors give optimal polynomial time algorithms for broadcasting a set of $n = 2$ items given arbitrary user profiles. In [6], the authors design lower bounds and constant factor approximation algorithms to design cyclic schedules (*i.e.*, schedules where each item is broadcast exactly once per cycle) for the case where correlations are restricted to dependencies between pairs of items.

Earlier theoretical work, including NP-hardness results and approximation algorithms, on the databroadcast problem with independent requests can be found in [2,5,17,19,16,7]. Related work on on-demand broadcast where request asks for a single item, can be found in [11,4]. Current results in this last setting include intractability results as well as competitive algorithms with resource augmentation. As far as we know, introducing explicit dependencies is still an unexplored question in this setting.

The Customized Newspaper Broadcast Problem. We adopt the following setting. n unit length news items are made available on a broadcast server, *e.g.*: Weather, Sport, Stock exchange, International,... Each user is interested in a given subset of the news items with some probability, *e.g.*: {Weather, Stock exchange, Sport}, {Weather, International}, or {International, Stock exchange},... He then connects at a random time, monitors the broadcast channel until he is served, *i.e.*, until all the news items he is interested in have been broadcast, *e.g.*: Weather and International. The goal for the server is to find a schedule of the news items that minimizes the average service time of the user, given the request probability for each available *subset* of items. These probabilities are naturally obtained by asking the user to check its topics of interests in the list of available items when subscribing to the news system (which is now a widespread practice over the Internet). In the terms of [6] and [15,14], our model of user requests is refered as "AND" and "unordered" requests, respectively. Note that the constraint that all news item have unit length is *not* restrictive, since any longer news item can be split into a set of unit length packets, which furthermore, fits the actual situation of a web server on Internet where all files are cut into equal size packets before been sent.

Our contribution. Our main contribution consists in providing a new lower bound for the cost of an optimum schedule (Propositions 3 and 4), which is shown to be tight up to a constant factor. This lower bound is expressed as a non-linear convex minimization problem and is solved to obtain polynomial time 4-approximation algorithms (Theorems 2 and 3). These are, as far as we know, the first algorithms with bounded guarantee when dependencies involve more than two items. We also use this lower bound to show that correlation-based schedulers can indeed improve the quality of the previously known solutions by an arbitrary factor with respect to the optimum cost (Example 1). Our algorithms use previously known heuristics as subroutines. In particular, perfectly periodic schedules introduced in [7,9] find here a new interesting application since their regularity can solve efficiently dependencies between items by forcing an order on them. Our algorithms yield significant improvements to system performances by managing correlations and are also not too complicated to implement (and would be in particular well adapted to time multiplexing environments such as Bluetooth networks). Interestingly enough, the classic randomized algorithm is shown to be inefficient in this setting (Theorem 1 and Example 2).

The next section gives a formal description of the problem and states its NP-hardness. Then, Section 3 presents our new lower bound, and shows that previous approaches ignoring correlations can lead to arbitrarily bad performances. Section 4 analyzes a classic randomized algorithm and shows that it achieves an approximation ratio of $2H_n$ exactly. Our deterministic 4-approximation algorithm is given in Section 5. Finally in Section 6, we extend our results to the setting where the broadcast of each item has a cost (*e.g.*, see [5]).

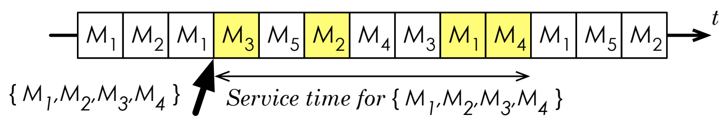
2 Notations and Preliminaries

The problem. The input consists of:

- n unit length news items $M_1, \ldots, M_n$,
- $\zeta = \{S_1, \ldots, S_k\}$ a set of k non-empty distinct sets of news items, $S_j \subseteq \{1, \ldots, n\}$, and
- positive request probabilities $(p_S)_{S\in\zeta}$ for each set S in ζ, such that $\sum_{S\in\zeta} p_S = 1$.

A *schedule* $\mathcal{S}$ is an infinite sequence $\mathcal{S} = \mathcal{S}(0)\mathcal{S}(1)\mathcal{S}(2)\mathcal{S}(3)\ldots$, where $\mathcal{S}(t) \in \{\perp, 1, \ldots, n\}$ for all t. If $\mathcal{S}(t) = i$, we say that *news item* M_i *is broadcast at time* t *in schedule* $\mathcal{S}$, *i.e.*, between time t and $t+1$; if $\mathcal{S}(t) = \perp$, no item is broadcast at time t. A schedule $\mathcal{S}$ is *periodic with period* T if $\mathcal{S}(t+T) = \mathcal{S}(t)$, for all time t. Such a periodic schedule is completely determined by its *cycle*, *i.e.*, the finite sequence $\mathcal{S}(0)\mathcal{S}(1)\ldots\mathcal{S}(T-1)$.

The *cost*, $\mathrm{COST}(\mathcal{S})$, of a schedule $\mathcal{S}$ is defined as the *average service time* to a random request, where the average is taken over the moments when requests occur *and* over the type S_j of news items subset requested. In our model, each client asks for news item set $S \in \zeta$ with probability p_S, connects at a random (integer) instant t according to some Poisson process, and is *served* when all the news items M_i, $i \in S$, have been broadcast. We denote by $\mathrm{ST}(\mathcal{S}, S, t)$ the *service time* of schedule $\mathcal{S}$ to a request for news items set S arriving at time t.



If $t_i \geqslant t$ is the first instant when news item M_i is broadcast on or after time t in $\mathcal{S}$, then $\mathrm{ST}(\mathcal{S}, S, t) = \max_{i \in S}(t_i + 1)$ (we consider that a request is served when the broadcast of the last requested item ends). Abusing the notation, $\mathrm{ST}(\mathcal{S}, M_i, t)$ will refer to $1 + t_i$. The average service time to a random request arriving at time t is then

$$\mathrm{AST}(\mathcal{S}, t) = \sum_{S \in \zeta} p_S \cdot \mathrm{ST}(\mathcal{S}, S, t).$$

Since the requests arrive according to some Poisson process, requests are uniformly distributed over any given time interval I (see [12]). Thus, the average service time $\mathrm{AST}(\mathcal{S}, I)$ of the schedule $\mathcal{S}$ during time interval $I = [t_1, t_2]$ is

$$\mathrm{AST}(\mathcal{S}, I) = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2-1} \mathrm{AST}(\mathcal{S}, t).$$

The cost of schedule $\mathcal{S}$ is then defined as the asymptotic value of this quantity as t goes to infinity:

$$\mathrm{COST}(\mathcal{S}) = \limsup_{t \to \infty} \mathrm{AST}(\mathcal{S}, [0, t]).$$

Note that if $\mathcal{S}$ is periodic with period T, its cost is simply defined as:

$$\mathrm{COST}(\mathcal{S}) = \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{AST}(\mathcal{S}, t).$$

Our goal is to compute a schedule with minimum cost. We denote by $\mathrm{OPT} = \inf_{\mathcal{S}} \mathrm{COST}(\mathcal{S})$ the optimum cost of a schedule for a given instance.

The customized newspaper problem is a generalization of the preemptive databroadcast setting which is shown to be strongly NP-hard in [19]. Thus,

Proposition 1 ([19]). *The customized newspaper problem is strongly NP-hard.*

Reduction to periodic schedules. General schedules are hard to handle since the frequency of each item may vary widely over time. The following lemma shows that we can restrict our study to periodic schedules which are much simpler to deal with. The following proposition follows the lines of [16].

Proposition 2. $\mathrm{OPT} = \inf_{\mathcal{S} \text{ periodic}} \mathrm{COST}(\mathcal{S})$.

3 Lower Bounding Optimal Cost

We present now a new lower bound on which our approximation algorithms rely. This lower bound takes into account the correlations between the requests, *i.e.*, the fact that a given item might be requested by different types of requests for different sets of news items. As opposed to previous approaches (*e.g.*, [2,5]), our lower bound cannot be solved by means of Lagrangian relaxation but can be expressed as a convex minimization program and solved by using the ellipsoid algorithm [13].

Section 3.2 shows that taking into account these correlations is indeed needed to estimate correctly the optimum cost, by showing that previously known methods (*e.g.*, [2,5,16,19]), that only used the probability that a given item is requested (ignoring possible correlations), can construct schedules with cost as large as $\Omega(\sqrt{n})$ times the optimum value.

3.1 Lower Bound

According to Proposition 2, any lower bound on the cost of periodic schedules is a lower bound on the optimum cost. We now focus on periodic schedules.

Lemma 1. *Let $\mathcal{S}$ be a periodic schedule with period T, such that each news item M_i is broadcast exactly n_i times per cycle. Then*

$$\mathrm{COST}(\mathcal{S}) \geqslant \mathrm{LB}(\tau),$$

where $\mathrm{LB}(\tau) = \frac{1}{2} + \frac{1}{2}\sum_{S\in\zeta} p_S \max_{i\in S} \tau_i$ *, with* $\tau_i = T/n_i$ *and* $1/\tau_1 + \cdots + 1/\tau_n \leqslant 1$.

Proof. Consider a request for a set $S \in \zeta$ arriving at time t. Since, this request waits for the broadcast of each news item M_i with $i \in S$, its service time is at least:

$$\mathrm{ST}(\mathcal{S}, S, t) \geqslant \mathrm{ST}(\mathcal{S}, M_i, t), \quad \text{for all } i \in S.$$

Then, by taking the average over t, the average service time to a request for set S is at least:

$$\mathrm{AST}(\mathcal{S}, S) \geqslant \mathrm{AST}(\mathcal{S}, M_i), \quad \text{for all } i \in S,$$

where $\mathrm{AST}(\mathcal{S}, M_i)$ denotes the average service time to a request that would ask for news item M_i alone. It is known from previous work (*e.g.*, [2,5]) that if M_i is broadcast n_i times in a periodic schedule with period T, then $\mathrm{AST}(\mathcal{S}, M_i) \geqslant$

$\frac{1}{2}+\frac{T}{2n_i}$. Consider indeed $t_1,\ldots,t_{n_i}$ the time elapsed between the ends of each of the n_i broadcasts of M_i during a cycle of $\mathcal{S}$. With probability t_j/T, the request for M_i falls in an interval of length t_j, waits on average $\left(\frac{1}{t_j}\sum_{t=1}^{t_j}(t_j-t)\right)=\frac{t_j-1}{2}$ time for the broadcast of M_i to begin, and is finally served one unit of time later, when the download of M_i is completed. Thus, $\mathrm{AST}(\mathcal{S},M_i)=1+\sum_{j=1}^{n_i}\frac{t_j(t_j-1)}{2T}$. But $t_1+\cdots+t_{n_i}=T$, then $\mathrm{AST}(\mathcal{S},M_i)=\frac{1}{2}+\sum_{j=1}^{n_i}\frac{t_j^2}{2T}$. Furthermore, the sum of the square of the t_js is classically minimized when they are all equal to T/n_i, which finally yields:

$$\mathrm{AST}(\mathcal{S},M_i)\geqslant\frac{1}{2}+\frac{T}{2n_i}=\frac{1}{2}+\frac{\tau_i}{2},$$

where $\tau_i=T/n_i$. Since, this holds for all $i\in S$,

$$\mathrm{AST}(\mathcal{S},S)\geqslant\max_{i\in S}\mathrm{AST}(\mathcal{S},M_i)\geqslant\frac{1}{2}+\frac{1}{2}\max_{i\in S}\tau_i.$$

Finally, summing over all $S\in\zeta$ gives

$$\mathrm{COST}(\mathcal{S})=\sum_{S\in\mathcal{S}}p_S\,\mathrm{AST}(\mathcal{S},S)\geqslant\frac{1}{2}+\frac{1}{2}\sum_{S\in\zeta}p_S\max_{i\in S}\tau_i=\mathrm{LB}(\tau).$$

As no more than T news items can be broadcast in a time interval of length T, we have $\sum_{i=1}^{n}n_i\leqslant T$, *i.e.*,

$$\sum_{i=1}^{n}\frac{1}{\tau_i}\leqslant 1.$$

We now state our lower bound, on which the algorithms presented in Sections 4 and 5 rely.

Proposition 3. *The following non-linear convex minimization problem* LB *is a lower bound on the optimum cost* OPT.

$$\mathrm{LB}=\begin{cases}\underset{\tau>0}{\text{Minimize}}\ \frac{1}{2}+\frac{1}{2}\sum_{S\in\zeta}p_S\max_{i\in S}\tau_i\\[2ex] \text{such that: }\frac{1}{\tau_1}+\cdots+\frac{1}{\tau_n}\leqslant 1\end{cases}\tag{1}$$

There exists a unique solution τ^* to the minimization problem LB*, and one can compute in polynomial time a feasible solution τ' such that* $\mathrm{LB}(\tau')\leqslant\mathrm{LB}+\frac{1}{4}$.

Proof. By Proposition 2 and Lemma 1, clearly $\mathrm{LB}\leqslant\mathrm{OPT}$. Now, the objective function $\mathrm{LB}(\tau)$ is a continuous convex function, and is minimized over a strictly convex closed domain $D=\{\tau\in(\mathbb{R}_+^*)^n:1/\tau_1+\cdots+1/\tau_n\leqslant 1\}$. Note that the Round Robin schedule that broadcasts cyclically M_1 to M_n, shows that $\mathrm{OPT}\leqslant n$ and then $\mathrm{LB}\leqslant n$. Since, $\mathrm{LB}(\tau)>n$ as soon as for some $S\in\zeta$ and some $i\in S$, $\tau_i>n/p_S$, the minimum of $\mathrm{LB}(\tau)$ is in fact obtained on the compact set $D' = D \cap (0, \frac{n}{\min_{S\in\zeta} p_S}]^n$. Since $\mathrm{LB}(\tau)$ is continuous, there exists $\tau^* \in D'$ such that $\mathrm{LB}(\tau^*) = \mathrm{LB}$. Assume that two such optimal solutions exist, say τ_1 and τ_2. Since the objective function and the domain are convex, $\tau' = \frac{\tau_1+\tau_2}{2}$ is also an optimal solution. But τ' lies in the interior of the domain, and then scaling it down by some factor $\lambda < 1$ allows to obtain a better feasible solution: $\mathrm{LB}(\lambda\tau') < \mathrm{LB}(\tau') = \mathrm{LB}$, contradiction. The optimal solution τ^* to LB is then unique. Furthermore, $1/\tau_1^* + \cdots + 1/\tau_n^* = 1$.

A feasible solution τ' within an additive error of $\frac{1}{4}$ of the optimum value can be computed in polynomial time using the ellipsoid method (*e.g.*, see [13]). The only ingredient needed is a separation oracle. Note that non-linear minimization problem LB can be restated as follows:

$$\mathrm{LB} = \begin{cases} \underset{\substack{\tau_1,\ldots,\tau_n > 0\\ \sigma_{S_1},\ldots,\sigma_{S_k} > 0}}{\text{Minimize}} & \frac{1}{2} + \sum_{S\in\zeta} p_S\sigma_S \\ \text{such that:} & \tau_i \leqslant \sigma_S \qquad \forall S, \forall i \in S \\ & \frac{1}{\tau_1} + \cdots + \frac{1}{\tau_n} \leqslant 1 \end{cases} \tag{2}$$

We just need to provide a separation oracle for each of these constraints. Only the last one is non-linear. But, for any solution $\tilde{\tau}$ violating this constraint, the tangent hyperplane to the differentiable convex surface $\partial D = \{\tau : \frac{1}{\tau_1} + \cdots + \frac{1}{\tau_n} = 1\}$ at the projection of $\tilde{\tau}$ on ∂D, provides in polynomial time a separation oracle for $\tilde{\tau}$.

3.2 Requests Correlations Mislead Previous Approaches

We show in this section that previously known algorithms (*e.g.*, [2,3,5,17,19,7]) can generate schedules with cost *arbitrarily larger* than the optimum value when requests are correlated.

In previous approaches, only the requests probability π_i for *each individual item* M_i are used. Since item M_i is requested for each request for a set S containing i, the probability π_i that an individual item M_i is requested by some user, is proportional to $\sum_{S:i\in S} p_S$, *i.e.*,

$$\pi_i = \sum_{S:i\in S} \frac{p_S}{\langle\zeta\rangle}, \quad \text{where } \langle\zeta\rangle \text{ denotes the average size of a set } \langle\zeta\rangle = \sum_{S\in\zeta} |S| p_S.$$

All previous approaches (*e.g.*, [2,5,17,19]) then construct a schedule such that each item M_i is broadcast every $\Theta(\vartheta_i)$ where ϑ is given by the "square-root rule" (*e.g.*, [2,5]):

$$\vartheta_i = \frac{\sum_{j=1}^n \sqrt{\pi_j}}{\sqrt{\pi_i}}.$$

The example bellow shows that previously known solutions using these ϑ_is can generate schedules with cost arbitrarily large with respect to the optimum.

Example 1. Consider $2n$ news items $A_1, \dots, A_n, B_1, \dots, B_n$ and n sets:

$$\zeta = \{\{A_1, \dots, A_n, B_1\}, \dots, \{A_1, \dots, A_n, B_n\}\},$$

where each set is requested with probability $\frac{1}{n}$. The request probability for each individual item is then $\pi_A = \frac{1}{n+1}$ for the A_is and $\pi_B = \frac{1}{n(n+1)}$ for the B_js. Thus, for all A_is, $\vartheta_A = \frac{n\sqrt{\pi_A}+n\sqrt{\pi_B}}{\sqrt{\pi_A}} = \Theta(n)$ and, for all B_js, $\vartheta_B = \frac{n\sqrt{\pi_A}+n\sqrt{\pi_B}}{\sqrt{\pi_B}} = \Theta(n\sqrt{n})$. Consider now a schedule that broadcasts each A_i every $\Theta(\vartheta_A)$ and each item B_j every $\Theta(\vartheta_B)$, as in previous approaches.

Assuming that the requests were independent, each item A_i and each item B_j are respectively requested with probabilities π_A and π_B, and the average service time of this schedule would be:

$$n \cdot \pi_A \cdot \Theta(\vartheta_A) + n \cdot \pi_B \cdot \Theta(\vartheta_B) = \Theta(n \cdot \frac{1}{n} \cdot n + n \cdot \frac{1}{n^2} \cdot n\sqrt{n}) = \Theta(n).$$

But, since the requests are not independent, the cost of any schedule that broadcasts each item every $\Theta(\vartheta)$, is at least, according to Lemma 1:

$$\mathrm{LB}(\Theta(\vartheta)) = \frac{1}{2} + \frac{1}{2} \cdot n \cdot \frac{1}{n} \cdot \max(\Theta(\vartheta_A), \Theta(\vartheta_B)) = \Omega(n\sqrt{n}).$$

In fact, the cost of the Round Robin schedule that broadcasts each item in turn cyclically, is only $\Theta(n)$ (every request is served after at most $2n$ time units).

We conclude that treating correlated requests individually can yield schedules with cost *as large as $\Omega(\sqrt{n})$ times the optimum*, where n is the number of items. □

As a consequence, considering the dependencies between requests for different items is essential to obtain good performances to data broadcast systems. The next sections show how we use our lower bound to compute efficient broadcast schedules.

4 Randomized Approximation

The lower bound in Proposition 3 suggests that each news item M_i should be broadcast every τ_i^*, which is approximated by τ_i' given by the ellipsoid method. We analyze here a classic randomized scheduler that chooses the next item M_i to be broadcast with probability $1/\tau_i'$. We show that it achieves a $2H_n$ factor approximation, and that our analysis is tight by providing a family of tight instances.

Lemma 2. *The expected cost of the random schedule $\mathcal{S}$ output by Algorithm 1 is at most*

$$\mathbb{E}[\mathrm{COST}(\mathcal{S})] \leqslant H_n \sum_{S \in \zeta} p_S \max_{i \in S} \tau_i',$$

where $H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$.

Algorithm 1 Randomized scheduler

Computes τ' given by Proposition 3.
for all $t \geqslant 0$ **do**
  Pick $i \in \{1, \ldots, n\}$ with probability $1/\tau'_i$.
  Broadcast M_i at time t.
end for

Proof. The proof relies on a classic coupon collector argument (*e.g.*, see [12]). Consider a request for a set S of size q. At any time t, each news item in S is broadcast with probability at least $\rho_S = \min_{i \in S} 1/\tau'_i$. Let t_j be the random variable for the time elapsed since the issue of the request until j distinct items of S have been downloaded. The expected service time to the request is clearly $\mathbb{E}[t_q]$. Let $T_j = t_j - t_{j-1}$ for $j = 1, \ldots, q$, with $t_0 = 0$. Clearly, $\mathbb{E}[t_q] = \sum_{j=1}^{q} \mathbb{E}[T_j]$. Since after time t_{j-1}, $j-1$ distinct items have been downloaded, the probability to get a new item from S in each time slot between t_{j-1} and t_j is at least $\rho_{S,j} = (q - j + 1)\rho_S$. Thus,

$$\mathbb{E}[T_j] \leqslant \sum_{t \geqslant 0} (t+1)(1 - \rho_{S,j})^t \rho_{S,j} = \frac{1}{(q - j + 1)\rho_S} = \frac{\max_{i \in S} \tau'_i}{q - j + 1}.$$

The expected service time to a request for S in the random schedule $\mathcal{S}$ is then at most:

$$\mathbb{E}[\mathrm{AST}(\mathcal{S}, S)] \leqslant \sum_{j=1}^{q} \frac{1}{q - j + 1} \max_{i \in S} \tau'_i \leqslant H_n \max_{i \in S} \tau'_i.$$

Then summing over all the sets $S \in \zeta$ yields the result.

Theorem 1. *Algorithm 1 is a polynomial time randomized* $2H_n$*-approximation for the customized newspaper problem.*

The next example shows that our analysis of the randomized scheduler is tight.

Example 2. Consider k disjoint sets $S_1, \ldots, S_k$ of q news items each, $S_j = \{M_{1,j}, \ldots, M_{q,j}\}$ for $j = 1, \ldots, k$. Each set is requested with probability $1/k$. By symmetry of the instance, the unique solution to LB is obviously $\tau^*_{i,j} = kq$ for all $1 \leqslant i \leqslant q$ and $1 \leqslant j \leqslant k$. Since each item of each set S_j is broadcast independently in each time slot with probability $1/kq$, by the classic coupon collector argument (*e.g.*, see [12]), the expected cost of the random schedule $\mathcal{S}$ is then exactly:

$$\mathbb{E}[\mathrm{COST}(\mathcal{S})] = H_q \cdot k \cdot q.$$

Now consider the Round Robin schedule $\mathcal{R}$ that broadcasts each item of each set cyclically as follows: $M_{1,1} \ldots M_{q,1} M_{1,2} \ldots M_{q,2} \ldots M_{1,k} \ldots M_{q,k} M_{1,1} \ldots$. Taking $k = \ln m$ and $q = m$, we get for $m \to \infty$:

$$\mathbb{E}[\mathrm{COST}(\mathcal{S})] \sim H_m \cdot m \ln m \quad \text{and} \quad \mathrm{COST}(\mathcal{R}) \sim \frac{m \ln m}{2},$$

Algorithm 2 Determistic 4-approximation algorithm

Computes τ' given by Proposition 3.
for all $i \in \{1, \ldots, n\}$ **do**
  $\beta_i \leftarrow 2^{\lceil \log_2 \tau'_i \rceil}$
end for
Broadcast the news items (M_i) according to a perfectly periodic schedule built on the periods (β_i).

which implies that the optimum cost is at least at a factor $2H_n$ of the optimum value where $n = m \ln m$ is the number of news items. □

5 Deterministic 4-Approximation

The main issue with the randomized algorithm above is that items of a given set appear in a random order which is inefficient and introduces a H_n factor to the cost (due to a "coupon collector phenomenon"). A special type of schedules, known as *perfectly periodic schedules*, is of particular interest here. These schedules were implicitly introduced in [3] in a very simple setting and studied in details by [7,9] in the context of bluetooth and sensor networks. In these networks, the clients access to a communication channel by means of time multiplexing. In these protocols, the ith client is given a period β_i and an offset o_i, and can emit only during time slots $o_i + k\beta_i$, with $k = 1, 2, \ldots$, each of these slots being different for each client. Such a schedule is said to be perfectly periodic since each client i gets access to the channel exactly every β_i time. These perfectly periodic schedules find an interesting new application here: their regularity solves efficiently the dependencies between items by forcing an order on them.

In [3,7], the authors show the following lemma that states that if the requested periods β_is are power of 2 and satisfies the maximum bandwidth constraint $1/\beta_1 + \cdots + 1/\beta_n \leqslant 1$, then one can construct in polynomial time a perfectly periodic schedule for this set of periods.

Lemma 3 (Perfectly periodic schedules [3,7]). *Given a set of periods* (β_i) *such that for all* i*,* $\beta_i = 2^{j_i}$ *for some integer* j_i*, and* $1/\beta_1 + \cdots + 1/\beta_n \leqslant 1$*, one can construct in polynomial time a perfectly periodic schedule that broadcasts each news item* M_i *exactly every* β_i*, for all* i*.*

Algorithm 2 follows the lines of the "power-of-two" heuristic given in [3]: first round up each period τ'_i (given by Proposition 3) to the closest power of 2, β_i, and then constructs a perfectly periodic schedules for the set of periods (β_i).

The next theorem shows that with our choice of τ' given by our lower bound, this algorithm achieves an approximation ratio of 4. This algorithm is, to our knowledge, the first to obtain a constant factor approximation for the data broadcast problem with dependencies involving more than two items.

Theorem 2. *Algorithm 2 is a polynomial time deterministic* 4*-approximation for the customized newspaper problem.*

Proof. Since the periods τ'_i are rounded up to the closest power of 2, β_i, we have $\tau'_i \leqslant \beta_i \leqslant 2\tau'_i$. Then $1/\beta_1 + \cdots + 1/\beta_n \leqslant 1$ and Lemma 3 gives in polynomial time a perfectly periodic schedule $\mathcal{S}$ that broadcasts each item M_i exactly every β_i. Every request for a set $S \in \zeta$ then waits at most β_i before downloading each M_i, with $i \in S$. The average service time to a request for S is then bounded by

$$\mathrm{AST}(\mathcal{S}, S) \leqslant 1 + \max_{i \in S} \beta_i \leqslant 1 + 2 \max_{i \in S} \tau'_i .$$

Summing over $S \in \zeta$ yields finally the following bound on the cost of $\mathcal{S}$:

$$\mathrm{COST}(\mathcal{S}) \leqslant 1 + 2 \sum_{S \in \zeta} p_S \max_{i \in S} \tau'_i \leqslant 1 + 4 \left(\mathrm{LB}(\tau') - \frac{1}{2}\right) \leqslant 1 + 4\left(\mathrm{LB} - \frac{1}{4}\right) \leqslant 4 \,\mathrm{OPT} .$$

6 Adding Broadcast Costs

A classic extension of data broadcast problem includes broadcast costs (*e.g.*, see [5]). An instance of the customized newspaper broadcast problem *with* broadcast costs associates a cost c_i to each news item M_i, which is applied every time M_i is broadcast. The goal is now to find a schedule $\mathcal{S}$ that minimizes to sum of two quantites: the average service time to a random request, $\mathrm{COST}(\mathcal{S})$, and the average broadcast cost, $\mathrm{BC}(\mathcal{S})$. The *average broadcast cost* of $\mathcal{S}$ over a time interval $I = [t_1, t_2]$ is defined as:

$$\mathrm{BC}(\mathcal{S}, I) = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2 - 1} c_{\mathcal{S}(t)}$$

The average broadcast cost of $\mathcal{S}$ is then defined as the asymptotic value of this quantity:

$$\mathrm{BC}(\mathcal{S}) = \limsup_{t \to \infty} \mathrm{BC}(\mathcal{S}, [0, t]).$$

The cost of a schedule is then defined as $\mathrm{COST}_{BC}(\mathcal{S}) = \mathrm{COST}(\mathcal{S}) + \mathrm{BC}(\mathcal{S})$. We denote by OPT_{BC} the optimum cost: $\mathrm{OPT}_{BC} = \inf_{\mathcal{S}} \mathrm{COST}_{BC}(\mathcal{S})$. As before, the following lower bound is obtained by bounding the cost of periodic schedules from which we derive a deterministic 4-approximation (proofs are omitted due to space constraints).

Proposition 4. LB_{BC} *is a lower bound on the optimum cost* OPT_{BC}*, where:*

$$\mathrm{LB}_{BC} = \begin{cases} \underset{\tau > 0}{\text{Minimize}} \ \dfrac{1}{2} + \dfrac{1}{2} \displaystyle\sum_{S \in \zeta} \left(p_S \max_{i \in S} \tau_i\right) + \sum_{i=1}^{n} \frac{c_i}{\tau_i} \\ \\ \text{such that: } \dfrac{1}{\tau_1} + \cdots + \dfrac{1}{\tau_n} \leqslant 1 \end{cases}$$

There exists a unique solution τ^*_{BC} to the convex minimization problem LB_{BC}*, and one can compute in polynomial time a feasible solution τ'_{BC} such that* $\mathrm{LB}_{BC}(\tau'_{BC}) \leqslant \mathrm{LB}_{BC} + \frac{1}{4}$.

Theorem 3. *Using periods τ'_{BC} instead of τ' in Algorithms 1 and 2 yields respectively a randomized $2H_n$-approximation and a deterministic 4-approximation for the customized newspaper problem with broadcast costs.*

References

1. S. Acharya. *Broadcast Disks: Dissemination-based Management for Assymmetric Communication Environments.* PhD thesis, Brown University, May 1998.
2. M. H. Ammar and J. W. Wong. The design of teletext broadcast cycles. In *Performance Evaluation*, volume 5(4), pages 235–242, 1985.
3. S. Anily, C. A. Glass, and R. Hassin. The scheduling of maintenance service. 1995.
4. N. Bansal, M. Charikar, S. Khanna, and J. S. Naor. Approximation the average response time in broadcast scheduling. In *Proc. of ACM-SIAM SODA*, 2005.
5. A. Bar-Noy, R. Bhatia, J. S. Naor, and B. Schieber. Minimizing service and operation costs of periodic scheduling. *Math. Of Op. Research*, 27(3):518–544, 2002.
6. A. Bar-Noy, J. S. Naor, and B. Schieber. Pushing dependent data in clients-providers-servers systems. *Wireless Networks*, 9:421–230, 2003.
7. A. Bar-Noy, A. Nisgav, and B. Patt-Shamir. Nearly optimal perfectly-periodic schedules. In *Proc. of ACM PODC'01*, pages 107–116, 2001.
8. A. Bar-Noy and Y. Shilo. Optimal broadcasting of two files over an asymmetric channel. *J. Parallel Distrib. Comput.*, 60(4):474–493, 2000.
9. Z. Brakerski, V. Dreizin, and B. Patt-Shamir. Dispatching in perfectly-periodic schedules. *J. Algorithms*, 49(2):219–239, 2003.
10. K. Cai, H. Lin, and C. Chen. Correlation-based data broadcasting in wireless networks. In *LNCS Proc. of 2nd BNCOD*, volume 3567, Jul. 2005.
11. J. Edmonds and K. Pruhs. Multicast pull scheduling: When fairness is fine. *Algorithmica*, 36(3):315–330, 2003.
12. W. Feller. *An Introduction to Probability Theory*, volume I. John Willey & Sons, 3rd edition, 1968.
13. M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization.* Springer, 1988.
14. J.-L. Huang and M.-S. Chen. Broadcast program generation for unordered queries with data replication. In *Proc. of the ACM SAC*, pages 866–870, Mar. 2003.
15. J.-L. Huang and M.-S. Chen. Dependent data broadcasting for unordered queries in a multiple channel mobile environment. *IEEE Trans. on Knowledge and Data Engineering*, 16(9):1143–1156, 2004.
16. C. Kenyon and N. Schabanel. The data broadcast problem with non-uniform transmission time. *Algorithmica*, 35:147–175, 2002.
17. C. Kenyon, N. Schabanel, and N. Young. Polynomial-time approximation scheme for data broadcast. In *Proc. of ACM STOC*, pages 659–666, 2000.
18. F. J. Ovalle-Martínez, J. S. González, and I. Stojmenović. A parallel hill climbing algorithm for pushing dependent data in clients-providers-servers systems. *Mobile Network and Applications*, 9:257–264, 2004.
19. N. Schabanel. The data broadcast problem with preemption. In *LNCS Proc. of the 17th STACS*, volume 1770, pages 181–192, 2000.