

Low Delay Scheduling in Wireless Network

Kyomin Jung
Mathematics, MIT
kmjung@mit.edu

Devavrat Shah
EECS, MIT
devavrat@mit.edu

Abstract—In a wireless network, a sophisticated algorithm is required to schedule simultaneous wireless transmissions while satisfying interference constraint that *two neighboring nodes can not transmit simultaneously*. The scheduling algorithm need to be excellent in performance while being simple and distributed¹ so as to be implementable. The result of Tassiulas and Ephremides (1992) imply that the algorithm, scheduling transmissions of nodes in the ‘maximum weight² independent set’ (MWIS) of network graph, is throughput optimal. However, algorithmically the problem of finding MWIS is known to be NP-hard and hard to approximate. This raises the following questions: is it even possible to obtain throughput optimal simple, distributed scheduling algorithm? if yes, is it possible to minimize delay of such an algorithm?

Motivated by these questions, we first provide a distributed throughput optimal algorithm for any network topology. However, this algorithm may induce exponentially large delay. To overcome this, we present an order optimal delay algorithm for any non-expanding³ network topology. Networks deployed in geographic area, like wireless networks, are likely to be of this type. Our algorithm is based on a novel distributed graph partitioning scheme which may be of interest in its own right. Our algorithm for non-expanding graph takes $O(n)$ total message exchanges or $O(1)$ message exchanges per node to compute a schedule.

I. INTRODUCTION

Wireless networks are becoming architecture of choice in ad-hoc networks and metro-area networks or mesh-networks. The tasks of resource allocation and scheduling are essential for good network utilization. Wireless medium being multi-access makes algorithm design for such network intrinsically different and more challenging than its wireline counterpart. Further, wireless architecture requires that algorithm be distributed and simple.

Despite these challenges, there has been an exciting recent progress based on optimization frame-work to characterize good resource allocation algorithm that combine resource allocation and scheduling (see [1], for example). However, these solutions either assume availability of good scheduling algorithm or use of imperfect scheduling (which will lead to poor performance). In this paper, we are interested in designing simple to implement, distributed and high-performance scheduling algorithms.

¹In this paper, by distributed we mean that algorithm operating at nodes of the network can only utilize local topological information.

²Weight is an appropriate function of queue-sizes and possibly other network parameters.

³See section IV for precise definition of the non-expanding graph.

A. Scheduling in wireless network

We consider an abstract model of wireless network given by graph $G = (V, E)$ with $|V| = n$ wireless nodes and edges represented by E . We consider the classical interference model for multi-access channel which imposes the constraint that two neighboring nodes can not transmit simultaneously. Subsequently, simultaneously transmitting nodes must correspond to *independent set* of G . When nodes are given weights, the weight of an independent set is the summation of weights of nodes in the independent set.

Based on results of Tassiulas and Ephremides [2] and optimization formulation of resource allocation, a throughput optimal algorithm for resource allocation and scheduling is equivalent to finding ‘maximum weight independent set’ (MWIS) in G every time, where weight is function of queue-size and other network parameters. We refer interested readers to a recent survey by Lin, Shroff and Srikant [1] where a detailed account of this development is given for wireless network with *node-exclusive* interference model (aka matching constraints).

B. Previous work

We present a brief summary of previous work on network scheduling algorithms. The result by Tassiulas and Ephremides [2] established that ‘max-weight scheduling’ policy is throughput optimal for a large class of scheduling problems. This result has been very influential in design of scheduling algorithms since then. Application to input-queued switches led to an excellent development of theory and practice of algorithms for scheduling under matching constraints: notably, the results of [3]–[9]. A recent interest in wireless network has led to proposal of distributed scheduling algorithms under matching constraints [10]–[13]. Most of these algorithms, based on finding maximal matching, guarantee only a constant fraction of throughput. Recently, Modiano, Shah and Zussman [14] exhibited a throughput optimal distributed scheduling algorithm with matching constraints. This algorithm, as discussed in [1], easily extends to provide throughput optimal algorithm for resource allocation and scheduling problem under matching constraints.

Apart from matching constraints, other scheduling constraints have received limited attention primarily due to inherent hardness of the other constraints. For example, Sharma, Mazumdar and Shroff [15] identify that max-weight scheduling with K –hop matching constraint becomes an instance of computationally hard combinatorial optimization problem.

They provide a centralized throughput optimal algorithm for unit disk graphs based on work by Hunt et. al. [16]. However, hardness of the max. wt. problem does not imply non-existence of throughput optimal algorithm. Specifically, in this paper we provide throughput optimal *distributed* algorithm (ALGO I) for hard independent set constraint (it will naturally extend to the K -hop matching model of [15] as well).

C. Contribution

The maximum weight independent set (MWIS) algorithm is throughput optimal for our setup. However, finding MWIS is NP-hard [17] and hard to approximate within $n^{1-o(1)}$ ($B/2^{O(\sqrt{\log B})}$ for degree B graph) factor [6]. This raises a challenging question: is it even possible to have any throughput optimal, polynomial (in n) time distributed algorithm? if yes, how does it's delay scale? more generally, is it possible to have both throughput and delay optimal polynomial time distributed algorithm for practical network topology? As the main contribution of this paper, we answer these tantalizing questions in affirmative.

First, we exhibit a distributed throughput optimal scheduling algorithm that takes $O(n^3)$ total operations to compute schedule (section III). By computing schedule once in $O(n^3)$ time, the cost per time is $O(1)$. Such *lazy* schedule is throughput optimal. That is, it is not difficult to have stable scheduling algorithm even when scheduling constraints are very hard. However, this algorithm is likely to induce exponentially large (in n) delay. This suggests that the complexity of algorithm trades off with delay rather than throughput.

Next, we present a delay (order) optimal scheduling algorithm that essentially finds excellent approximation to MWIS in $O(n)$ operations in total or $O(1)$ operations per node for 'practical networks' modeled as non-expanding graphs (section IV). The algorithm is distributed and simple. It is based on a new randomized distributed graph partitioning with certain properties. Next, we provide definition and examples of non-expanding graph without making them mysterious for reader till later in the paper⁴.

Non-expanding graphs. Given a graph $G = (V, E)$, let $D : V \times V \rightarrow \mathbb{R}_+$ be a metric on nodes of V . A special metric induced by G is the shortest-path distance metric, $D_G : V \times V \rightarrow \mathbb{R}_+$ where $D_G(u, v)$ is the length of shortest path connecting u, v (∞ if u, v are not connected). With respect to a given metric D (not necessarily D_G), for a given vertex $v \in G$ and $i \in \mathbb{N}$, let $f_v(i) = |\{w \in V : i-1 < D(v, w) \leq i\}|$, and $F_v(i) = |\{w \in V : D(v, w) \leq i\}| = \sum_{j=1}^i f_v(j)$.

Definition 1 A graph G is said to be "non-expanding" if there exists a metric $D : V \times V \rightarrow \mathbb{R}_+$, constants Δ, β such that

- (0) (Contracting) $D \leq D_G$, i.e. $D(u, v) \leq D_G(u, v)$, $\forall (u, v) \in V \times V$.
- (1) (Bounded neighbors) $F_v(1) \leq \Delta$ for all $v \in V$.
- (2) (Polynomial-growth)⁵ $F_v(3i) \leq \beta F_v(i) \forall i$.

⁴Alternatively, reader may skip this definition and come back to it on reaching section IV.

⁵The condition immediately implies that $F_v(k) \leq k^{\log_3 \beta}$. Hence the name polynomial-growth.

- (3) (Absence-of-thick-boundary)⁶ For any $\varepsilon > 0$, there exists constant $\ell(\varepsilon)$ such that

$$\left[\sum_{i=1}^{\ell(\varepsilon)} f_v(i)^2 \right] \leq \varepsilon F_v^2(\ell(\varepsilon)).$$

Example 1. Consider a $\sqrt{n} \times \sqrt{n}$ grid graph of n nodes in two-dimension. Then, for $D = D_G$ it is non-expanding as we have $f_v(i) = \Theta(i)$ (with $\ell(\varepsilon) = O(1/\varepsilon)$).

Example 2: Suppose there are infinitely many nodes placed in a plane (or even three dimension) so that for some $R > 0$, (a) nodes are connected to each other if they are within distance R of each other, and (b) number of nodes in any disc of radius αR is bounded above by γ and below by 1 where $\alpha \in (0, 1/2)$, $\gamma \geq 1$ are constants. Now consider any square of side-length N in plane. Let G be the graph formed by nodes within this square. Such a graph captures the characteristics of wireless network deployed in practice.

Lemma 1 The graph G , with respect to $D = d/R$ where d is the Euclidian distance, is non-expanding.

Proof: We need to show with respect to the metric $D = d/R$, graph G has the desired properties (0)-(3). Consider any two nodes $u, v \in V$. If $D_G(u, v) = \infty$, then (0) is trivially satisfied since $D(u, v) = d(u, v)/R < \infty$ for all u, v lying in a square of side N . If $D_G(u, v) < \infty$, since each edge under the property (a), is at most of length R , the Euclidian distance between u, v , $d(u, v)$ can be at most $R D_G(u, v)$ (by standard triangular inequality of metric). That is, $D(u, v) = d(u, v)/R \leq D_G(u, v)$.

To prove properties (1)-(3), it is sufficient to show that for any $v \in V$,

$$f_v(i) = |\{w : D(v, w) \in (i-1, i]\}| = \Theta(i),$$

ignoring boundary effect for large i . Now, $D(v, w) \in (i-1, i]$ is equivalent to $d(v, w) \in (R(i-1), R]$. For any $v \in V$, this corresponds to nodes that lie in certain (at least quarter) ring of width R and area $\Theta(Ri)$. This region contains $\Theta(i)$ complete discs of radius αR for any $\alpha \leq 1/2$. Therefore, we have $f_v(i) = \Omega(i)$. For similar reasons, we have $f_v(i) = O(i)$. That is, $f_v(i) = \Theta(i)$. This completes the proof of the claim and Lemma 1. ■

Remark: Finally, some remarks on our results: (a) We consider the single-hop model. However, it should be clear to an informed reader that exactly the same algorithms with different weights will provide desired *optimal* performance: weights being "difference of queue-sizes" under multi-hop model of [2] and weights being appropriate Lagrange parameters for resource allocation in multi-hop network as explained in [1]. (b) The independent set constraint is general enough abstract model to capture any combinatorial scheduling constraint. Thus, our results should extend to a large class of scheduling problem. For example, a natural adaptation of ALGO I for

⁶The condition says that no 'boundary' formed by nodes at a particular distance should have most of the nodes till range $\ell(\varepsilon)$.

K -hop matching model will provide distributed throughput optimal algorithm (thus, answering the question implicitly raised in [15]). (c) We note conceptual similarity of ALGO II with that of [16]. However, inherently the algorithm of [16] is centralized (uses dynamic programming and centralized graph partition) while ours is distributed.

II. NOTATIONS AND MODEL

As before, let $G = (V, E)$ be the undirected network graph with $|V| = n$. Let $\mathcal{N}(v) = \{u \in V : (u, v) \in E\}$ denote the set of all neighbors of $v \in V$. The time is assumed to be slotted and $\tau \in \mathcal{Z}_+$ denote the time. Each node $v \in V$ is capable of wireless transmission at unit rate to any of its neighbor. We ignore the power control for simplicity but as reader may notice, it can be easily included in the model. At each node, packets (of unit size) are arriving according to an external arrival process. Let $\bar{A}(\tau) = [\bar{A}_v(\tau)]$ denote the cumulative arrival process until time $\tau \in \mathcal{Z}_+$, i.e. $\bar{A}_v(\tau)$ be the total number of packet arrived at node v in the time interval $[0, \tau]$; $\bar{A}(\tau) = \mathbf{0}$. Let $A_v(\tau) = \bar{A}_v(\tau) - \bar{A}_v(\tau - 1)$ be the number of packets arriving at node v in time slot τ . We assume that at most one packet can arrive at a node v in a time slot, i.e. $A_v(\tau) \in \{0, 1\}$. Finally, we assume that $A_v(\cdot)$ are Bernoulli i.i.d. random variable with $\Pr(A_v(\tau) = 1) = \lambda_v$. Let $\lambda = [\lambda_v]$ denote the arrival rate vector.

For simplicity and ease of explanation, we assume that network is a single-hop⁷, i.e. data arriving at a node v is to be sent to one of its neighbors. Let $Q_v(\tau)$ denote the queue-size at node v at time τ with $Q(\tau) = [Q_v(\tau)]$. We assume the system starts empty, i.e. $Q(0) = \mathbf{0}$. Let $\bar{D}(\tau) = [\bar{D}_v(\tau)]$ denotes the cumulative departure process from $Q(\tau)$; $D(\tau) = [D_v(\tau)]$ denote the number of departures in time slot τ . Then,

$$\begin{aligned} Q(\tau) &= Q(0) + \bar{A}(\tau) - \bar{D}(\tau) = \bar{A}(\tau) - \bar{D}(\tau) \\ &= Q(\tau - 1) + A(\tau) - D(\tau). \end{aligned} \quad (1)$$

Departure happens according to the scheduling algorithm which need to satisfy interference constraint that no two neighboring nodes are transmitting data in the same time slot. To this end, let \mathcal{I} denote the set of all independent set of G . Then, at each time the scheduling algorithm schedules nodes of an independent set $I \in \mathcal{I}$ to transmit packets. In what follows, we will denote independent set I as vector $I = [I_v]$ with $I_v \in \{0, 1\}$ and $I_v = 1$ indicates that node v is in I .

We say that a system is *stable* for given λ under the particular scheduling policy if

$$\limsup_{\tau \rightarrow \infty} \mathbb{E}[Q_v(\tau)] < \infty, \quad \forall v \in V.$$

From [2], it is clear that the set of all λ for which there exists a scheduling policy so that the system is stable is given by $\Lambda = \text{Co}(\mathcal{I})$, where $\text{Co}(\mathcal{I})$ is the convex hull of \mathcal{I} in \mathbb{R}^n . Hence, we call $\text{Co}(\mathcal{I})$ the *throughput region* of the system.

⁷The model ignores multi-hop situation. However, as explained in [2], the scheduling algorithm remains maximum weight independent set with weights being ‘‘difference of queue-sizes’’. Similarly, in the context of resource allocation as explained in [1], the weights are based on ‘‘Lagrange multipliers’’.

In [2], it was shown that a ‘maximum weight independent set’ scheduling algorithm is stable for all $\lambda \in \text{Co}(\mathcal{I})$, where the schedule or independent set $I^*(\tau)$ chosen at time τ is

$$I^*(\tau) = \arg \max_{I \in \mathcal{I}} \langle I, Q(\tau - 1) \rangle,$$

with notation that $\langle A, B \rangle = \sum_{v \in V} A_v B_v$. Such an algorithm will be called to provide 100% throughput or through optimal.

As discussed before, finding max. wt. independent set can be computationally hard. In the rest of the paper, we will be interested in designing scheduling algorithms that are: (a) stable, (b) induce low average queue-size (equivalently low delay due to Little’s Law) and (c) simple and distributed.

A. Relation to other models

The above described model ignores the multi-hop setup. However, we have done so to keep exposition simple. The scheduling algorithm of interest with the independent set interference constraint remain the same as max. wt. independent set with weights being some-what different. To explain this, we give example of two such well-known scenarios as follows.

1. *Multi-hop queuing network*. This was considered in [2]. For given network G , let S be set of data-flows with arrival rate λ_s for flow $s \in S$. Let f_s, d_s denote source and destination node respective for flow $s \in S$. The routing is assumed to be pre-determined in the network. If s passes through $v \in V$ then let $h(v, s) \in V$ denote its next hop unless $v = d_s$ in which case it’s data departs from G . Let $Q_{vs}(\tau)$ denote queue-size of flow s at node v at time τ . Define

$$W_{vs}(\tau) = \begin{cases} Q_{vs}(\tau) - Q_{h(v,s)s}(\tau) & \text{if } v \neq d_s \\ 0 & \text{if } v = d_s. \end{cases}$$

Define $W_v(\tau) = \max_{s \in S} W_{vs}(\tau)$ and $W(\tau) = [W_v(\tau)]$. Then the throughput optimal (stable) algorithm of [2] chooses $I^*(\tau)$ as the schedule which is a max. wt. independent set with respect to $W(\tau - 1)$, i.e.

$$I^*(\tau) = \arg \max_{I \in \mathcal{I}} \langle I, W(\tau - 1) \rangle.$$

2. *Joint resource allocation & scheduling*. In [1], it is very well-explained that the problem of congestion control and scheduling decomposes into weakly coupled two sub-problems: (i) congestion control, and (ii) scheduling. We describe the link-level scheduling problem. We urge an interested reader to go through [1] for details. The setup of the problem is same as in example 1 described above with difference that routing is not pre-determined. The coupling of congestion control and scheduling happens via Lagrange multipliers $\mathbf{q}(\tau) = [q_e(\tau)]_{e \in E}$. With the interference model of this paper, the scheduling problem boils down to selection of max. wt. independent set $I^*(\tau)$ with respect to weight $W(\tau - 1) = [W_v(\tau - 1)]$, where $W_v(\tau - 1) = \max_{e: e=(u,v) \in E} q_e(\tau - 1)$.

We re-iterate that due to our interest in algorithm design for the max. wt. independent set, we will restrict our discussion the simple model described earlier in this paper. However, it should be clear that our algorithms apply in more general setup.

B. Technical preliminaries

We present useful technical preliminaries here. Consider a discrete time Markov chain on countable state space $S = \mathbb{N}^M$ for some finite integer M . Let $X(\tau)$ denote the random state of Markov chain at time $\tau \in \mathbb{Z}_+$. Let $X(0) = \mathbf{0}$ (can be any arbitrary good state). Let $L : S \rightarrow [0, \infty)$ and $f : S \rightarrow [0, \infty)$ be any non-negative valued functions with $L(\mathbf{0}) = 0$. The following is a well-known result and can be found in book by Meyn and Tweedie [19].

Proposition 1 *Let Markov chain be aperiodic and irreducible. Let there exist a closed and bounded set $C \subset S$ such that Markov chain satisfies the following condition: $\forall \tau \in \mathbb{Z}_+$,*

$$\mathbb{E}[L(X(\tau+1)) | X(\tau)] \leq L(X(\tau)) - f(X(\tau)) + B \mathbf{1}_{\{X(\tau) \in C\}},$$

with $B > 0$ a constant, $\sup_{x \in C} L(x) < \infty$. Then,

- (a) *Markov chain is recurrent with unique stationary distribution $\pi = [\pi(x)]_{x \in S}$ such that*

$$\pi(f) = \sum_{x \in S} \pi(x) f(x) < \infty.$$

- (b) *Further,*

$$\lim_{\tau \rightarrow \infty} \mathbb{E}[f(X(\tau))] \rightarrow \pi(f).$$

Similar result (also known as Foster's criteria) is used in most of the previous work to prove stability and obtain bound on average queue-size.

III. DISTRIBUTED STABLE ALGORITHM

We describe a simple and distributed stable algorithm, denoted by ALGO I. The algorithm uses two distributed sub-routines, RANDOM and APRX-CNT, with the following properties, explained later in this section:

- P1.** RANDOM samples independent sets of graph G in distributed manner so that each independent set has get sampled with probability at least 2^{-n} . It takes total $O(|E| + n) \leq O(n^2)$ distributed operations for any G .
- P2.** APRX-CNT(ϵ) takes given independent set and node weights W and produces an estimate of $w_I = \langle I, W \rangle$, say \hat{w} , so that $\hat{w} \in ((1-\epsilon)w_I, (1+\epsilon)w_I)$ with probability at least $1 - 3^{-n}$ in total $O(n^3)$ distributed operations for any G .

ALGO I

- o Let $I(\tau)$ be independent set schedule chosen by algorithm at time τ .
- o At time $\tau + 1$, choose $I(\tau + 1)$ as follows:
 - Generate a random independent set $R(\tau + 1)$ using RANDOM.
 - Obtain estimates \hat{w}_I, \hat{w}_R of weights of $I(\tau), R(\tau + 1)$ with respect to $Q(\tau)$ respectively using APRX-CNT($\epsilon/8$).
 - If $\hat{w}_R > \frac{(1+\epsilon/8)}{(1-\epsilon/8)} \hat{w}_I$, then set $I(\tau + 1) = R(\tau + 1)$. Else, set $I(\tau + 1) = I(\tau)$.
- o Repeat the above algorithm every time.

Before we establish that algorithm is stable (or throughput optimal), we will describe the sub-routines RANDOM, APRX-CNT and their properties useful in the analysis.

A. RANDOM

The algorithm RANDOM is described as follows.

RANDOM

- o Each node $v \in V$ chooses $I_v = 0$ or 1 with probability $1/2$ independently.
- o If node v finds any $u \in \mathcal{N}(v)$ such that $I_u = 1$, it immediately sets $I_v = 0$.
- o Now, output $I = [I_v]$ as a sampled independent set.

Proof: [Property **P1**] Since each node selects 0 or 1 independently with probability $1/2$, each one of the 2^n assignment of $\{0, 1\}^n$ is equally likely (i.e. probability 2^{-n}). Each independent set corresponds to one the assignment in $\{0, 1\}^n$. As part of the algorithm, if random node assignment is by itself an independent set, then the final output is this independent set. Thus, each independent set has at least 2^{-n} probability of being selected.

Now, the number of operations done by algorithm are n random coin tosses and at most $2|E|$ operations for nodes to check values of their neighbors. These are all extremely simple distributed operations and total are $O(|E| + n)$. For any graph $|E| = O(n^2)$. Hence, it is $O(n^2)$. ■

B. APRX-CNT and its properties

The purpose of algorithm is to compute summation of node weights (approximately) for given independent set. A useful property of this algorithm is that all nodes obtain the same estimate and hence allows for distributed decision in ALGO I. Now, some useful probabilistic facts:

- F1.** Let X_1, \dots, X_k be independent random variables with exponential distribution and parameters r_1, \dots, r_k . Then, $X_* = \min_{1 \leq i \leq k} X_i$ has exponential distribution with parameter $\sum_{i=1}^k r_i$.
- F2.** Let Y_1, \dots, Y_m be independent exponential random variables with parameter r . Let $S_m = \frac{1}{m} \sum_{i=1}^m Y_i$. Then, for $\gamma \in (0, 1/2)$

$$\Pr(S_m \notin (1-\gamma)r^{-1}, (1+\gamma)r^{-1}) \leq 2 \exp(-\gamma^2 m/2).$$

The **F1** is well-known about exponential distribution; the **F2** follows from Cramer's Theorem [20] about large deviation estimation for exponential distribution.

Given an independent set $I = [I_v]$ and node weights $W = [W_v]$, **F1** and **F2** can be used to compute this weight approximately as follows: each node $v \in V$ draws an independent exponential random variable with parameter W_v (nodes with $W_v = 0$ or $I_v = 0$ do not participate); then they compute minimum, say X_* of these random numbers in distributed fashion by iteratively asking their neighbors for their estimates of minimum. Nodes should terminate this process after $\Theta(n)$ transmissions. Repeat this for m times to obtain minimums $X_*(i), 1 \leq i \leq m$.

Now set $S_m = \frac{1}{m} \sum_{i=1}^m X_*(i)$ and declare $Z_m = 1/S_m$ as an estimate of weight of independent set, i.e. $\langle I, W \rangle$.

Now, given small enough ε it follows from **F1**, **F2** that by selecting $m = O(\varepsilon^{-2}n)$, we obtain estimate of weight of an independent set, say $\hat{w}(I)$ such that

$$\Pr(\hat{w}(I) \notin ((1-\varepsilon)\langle I, W \rangle, (1+\varepsilon)\langle I, W \rangle)) \leq 3^{-n}. \quad (2)$$

Computation of a single minimum over the network can be done in a distributed manner in many ways. We skip the details here in interest of space. However, we refer an interested reader to see [21] for interesting account on such algorithms. The minimum computation takes total $O(n^2)$ exchanges or $O(n)$ per node. This provides property **P2**.

C. ALGO I: stability and complexity

Complexity. The algorithm ALGO I uses two sub-routines, RANDOM and APRX-CNT at every time-slot. Properties **P1** and **P2** imply that these two algorithms take $O(n^2)$ and $O(n^3)$ total (network-wide) operations for fixed ε . Thus, algorithm ALGO I performs $O(n^3)$ net operations to compute a schedule. The complexity burden can be reduced by making the algorithm *lazy* as follows: find schedule every T time-steps and use the same schedule for in between the T steps. As long as T is finite, the algorithm remains stable and the average queue-size increases only by $O(nT)$. Thus, by choice of $T = \Theta(n^3)$, the same conclusion as in Theorem 2 can be obtained with amortized cost of $O(1)$ operation per time-step.

Stability. Now, throughput optimality of ALGO I.

Theorem 2 *The algorithm ALGO I based on RANDOM and APRX-CNT is stable as long as $\lambda \in (1-\varepsilon)\text{Co}(\mathcal{I})$ for any $\varepsilon > 0$. Further,*

$$\lim_{\tau \rightarrow \infty} \mathbb{E}[\langle Q(\tau), \mathbf{1} \rangle] = O(6^n).$$

Proof: At time τ , define Lyapunov function

$$L(\tau) = \langle Q(\tau), Q(\tau) \rangle = \sum_v Q_v^2(\tau).$$

We will study the average drift in $L(\cdot)$ at time-slots $\tau_k = kT$ for large enough T (will be 2.2^n) so that for $\lambda \in (1-\varepsilon)\text{Co}(\mathcal{I})$

$$\mathbb{E}[L(\tau_{k+1})|Q(\tau_k)] \leq L(\tau_k) - \frac{0.4\varepsilon}{n} \langle Q(\tau_k), \mathbf{1} \rangle + B, \quad (3)$$

for some large enough (exponentially dependent on n) B . This will immediately imply that

$$\mathbb{E}[L(\tau_{k+1})|Q(\tau_k)] \leq L(\tau_k) - \phi \langle Q(\tau_k), \mathbf{1} \rangle + B \mathbf{1}_{\{Q(\tau_k) \in C\}}, \quad (4)$$

for some closed bounded set C , constant B and $\phi > 0$. By Proposition 1 and fact that number of arrivals in a time interval of length T is at most nT , we will obtain the desired conclusion that

$$\limsup_{\tau \rightarrow \infty} \mathbb{E}[Q_v(\tau)] < \infty.$$

Next, we proceed towards proving (3). Given $Q(\tau_k) = Q(kT)$, we wish to study $L(\tau_{k+1}) - L(\tau_k)$: let $I(\tau)$ be independent set schedule chosen by ALGO I at τ . Define

$$\Delta_v(\tau+1) = A_v(\tau+1) - D_v(\tau+1).$$

From queueing dynamics 1,

$$\begin{aligned} L(\tau+1) - L(\tau) &= \langle Q(\tau+1), Q(\tau+1) \rangle - \langle Q(\tau), Q(\tau) \rangle \\ &= \sum_v (Q_v(\tau+1) - Q_v(\tau))(Q_v(\tau+1) + Q_v(\tau)) \\ &= \sum_v \Delta_v(\tau+1)(2Q_v(\tau) + \Delta_v(\tau+1)) \\ &= \sum_v \Delta_v^2(\tau+1) + 2Q_v(\tau)\Delta_v(\tau+1). \end{aligned} \quad (5)$$

We will use the following facts: for all τ , (1) $Q_v(\tau)D_v(\tau+1) = Q_v(\tau)I_v(\tau+1)$, (2) $\Delta_v^2(\tau+1) \leq 1$. By telescopic summation of (5) for $\tau = \tau_k, \dots, \tau_{k+1}-1$, we obtain

$$L(\tau_{k+1}) - L(\tau_k) \leq nT + 2 \sum_{\tau=\tau_k}^{\tau_{k+1}-1} \langle Q(\tau), \Delta(\tau+1) \rangle. \quad (6)$$

Since arrival process is Bernoulli i.i.d. with rate vector λ ,

$$\begin{aligned} \mathbb{E}[L(\tau_{k+1}) - L(\tau_k)|Q(\tau_k)] &\leq nT + 2 \sum_{\tau=\tau_k}^{\tau_{k+1}-1} \mathbb{E}[\langle Q(\tau), \lambda - I(\tau+1) \rangle | Q(\tau_k)]. \end{aligned} \quad (7)$$

We know that $\lambda \in (1-\varepsilon)\text{Co}(\mathcal{I})$, i.e.

$$\lambda \leq \sum_j \alpha_j I_j, \quad \alpha_j \geq 0, \quad I_j \in \mathcal{I}, \quad \sum_j \alpha_j = 1 - \varepsilon.$$

Define

$$I^*(\tau) = \arg \max_{I \in \mathcal{I}} \langle Q(\tau-1), I \rangle, \quad W^*(\tau) = \langle Q(\tau-1), I^*(\tau) \rangle,$$

$$W(\tau) = \langle Q(\tau-1), I(\tau) \rangle, \quad \delta(\tau) = W^*(\tau) - W(\tau).$$

Now, since at most n arrival and n departures can happen in a time-slot, we have $|W^*(\tau+s) - W^*(\tau)| \leq 2ns$ for every τ, s . Above discussions, definitions and some re-arrangement yields the following.

$$\begin{aligned} \mathbb{E}[L(\tau_{k+1}) - L(\tau_k)|Q(\tau_k)] &\leq nT + 2 \sum_{\tau=\tau_k}^{\tau_{k+1}-1} \mathbb{E}[\delta(\tau+1) - \varepsilon W^*(\tau+1)|Q(\tau_k)] \\ &\leq nT + 4nT^2 - \varepsilon T W^*(\tau_k) + 2 \sum_{\tau=\tau_k}^{\tau_{k+1}-1} \mathbb{E}[\delta(\tau)|Q(\tau_k)]. \end{aligned} \quad (8)$$

Now, we bound term $\sum_{\tau} \mathbb{E}[\delta(\tau)|Q(\tau_k)]$. For this we will use property of ALGO I. Note that so far, the derivation was independent of algorithm and will be used in proof of Theorem 5 later. To this end, first some useful definition and facts. Define

$$Z = \inf_{m \geq 1} \{R(\tau_k + m) = I^*(\tau_k + m)\},$$

$$Z_1 = \inf_{m \geq 0} \{\mathbf{P2} \text{ of APRX-CNT does not hold at } \tau_k + Z + Z_1\}.$$

By **P1**, we know that

$$\mathbb{E}[\min\{Z, T\}|Q(\tau_k)] \leq \mathbb{E}[Z|Q(\tau_k)] = \mathbb{E}[Z] \leq 2^n.$$

Define $\hat{T} = T - \min T, Z_1$. By **P2** and union bound suggest that $Z_1 \leq T$ with probability at most $T3^{-n}$. Hence,

$$\mathbb{E}[\hat{T}|Q(\tau_k)] = \mathbb{E}[\hat{T}] \leq T^2 3^{-n}.$$

Now, we are ready to bound $\sum_{\tau} \mathbb{E}[\delta(\tau)|Q(\tau_k)]$: let $A \subset [\tau_k + 1, \tau_{k+1}]$ be defined as $A = [\tau_k + Z, \tau_k + Z + Z_1]$. Let $B = [\tau_k + 1, \tau_{k+1}] - A$. On the starting time of A , the RANDOM picks the max. wt. ind. set of that time, i.e. $I^*(\tau_k + Z)$. If $Z_1 > 0$ (i.e. $A \neq \emptyset$), then by property of APRX-CNT and ALGO I, we will have a schedule $I(\tau_k + Z)$ so that

$$W(\tau_k + Z) \geq \left(\frac{1 - \varepsilon/8}{1 + \varepsilon/8} \right)^2 W^*(\tau_k + Z) \approx (1 - \varepsilon/2)W^*(\tau_k + Z).$$

Now, for $\tau_k + Z < \tau \in A$, due to **P2** holding (by definition of Z_1) and ALGO I, we have that

$$W(\tau) \geq W(\tau - 1) - n.$$

Putting above discussion together, some re-arrangement and some bounds discussed above gives:

$$\sum_{\tau \in A} \mathbb{E}[\delta(\tau)|Q(\tau_k)] \leq 2nT^2 + \frac{\varepsilon T}{2} W^*(\tau_k). \quad (9)$$

For $\tau \in B$, note that the length of B is upper bounded by $\min\{Z, T\} + \hat{T}$. Using the bound on them as discussed above and obvious bound $\delta(\tau) \leq W^*(\tau)$ we have

$$\sum_{\tau \in B} \mathbb{E}[\delta(\tau)|Q(\tau_k)] \leq 2nT^2 + (2^n + T^2 3^{-n})W^*(\tau_k). \quad (10)$$

For $T = 2.2^n$ and n large enough, it is clear that

$$(2^n + T^2 3^{-n}) \leq 0.1\varepsilon T. \quad (11)$$

Replacing (9)-(11) in (8), we have

$$\begin{aligned} \mathbb{E}[L(\tau_{k+1}) - L(\tau_k)|Q(\tau_k)] &\leq (nT + 8nT^2) - 0.4\varepsilon W^*(\tau_k) \\ &= -\frac{0.4\varepsilon}{n} \langle Q(\tau_k), \mathbf{1} \rangle + O(5^n), \end{aligned} \quad (12)$$

since $T = 2.2^n$ and $W^*(\tau_k) \geq \langle Q(\tau_k), \mathbf{1} \rangle/n$ for any graph G . The (12) is the same as (3), hence we proved the stability.

Next, we prove claim for average queue-size, $\mathbb{E}[\langle Q(\tau), \mathbf{1} \rangle]$ as $\tau \rightarrow \infty$. Taking expectation w.r.t. $Q(\tau_k)$ in (12),

$$\mathbb{E}[L(\tau_{k+1}) - L(\tau_k)] \leq -\frac{0.4\varepsilon}{n} \mathbb{E}[\langle Q(\tau_k), \mathbf{1} \rangle] + O(5^n). \quad (13)$$

Telescopically sum (13) for $k = 0, \dots, K-1$; use fact $L(\cdot) \geq 0$ and some re-arrangement yields

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\langle Q(\tau_k), \mathbf{1} \rangle] \leq O\left(\frac{n5^n}{\varepsilon}\right).$$

Using the fact that for $\tau \in (\tau_k, \tau_{k+1}]$, $\langle Q(\tau), \mathbf{1} \rangle \leq \langle Q(\tau_k), \mathbf{1} \rangle + nT$, taking $K \rightarrow \infty$ and above inequality gives us

$$\limsup_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{s=0}^{\tau-1} \mathbb{E}[\langle Q(s), \mathbf{1} \rangle] = O(6^n). \quad (14)$$

Given (12), the implication of Proposition 1(b) and (14), relation between cesaro limit and limit of sequence implies that

$$\lim_{\tau \rightarrow \infty} \mathbb{E}[\langle Q(\tau), \mathbf{1} \rangle] = O(6^n). \quad (15)$$

This completes the proof of Theorem 2. \blacksquare

IV. DISTRIBUTED STABLE ALGORITHM: LOWER DELAY

The ALGO I shows that scheduling problems with hard constraint such as independent set can have extremely simple, distributed and stable algorithms. The ALGO I essentially finds independent set schedule whose average weight at any time τ is $(1 - \varepsilon)W^*(\tau) - B_n$, where $W^*(\tau)$ is weight of max. wt. independent set and B_n is some exponentially large constant. The stability follows due to small multiplicative approximation loss of $1 - \varepsilon$, but the average queue-size suffers due to large constant B_n . This suggests that we need an algorithm that has average weight at least $(1 - \varepsilon)W^*(\tau)$.

As noted earlier, finding approximation to max. wt. independent set is computationally hard. That is, there exists graph instances for which finding such approximation will require exponential time, unless $P = NP$. However, the question is: are graphs arising in practice are of this type? Next, we present algorithm for practical graphs modeled as non-expanding to obtain approximate max. wt. ind. set. schedule.

A. GRAPH-PARTITION and its properties

Given a non-expanding graph G (with a metric D) and $\varepsilon > 0$, let $\mathbf{L}(\varepsilon) \geq 3$ be such that for any $v \in V$ and $i \leq \mathbf{L}(\varepsilon)$,

$$\sum_{i \leq \mathbf{L}(\varepsilon)} f_v^2(i) \leq \frac{\varepsilon}{3\beta^4(\log \varepsilon^{-1})^2} F_v^2(\mathbf{L}(\varepsilon)). \quad (16)$$

Let $N(\varepsilon) = \max_v F_v(\mathbf{L}(\varepsilon))$ and define

$$p_v(\varepsilon) = \frac{\beta^2 \log \varepsilon^{-1}}{F_v(2\mathbf{L}(\varepsilon))}. \quad (17)$$

The GRAPH-PARTITION algorithm that partitions graph in good clusters and boundary is described as follows.

GRAPH-PARTITION

- (0) Each $v \in V$ becomes *cluster-center* independently with probability p_v .
- (1) If v becomes a center, v sends notifying messages to nodes within distance $\mathbf{L}(\varepsilon)$ w.r.t. D . This can be implemented distributively by setting clock on message and spreading it around.
- (2) A node w takes decision as follows. w is in boundary if either of the following is true:
 - (a) w does not receive message from any node at distance $\leq \mathbf{L}(\varepsilon) - 1$.
 - (b) If w receives messages from two or more vertices, say v_1, \dots, v_k so that $D(v_i, w) \leq D(v_{i+1}, w)$ with $1 \leq i \leq k-1$, and $|D(v_1, w) - D(v_2, w)| \leq 2$.
- (3) If none of (2)(a)-(2)(b) is satisfied then w is in the cluster of a node v that is closet to w among all nodes from which w has received a message.

The GRAPH-PARTITION algorithm has the following property.

Lemma 3 *Under GRAPH-PARTITION, each node $v \in V$ is in boundary set B with probability at most 2ε .*

Proof: To prove lemma 3, by union bound it is sufficient to show that the probability of any node $w \in V$ becoming boundary due to condition (2)(a) and (2)(b) is bounded above by ε each. Before proving these bounds, we note that the size of a cluster centered at a node v is at most $N(\varepsilon)$. Hence, each cluster has size at most $N(\varepsilon)$.

Prob. of w in B under 2(a). This event happens when no node v with $D(v, w) \leq \mathbf{L}(\varepsilon) - 1$ has become center. Call this event $G(w)$. Then,

$$\begin{aligned} \Pr(G(w)) &= \prod_{v:D(v,w) \leq \mathbf{L}(\varepsilon)-1} (1 - p_v) \\ &\leq \exp\left(-\sum_{v:D(v,w) \leq \mathbf{L}(\varepsilon)-1} p_v\right), \end{aligned} \quad (18)$$

where the last inequality uses fact: $1 - x \leq e^{-x}$, $\forall x \geq 0$. Now, for any v such that $D(v, w) \leq \mathbf{L}(\varepsilon)$,

$$F_v(2\mathbf{L}(\varepsilon)) \leq F_w(3\mathbf{L}(\varepsilon)) \leq \beta F_w(\mathbf{L}(\varepsilon)),$$

where we used property of non-expanding G . From this, we have

$$\begin{aligned} \sum_{v:D(v,w) \leq \mathbf{L}(\varepsilon)} p_v &= \sum_{v:D(v,w) \leq \mathbf{L}(\varepsilon)-1} \frac{\beta^2 \log \varepsilon^{-1}}{F_v(2\mathbf{L}(\varepsilon))} \\ &= \log \varepsilon^{-1} \left(\sum_{v:D(v,w) \leq \mathbf{L}(\varepsilon)-1} \frac{\beta^2}{F_v(2\mathbf{L}(\varepsilon))} \right) \\ &\geq \log \varepsilon^{-1} \left(\sum_{v:D(v,w) \leq \mathbf{L}(\varepsilon)-1} \frac{\beta^2}{\beta F_w(\mathbf{L}(\varepsilon))} \right) \\ &= \beta \log \varepsilon^{-1} \frac{F_w(\mathbf{L}(\varepsilon) - 1)}{F_w(\mathbf{L}(\varepsilon))} \\ &\geq \log \varepsilon^{-1}, \end{aligned} \quad (19)$$

where the last inequality follows by noticing that for $\mathbf{L}(\varepsilon) \geq 3$, $3(\mathbf{L}(\varepsilon) - 1) \geq \mathbf{L}(\varepsilon)$ and hence

$$\beta F_w(\mathbf{L}(\varepsilon) - 1) \geq F_w(3(\mathbf{L}(\varepsilon) - 1)) \geq F_w(\mathbf{L}(\varepsilon)).$$

Replacing (19) in (18), we have

$$\Pr(G_w) \leq \exp(-\log \varepsilon^{-1}) = \varepsilon. \quad (20)$$

Prob. of w in B under 2(b). For this consider the following events: (i) $E_i(w)$, the event that w receives message from two nodes at distance $(i - 1, i]$ and no messages from nodes at distance $\leq i - 1$; (ii) $F_i(w)$, the event that w receives message from two nodes at distances in $(i - 1, i]$, $(i, i + 1]$ respectively and no messages from any other nodes at distance $\leq i - 1$; and (iii) $H_i(w)$, the event that w receives message from two nodes

at distance $(i - 1, i]$, $(i - 1, i + 2]$ respectively and no messages from any other nodes at distance $\leq i - 1$. Also define

$$p_*(w) = \max_{v:D(v,w) \leq \mathbf{L}(\varepsilon)} p_v.$$

Note that for any $v \in V$ such that $D(v, w) \leq \mathbf{L}(\varepsilon)$, we have $F_v(2\mathbf{L}(\varepsilon)) \geq F_w(\mathbf{L}(\varepsilon))$. Hence, by definition of p_v in (17) we have

$$p_*(w) \leq \frac{\beta^2 \log \varepsilon^{-1}}{F_w(\mathbf{L}(\varepsilon))}. \quad (21)$$

$\Pr(w \in B$ due to (2)(b))

$$\begin{aligned} &\leq \sum_{i=1}^{\mathbf{L}(\varepsilon)-1} \Pr(E_i(w)) + \Pr(F_i(w)) + \Pr(G_i(w)) \\ &\stackrel{(a)}{\leq} \sum_{i=1}^{\mathbf{L}(\varepsilon)-1} \binom{f_w(i)}{2} p_*^2(w) + \sum_{i=1}^{\mathbf{L}(\varepsilon)-2} f_w(i) f_w(i+1) p_*^2(w) \\ &\quad + \sum_{i=1}^{\mathbf{L}(\varepsilon)-3} f_w(i) f_w(i+2) p_*^2(w) \\ &\stackrel{(b)}{\leq} p_*^2(w) \left(3 \sum_{i=1}^{\mathbf{L}(\varepsilon)} f_w^2(i) \right) \\ &\stackrel{(c)}{\leq} p_*^2(w) \frac{\varepsilon}{\beta^4 (\log \varepsilon^{-1})^2} F_w^2(\mathbf{L}(\varepsilon)) \stackrel{(d)}{\leq} \varepsilon. \end{aligned} \quad (22)$$

Some justification: (a) follows by simple counting; (b) follows from the fact that $f_v(i) f_v(i+1) \leq 0.5(f_v(i)^2 + f_v(i+1)^2)$, $f_v(i) f_v(i+2) \leq 0.5(f_v(i)^2 + f_v(i+2)^2)$; (c) follows from (16) and (d) follows from (21). Now, the (20) and (22) completes the proof of Lemma 3. ■

Lemma 4 *Under GRAPH-PARTITION algorithm, each cluster is of size at most $N(\varepsilon)$. Further, nodes that belong to different clusters are not connected to each other in G .*

Proof: As explained in (3) of GRAPH-PARTITION, each cluster can be identified by its cluster center. All nodes in a given cluster must be at distance at most $\mathbf{L}(\varepsilon) - 1$ from the cluster center. Hence, by definition the size of any cluster is at most $N(\varepsilon)$.

Now, contrary to the claim of Lemma suppose there are two non-empty clusters or partitions S_1 and S_2 , with centers v_1 and v_2 respectively, such that there are nodes $u \in S_1$ and $w \in S_2$ such that u and w are connected in G , i.e. $D_G(u, w) = 1$. By definition of D , $D(u, w) \leq D_G(u, w) = 1$. Since u and w are not part of boundary, we have that $D(u, v_1) \leq \mathbf{L}(\varepsilon) - 1$, $D(w, v_2) \leq \mathbf{L}(\varepsilon) - 1$. Therefore,

$$D(u, v_2) \leq D(u, w) + D(w, v_2) \leq \mathbf{L}(\varepsilon).$$

Similarly, $D(w, v_1) \leq \mathbf{L}(\varepsilon)$. Thus, u and w must have received messages from both v_1 and v_2 . Now

$$D(u, v_2) \leq D(u, w) + D(w, v_2) = 1 + D(w, v_2).$$

Similarly,

$$D(w, v_1) \leq 1 + D(u, v_1).$$

Recall that, by (3) of GRAPH-PARTITION, it must be that $D(u, v_1) \leq D(u, v_2)$ and $D(w, v_2) \leq D(w, v_1)$. Therefore, we obtain that

$$|D(u, v_2) - D(u, v_1)| \leq 2.$$

But this will imply that u should be in boundary according to 2(b) of GRAPH-PARTITION. This is a contradiction. That is, it is not possible to have nodes in two difference clusters that are neighbor of each other. This completes the proof of Lemma 4. ■

B. Algorithm ALGO II

Now, we present a randomized algorithm that essentially finds an ind. set with average weight $(1 - \varepsilon)W^*(\tau)$ in total $O(n)$ message exchanges for any *non-expanding* graph.

The basic idea behind algorithm is as follows: given partition of nodes of V by GRAPH-PARTITION into disjoint sets S_1, \dots, S_K each of size $O(1)$ and boundary set B so that S_i form connected components of G (i.e. no two vertices in different S_i s are connected to each other), separated by nodes in B . We find exact max. wt. independent set, say I^i , restricted to each of S_i using essentially ALGO I multiple (constant) times. Then, form an independent set as $I = \cup I^i$, i.e. set all nodes in B to 0. The I is a valid independent set in G . Now, if sum of weights of nodes in B is *small* compared the weight of I , then I is a good approximation of max. wt. independent set. This is guaranteed by Lemmas 3 and 4.

ALGO II

- (0) At each time τ , algorithm performs steps (1)-(3).
- (1) Given $\varepsilon > 0$, partition the graph into $\Theta(n)$ clusters $S_1, \dots, S_{\phi n}$, $\phi \in (0, 1)$ and boundary set B using GRAPH-PARTITION. Each S_i has $N_i(\varepsilon)$ nodes, which is at most $N(\varepsilon)$.
 - each node knows whether it is in a cluster or in B .
- (2) Each cluster S_i do: for $k = 1, \dots, 2.5^{N_i(\varepsilon)}$
 - (o) Initially, set $k = 1$ and $I^i(0) = \mathbf{0}$ (i.e. empty set).
 - (a) Generate a random independent set $R^i(k)$ using RANDOM.
 - (b) Find weight estimate $w_R^i(k), w_I^i(k - 1)$ of $R^i(k), I^i(k - 1)$ using APRX-CNT($\varepsilon/8$).
 - (c) Set $I^i(k) = R^i(k)$ if $w_R^i(k) > \left(\frac{1+\varepsilon/8}{1-\varepsilon/8}\right) w_I^i(k - 1)$; else $I^i(k) = I^i(k - 1)$.
 - (d) Set $k = k + 1$ and repeat (a)-(c) till $k = 2.5^{N_i(\varepsilon)}$. When $k = 2.5^{N_i(\varepsilon)}$ call $I^i(k)$ as $I^i(\tau)$ (with abuse of notation).
- (3) Declare schedule at time τ , $I(\tau) = \cup I^i(\tau)$, i.e. nodes use their distributively learnt assignment.

Remark. The Figure 1 shows an idealized execution of GRAPH-PARTITION with 9 partitions and line-boundaries for grid-style graph. Solving the max weight independent set problem in each of these partitions separately and using them as a solution for the whole graph can give a good approximation. However, such partitions are not easy to obtain

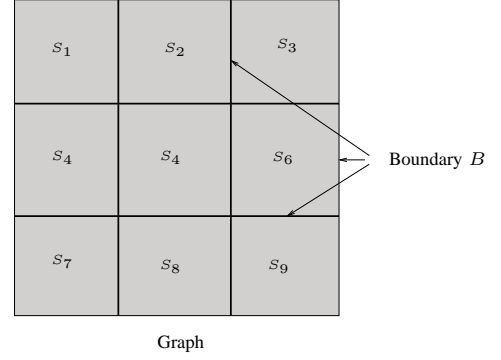


Fig. 1. An example of GRAPH-PARTITION and ALGO II for grid-style graph. The figure shows an idealized execution of the algorithm.

for any such graph in a distributed manner, and static partition does not work. Our algorithm ALGO II overcomes this non-triviality in a simple and elegant manner as explained later. Next, we state the main theorem about performance of ALGO II.

Theorem 5 *Given non-expanding graph G , the algorithm ALGO II is stable as long as $\lambda \in (1 - \delta(\varepsilon))\text{Co}(\mathcal{I})$ for any small enough $\varepsilon > 0$ and*

$$\lim_{\tau \rightarrow \infty} \mathbb{E}[\langle Q(\tau), \mathbf{1} \rangle] = O(n),$$

where $\delta(\varepsilon) = 4\varepsilon(\Delta + 2)$. This average queue-size is order optimal, i.e. there exists $\lambda \in (1 - \varepsilon)\text{Co}(\mathcal{I})$ for non-expanding G such that for any algorithm

$$\liminf_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{s=0}^{\tau-1} \mathbb{E}[\langle Q(s), \mathbf{1} \rangle] = \Omega(n).$$

Further, ALGO II takes $O(n)$ total operations with constant dependent on ε .

C. ALGO II : Complexity, Stability and Delay

Complexity. The algorithm ALGO II uses GRAPH-PARTITION, RANDOM and APRX-CNT. The algorithm GRAPH-PARTITION takes $O(n)$ distributed operations since message generated by each node can traverse at most $O(N^2(\varepsilon))$ times. The ALGO II calls RANDOM and APRX-CNT for $O(2.5^{N(\varepsilon)})$ times for each of the $\Theta(n)$ partitions. But since each partition is of size at most $N(\varepsilon)$, the net operation done by RANDOM and APRX-CNT for each partition is constant (dependent on $N(\varepsilon)$). Subsequently, the total operations performed by ALGO II is $O(n)$ with constant dependent of ε .

Stability and Delay. These properties are stated in Theorem 5 which essentially uses Lemma 6 (proved in [18]). Let $I(\tau)$ be schedule chosen by ALGO II at time τ , and the max. wt. independent set be $I^*(\tau)$, i.e. $I^*(\tau) = \arg \max_{I \in \mathcal{I}} \langle Q(\tau - 1), I(\tau) \rangle$. Let $W(\tau) = \langle Q(\tau - 1), I(\tau) \rangle$, and let $W^*(\tau) = \langle Q(\tau - 1), I^*(\tau) \rangle$.

Lemma 6 For non-expanding graph G with maximum vertex degree Δ , for any time τ , $\mathbb{E}[W(\tau)] \geq (1 - 0.5\delta(\varepsilon))W^*(\tau)$, where expectation is with respect to randomness of the algorithm GRAPH-PARTITION.

Proof: As per Lemma 4, GRAPH-PARTITION divides graph into partitions (clusters) $S_1, \dots, S_{\phi n}, \phi \in (0, 1)$ and boundary B . These have property : if $u \in S_i, v \in S_j$ with $j \neq i$, then $(u, v) \notin E$ (recall E is edge-set of G). Now ALGO II through RANDOM and APRX-CNT finds independent set $I^i(\tau)$ for S_i and $I(\tau) = \cup_i I^i(\tau)$. Given the above stated property, clearly $I(\tau)$ is an independent set of G . Let $W^i(\tau)$ be weight of $I^i(\tau)$. Then,

$$W(\tau) = \langle Q(\tau - 1), I(\tau) \rangle = \sum_i W^i(\tau).$$

Suppose $I^{i,*}(\tau)$ be max. wt. ind. set restricted to G_i which has vertices S_i and edges between nodes of S_i only; let $W^{i,*}(\tau)$ be its weight. Now consider $I^*(\tau)$ and its restriction to S_i . It is an independent set for G_i and hence the weight of that restriction is at most $W^{i,*}$. Using this argument over all S_i and B , we obtain

$$W^*(\tau) \leq W^B(\tau) + \sum_i W^{i,*}(\tau), \quad (23)$$

where $W^B(\tau)$ is the sum of weights of all nodes in B , i.e. $W^B(\tau) = \sum_{v \in B} Q_v(\tau - 1)$. By property of GRAPH-PARTITION as proved in Lemma 3, we have that each node is in B with probability at most 2ε . Hence,

$$\mathbb{E}[W^B(\tau)] \leq 2\varepsilon \langle Q(\tau - 1), \mathbf{1} \rangle. \quad (24)$$

We quickly remind ourselves that for any graph with max. vertex degree Δ ,

$$W^*(\tau) \geq \frac{1}{\Delta + 1} \langle Q(\tau - 1), \mathbf{1} \rangle. \quad (25)$$

To see (25), note that any such graph can be vertex colored by at most $\Delta + 1$ colors. Each such color gives rise to an independent set of G . The weight of these $\Delta + 1$ independent set adds up to $\langle Q(\tau - 1), \mathbf{1} \rangle = \sum_v Q_v(\tau - 1)$. Hence, it follows that one of them must be of weight at least $\frac{1}{\Delta + 1} \langle Q(\tau - 1), \mathbf{1} \rangle$, and hence $W^*(\tau)$.

Finally, to complete proof of lemma, we will show that

$$\mathbb{E}[W^i(\tau)] \geq (1 - 2\varepsilon)W^{i,*}(\tau). \quad (26)$$

Before establishing (26), we will complete the proof of Lemma 6 based on it. To this end, given (23)-(26), we have

$$\begin{aligned} \mathbb{E}[W(\tau)] &= \sum_i \mathbb{E}[W^i(\tau)] \geq (1 - 2\varepsilon) \left[\sum_i W^{i,*}(\tau) \right] \\ &\geq (1 - 2\varepsilon) [W^*(\tau) - \mathbb{E}[W^B(\tau)]] \\ &\geq (1 - 2\varepsilon)(1 - 2\varepsilon(\Delta + 1))W^*(\tau) \\ &\geq (1 - 2\varepsilon(\Delta + 2))W^*(\tau). \end{aligned} \quad (27)$$

Thus, only thing that remains to be established in (26). Now, in ALGO II an iteration of ALGO I is repeated $2.5^{N_i(\varepsilon)}$ times

in S_i at a given time τ (i.e. $Q(\tau)$ remains the same unlike in ALGO I where it changes between two iterations). The property **P1** of RANDOM suggests at least once in these $2.5^{N_i(\varepsilon)}$ iteration, it will sample the $I^{i,*}(\tau)$ with probability at least $1 - O(\exp(-1.25^{N_i(\varepsilon)}))$, which is at least $1 - 0.1\varepsilon$ for large enough $N_i(\varepsilon)$. Now, by **P2** of APRX-CNT, the weight of the schedule retained till the end of $2.5^{N_i(\varepsilon)}$ iterations is at least $(1 - \varepsilon)W^{i,*}(\tau)$ with probability at least $1 - O\left(\left(\frac{2.5}{3}\right)^{N_i(\varepsilon)}\right)$, which is at least $1 - 0.1\varepsilon$ for large enough $N_i(\varepsilon)$ or by proper choice of parameters⁸. Hence, using union bound we obtain that

$$\mathbb{E}[W^i(\tau)] \geq (1 - \varepsilon)(1 - 0.2\varepsilon)W^{i,*}(\tau) \geq (1 - 2\varepsilon)W^{i,*}(\tau). \quad \blacksquare$$

Proof: [Theorem 5] In interest of space, we provide a sketch of proof. As in the proof of Theorem 2, consider Lyapunov function

$$L(Q(t)) = Q(t) \cdot Q(t) = \sum_v Q_v^2(t).$$

Then, using the fact that $\lambda \in (1 - \delta(\varepsilon))\text{Co}(\mathcal{I})$ and by Lemma 6, arguments in Theorem 2 to obtain (8) with appropriate rearrangement will give us the following:

$$\mathbb{E}[L(\tau + 1) - L(\tau) | Q(\tau)] \leq n - \delta(\varepsilon)W^*(\tau + 1). \quad (28)$$

Using (25) in (28), we obtain

$$\mathbb{E}[L(\tau + 1) - L(\tau) | Q(\tau)] \leq n - \frac{\delta(\varepsilon)}{\Delta + 1} \langle Q(\tau), \mathbf{1} \rangle. \quad (29)$$

Invoking Proposition 1 along with (29) implies the stability of the system.

To prove bound on average queue-size, we can repeat the argument at the end of proof of Theorem 2 to obtain

$$\limsup_{\tau \rightarrow \infty} \mathbb{E}[\langle Q(\tau), \mathbf{1} \rangle] = O\left(\frac{n\Delta}{\delta(\varepsilon)}\right) = O(n).$$

Finally to prove the lower bound, consider the following. Every packet stays in a queue at least one time-step. Hence, the time-average of net queue-size, $\langle Q(\tau), \mathbf{1} \rangle$ is lower bounded by $\langle \lambda, \mathbf{1} \rangle$. For non-expanding graph G , the degree of each vertex is bounded by Δ . So using coloring argument, it follows that for $\lambda^u = \frac{1}{2(\Delta + 1)}\mathbf{1}$, we have $\lambda^u \in 0.5\text{Co}(\mathcal{I})$. But,

$$\langle \lambda^u, \mathbf{1} \rangle = \frac{1}{2(\Delta + 1)} \langle \mathbf{1}, \mathbf{1} \rangle = \frac{n}{2(\Delta + 1)}.$$

Thus, we have established that there is arrival rate for which

$$\liminf_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{s=0}^{\tau-1} \mathbb{E}[\langle Q(s), \mathbf{1} \rangle] = \Omega(n).$$

This completes the proof of Theorem 5. \blacksquare

⁸That is, if $N_i(\varepsilon)$ is not large enough, RANDOM and APRX-CNT can be run long enough constant time to obtain desired prob. estimates.

REFERENCES

- [1] X. Lin, N. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *Submitted, available through* csl.uiuc.edu/rsrikant, 2006.
- [2] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, pp. 1936–1948, 1992.
- [3] N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Proceedings of IEEE Infocom*, 1996, pp. 296–302.
- [4] N. McKeown, "iSLIP: a scheduling algorithm for input-queued switches," *IEEE Transaction on Networking*, vol. 7, no. 2, pp. 188–201, 1999.
- [5] J. Dai and B. Prabhakar, "The throughput of switches with and without speed-up," in *Proceedings of IEEE Infocom*, 2000, pp. 556–564.
- [6] L. Trevisan, "Non-approximability results for optimization problems on bounded degree instances," in *ACM STOC*, 2001.
- [7] P. Giaccone, B. Prabhakar, and D. Shah, "Randomized scheduling algorithms for high-aggregate bandwidth switches," vol. 21, no. 4, 2003, pp. 546–559.
- [8] D. Shah, "Stable algorithms for input queued switches," in *Proceedings of Allerton Conference on Communication, Control and Computing*, 2001.
- [9] D. Shah and D. J. Wischik, "Optimal scheduling algorithm for input queued switch," in *IEEE INFOCOM*, 2006.
- [10] B. Hajek and G. Sasaki, "Link scheduling in polynomial time," *IEEE Trans. Inf. Theory*, vol. 34, 1988.
- [11] P. Chaporkar, K. Kar, and S. Sarkar, "Throughput guarantees through maximal scheduling in wireless networks," in *43rd Allerton conference on Comm. Control and computing*, 2005.
- [12] L. Chen, S. H. Low, M. Chang, and J. C. Doyle, "Optimal cross-layer congestion control, routing and scheduling design in ad-hoc wireless networks," in *IEEE INFOCOM*, 2006.
- [13] X. Lin and N. B. Shroff, "Impact of imperfect scheduling in wireless networks," in *IEEE INFOCOM*, 2005.
- [14] E. Modiano, D. Shah, and G. Zussman, "Maximizing throughput in wireless network via gossiping," in *ACM SIGMETRICS/Performance*, 2006.
- [15] G. Sharma, R. Mazumdar, and N. Shroff, "On the complexity of scheduling in wireless networks," in *ACM Mobicom*, 2006.
- [16] H. B. Hunt-III, M. V. Marathe, V. Radhakrishnan, S. S. Ravi, D. J. Rosenkrantz, and R. E. Stearns, "NC-approximation schemes for NP- and PSPACE-hard problems for geometric graphs," *J. Algorithms*, vol. 26, no. 2, pp. 238–274, 1998.
- [17] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*.
- [18] K. Jung and D. Shah, "Low delay scheduling in wireless network," *Preprint, available at* <http://web.mit.edu/devavrat/www/delay.pdf>.
- [19] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993. [Online]. Available: <http://probability.ca/MT/>
- [20] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Jones and Barlett Publishers, 2003.
- [21] D. Mosk-Aoyama and D. Shah, "Computing separable functions via gossip," in *ACM PODC*, 2006.