Age of Information: An Introduction and Survey

Roy D. Yates, Yin Sun, D. Richard Brown III, Sanjit K. Kaul, Eytan Modiano and Sennur Ulukus

Abstract—**We summarize recent contributions in the broad area of age of information (AoI). In particular, we describe the current state of the art in the design and optimization of low-latency cyberphysical systems and applications in which sources send time-stamped status updates to interested recipients. These applications desire status updates at the recipients to be as timely as possible; however, this is typically constrained by limited system resources. We describe AoI timeliness metrics and present general methods of AoI evaluation analysis that are applicable to a wide variety of sources and systems. Starting from elementary single-server queues, we apply these AoI methods to a range of increasingly complex systems, including energy harvesting sensors transmitting over noisy channels, parallel server systems, queueing networks, and various single-hop and multi-hop wireless networks. We also explore how update age is related to MMSE methods of sampling, estimation and control of stochastic processes. The paper concludes with a review of efforts to employ age optimization in cyberphysical applications.**

I. Introduction

Low-latency cyberphysical system applications continue to grow in importance. Camera images from vehicles are used to generate point clouds that describe the surroundings. Video streams are augmented with informative labels. Sensor data needs to be gathered and analyzed to detect anomalies. A remote surgery system needs to update the positions of the surgical tools. From a system perspective, these examples share a common description: a source generates time-stamped status update messages that are transmitted through a network to one or more monitors. Awareness of the state of the remote sensor or system needs to be as timely as possible.

Research efforts directed toward low-latency networks are underway. Machine-to-machine communication and the tactile internet, each requiring link delays of just a few milliseconds, were key drivers for the 5G cellular standard [1]–[3]. Edge cloud computing that will eliminate transcontinental round-trip propagation delays on the order of 40 ms is another essential ingredient. However, while new systems supporting low-latency communication are necessary, they are also not sufficient for timely operation. Packet congestion in networks and backlogged jobs in edge-cloud processing centers may preclude the timely delivery of updates.

From these observations, timeliness of status updates has emerged as a new field of network research. It has been shown, even in the simplest queueing systems, that timely updating is not the same as maximizing the utilization of the system that delivers these updates, nor the same as ensuring that updates are received with minimum delay [4]. While utilization is maximized by sending updates as fast as possible, this strategy will lead to a monitor receiving delayed updates that were backlogged in the communication system. In this case, the timeliness of status updates at the receiver can be improved by *reducing* the update rate. On the other hand, throttling the update rate will also lead to a monitor having unnecessarily outdated status information because of a lack of updates.

This has led to new interest in *Age of Information (AoI)* performance metrics that describe the timeliness of a monitor's knowledge of an entity or process. AoI is an end-to-end metric that can be used to characterize latency in status updating systems and applications. An update packet with timestamp u is said to have age $t-u$ at a time $t \geq u$. An update is said to be *fresh* when its timestamp is the current time t and its age is zero. When the monitor's freshest[1] received update at time t has time-stamp $u(t)$, the age is the random process $\Delta(t) = t - u(t)$.

While this AoI survey focuses on recent work on age, we note that data freshness has been a recurring research theme. Periodic transactions updating real time databases [5], [6] was perhaps the earliest use of freshness. In [5] sensors wrote time-stamped fresh measurements into a real-time database and the age of an update was used to enforce concurrency of computations based on multiple measurements. Other early studies of timeliness include modeling and maximizing the freshness of query responses from a data warehouse [7], distributed QoS routing based on aged and imprecise network state information [8], network architectures that limit the "*degree of staleness*" of a cache [9], and ad hoc networking mechanisms that avoid propagation of stale route information [10] and that balance network congestion against nodes having stale information [11]. For web-caching, page-refresh policies have been tuned to maximize the freshness of cached pages [12] using an age metric in which age accumulated once the cached copy became outdated. Also noteworthy is [13] in which updates from a source are distributed over a graph by a gossip network.

Roy D. Yates is with WINLAB and the ECE Department, Rutgers University, NJ, USA, e-mail: ryates@winlab.rutgers.edu.

Yin Sun is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849, USA. e-mail: yinsun@auburn.edu.

D. Richard Brown III is with the Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester MA 01609 USA, e-mail: drb@wpi.edu.

Sanjit K. Kaul is with Wireless Systems Lab, IIIT-Delhi, India, e-mail: skkaul@iiitd.ac.in.

Eytan Modiano is with the Laboratory for Information and Decision Systems (LIDS) at the Massachusetts Institute of Technology (MIT), Cambridge, MA, e-mail: modiano@mit.edu.

Sennur Ulukus is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, e-mail: ulukus@umd.edu.

This work was supported by NSF grant CNS-1717041 (Yates), NSF grant CCF-1813050 (Sun), CNS-1836690 (Brown), and NSF grant CNS-1713725 and Army Research Office (ARO) grant number W911NF-17-1-0508 (Modiano).

[1] One update is fresher than another if its age is less.

The initial motivation for [4] came from the study of vehicular safety messaging over a CSMA network, initiated in [14] and continued in [15]. Over a random graph of vehicular nodes in a DSRC network, a round robin schedule was shown to lead to an average status-age that is smaller under the condition that nodes' updates piggyback each others' updates [15]. These simulation studies of vehicular updating [14], [15] prompted the AoI analysis in single-source single-server queues [4]. In contrast to the prior work [5], [12], [13] based on status update age, [4] focused on the impact of random service times on the age of delivered updates and showed that minimizing age required balancing the rate of updates against congestion. The takeaway message was that both the update arrivals and the service system could be designed, tuned, and even controlled to minimize the age.

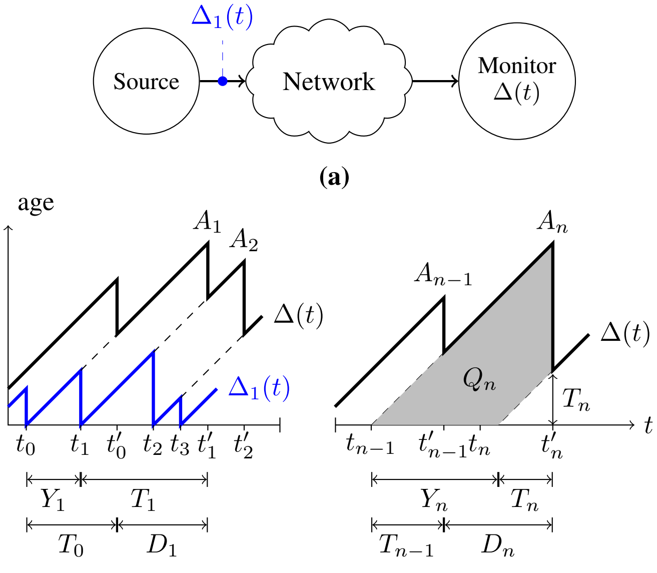
This survey focuses on recent contributions to AoI analysis. Section II introduces the age process and associated age metrics, and basic methods for the analysis of AoI. Section III summarizes AoI results in single-server queues, in order to demonstrate how AoI is influenced by the update arrival rate, the queue discipline, and packet management schemes designed explicitly to optimize freshness. This leads to a review of queueing networks, with a focus on scheduling updates of multiple sources at multiple servers in Section IV. This is followed in Section V by the study of energy-constrained updating. Here the emphasis is on energy harvesting systems in which a sensor's ability to transmit an update is constrained by its harvesting process. In this area, we examine *generate-at-will* sources that can generate a fresh update whenever they wish. Generate-at-will models are further explored in the context of sampling, estimation and control in Section VI. This is followed by a study of wireless networks in Section VII and a discussion of various applications of AoI in Section VIII. Finally, the conclusion in Section IX discusses potential application areas of AoI.

II. AoI Metrics and Analysis

As depicted in Figure 1(a), the canonical updating model has a source that submits fresh updates to a network that delivers those updates to a destination monitor. In a complex system, there may be additional monitors/observers in the network that serve to track the ages of updates in the network. For example, Figure 1(a) depicts an additional monitor that observes fresh updates as they enter the network.

These fresh updates are submitted at times $t_1, t_2, \ldots$ and this induces the AoI process $\Delta_1(t)$ shown in Figure 1(b). Specifically, $\Delta_1(t)$ is the age of the most recent update seen by a monitor at the input to the network. Because the updates are fresh, $\Delta_1(t)$ is reset to zero at each t_i. However, in the absence of a new update, the age $\Delta_1(t)$ grows at unit rate. If the source in Fig. 1 submits fresh updates as a renewal point process, the AoI $\Delta_1(t)$ is simply the age (also known as the backwards excess) [16], [17] of the renewal process.

These updates are delivered to the destination monitor at corresponding times $t'_1, t'_2, \ldots$. Consequently, the AoI process $\Delta(t)$ at the destination monitor is reset at time t'_j to $\Delta(t'_j) = t'_j - t_j$, the age of the jth update when it is delivered. Once again, absent the delivery of a newer update, $\Delta(t)$ grows at unit rate. Hence the age processes $\Delta_1(t)$ and $\Delta(t)$ have the characteristic sawtooth patterns shown in Figure 1(b). Furthermore, any other monitor in the network that sees updates arrive some time after they are fresh, will have an age process resembling $\Delta(t)$.



(a)

(b)

Fig. 1. (a) Fresh updates from a source pass through the network to a destination monitor. Monitor 1 (marked by ●) sees fresh update packets at the network access link. (b) Since monitor 1 sees fresh updates as a point process at times t_i, its age process $\Delta_1(t)$ is reset to zero at times t_j. Since the destination monitor sees updates that are delivered at times t'_j after traveling through the network, its age process $\Delta(t)$ is reset to $\Delta(t'_i) = t'_j - t_j$, which is the age of update j when it is delivered. For the nth delivered update of the $\Delta(t)$ process, Y_n, T_n and D_n are the interarrival, system and inter-departure times, and A_n is the corresponding age peak. The shaded area Q_n is employed for average age analysis

In the rest of Section II, we describe three approaches to AoI analysis. We start with with methods that analyze the limiting time-average age by graphical decomposition of the area under the sawtooth function $\Delta(t)$. We next introduce the average peak age metric and then the stochastic hybrid systems (SHS) approach to AoI analysis. This is followed by a discussion of nonlinear age penalty functions and functionals of the age process that are designed to capture the role of age in different classes of applications.

A. Time-Average Age

Initial work applied graphical methods to sawtooth age waveforms $\Delta(t)$ to evaluate the time-average AoI

$$\langle \Delta \rangle_{\mathcal{T}} = \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \Delta(t)dt. \tag{1}$$

in the limit of large $\mathcal{T}$. While this time average is often called the AoI, this survey employs AoI and age as synonyms that refer to the process $\Delta(t)$.

Figure 1(b) shows a sawtooth sample path of an age process $\Delta(t)$ in greater detail. For the nth delivered update, $Y_n = t_n - t_{n-1}$ and $T_n = t'_n - t_n$ denote the interarrival time and system time of the update. As shown on the right side of Figure 1(b), the key idea in the graphical method of age analysis is to decompose the area defined by the integral in (1) into a sum of trapezoidal areas

$$Q_n = \frac{1}{2}(T_n + Y_n)^2 - \frac{1}{2}T_n^2 = Y_n T_n + Y_n^2/2. \tag{2}$$

When (Y_n, T_n) is a stationary ergodic process, the time-average AoI $\Delta = \lim_{\mathcal{T}\to\infty} \langle\Delta\rangle_{\mathcal{T}}$ satisfies[2]

$$\Delta = \frac{\mathrm{E}[Q_n]}{\mathrm{E}[Y_n]} = \frac{\mathrm{E}[Y_n T_n] + \mathrm{E}\left[Y_n^2\right]/2}{\mathrm{E}[Y_n]}. \tag{3}$$

Equation (3) can be applied to a broad class of service systems, including both lossless FCFS systems as well as lossy last-come-first-served (LCFS) systems in which updates are preempted and discarded. Furthermore, it makes no specific assumptions regarding other traffic that might share the system with the update packets of interest.

However, AoI analysis can be challenging. With respect to (3), a large interarrival time Y_n allows the queue to empty, yielding a small waiting time and typically a small system time T_n. That is, Y_n and T_n tend to be negatively correlated and this complicates the evaluation of $\mathrm{E}[T_n Y_n]$.

B. Peak Age

The challenge in evaluating $\mathrm{E}[T_n Y_n]$ prompted the introduction of peak age of information (PAoI) [19], an alternate (and often more tractable) age metric. Referring to Fig. 1(b), we observe that the age process $\Delta(t)$ reaches a peak

$$A_n = T_{n-1} + D_n \tag{4}$$

the instant before the service completion at time t'_n. As an alternative to the computation of the average age, [19] proposed the average peak age of information (PAoI)

$$\Delta^{(p)} = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} A_n. \tag{5}$$

Under mild ergodicity assumptions, the PAoI is[3]

$$\Delta^{(p)} = \mathrm{E}[A] = \mathrm{E}[T_{n-1}] + \mathrm{E}[D_n]. \tag{6}$$

Hence PAoI avoids the computation of $\mathrm{E}[T_n Y_n]$, but, like the average age, the peak age captures the key characteristics of the age process. If the system is lightly loaded, then the average inter-departure time $\mathrm{E}[D]$ will be large; conversely as the system load gets heavy, the average system time $\mathrm{E}[T]$ will become large.

For single-server queues, it has been observed [20] that by defining $t'_{-1} = 0$ and $T_{-1} = \Delta(0)$ that

$$\Delta(t) = T_{n-1} + (t - t'_{n-1}), \quad t \in (t'_{n-1}, t'_n). \tag{7}$$

for $n = 0, 1, 2, \ldots$. Thus the sample path of $\Delta(t)$ is completely determined by the point process $\{(t'_n, T_n)\colon n = 0, 1, \ldots\}$. Since the departure times t'_n can be reconstructed from the inter-departure sequence D_n and (4) implies $D_n = A_n - T_{n-1}$, the sequence of pairs (T_{n-1}, A_n) is also sufficient to reconstruct the age process $\Delta(t)$. This shows that the age peaks A_n are a fundamental characterization of the age process [20].

C. Stochastic Hybrid Systems for AoI Analysis

An alternate approach to average age analysis employing a stochastic hybrid system (SHS) [21] was introduced in [22]. It was shown that age tracking can be implemented as a simplified SHS with non-negative linear reset maps in which the continuous state is a piecewise linear process [23]–[25]. For finite-state systems, this led to a set of age balance equations and simple conditions [22, Theorem 4] under which $\mathrm{E}[\mathbf{x}(t)]$ converges to a fixed point. A description of this simplified SHS for AoI analysis now follows.

In the SHS approach, the network shown in Fig. 1(a) has a hybrid state $[q(t), \mathbf{x}(t)]$ such that $\mathbf{x}(t) \in \mathbb{R}^{1\times n}$ and $q(t) \in \mathcal{Q} = \{0, \ldots, M\}$ is a continuous-time Markov chain. For AoI analysis, $q(t)$ describes the discrete state of a network while the real-valued age vector $\mathbf{x}(t)$ describes the continuous-time evolution of a collection of age-related processes. One of the components of $\mathbf{x}(t)$ is the age $\Delta(t)$ at a monitor of interest.

In the graph representation of the Markov chain $q(t)$, each state $q \in \mathcal{Q}$ is a node and each transition $l \in \mathcal{L}$ is a directed edge (q_l, q'_l) with transition rate $\lambda^{(l)}$ from state q_l to q'_l. Associated with each transition l is a transition reset mapping $\mathbf{A}_l \in \{0,1\}^{n\times n}$ that induces a jump $\mathbf{x}' = \mathbf{x}\mathbf{A}_l$ in the continuous state $\mathbf{x}(t)$. Unlike an ordinary continuous-time Markov chain, the SHS Markov chain may include self-transitions in which the discrete state is unchanged because a reset occurs in the continuous state. Furthermore, for a given pair of states $q, q' \in \mathcal{Q}$, there may be multiple transitions l and $\hat{l}$ in which $q(t)$ jumps from q to q' but the transition maps $\mathbf{A}_l$ and $\mathbf{A}_{\hat{l}}$ are different.

For each state $\bar{q}$, we denote the respective sets of incoming and outgoing transitions by

$$\mathcal{L}'_{\bar{q}} = \{l \in \mathcal{L} : q'_l = \bar{q}\}, \quad \mathcal{L}_{\bar{q}} = \{l \in \mathcal{L} : q_l = \bar{q}\}. \tag{8}$$

Assuming the discrete state Markov chain is ergodic, $q(t)$ has unique stationary probabilities $\bar{\boldsymbol{\pi}} = [\bar{\pi}_0 \cdots \bar{\pi}_M]$ satisfying

$$\bar{\pi}_{\bar{q}} \sum_{l\in\mathcal{L}_{\bar{q}}} \lambda^{(l)} = \sum_{l\in\mathcal{L}'_{\bar{q}}} \lambda^{(l)} \bar{\pi}_{q_l}, \quad \bar{q} \in \mathcal{Q}; \quad \sum_{\bar{q}\in\mathcal{Q}} \bar{\pi}_{\bar{q}} = 1. \tag{9}$$

The limiting average age vector $\mathrm{E}[\mathbf{x}] = \lim_{t\to\infty} \mathrm{E}[\mathbf{x}(t)]$ can be found by solving a set of linear equations [22, Theorem 4]. Specifically, if the discrete-state Markov chain $q(t)$ is ergodic with stationary distribution $\bar{\boldsymbol{\pi}} > 0$ and there exists a non-

[2] This decomposition is not unique. As can be seen in Fig. 1(b), an alternate approach [18] shows $Q_n = D_n T_{n-1} + D_n^2/2$ where $D_n = t'_n - t'_{n-1}$ is the nth inter-*departure* time. This implies an updating system has average age

$$\Delta = \frac{\mathrm{E}[Q_n]}{\mathrm{E}[D_n]} = \frac{\mathrm{E}[D_n T_{n-1}] + \mathrm{E}\left[D_n^2/2\right]}{\mathrm{E}[D_n]}.$$

[3] There is also more than one way to calculate PAoI. Fig. 1(b) reveals that $A_n = Y_n + T_n$. It follows that PAoI is also $\Delta^{(p)} = \mathrm{E}[Y_n] + \mathrm{E}[T_n]$, which is the decomposition in [18].

negative vector $\bar{\mathbf{v}} = [\bar{\mathbf{v}}_0 \cdots \bar{\mathbf{v}}_M]$ such that

$$\bar{\mathbf{v}}_{\bar{q}} \sum_{l \in \mathcal{L}_{\bar{q}}} \lambda^{(l)} = \mathbf{1}\bar{\pi}_{\bar{q}} + \sum_{l \in \mathcal{L}'_{\bar{q}}} \lambda^{(l)} \bar{\mathbf{v}}_{q_l} \mathbf{A}_l, \quad \bar{q} \in \mathcal{Q}, \tag{10}$$

then the average age is $\mathrm{E}[\mathbf{x}] = \sum_{\bar{q} \in \mathcal{Q}} \bar{\mathbf{v}}_{\bar{q}}$. This SHS method is extended in [26, Theorem 1] to provide the stationary age moments $\lim_{t\to\infty} \mathrm{E}\big[[x_1^m(t) \cdots x_n^m(t)]\big]$ and stationary age MGF $\lim_{t\to\infty} \mathrm{E}\big[[e^{sx_1(t)} \cdots e^{sx_n(t)}]\big]$ of the age process $\mathbf{x}(t)$.

D. Nonlinear Age Functions

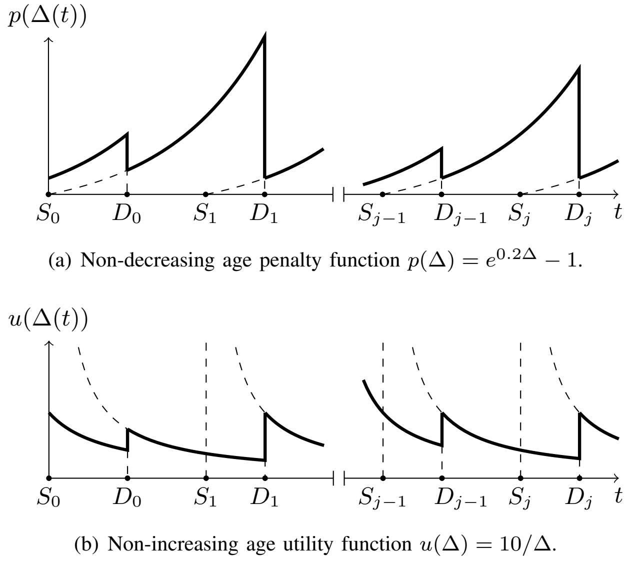
Although the AoI $\Delta(t)$ grows at unit rate, the performance degradation caused by information aging may not be a linear function of time. For instance, consider the problem of estimating the state of a Gaussian Linear Time-Invariant (LTI) system: If the system is stable, the state estimation error is a sub-linear function of $\Delta(t)$ that converges to a finite constant as $\Delta(t) \to \infty$ [27]; if the system is unstable, the state estimation error grows exponentially with $\Delta(t)$ [28], [29].

One approach to characterizing this nonlinear behavior is to define *freshness* and *staleness* as nonlinear functions of the AoI [30]–[33]. Since stale data is usually less desirable than fresh data [12], [13], [34]–[38], dissatisfaction with information staleness (or need for data refreshment) can be represented by a *non-decreasing* penalty function $p(\Delta(t))$ of the age $\Delta(t)$. Similarly, information freshness can be characterized by a *non-increasing* utility function $u(\Delta(t))$ [13], [35]. Notice that because the AoI $\Delta(t)$ is a function of time t, $p(\Delta(t))$ and $u(\Delta(t))$ are both time-varying, as shown in Fig. 2. In practice, the choice of monotonic AoI functions $p(\cdot)$ and $u(\cdot)$ is application-specific, as illustrated by the examples provided below and in [12], [13], [35]–[38]. In addition, applications of non-monotonic AoI functions were explored in [39], [40].

Auto-correlation Function: The auto-correlation function $\mathbb{E}[X_t^* X_{t-\Delta(t)}]$ of a source signal X_t can be used to evaluate the freshness of the sample $X_{t-\Delta(t)}$ [31]. For stationary sources, $|\mathbb{E}[X_t^* X_{t-\Delta(t)}]|$ is a nonlinear function of the AoI $\Delta(t)$. For example, in stationary ergodic Gauss-Markov block fading channels, the impact of channel aging can be characterized by the auto-correlation function of fading channel coefficients. When the AoI $\Delta(t)$ is small, the auto-correlation function and the data rate both decay with respect to $\Delta(t)$; when $\Delta(t)$ is large, the auto-correlation function and the data rate are not monotonic on the AoI [39].

Real-time Signal Estimation Error: Suppose samples of a Markov source X_t are forwarded to a remote estimator that reconstructs a causal estimate $\hat{X}_t$. Let S_i and D_i denote the generation time and delivery time of the i-th sample, respectively. If the sampling times $\{S_i\}$ are independent of the observed source $\{X_t \colon t \geq 0\}$, one can show that the mean-squared estimation error at time t is an increasing age function $p(\Delta(t))$ [22], [27], [41]. However, if the $\{S_i\}$ are chosen based on causal knowledge of the source, the estimation error is not necessarily a function of $\Delta(t)$ [27], [41].

The above result can be generalized to the state estimation error of feedback control systems [28], [29]. Consider a linear control system, where a controller observes the state $X_t \in \mathbb{R}^n$ of a plant and generates a control signal $U_t \in \mathbb{R}^m$ to adjust the operation of the plant, where the plant is disturbed by Gaussian noise N_t. The state evolution of the plant is governed by a linear equation



(a) Non-decreasing age penalty function $p(\Delta) = e^{0.2\Delta} - 1$.

(b) Non-increasing age utility function $u(\Delta) = 10/\Delta$.

Fig. 2. Two examples of non-linear age functions, where S_i and D_i are the generation time and delivery time of the i-th sample, respectively.

$$X_{t+1} = AX_t + BU_t + N_t, \tag{11}$$

where the $\{N_t\}$ are i.i.d. zero-mean Gaussian noise vectors with covariance matrix Σ. The constant matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are the system and input matrices, respectively, where (A, B) is assumed to be controllable. Under some assumptions, the state estimation error can be proven to be independent of the adopted control policy for determining U_t [42]. If the sampling times $\{S_i\}$ are independent of the state process X_t, the mean-square state estimation error is an increasing age penalty function $p(\Delta(t)) = \sum_{k=0}^{\Delta(t)-1} \mathrm{tr}[A^k \Sigma (A^T)^k]$ [28], [29]. On the other hand, if the $\{S_i\}$ are based on past values of the system state, the estimation error is not necessarily a function of the AoI.

Information Theoretic Freshness Metrics: Let

$$W^t = \{(X_{S_i}, S_i) \colon D_i \leq t\} \tag{12}$$

denote the samples that have been delivered to the receiver by time t. One can use the mutual information [43]

$$I(X_t; W^t) = H(X_t) - H(X_t|W^t), \tag{13}$$

i.e., the information the received samples W^t carry about the current source value X_t, to evaluate the freshness of W^t. If $I(X_t; W^t)$ is close to $H(X_t)$, the sample W^t contains a lot of information about X_t and is considered to be fresh; if $I(X_t; W^t)$ is near 0, W^t provides little information about X_t and is deemed to be obsolete.

If X_t is a stationary Markov chain and the sampling times $\{S_i\}$ are independent of the source $\{X_t \colon t \geq 0\}$, one can use the data processing inequality [44, Theorem 2.8.1] to show that $I(X_t; W^t) = I(X_t; X_{t-\Delta(t)})$ is a non-negative non-increasing function $u(\Delta(t))$ of $\Delta(t)$ [32]. A similar result holds in the case that X_t is a stationary discrete-time Markov chain with memory k, where each sample $V_t = (X_t, X_{t-1}, \ldots, X_{t-k+1})$

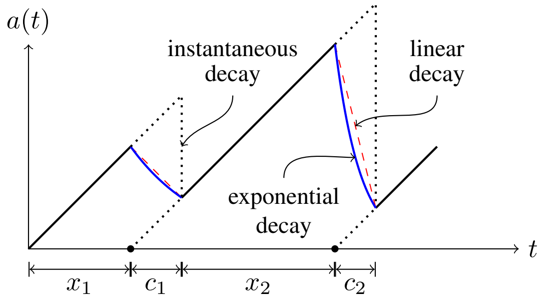


Fig. 3. Hard and soft updates: Hard updates take effect after a service time but yield instantaneous decay. Soft updates start taking effect right away, but gradually over time, giving rise to exponential or linear decay depending on the model.

should contain the source values at k successive time instants. Let $W^t = \{(V_{S_i}, S_i) : D_i \leq t\}$, then $V_{t-\Delta(t)}$ is a sufficient statistic of W^t for inferring X_t and $I(X_t; W^t) = I(X_t; V_{t-\Delta(t)})$ is a non-negative non-increasing function of $\Delta(t)$. One can also use the conditional entropy $H(X_t|W^t)$ to represent the staleness of W^t [45]–[47]. If the sampling times $\{S_i\}$ are independent of $\{X_t \colon t \geq 0\}$ and X_t is a stationary Markov chain, then $H(X_t|W^t) = H(X_t|X_{t-\Delta(t)})$ is a non-decreasing penalty function $p(\Delta(t))$ of the AoI $\Delta(t)$.

If (i) X_t is not a Markov chain or (ii) the $\{S_i\}$ are determined based on causal knowledge of X_t, then $I(X_t; W^t)$ and $H(X_t|W^t)$ are not necessarily monotonic functions of the AoI. Recently, conditional entropy and its variations were found useful for characterizing the minimum training error of supervised learning based time-series forecasting [40]. In these applications, the Markovian property may not hold and the training error is not always monotonic in $\Delta(t)$.

Age Violation Probability: If $p(\Delta(t))$ is chosen as the indicator function [48]

$$p(\Delta(t)) = 1_{\{\Delta(t)>d\}} = \begin{cases} 1, & \text{if } \Delta(t) > d; \\ 0, & \text{if } \Delta(t) \leq d, \end{cases} \tag{14}$$

then the fraction of time such that $\Delta(t)$ exceeds a threshold d is given by

$$\Pr\{\Delta(t) > d\} = \lim_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} 1_{\{\Delta(t)>d\}} dt. \tag{15}$$

Therefore, the age violation probability is a time-average of the indicator age penalty function in (14).

Soft Updates: Another instance where nonlinear age metrics appear is in soft updates [49], [50]. This setting models human and social media interactions where an update consists of viewing and digesting many small pieces of information posted, that are of varying importance, relevance, and interest to the monitor. Most of the AoI literature considers *hard* updates, which are contained in information packets. A hard update takes effect and reduces the age instantaneously when the update is delivered to the monitor. This is referred to as *instantaneous decay* in Fig. 3. The time for the update to take effect (denoted by c_1 for the first update in the figure) is the service time. In contrast, a soft update gradually reduces the age while the source is delivering the update. Depending on the updating model, this gradual decrease may yield nonlinear instantaneous age functions.

References [49], [50] consider two models for the age function $a(t)$ of the soft update process: In the first model, the rate of decrease in age is proportional to the current age: $da(t)/dt = -\alpha a(t)$, where α is a fixed constant. This is motivated by new information being most valuable and innovative when the current information is most aged. This model leads to an exponential decay in the age in Fig. 3. This exponential decay is consistent with information dissemination in human interactions where the most important information is conveyed/displayed first, reducing the age faster initially, and the subsequent less important information follows, reducing the age more slowly. In the second model, the rate of decrease in age is not a function of the current age, rather it is constant: $da(t)/dt = -\alpha$. In this case, the age decreases linearly, as shown in Fig. 3.

E. Functionals of Age Processes

One can use a non-decreasing functional[4] $f(\{\Delta(t) \colon t \geq 0\})$ of the age process $\{\Delta(t) \colon t \geq 0\}$, also termed *age penalty functional*, to measure the dissatisfaction for having aged information at the monitor [48], [52]–[54]. Examples of age penalty functionals include the time-average age (1) and the *time-average age penalty*

$$f_{\text{avg-penalty}}(\{\Delta(t) \colon t \geq 0\}) = \lim_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} p(\Delta(t)) dt, \tag{16}$$

where $p \colon [0, \infty) \mapsto \mathbb{R}$ can be any non-decreasing function.

We note that the peak age (5) is not an age penalty functional.[5] In addition, neither the peak age violation probability $\lim_{\mathcal{T}\to\infty} 1/N(\mathcal{T}) \sum_{n=1}^{N(\mathcal{T})} 1_{\{A_n>d\}}$ nor the peak age penalty $\lim_{\mathcal{T}\to\infty} 1/N(\mathcal{T}) \sum_{n=1}^{N(\mathcal{T})} p(A_n)$, with $p(\cdot)$ being a non-decreasing penalty function, are age penalty functionals.

III. Age in Elementary Queues

In this section, we examine AoI when the network in Fig. 1(a) is an elementary queue and sources submit updates as a stochastic process, independent of the queue state. This model includes the M/M/1, M/D/1 and D/M/1 queues and versions of those queues that incorporate preemption (in service or in waiting) or blocking of new arrivals.

While space considerations preclude an in-depth discussion of the entire literature of AoI in queues, there are a number of other notable contributions. For a single updating source, distributional properties of the age process were analyzed for the D/G/1 queue under first-come-first-served (FCFS) service [57]. General queueing systems in the form of G/G/1/1 queues were analyzed in [58], [59]. Non-i.i.d. service times modeled

[4] A functional is a mapping from functions to real numbers. A functional f is *non-decreasing* if $f(\{\Delta_1(t) \colon t \geq 0\}) \leq f(\{\Delta_2(t) \colon t \geq 0\})$ whenever $\Delta_1(t) \leq \Delta_2(t)$ for all $t \geq 0$ [51, p. 281].

[5] The claim that "the peak age is a non-decreasing functional of the age process" in [48], [52]–[56] is wrong. In fact, one can increase an age process $\Delta_1(t)$ to create another age process $\Delta_2(t)$ satisfying (i) all peaks of $\Delta_1(t)$ are also peaks of $\Delta_2(t)$ and (ii) $\Delta_2(t)$ has some additional smaller peaks that do not exist in $\Delta_1(t)$. Even though $\Delta_2(t) \geq \Delta_1(t)$, these additional low peaks can be chosen so that the average of peaks in $\Delta_2(t)$ is smaller than that in $\Delta_1(t)$. Therefore, the average peak age in (5) may drop even though the age process is increased.

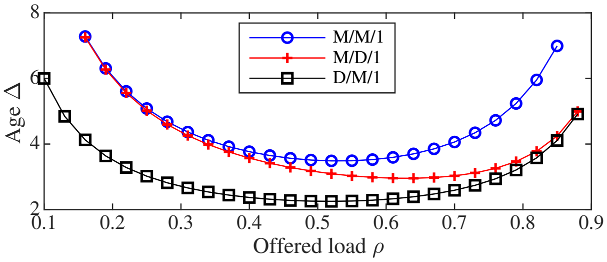


Fig. 4. Time-average age as a function of offered load $\rho = \lambda/\mu$ for the FCFS M/M/1, M/D/1 and D/M/1 queues. The expected service time is $1/\mu = 1$.

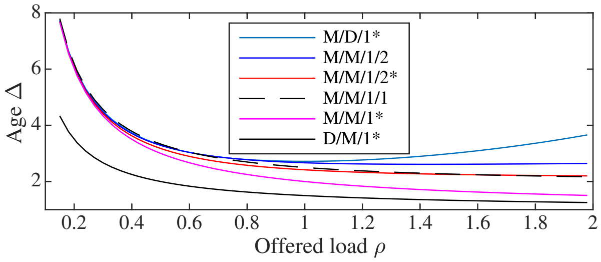


Fig. 5. Time-average age of the FCFS M/M/1/2 and M/M/1/1 blocking queues, the LCFS M/D/1*, M/M/1*, and D/M/1* queues with preemption in service, and the M/M/1/2* queue with preemption in waiting. The expected service time is $1/\mu = 1$.

as a Gilbert-Elliott process with *good* and *bad* serving states were studied in [60]. Packet deadlines were found to improve AoI [61]. Age-optimal preemption policies were identified for updates with deterministic service times [62]. AoI was evaluated in the presence of packet erasures at the M/M/1 queue output [63] and for memoryless arrivals to a two-state Markov-modulated service process [64].

In Section III-A, we examine average age for a single source sending updates through a queue. We focus on arrival and services processes that are either deterministic or memoryless in order to characterize elementary properties of the average age. This is followed by Section III-B, which uses zero-wait systems to derive age lower bounds, and Section III-C. which examines age in queues that serve multiple sources.

A. Age in Single-source Single-server Queues

This review is based chiefly on AoI results in [4], [18], [20]. We start with variations on non-preemptive and preemptive single server queues, for which the representation in [20] of the age process $\Delta(t)$ by the point process $\{(t'_n, T_n)\colon n = 0, 1, \ldots\}$ has led to a panoply of results, including the extension to distributional results for the stationary age $\Delta(t)$ and the peak age A_n and also generalization to GI/GI/1 queues.

In this discussion, each server has expected service time $\mathrm{E}[S]$ and each service system has i.i.d. interarrival times with expected value $\mathrm{E}[Y]$. For consistency of presentation, $\lambda = 1/\mathrm{E}[Y]$ is the arrival rate, $\mu = 1/\mathrm{E}[S]$ is the service rate, and the system has offered load $\rho = \lambda/\mu$. Numerical comparisons will be presented in terms of the load ρ with $1/\mu = 1$.

For the FCFS M/M/1 queue with offered load ρ, it was shown [4] using (3) that the average age is

$$\Delta_{\text{M/M/1}} = \frac{1}{\mu}\left(1 + \frac{1}{\rho} + \frac{\rho^2}{1-\rho}\right). \tag{17}$$

For fixed service rate μ, the age-optimal utilization ρ^* satisfies $\rho^4 - 2\rho^3 + \rho^2 - 2\rho + 1 = 0$ and thus $\rho^* \approx 0.53$. The optimal age is achieved by choosing a λ that biases the server towards being busy only slightly more than being idle. Note that we would want ρ close to 1 if we wanted to maximize the throughput. If we instead wanted to minimize packet delay, we would want ρ to be close to 0. Analysis of the M/D/1 queue [20] and the D/M/1 queue [22] yielded

$$\Delta_{\text{M/D/1}} = \frac{1}{\mu}\left(\frac{1}{2(1-\rho)} + \frac{1}{2} + \frac{(1-\rho)\exp(\rho)}{\rho}\right), \tag{18}$$

$$\Delta_{\text{D/M/1}} = \frac{1}{\mu}\left(\frac{1}{2\rho} + \frac{1}{1+\rho\mathcal{W}(-\exp[-1/\rho]/\rho)}\right), \tag{19}$$

where $\mathcal{W}(\cdot)$ denotes the Lambert-W function.

Fig. 4 presents age comparisons of the M/M/1, M/D/1, and D/M/1 queues from [4]. For each queue there is an age-minimizing offered load ρ. Among these FCFS queues, we observe that for each value of system load, D/M/1 is better than M/D/1, which is better than M/M/1. At low load, randomness in the interarrivals dominates the average status-age. At high load, M/D/1 and D/M/1 substantially outperform M/M/1 because the determinism in either arrivals or service helps to reduce the average queue length.

What these FCFS queues make apparent is that the arrival rate can be optimized to balance update frequency against the possibility of congestion. This prompted the study of lossy queues that may discard an arriving update while the server was busy or replace an older waiting update with a fresher arrival [18], [19], [65]. These strategies, identified as *packet management* [18], [19], include the M/M/1/1 queue that blocks and clears a new arrival while the server is busy, the M/M/1/2 queue that will queue one waiting packet but blocks an arrival when the waiting space is occupied, and the M/M/1/2* queue that will preempt a waiting packet with a fresh arrival.[6]

Another system in this category is the LCFS queue with

[6] While Kendall notation $A/S/c$ is consistently used to signify the *A*rrival process, the *S*ervice time, and the number of servers c, there is no consensus on a fourth entry for these systems. Here we (mostly) follow [18], with the fourth entry classifying how arrivals access the servers: $\cdot/\cdot/c$ is a c server system with an unbounded queue; $\cdot/\cdot/c/m$ indicates a system capacity of m updates (i.e. an FCFS waiting room of size $m - c$) with new arrivals blocked when the waiting room is full, and $\cdot/\cdot/c/m^*$ with $m = c+1$, indicates a single packet waiting room with preemption in waiting. We then add the convention $\cdot/\cdot/c^*$ to signify that a new arrival preempts the oldest update in service. (Since preempted packets are discarded, the waiting room becomes irrelevant.) Note that in [22], the M/M/1* and M/M/1/2* queues were called LCFS-S and LCFS-W, with S and W denoting preemption, S in Service and W in Waiting. In both [18], [22], it was assumed that obsolete updates were discarded. In [20], the fourth entry was the size of the waiting room, LCFS designated queues in which a new arrival moved in front of any waiting updates and the prefixes P and NP indicated whether the service was preemptive (P) or Non-Preemptive (NP), i.e. does the new arrival go immediately into service or simply to the head of the waiting line. [20] also used suffixes (C) and (D) to indicate whether the queue was work Conserving or whether obsolete updates were Discarded. Thus the M/M/1/2* queue [18], the M/M/1 LCFS-W [22] and the M/M/1/1 NP-LCFS (D) queue [20] were all the same queue.

preemption in service that permits a new arrival to preempt an update in service. Extending the notation introduced in [18], we call this an M/M/1* queue. These systems were shown [18] to achieve average ages

$$\Delta_{\text{M/M/1}^*} = \frac{1}{\mu}\left(1+\frac{1}{\rho}\right), \tag{20a}$$

$$\Delta_{\text{M/M/1/1}} = \frac{1}{\mu}\left(1+\frac{1}{\rho}+\frac{\rho}{1+\rho}\right), \tag{20b}$$

$$\Delta_{\text{M/M/1/2}^*} = \frac{1}{\mu}\left(1+\frac{1}{\rho}+\frac{\rho^2(1+3\rho+\rho^2)}{(1+\rho+\rho^2)(1+\rho)^2}\right), \tag{20c}$$

$$\Delta_{\text{M/M/1/2}} = \frac{1}{\mu}\left(1+\frac{1}{\rho}+\frac{2\rho^2}{1+\rho+\rho^2}\right). \tag{20d}$$

From (20), simple algebra will verify $\Delta_{\text{M/M/1}^*} \leq \Delta_{\text{M/M/1/1}}$ and $\Delta_{\text{M/M/1}^*} \leq \Delta_{\text{M/M/1/2}^*} \leq \Delta_{\text{M/M/1/2}}$. The age performance of the M/M/1/1 system is less easy to classify, although the relative performance of the M/M/1/1 system improves as ρ increases.

Figure 5 compares the average age for these queues. At low load, all of these queues achieve essentially the same average age $(1 + 1/\rho)/\mu$ as the M/M/1 and M/D/1 FCFS queues evaluated in Fig. 4. When the queue is almost always empty, the LCFS ability for the freshest update to jump ahead of older update packets is negated. However, at high loads, packet management ensures the M/M/1/2*, M/M/1/1 and M/M/1* queues have average age that decreases with offered load.[7] In fact, as the arrival rate $\lambda \to \infty$, the age $\Delta_{\text{M/M/1}^*}$ will approach the $2/\mu$ lower bound for exponential service systems because bombarding the server with new update packets ensures that a fresh status update packet will enter the waiting room the instant before each service completion.

We can conclude for memoryless service that preemption of old updates by new always helps. However, the comparisons are muddier between buffering and discarding. This is particularly true when we compare $\Delta_{\text{M/M/1}}$, which buffers every update, against the ages $\Delta_{\text{M/M/1/1}}$ and $\Delta_{\text{M/M/1/2}}$.

We also note that the apparent superiority of preemption in service is somewhat misleading; this property holds for memoryless service times, but not in general. For example, the M/D/1* and D/M/1* queues, both of which support preemption in service, have average ages [20, pp. 8318]

$$\Delta_{\text{M/D/1}^*} = \frac{1}{\mu}\frac{\exp(\rho)}{\rho}, \qquad \Delta_{\text{D/M/1}^*} = \frac{1}{\mu}\left(1+\frac{1}{2\rho}\right). \tag{21}$$

In (21), and also in Fig. 2, we see that the average age $\Delta_{\text{D/M/1}^*}$ of the D/M/1 preemptive server is monotonically decreasing in the offered load ρ. This is because no matter how high the arrival rate λ is (and thus how fast packets are being preempted), the departure rate is μ as long as an update is in service. By contrast, in the preemptive M/D/1* queue, $\Delta_{\text{M/D/1}^*}$ has a minimum at $\rho = 1$ and increases without bound for $\rho > 1$. With deterministic unit-time service and arrival rate $\lambda = \rho$, an update completes service with probability $e^{-\rho}$. As ρ becomes large, too many updates are preempted, and the system thrashes, with updates being preempted before they can complete service and be delivered to the monitor.

B. Zero-wait updates

When the update generator (source) can neither observe nor control the state of the packet update queue, the optimal load ρ^* strikes a balance between overloading the queue and leaving the queue idle. Here we derive lower bounds to the age Δ by considering a system in which the update generator observes the state of the packet update queue so that a new status update arrives just as the previous update packet departs the queue. Since each delivered update packet is as fresh as possible, the average age for this system is a lower bound to the age for any queue in which updates are generated as a stochastic process independent of the current queue state.

In this *zero-wait* system, the update service times S_n are i.i.d. with moments $\mathrm{E}[S]$ and $\mathrm{E}\big[S^2\big]$. Referring to the age $\Delta(t)$ in Figure 1(b), $t_n = t'_{n-1}$. This implies update n has interarrival time $Y_n = S_{n-1}$, zero waiting time, and system time $T_n = S_n$. Further, $\mathrm{E}[YT] = \mathrm{E}[Y_nT_n] = \mathrm{E}[S_{n-1}S_n] = (\mathrm{E}[S])^2$. From Equation (3), the average age becomes

$$\Delta^* = \frac{1}{\mathrm{E}[S]}\left[\frac{E[S^2]}{2} + (\mathrm{E}[S])^2\right]. \tag{22}$$

Thus, for a system with memoryless service times with $\mathrm{E}[S] = 1/\mu$, the minimum average age is $\Delta^*_{\cdot/M/1} = 2/\mu$. Moreover, since $\mathrm{E}\big[S^2\big] \geq (\mathrm{E}[S])^2 = 1/\mu^2$, (22) yields the lower bound

$$\Delta^* \geq \frac{3\,\mathrm{E}[S]}{2} = \frac{3}{2\mu}. \tag{23}$$

This bound holds for all service time distributions and is tight when the service times are deterministic. However, it bears repeating that (23) is an age lower bound for service systems in which updates are generated as a stochastic process *independent* of the current queue state. We will see in Section VI that a generate-at-will source can exploit the age and queue state to achieve even lower AoI.

C. Multiple Sources at a Single-server Queue

When updates have stochastic service times, AoI analysis of multiple updating sources sharing a simple queue has proven challenging and there have been relatively few contributions [22], [66]–[69]. In these papers, each source i generates updates as an independent Poisson process of rate λ_i and the update service time S has expected value $1/\mu$. Thus source i has offered load $\rho_i = \lambda_i/\mu$ and the total load is $\rho = \sum_{i=1}^{N}\rho_i$.

Extending the single-source age analysis in [70], reference [68] derived the average age and average peak age of each source. With $P_\lambda = \mathrm{E}\big[e^{-\lambda S}\big]$ denoting the Laplace transform of the service time S at $\lambda = \sum_{i=1}^{N}\lambda_i$, [68] showed that user i in the M/G/1/1 queue has average age Δ_i and average peak age $\Delta_i^{(p)}$ given by

$$\Delta_i = \frac{1}{\lambda_i P_\lambda}, \qquad \Delta_i^{(p)} = \frac{1}{\lambda_i P_\lambda} + \frac{\mathrm{E}\big[Se^{-\lambda S}\big]}{P_\lambda}. \tag{24}$$

[7] Because congestion is avoided by blocking packets, $\Delta_{\text{M/M/1/2}}$ avoids blowing up as $\rho \to \infty$. However, it achieves a minimum age of $\Delta = 2.61$ at $\rho = 1.427$ and then becomes an increasing function for $\rho > 1.427$. For large ρ, the M/M/1/2 queue admits its next update too quickly.

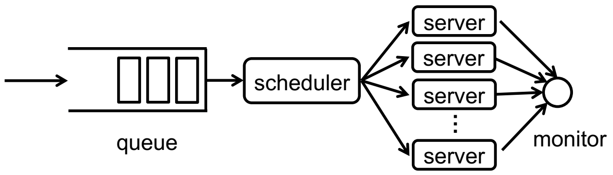


Fig. 6. Status updates in a single-hop, multi-server queueing network.

The first age analysis of the multi-source FCFS M/M/1 queue appeared in [66], which propagated to [22]. Unfortunately this analysis had an error, as observed in [69]. In a corrected analysis using SHS [71], it was shown with $\rho_{-i} \equiv \rho - \rho_i$ and

$$\mathcal{E}_i \equiv \frac{1+\rho-\sqrt{(1+\rho)^2-4\rho_{-i}}}{2\rho_{-i}}, \tag{25}$$

that source i has average age

$$\Delta_i = \frac{1}{\mu}\left[\frac{1-\rho}{(\rho-\rho_{-i}\mathcal{E}_i)(1-\rho\mathcal{E}_i)} + \frac{1}{1-\rho} + \frac{\rho_{-i}}{\rho_i}\right]. \tag{26}$$

This result is numerically identical to the independently derived result in [69, Theorem 1] and numerically close enough to the original claim in [22] that the qualitative observations in [22] remain valid.

Heterogeneous users sharing a single queue have been analyzed with different service time distributions [67] and with different priorities [72]–[74]. Different queueing disciplines, specifically, FCFS for the lower priority stream and LCFS with preemption allowed in service for the higher priority stream, have also been explored [72].

IV. Age in Queueing Networks

We now examine updates from one or more sources traversing a network of queues. Our starting point is the single-source, single-hop parallel-server network depicted in Fig. 6, consisting of one queue with buffer size $B \geq 0$ feeding c servers. If B is finite, each packet arriving to a full buffer is either dropped or replaces an existing packet in the system. If $B = 0$, the system can keep at most c packets that are being processed by the servers. For this class of systems, the challenge of AoI analysis is out-of-order packet delivery; a packet in service can be rendered obsolete by a delivery of another server.

Section IV-A considers elementary versions of the parallel server system, specifically, the M/M/2 queue with $c = 2$ servers and $B = \infty$ buffer space, the M/M/∞ queue in which an update immediately goes into service, and the M/M/c^* queue with a scheduler that preempts the oldest update in service if a new update arrives when all servers are busy. This is followed by Section IV-B, which examines of age-optimal scheduling for the parallel server system, and Section IV-C, which presents AoI results for update scheduling in other queueing networks.

A. The Parallel-Server Queue

The observation in [75] that a single server queue is not representative of networks in which packets may be delivered via multiple paths prompted AoI analysis of a queue with c parallel servers depicted in Fig. 6. Initial work [76], [77] addressed the extreme cases, namely the M/M/2 and M/M/∞ queues. Since the M/M/2 queue has infinite buffers, its average age can suffer from congestion induced by waiting updates. The M/M/2 queue age performance is also penalized by obsolete updates remaining in service. Nevertheless, it was found that M/M/2 service still could reduce the AoI by an approximate factor of 2 relative to M/M/1 [77].

The M/M/∞ age analysis in [77] is complex and does not reduce to a simple formula. However, with arrival rate λ and per-server rate μ service, the exact AoI $\Delta_{\text{M/M/}\infty}$ was found to be subject to the reasonably tight bounds

$$\frac{1}{\mu}\left(\frac{1}{\rho} + \frac{1+\rho+\rho^2}{(1+\rho)^3}\right) \leq \Delta_{\text{M/M/}\infty} \leq \frac{1}{\mu}\left(1+\frac{1}{\rho}\right). \tag{27}$$

It is not surprising that the upper bound in (27) equals the age $\Delta_{\text{M/M/1}^*}$ in (20a) since the monitor in the M/M/∞ system can choose to mimic the M/M/1* update delivery process by discarding all service completions except those by the freshest update in the system.

In [78], the M/M/c^* preemptive parallel server system was analyzed using SHS. In this system, a fresh update arriving when all servers are busy preempts the oldest update in service. For this system, the average age was found to be

$$\Delta_{\text{M/M/}c^*} = \frac{1}{\mu}\left[\frac{1}{c}\prod_{i=1}^{c-1}\frac{\rho}{i+\rho} + \frac{1}{\rho} + \frac{1}{\rho}\sum_{l=1}^{c-1}\prod_{i=1}^{l}\frac{\rho}{i+\rho}\right]. \tag{28}$$

When $\rho = \lambda/\mu \ll c$, one can expect that M/M/c^* system will approximate the infinite server system. Once again, since the M/M/∞ system can mimic this system by discarding service completions other than those of the c most recent arrivals, $\Delta_{\text{M/M/}\infty} \leq \Delta_{\text{M/M/}c^*}$.

B. Scheduling for Parallel Servers

We now examine update scheduling in the parallel server system of Fig. 6. In [48], [52]–[54], (near) age-optimal scheduling results were established using sample-path arguments. These results hold for out-of-order packet arrivals and quite general AoI metrics (e.g., age penalty functions and functionals), for which AoI analysis is a challenge. Therefore, AoI analysis and age-optimal scheduling provide complementary perspectives on status update systems.

Let us consider a *Last-Generated, First-Served (LGFS)* scheduling policy [55] in which the last generated packet is served first, with ties broken arbitrarily. In the Preemptive LGFS (P-LGFS) policy, a fresh packet can preempt an old packet that was in service. The preempted packets can be dropped or stored back to the queue; whether the preempted packets are dropped or stored back to the queue does not affect the age performance of the P-LGFS policy. In the Non-Preemptive LGFS (NP-LGFS) policy, each server must complete sending the current packet before starting to serve a fresher packet; in order to reduce the AoI, the freshest packet should be kept in the queue when packet dropping/replacement occurs.

If the packet service times are exponentially distributed, i.i.d. across servers and time, then the P-LGFS policy is age-optimal in a strong sense. Specifically, for *arbitrarily given* packet generation times $(S_1, S_2, \ldots)$, packet arrival times $(C_1, C_2, \ldots)$, buffer size B, and number of servers c, the P-LGFS policy minimizes *any* non-decreasing functional of the age process, including the time-average age (1) and time-average age penalty (16) [48].[8]

In addition, near age-optimal scheduling results can be established for a class of New-Better-than-Used (NBU) service time distributions [51], which include exponential distribution as a special case.[9] If the service times are i.i.d. NBU and the queue can store at least one packet ($B \geq 1$), then the expected time-average age of the NP-LGFS policy is within a small additive gap from the optimum, where the gap is invariant of the packet generation and arrival times, the number of servers c, and the buffer size B [48].

If packets arrive at the queue in the order of their generation times (i.e., $C_i \leq C_j$ whenever $S_i < S_j$), then the LGFS policy reduces to the LCFS policy. Hence, the P-LCFS policy is age-optimal and the NP-LCFS policy is near age-optimal for in-order arrivals.

C. *Scheduling for Multiple Hops and Multiple Sources*

These scheduling results have been extended to a few other network settings. In [52], the scheduling of a single packet flow in multi-hop queueing networks was studied. For i.i.d. memoryless service times, the P-LGFS policy is optimal for minimizing the age processes at all nodes of the network. In addition, the NP-LGFS policy is near age-optimal for i.i.d. NBU service times.

Age-optimal scheduling of multiple flows with synchronized arrivals in a single-hop queue was investigated in [53]. The authors first proposed a Preemptive, Maximum Age First, Last Generated First Served (P-MAF-LGFS) scheduling policy in which the last generated packet from the flow with the maximum age is served first among all packets of all flows, with ties broken arbitrarily. When the packet service times are i.i.d. exponentially distributed and the queue has one server, the P-MAF-LGFS policy minimizes the stochastic process $\{p_t(\boldsymbol{\Delta}(t))\colon t \geq 0\}$ in terms of stochastic ordering, where $p_t(\cdot)$ is a time-dependent, symmetric, and non-decreasing penalty function of the flow age vector $\boldsymbol{\Delta}(t) = [\Delta_1(t), \ldots, \Delta_N(t)]$.

In addition, a Non-Preemptive, Maximum Age of Served Information First, Last Generated First Served (NP-MASIF-LGFS) policy was introduced in [53]. In the NP-MASIF-LGFS policy, the last generated packet from the flow with the maximum age of served information is served first among all packets of all flows, with ties broken arbitrarily. When the packet service times are i.i.d. NBU and the queue has multiple servers, the NP-MASIF-LGFS policy is within a small additive gap from the optimum for minimizing the total time-average age. If multiple servers are idle, these servers are assigned to process different flows. Therefore, the behavior of the NP-MASIF-LGFS policy is similar to the maximum age matching approach [79] for orthogonal channel systems.

Motivated by [80], a notion of lexicographic age optimality, or simply lex-age-optimality, was introduced in [54] for scheduling multiple flows with diverse priority levels. A lex-age-optimal scheduling policy first minimizes the AoI of high-priority flows, and then, within the set of optimal policies for high-priority flows, achieves the minimum AoI metrics for low-priority flows. When the packet service times are i.i.d. exponentially distributed, a Preemptive Priority, Maximum Age First, Last-Generated, First-Served (PP-MAF-LGFS) scheduling policy was shown to be lex-age-optimal in a single-hop, single-server queue. In the PP-MAF-LGFS policy, the system will serve an informative packet (i.e., a packet that is fresher than any delivered packet) that is selected (with ties broken arbitrarily at each step) as follows: (i) among flows with informative packets, pick the highest priority class of flows; (ii) from the selected priority class, pick the flow with the maximum age; (iii) from the selected flow, pick the last generated informative packet. The scheduling result in [54] complements the AoI analysis of preemptive priority service systems for multiple sources [72]–[74].

V. Resource Constrained Updating

In this section, we focus on systems where the ability of a sensor to make an update is further constrained by external factors. A prominent example is an energy harvesting transmitter. A sensor that relies on energy harvested from nature cannot transmit continuously; otherwise it may run out of energy and risk having overly stale updates at the monitor. Therefore, the fundamental question is how to manage the harvested energy to send timely status updates.

We will start with an overview in Section V-A and then go on to summarize [81]–[85] that focus on online generate-at-will policies with various battery capacities over noiseless and erasure channels in Section V-B and Section V-C, respectively. In these sections, there is no server and the randomness in updating is purely due to uncertain energy arrivals. In Section V-D, we summarize works [86], [87], where there is an additional server at which the updates arrive exogenously, as opposed to being generated at-will.

A. *Overview*

There have been a number of works studying AoI with energy harvesting under various assumptions [81]–[111]. With the exception of [91], [109], [110], an underlying assumption in these works is that energy expenditure is normalized, i.e., sending one status update consumes one energy unit. For sensors with an unlimited battery capacity, [93] examines online policies under stochastic service times, [88] focuses on both offline and online policies with zero service times, i.e., with updates being transmitted instantly,[10] and [89] studies

[8] However, the P-LGFS policy may not be optimal for minimizing the peak age in (5), which is not an age penalty functional.

[9] A non-negative random variable X is said to be *New-Better-than-Used (NBU)* [51] if $\bar{F}(\tau + t) \leq \bar{F}(\tau)\bar{F}(t)$ holds for all $t, \tau \geq 0$, where $\bar{F}(t) = \Pr[X > t]$. Examples of NBU distributions include exponential, shifted exponential, geometric, gamma, and negative binomial distributions.

[10] In fact, most studies on AoI with energy harvesting sensors focus on the zero service time model in which transmission times are negligible relative to the large inter-transmission times induced by energy constraints.

the effect of sensing costs on AoI when transmitting through a noisy channel. Using a harvest-then-use protocol, [89] presents a steady state analysis of AoI under both deterministic and stochastic energy arrivals. The offline policy in [88] is extended to non-zero, but fixed, service times in [90] for both single and multi-hop settings, and in [91] to energy-controlled variable service times.

The online policy in [88] is analyzed through a dynamic programming approach in a discrete time setting, and is shown to have a threshold structure, i.e., an update is sent only if the age grows above a certain threshold and energy is available for transmission. Motivated by such results for the infinite battery case, [92] then studies the performance of online threshold policies for the finite battery case under zero service times. Reference [81] proves the optimality of online threshold policies under zero service times for the special case of a unit-sized battery, via tools from renewal theory. It also shows the optimality of best-effort online policies, where updates are sent over uniformly-spaced time intervals if energy is available, for the infinite battery case. Reference [94] shows that such a best-effort policy is optimal in the online case of multihop networks, thereby extending the offline work in [90]. Best-effort is also shown to be optimal, for the infinite battery case, when updates are subject to erasures, with and without erasure feedback, in [83], [95], [96].

Under the same system model of [95], reference [97] analyzes the best-effort online policy as well as the save-and-transmit online policy in which the sensor saves some energy in its battery before attempting transmission, for the purpose of coding to combat channel erasures. A slightly different system model is considered in [86], in which update arrivals are exogenous, i.e., the generation of fresh updates is not controlled by the sensor. With a finite battery, and stochastic service times, reference [86] employs tools from stochastic hybrid systems to analyze the long-term average AoI. The work in [87] considers a similar queuing framework as in [86] and studies the value of preemption in service on AoI. Reference [98] also considers a similar approach as in [86], [87] under general energy and data buffer sizes. An interesting approach is followed in [99] where the idea of sending extra information, on top of the measurement status updates, is introduced and analyzed for unit-sized batteries and zero service times.

The optimality of threshold policies for finite batteries with online energy arrivals has been shown in [82], [100], [101] using tools from renewal theory and a Lagrangian framework, which provide closed-form solutions for the optimal thresholds. This has also been shown independently and concurrently in [102] using tools from optimal stopping theory; see Section VI-A for a detailed discussion. Reference [103] shows the optimality of threshold policies under general age-penalty functions. Online policies for unit-sized batteries with update erasures also have been shown to have a threshold structure in [84], [85].

In addition to the studies to be detailed in Sections V-B, V-C and V-D, several other contributions have explored the interplay of age, energy, and damaged or lost updates. Erasures without feedback were also considered in [112], where a truncated automatic repeat request (TARQ) scheme was developed to retransmit the current status update until a time threshold is exceeded or a new update is available. The proposed TARQ scheme was shown to achieve a lower average AoI than classical ARQ.

In [113], [114], optimal transmission policies were derived for a generate-at-will source subject to an energy constraint. Some updates are damaged or lost and the monitor provides ACK/NACK feedback for each packet. For a monitor using classic ARQ, fresh updates are always transmitted after a failed update since the probability of a failed update is assumed to be constant. For a monitor using Hybrid-ARQ, however, it can be optimal to retransmit a failed update since the receiver can combine packets so that each retransmission has a lower error probability. In unknown environments, [114] went on to use reinforcement learning techniques. In [106], a setting was considered in which a sensor node must decide when to sleep to save energy but miss updates or wake and use energy to receive updates. An age-threshold based ON-OFF scheme was developed to minimize age subject to energy harvesting and battery capacity constraints.

In addition to the previously discussed settings with exogenous sources of harvested energy, another line of research has considered systems in which source nodes harvest energy through wireless energy transfer (WET) from an access point with a stable energy supply. In [115], a time-slotted system is considered. In each time slot, based on the available energies at the source nodes, the AoI values of different processes at the destination node, and the channel state information, the system must determine whether the time slot should be used for WET or for an update from a source node. A finite-state finite-action Markov decision process (MDP) is formulated and deep reinforcement learning techniques are used to find a solution. In a similar setting except with two-way data exchange [116], the downlink is assumed to use power splitting between WET and data and various tradeoffs between uplink and downlink AoI are analyzed.

Like [115], several studies have developed online policies for resource-constrained AoI using a MDP framework. In [117], a setting is considered where multiple sensors monitor the same process. Each sensor has a different age distribution and energy cost and the goal is to choose sensors to minimize the age at the destination subject to an energy constraint. Tradeoffs between transmit energy and error probability have also been examined in an AoI context [109], [110]. While their settings are slightly different, both papers explore the fundamental tradeoff between using more energy in each transmission to improve the probability of successfully delivering an update and reducing age and the potential for depleting the battery and consequently increasing age. In [111], a cognitive radio setting is assumed and a secondary user must decide whether to use energy to sense the channel or send updates.

Other frameworks that combine AoI with energy harvesting include multiple access channels [104], monitoring with priority [105], operational and sensing costs [107], and trade-offs between AoI and distortion [108].

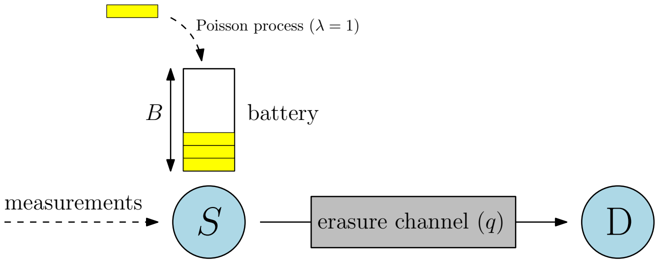


Fig. 7. System model for status updating over erasure channels. Updates packets are erased with probability q. The channel is noiseless when $q = 0$.

B. Energy Harvesting Noiseless Channels

In this section, we discuss the results reported in [81], [82]. In these works, the channel is noiseless with packet erasure probability $q = 0$ and energy arrives in single units according to a Poisson process with normalized rate $\lambda = 1$ arrival per unit time; see Fig. 7. The energy expenditure is normalized so that one update transmission consumes one energy unit. In addition, the transmission time of an update is negligible, i.e, zero.[11] Hence, updates are sent as a point process $s_1, s_2, \ldots$ such that s_i is the time the sensor acquires (and transmits) the ith measurement update.

At time s_i^-, the instant before the ith transmission, the sensor must have an energy unit. Thus, with $\mathcal{E}(t)$ denoting the energy in the battery at time t, we have the energy causality constraint

$$\mathcal{E}\left(s_i^-\right) \geq 1, \quad \forall i. \tag{29}$$

We assume the system starts with an empty battery at time 0. The battery evolves over time as

$$\mathcal{E}\left(s_i^-\right) = \min\left\{\mathcal{E}\left(s_{i-1}^-\right) - 1 + \mathcal{A}\left(x_i\right), B\right\}, \tag{30}$$

where $x_i \triangleq s_i - s_{i-1}$, and $\mathcal{A}(x_i)$ denotes the number of energy arrivals in $[s_{i-1}, s_i)$, and B is the battery capacity. Note that $\mathcal{A}(x_i)$ is a Poisson random variable with expected value x_i. We denote by $\mathcal{F}$, the set of feasible transmission times $\{s_i\}$ described by (29), (30), and $\mathcal{E}(0) = 0$.

Let $n(t)$ denote the total number of updates sent by time t. We wish to minimize the average AoI. Referring to Fig. 8, the area under the age evolution curve by time t is

$$Q(t) \equiv \frac{1}{2}\sum_{i=1}^{n(t)} x_i^2 + \frac{1}{2}\left(t - s_{n(t)}\right)^2. \tag{31}$$

The goal is to choose a set of feasible transmission times $s_i \in \mathcal{F}$ such that the long-term average AoI is minimized. Equivalently, one can optimize the inter-update times $\{x_i\}$. Therefore, the goal is to characterize the optimal long-term average AoI $\Delta^*(B)$ as a function of the battery size B by solving

$$\Delta^*(B) \equiv \min_{\{x_i\}\in\mathcal{F}} \limsup_{T\to\infty} \frac{1}{T}\,\mathrm{E}[Q(T)]. \tag{32}$$

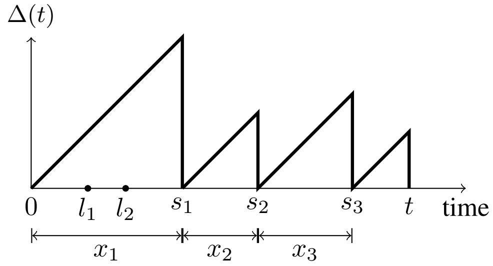


Fig. 8. The age evolution with $n(t) = 3$ updates at times s_i, with inter-update times x_i. When updates are successful with probability $q < 1$, there may be multiple transmissions for each update. In this example, updates are sent at times l_i and the first successful update at time $s_1 = l_3$ was preceded by failed transmissions at times l_1 and l_2.

When $B = \infty$, the battery size is infinite and no energy overflow can occur. Let us define the following policy:

- *Best-Effort Uniform (BU) Updating*: [81] The sensor is scheduled to send a new update at $s_n = n$, $n = 1, 2, 3, \ldots$. The sensor performs the task as scheduled if $\mathcal{E}(s_n^-) \geq 1$. Otherwise, it stays silent until the next scheduled update.

The best-effort uniform (BU) updating policy is always feasible. One of the main results of [81] is the following: The best-effort uniform updating policy is optimal for $B = \infty$, with $\Delta^*(\infty) = 1/2$.

When B is finite, the status update policy should try to prevent battery overflows, since wasted energy leads to performance degradation. On the other hand, owing to the nature of AoI, one should also try to send updates as uniformly as possible (as seen in the $B = \infty$ case). The optimal policy would then strike a balance between these objectives.

One main attribute of the optimal policy's behavior is that it has a *renewal* structure. In particular, defining $(\mathcal{E}(t), \Delta(t))$ as the system state at time t, [82] develops the following structural result: For $B < \infty$, the optimal update policy for problem (32) is a renewal policy for which visits to state $(0, 0)$ form a renewal process.

Based on this result, reference [82] shows that the optimal renewal-type status update policy has a *multi threshold* structure: a new status update is transmitted only if the AoI grows above a certain threshold that depends on the available battery energy. Such thresholds are found via a Lagrangian approach in closed-form. Further, it is shown that the thresholds are monotonically decreasing in the available energy.

C. Energy Harvesting Erasure Channels

In this section, we discuss the results reported in [83]–[85]. The system model is similar to that described in Section V-B, except *status updates are subject to erasures.* Specifically, the communication channel between the sensor and the destination is modeled as a time-invariant noisy channel, in which each update transmission gets erased with probability $q \in (0, 1)$, independently from other transmissions.

Since each update transmission is not necessarily successful, we denote the set of update transmission times by $\{l_i\}$ and

[11] Normalized arrival rates and zero transmission times are without loss of generality. Extensions to non-normalized arrival rates and fixed nonzero transmission times can be directly derived, at the expense of increased AoI as the arrival rate decreases and/or the transmission time increases.

the set of times of that are *successful* by $\{s_i\} \subseteq \{l_i\}$. The energy causality constraint in (29) now becomes

$$\mathcal{E}\left(l_i^-\right) \geq 1, \quad \forall i, \tag{33}$$

and the battery evolution in (30) becomes

$$\mathcal{E}\left(l_i^-\right) = \min\{\mathcal{E}\left(l_{i-1}^-\right) - 1 + \mathcal{A}(x_i), B\}, \quad \forall i, \tag{34}$$

where $x_i \triangleq l_i - l_{i-1}$ now denotes the inter-update attempt delay. We assume $s_0 = l_0 = 0$ without loss of generality, i.e., the system starts with fresh information at time 0. We denote by $\mathcal{F}_q$ the set of feasible transmission times $\{l_i\}$ satisfying (33), (34), and $\mathcal{E}(0) = 0$.[12]

Referring again to Fig. 8, the area under the age evolution curve during $[0, t]$ is still given by $Q(t)$ in (31) and the goal is to choose a set of feasible transmission times $\{l_i\} \in \mathcal{F}_q$ such that the long-term average AoI is minimized. Therefore, as a function of the battery size B, the goal is to solve

$$\Delta_q^\omega(B) \equiv \min_{\{l_i\} \in \mathcal{F}_q} \limsup_{T \to \infty} \frac{1}{T} \operatorname{E}[Q(T)]. \tag{35}$$

In (35), the superscript ω is used to differentiate between two main cases in our treatment:

1) *No updating feedback*: ($\omega \equiv$ noFB) The sensor has no knowledge of whether an update is successful. It can only use the up-to-date energy arrival profile and status updating decisions, as well as statistical information, such as the energy arrival rate and the erasure probability of the channel, to decide the upcoming update times.
2) *Perfect updating feedback*: ($\omega \equiv$ wFB) The sensor receives instantaneous, error-free feedback when an update is transmitted. Therefore, it can decide when to update next based on the feedback, along with the information available in the no-feedback case.

In view of the two feedback cases, we now present the solution of (35) for $B = \infty$ followed by the special case of $B = 1$.

With $B = \infty$ battery capacity but no updating feedback, [83] shows that the best-effort uniform (BU) updating policy is optimal. While for the perfect feedback scenario, [83] proposes this retransmission-based policy:

- *Best-effort Uniform with Retransmission (BUR) Updating*: [83] The sensor is scheduled to send new updates at $s_n = n/(1-q)$, $n = 1, 2, \ldots$. At time s_n, the sensor keeps sending updates until an update is successful or until it runs out of battery.

Independent of whether an update is successful at time s_n, the BUR updating sensor will keep silent until the next scheduled update time s_{n+1}. The optimality of BUR updating is shown in [83]. We summarize the results in [83] as follows: For the AoI minimization problem (35) with a size $B = \infty$ battery:

(a) With no updating feedback, BU updating is AoI-optimal with

$$\Delta_q^{\text{noFB}}(\infty) = \frac{1+q}{2(1-q)}. \tag{36}$$

(b) With perfect updating feedback, BUR updating is AoI-optimal with

$$\Delta_q^{\text{wFB}}(\infty) = \frac{1}{2(1-q)}. \tag{37}$$

To prove the above result, reference [83] proposes a novel *virtual policy* based approach that employs energy removals. Specifically, for both BU and BUR updating policies, a sequence of virtual policies defined by a time parameter T_0 is constructed. In both cases, the virtual policy initially follows the original BU or BUR policy. However, the virtual policy guarantees the sensor enters the zero energy state by time T_0. In particular, if zero energy has not previously been reached by time T_0, the virtual policy will intentionally discharge the battery. While these virtual policies are strictly suboptimal for finite T_0, they enable the effects of battery outages and updating errors to be decoupled in the performance analysis. Moreover, as $T_0 \to \infty$, these virtual BU and BUR policies are shown to become AoI-optimal and also converge to their original counterparts.

We now focus on the special case of a finite battery with $B = 1$ in which one update completely depletes the battery. Similar to the finite battery analysis of Section V-B, it will be shown that an *erasure-dependent threshold* policy is optimal for the case without feedback. For the case with perfect feedback, the focus will be on a class of *threshold-greedy* policies. References [84], [85] developed the following structural result: With or without updating feedback, the optimal updating policy for problem (35) with a size $B = 1$ battery is a renewal policy with i.i.d. inter-update times $\{y_i\}$. The renewal structure in this result greatly reduces the complexity of the problem. Now we need only to optimize the updating policy over one renewal interval. How this is done depends on whether there is feedback.

For updating without feedback, reference [84] shows that the optimal policy is given by an erasure-dependent threshold policy in which a new status update is sent only if the AoI grows above a certain threshold that depends on the erasure probability q. It is also shown that the optimal threshold is non-increasing in q, which is quite intuitive, since the sensor should be more eager to send new updates if the erasure probability is high.

For updating with perfect feedback, reference [85] focuses on a class of policies in which the first update attempt has a threshold structure, and the subsequent attempts, if the first is not successful, follow a greedy structure. This class is intuitive because if the first update is unsuccessful, then the AoI has already grown relatively high, which urges the sensor to transmit its subsequent updates as soon as energy is available. It is then shown that this class of *threshold-greedy* policies represent a fixed-point equilibrium in the sense that if the first update attempt is threshold-based then the following attempts should be greedy, and, conversely, if the second and subsequent update attempts are all greedy then the first should be threshold-based. The optimal threshold-greedy policy is then fully characterized.

While many useful takeaway points can be drawn in Sections V-B and V-C, where the focus has been on generate-

[12] In [83], it is assumed that $\mathcal{E}(0) = 1$ to simplify the analysis. For $\mathcal{E}(0) = 0$, the same results would follow after slightly modifying the proofs. We set $\mathcal{E}(0) = 0$ for consistency.

at-will policies, a crucial one is that *greedy status updating, whenever energy is available, is not always optimal.* Rather, it is optimal to evenly spread out the status updates over time, to the extent allowed by energy availability and energy causality. This is achieved by best effort-based policies when $B = \infty$ and threshold-based policies when B is finite.

D. Energy Harvesting Channels with Servers

While the previously discussed studies largely focused on a generate-at-will source and zero service times, several papers, including the original work [93], consider a setting with stochastic service delays. In [93], the energy harvesting process $H(t)$ is assumed to be ergodic with rate η, the source is assumed to have infinite battery capacity.

The source is also assumed to know the state of the server and can time its updates relative to the service completions. A β-minimum update policy was developed to minimize the average age. This policy avoids transmitting an update immediately after an update with short service time since the payoff (in age reduction) from this subsequent update will be small whereas the cost is fixed. On the other hand, if an update has a long service time, the next update is transmitted immediately since the payoff is higher. This policy, which counterintuitively can leave the server idle even if there is sufficient energy to transmit an update, was shown to outperform "best effort" and "fixed delay" policies.

In a similar setting [86], the average age was characterized as a function of the information and energy arrival rates, λ and η, respectively, as well as the battery capacity B. In this study, the source was assumed to always submit new updates immediately to the server. New updates enter service if the server is idle and has sufficient energy to service the packet. If the server is busy or does not have sufficient energy, the update is dropped. This paper leveraged SHS to determine the average AoI for two cases corresponding to whether the server is able or unable to harvest energy during service. If the server is unable to harvest energy while a packet is in service, it was shown that the average age satisfied

$$\Delta^*(B) = \begin{cases} \frac{2B\rho^2+(2B+2)\rho+B+2}{\mu[B\rho^2+(B+1)\rho]} & \beta = \rho, \\ \frac{(2\rho^2+2\rho+1)\beta^{B+2}-(2\beta^2+2\beta+1)\rho^{B+2}}{\mu[(\rho^2+\rho)\beta^{B+2}-(\beta^2+\beta)\rho^{B+2}]} & \beta \neq \rho, \end{cases} \tag{38}$$

where $\beta = \eta/\mu$ is the normalized energy arrival rate, $\rho = \lambda/\mu$ is the normalized packet arrival rate, μ is the service rate, and B is the battery capacity. Note the somewhat surprising result that the average age in this setting is invariant to exchanging β and ρ even though energy and packets are handled in different manners by the server.

A subsequent study [87] extended this work to consider preemption in service under the assumption that the energy expended on an update in service is lost if the update is preempted. Preemption was shown to decrease average age only in energy-rich operating regimes, i.e., regimes in which the server typically has a full battery and energy lost to preempted packets is inconsequential. In energy-starved operating regimes, preemption was shown to increase average age since preemption led a higher probability of battery depletion. Another study [98] extended these results to queues of arbitrary length, FCFS and LCFS queue disciplines, and nonlinear age penalty functions.

VI. Sampling, Estimation and Control

One method for reducing the AoI is to design a *sampling* policy that progressively determines when to generate update packets at the source. Consider the status update system illustrated in Fig. 9. A generate-at-will source decides when to sample a signal X_t. These samples are sent one-by-one to the monitor through a FCFS queue. Once a sample/update is delivered, an acknowledgement (ACK) is fed back to the sampler with no delay. By these ACKs, the sampler has access to the idle/busy state of the server in real-time.

In the event of queueing, the sampled packets would need to wait in the queue for their transmission opportunity and would become stale during the waiting time. Hence, it is better to suspend sampling when the channel is busy, and reactivate it when the channel becomes idle. A reasonable sampling policy is the *zero-wait* policy introduced in Section III-B that submits a new sample once the previous sample is delivered. The zero-wait policy would appear to be quite good, as it simultaneously achieves the maximum throughput and the minimum delay: Because the server is busy at all time, the maximum possible throughput is achieved; meanwhile, since the waiting time in the queue is zero, the delay is equal to the mean service time, which is the minimum possible delay.

Surprisingly, this zero-wait policy does **not** always minimize the AoI [30], [93]. Instead it can be better to wait before submitting an update. From an AoI perspective, the novelty of an update corresponds with how much it reduces the age at the monitor; when channel uses are a precious resource, they should not be wasted on updates that lack sufficient novelty. Since this may be unclear, the reader is encouraged to study the sample path example in [30].

In fact, the zero-wait sampling policy can be far from the optimum if (i) the goal is to minimize a nonlinear age function that grows quickly with respect to the AoI, and/or (ii) when the service times follow a heavy-tail distribution [30], [32]. This highlights a key difference between data communication systems and status update systems: In data communication, all packets are equally important; however, in status updating, a sample packet is useful only if it carries fresh information to the monitor.

Here we examine optimal sampling policies from age and estimation error perspectives. We start in Section VI-A with a study of sampling for AoI minimization. In Section VI-B, AoI minimization is then compared against sampling approaches that aim to minimize signal reconstruction error at the monitor. In these sections, $\pi = (S_1, S_2, \ldots)$ represents a sampling policy where S_i is the generation time of sample/update i, and Π denotes the set of causal sampling policies. The service time of update i is Y_i and the corresponding delivery time is

$$D_i = S_i + Y_i. \tag{39}$$

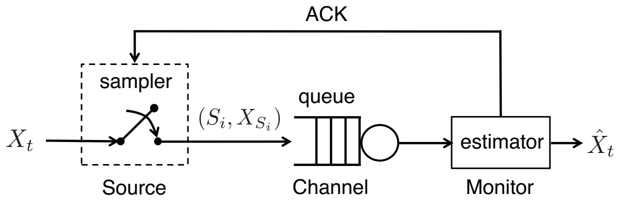


Fig. 9. A status update system with a sampler and an estimator.

The $\{Y_i\}$ are assumed to be i.i.d. with $0 < \mathrm{E}[Y_i] < \infty$. The age at time t is

$$\Delta(t) = \min\{t - S_i \colon D_i \le t\}. \tag{40}$$

In various settings, a sampling policy π will depend on a parameter such as β and the sampling and delivery times will be denoted $S_i(\beta)$ and $D_i(\beta)$ respectively.

A. Sampling for AoI Minimization

The optimal sampling problem for minimizing the time-average age penalty is formulated as

$$\bar{p}_{\text{opt}} = \inf_{\pi\in\Pi} \limsup_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \ \mathrm{E}\left[\int_0^{\mathcal{T}} p(\Delta(t))\, dt\right]. \tag{41}$$

It has been shown [32] that if $p(\cdot)$ is non-decreasing then the sampling policy $(S_1(\beta), S_2(\beta), \ldots)$ defined by

$$S_{i+1}(\beta) = \inf\{t \ge D_i(\beta) \colon \ \mathrm{E}[p(\Delta(t+Y_{i+1}))] \ge \beta\}, \tag{42a}$$

with threshold β satisfying

$$\mathrm{E}\left[\int_{D_i(\beta)}^{D_{i+1}(\beta)} p(\Delta(t))\, dt\right] = \beta\, \mathrm{E}[D_{i+1}(\beta) - D_i(\beta)], \tag{42b}$$

is an optimal solution to (41). Further, $\beta = \bar{p}_{\text{opt}}$.

The optimal sampling policy in (42) has a nice structure. Specifically, sample $i+1$ is generated at the earliest time t satisfying two conditions:

(i) $t \ge D_i(\beta)$, i.e., sample i has already been delivered by time t, and

(ii) the expected age penalty $\mathrm{E}[p(\Delta(t+Y_{i+1}))]$ has grown to be no smaller than the threshold $\beta = \bar{p}_{\text{opt}}$.

The optimal threshold β satisfying (42b) is unique [27] and can be calculated via bisection search, Newton's method, or the following fixed-point iteration method

$$\beta_{k+1} = \frac{\mathrm{E}\big[\int_{D_i(\beta_k)}^{D_{i+1}(\beta_k)} p(\Delta(t))\, dt\big]}{\mathrm{E}\big[D_{i+1}(\beta_k) - D_i(\beta_k)\big]}. \tag{43}$$

Note that Newton's method and the fixed-point iteration (43) converge faster than bisection search [27], [118].

Optimal sampling policies were also derived in other settings, including discrete-time sampling and sampling subject to a maximum sampling rate constraint [32], two-way communication delay [118], as well as multi-source updates for minimizing the total time-average age [119] and total time-average of nonlinear age penalty functions [120].

In [103], an age-optimal sampler design was obtained for an energy harvesting source with a finite battery size B and zero service time $Y_i = 0$. If energy units arrive according to a Poisson process, the optimal sampler is given by a multi-threshold sampling policy

$$S_{i+1}(\beta) = \inf\{t \ge S_i(\beta) \colon p(\Delta(t)) \ge \beta(E(t))\}, \tag{44}$$

where the threshold $\beta(E(t))$ is a decreasing function of the instant battery level $E(t) \in \{0, 1, \ldots, B\}$. Hence, samples are taken more frequently when the battery level is high, and less frequently when the battery level is low. Interestingly, the threshold $\beta(B)$ associated with a full battery level $E(t) = B$ equals the optimal objective value for this setting [103].

B. Sampling and Remote Estimation

The states of many cyberphysical systems, such as UAV mobility trajectory and sensor measurements, are in the form of a signal X_t. A natural task in these systems is to reconstruct the signal X_t at the remote monitor, based on samples that are causally received from the source. This requires an extension of Nyquist sampling theory to causal sampling and reconstruction. In the non-causal case, periodic sampling can achieve perfect reconstruction of bandlimited signals with no error; however, a non-zero reconstruction error is inevitable in causal signal processing and the design goal should be minimizing the reconstruction error. The problem of causal sampling and reconstruction is called *remote estimation* in the control literature; see [121] for a recent survey.

Recently, a connection between AoI and remote estimation was revealed in [27], [41]. To understand this connection, let $\hat{X}_t$ denote the MMSE estimate of X_t, based on the samples that have been delivered by time t, so that the signal reconstruction error is

$$\varepsilon_X(t) = X_t - \hat{X}_t. \tag{45}$$

Now consider the following optimal sampling problem for minimizing the mean-squared signal reconstruction error

$$\overline{\mathsf{mse}}_{\text{opt}} = \inf_{\pi\in\Pi} \ \limsup_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \mathrm{E}\left[\int_0^{\mathcal{T}} |\varepsilon_X(t)|^2\, dt\right]. \tag{46}$$

Problem (46) belongs to the class of continuous-time MDPs with continuous state space, which are usually quite challenging to solve due to the curse of dimensionality. Nonetheless, an exact solution to (46) has been found for two Gauss-Markov signals: the Wiener process and the Ornstein-Uhlenbeck process, which is the continuous-time analogue of the first-order autoregressive, i.e., AR(1), process.

If X_t is a Wiener process or an Ornstein-Uhlenbeck process, then $(S_1(\beta), S_2(\beta), \ldots)$ is an optimal solution to (46), where [27], [41]

$$S_{i+1}(\beta) = \inf\left\{t \ge D_i(\beta) \colon \left|\varepsilon_X(t)\right| \ge v(\beta)\right\}, \tag{47a}$$

with β satisfying

$$\mathrm{E}\left[\int_{D_i(\beta)}^{D_{i+1}(\beta)} |\varepsilon_X(t)|^2\, dt\right] = \beta\, \mathrm{E}[D_{i+1}(\beta) - D_i(\beta)]. \tag{47b}$$

For the Wiener process, $v(\beta) = \sqrt{3(\beta - \mathrm{E}[Y_i])}$ is the threshold function; for the Ornstein-Uhlenbeck process, $v(\cdot)$ is given by [27, Equation (18)]. Further, β is exactly the optimal value to (46), i.e., $\beta = \overline{\mathsf{mse}}_{\text{opt}}$.

The structure of the optimal sampling policy in (47) is similar to that in (42). Specifically, sample $i+1$ is generated at the earliest time t satisfying two conditions: (i) sample i has already been delivered by time t, i.e., $t \geq D_i(\beta)$, and (ii) the instantaneous estimation error $|\varepsilon_X(t)|$ is no smaller than a pre-determined threshold $v(\beta)$, where β equals the optimum value of (46) and the function $v(\cdot)$ is determined by the signal and the service time distribution.

One can add a maximum sampling rate constraint in the optimal sampling problem (46); its solution appeared in [27], [41]. Recently, remote estimation of a Wiener process with two-way random communication delay was studied in [122].

In remote estimation systems, the optimal sampler is affected by the selected estimator; and conversely, the optimal estimator is also influenced by the selected sampling policy. This "chicken and egg" dilemma can be resolved by jointly optimizing the sampler and estimator. In [123]–[127], tools from majorization theory were utilized to show that threshold-type samplers remain optimal in the joint sampling and estimation problem for several discrete-time remote estimation systems. It was also pointed out [123, p. 619] that similar results can be also established for continuous-time systems. Recently, it was shown that the sampling policy (47) and the MMSE estimator are indeed jointly optimal for the remote estimation of a class of continuous-time Markov signals [128].[13]

AoI and Signal-agnostic Sampling: Next, we consider a variation of problem (46) that is tightly related to the AoI. We say a sampling policy $\pi \in \Pi$ is *signal-aware* (*signal-agnostic*), if the sampling times S_i are determined with (without) causal knowledge of the signal X_t. Hence, the S_i are independent of the signal X_t in signal-agnostic sampling policies. Let $\Pi_{\text{agnostic}} \subset \Pi$ denote the set of signal-agnostic sampling policies. For every policy $\pi \in \Pi_{\text{agnostic}}$ and time-homogeneous Markov chain X_t, there exists an increasing function $p(\Delta(t))$ of the AoI such that

$$\mathrm{E}\left[\int_0^{\mathcal{T}} |\varepsilon_X(t)|^2 \, dt\right] = \mathrm{E}\left[\int_0^{\mathcal{T}} p(\Delta(t)) \, dt\right]. \tag{48}$$

By reducing the policy space Π to Π_{agnostic}, (46) becomes the following signal-agnostic sampler design problem

$$\overline{\mathsf{mse}}_{\text{age-opt}} = \inf_{\pi \in \Pi_{\text{agnostic}}} \limsup_{\mathcal{T} \to \infty} \frac{1}{\mathcal{T}} \mathrm{E}\left[\int_0^{\mathcal{T}} |\varepsilon_X(t)|^2 \, dt\right]. \tag{49}$$

For signal-agnostic policies, it follows from (48) that (49) is an instance of problem (41). It was shown in [27], [41] that $(S_1(\beta), S_2(\beta), \ldots)$ defined by

$$S_{i+1}(\beta) = \inf\left\{t \geq D_i(\beta)\colon\ \mathrm{E}\left[|\varepsilon_X(t + Y_{i+1})|^2\right] \geq \beta\right\}, \tag{50a}$$

with threshold β satisfying

$$\mathrm{E}\left[\int_{D_i(\beta)}^{D_{i+1}(\beta)} |\varepsilon_X(t)|^2 \, dt\right] = \beta \, \mathrm{E}[D_{i+1}(\beta) - D_i(\beta)], \tag{50b}$$

is an optimal solution to (49). Further, $\beta = \overline{\mathsf{mse}}_{\text{age-opt}}$.

Let us compare the optimal designs of signal-aware and signal-agnostic samplers: In the signal-aware sampling policy (47), the sampling time is determined by the *instantaneous* estimation error $|\varepsilon_X(t)|$, and the threshold function $v(\cdot)$ varies with the signal model. In the signal-agnostic sampling policy (50), the sampling time is determined by the *expected* estimation error $\mathrm{E}\left[|\varepsilon_X(t + Y_{i+1})|^2\right]$ at the delivery time $t + Y_{i+1}$ of the new sample. Hence, (50) requires that the expected estimation error upon the delivery of the new sample is no less than β. In both cases, β equals the optimal objective value.

Age-based sampler design is of particular interest in discrete-time feedback control systems [28], [29]. According to [32, Theorem 3], the optimal signal-agnostic sampler for such discrete-time systems can be obtained by (i) requiring t in (50a) to be discrete, i.e., $t \in \{0, T_s, 2T_s, \ldots\}$, and (ii) changing the integral in (50b) to a sum over discrete time. These results are not limited to the queue model; in fact, the optimal signal-agnostic sampler for remote estimation over a packet erasure channel has a quite similar structure, as reported independently in [29]. The scheduling of channel resources among multiple feedback control loops for reducing state estimation error was considered in [129], [130]. When the AoI is high, it is difficult to maintain a small control error. The optimal tradeoff between the AoI and control performance was studied in [42].

VII. Wireless Networks

In many applications, timely information is disseminated through wireless networks, where interference is one of the primary limitations to system performance. In this section, we start with an overview of recent contributions to AoI in wireless networks, followed in Section VII-B by works on updates over erasure channels that use ARQ/HARQ. Further in Section VII-C we discuss AoI results for conventional distributed multiaccess protocols. Section VII-D then reviews the main results from [131] on AoI optimization in broadcast wireless networks, followed in Section VII-E by results from [132] on wireless networks under general interference constraints.

A. Overview

Over the past few years, there has been a growing body of work on AoI minimization in wireless settings. At the physical layer, updates through channels with bit erasures have been studied in [70], [113], [133]–[135]. The problem of scheduling finite number of update packets under physical interference constraint for age minimization was shown to be NP-hard in, e.g., [136], [137]. The impact of information freshness on collision avoidance in a network of UAVs was studied in [138]. Age for a wireless network where only a single link can be activated at any time was studied in, e.g., [139]–[143], and index policies were proposed. Threshold policies were proven to be optimal in, e.g., [114], [144].

Various works have considered decentralized access of a shared medium by nodes sending updates. In [14], [145]–[149] authors analyzed ALOHA and CSMA like random access. In [150] the authors showed asymptotic optimality of a decentralized round robin scheduling policy. In [151] the

[13] This result was shown for a remote estimation system with zero service time, i.e., $Y_i = 0$ [128]. The treatment in [128] may also apply to the case of non-zero service time.

authors considered multiaccess in which a node chooses its access probability as a function of the age of its updates. In [152] a sleep-wake strategy is designed for a network of low-powered battery constrained nodes, which use carrier sensing based access.

In [131], the authors addressed the problem of scheduling transmissions in broadcast wireless networks in order to minimize AoI. Age minimization with minimum throughput requirements was studied in, e.g., [153], [154]. Stochastic arrival processes were considered in [155]. In [132], [156], the authors considered the problem of optimizing AoI in wireless networks under general interference constraints. Several extensions have recently appeared, including the case of multi-source multi-hop wireless networks [157], [158], distributed policies for age minimization [159], and age-based [160] and virtual-queue-based [161] policies for age minimization. Scaling of age as a function of the number of nodes in a large multi-hop wireless network has been studied [162]–[165].

B. Updates Through Erasure Channels

In [134] age is analyzed under different approaches to coded redundancy. While increasing redundancy improves the probability that an update will be successfully delivered, it comes at the expense of longer update transmission times. The authors analyze finite redundancy (FR), wherein a k symbol update is transmitted as a fixed number n of coded symbols, and infinite incremental redundancy (IIR), wherein coded symbols are sent until k are successfully received. In [70] the HARQ protocols of FR and IIR are considered. The HARQ FR is different from FR in [134] in that it retransmits until an update is successfully transmitted. Further, a fresh update is generated at the beginning of a transmission, unlike [134] where updates arrive as a Poisson process. A new arrival must either preempt the currently transmitting update or be discarded; the latter is found to be the better strategy.

In [135] ARQ and HARQ are analyzed for when an update is coded as n symbols for transmission. Fresh updates arrive to a FCFS queue as a Poisson process. The codeword length n that optimizes age is of interest. In [113], [114] ARQ and HARQ are studied with the goal of minimizing age when there is a constraint on the time-average of the number of transmissions at the source. Last but not the least, [133] compares HARQ and a transmit scheme that encodes an update into n symbols but transmits it only once. Random linear codes are assumed for forward error correction.

C. Decentralized Multiaccess

In a vehicular network setting where each vehicle sent updates using carrier sense based 802.11 medium access, different packet management strategies were simulated and a heuristic gradient-descent based algorithm was empirically shown to minimize the average age [14]. In [146], the authors compared ALOHA with scheduled access over unreliable channels. When all network nodes desired the same average age, it was shown that ALOHA led to age that was worse than scheduled access by a factor of $2e$.

In [145] nodes use CSMA to access the channel. The authors suitably approximate practical CSMA in developing a SHS model and the corresponding expression for average age. The age is optimized over the back-off rates of the nodes. They show that these rates are independent of the arrival rates of fresh updates at the nodes. In [149] it is shown that one can reduce average AoI by a factor of 2 in comparison to the minimum that can be achieved with typical slotted ALOHA. Nevertheless, the authors propose using ALOHA, but with large offered loads together with a load thinning mechanism that discards fresh packets that would lead to age reductions smaller than a threshold.

All the above works consider slotted access, with [14] and [145] also assuming carrier sensing. Unlike these, [148] considers a network of transmit-only network of nodes. The setting is unslotted, precludes carrier sensing, and allows for channel error. Updates are assumed to arrive as a Poisson process and transmission times are exponentially distributed; SHS is used to derive the average AoI.

In [141], [150] the authors consider policies that schedule transmissions in a network of nodes and provide packet management at each node. The transmissions must be scheduled in a non-interfering manner. Also, it should be possible to implement the policies in a decentralized manner. The authors show that the RR-ONE policy, which enforces a round robin schedule and only retains the latest packet at any node is asymptotically optimal (minimizes average age) among all policies. Specifically, it achieves an optimal asymptotic scaling factor (average age of a node normalized by the number of nodes) of 0.5. In comparison, the scaling factor of CSMA is shown to be at least 1. While asymptotically optimal, RR-ONE may not perform well with fewer nodes and low packet arrival rates.

In [142], [143] the authors formulate age minimization as a restless multi-armed bandit problem. They derive the Whittle index and show that each source can calculate its own index independently of the others. In a centralized setup, access to indices of all nodes is available and the resulting index policy has near-optimal performance. They propose a decentralized scheme that has nodes access the channel with probabilities that are a function of their index.

In [151] the nodes don't transmit if the age of their update is smaller than a certain threshold. All nodes whose age exceeds the threshold access the medium with the same probability. The average age for the network is derived for such a policy and the same is optimized over the choices of threshold and access probability.

Unlike all the above works on shared medium access, [152] considers a network of battery-constrained nodes. The goal is to design a sleep-wake strategy that minimizes the weighted average peak age while ensuring that the energy constraints are satisfied.

D. Broadcast Wireless Networks

Consider a single-hop wireless network with N nodes sharing time-sensitive information through unreliable communication links to a base station (BS). Let the time be slotted, with slot duration normalized to unity and slot index $t \in \{1, 2, \cdots, T\}$, where T is the time-horizon of this discrete-time system. The broadcast wireless channel allows at most one packet transmission per slot. In each time-slot t, the BS either idles or schedules a transmission in a selected link $i \in \{1, 2, \cdots, N\}$. Let $v_i(t) \in \{0, 1\}$ be the indicator function that is equal to 1 when the BS selects link i during slot t, and $v_i(t) = 0$ otherwise. *When $v_i(t) = 1$ the corresponding source samples fresh information, generates a new packet and transmits this packet over link i.* Notice that packets are not enqueued. Since the BS can select at most one link at any given slot t, we have

$$\sum_{i=1}^{N} v_i(t) \leq 1, \qquad t \in \{1, \ldots, T\}. \tag{51}$$

The transmission scheduling policy governs the sequence of decisions $\{v_i(t)\}_{i=1}^{N}$ of the BS over time.

Let $c_i(t) \in \{0, 1\}$ represent the channel state associated with link i during slot t. When the channel is *ON*, we have $c_i(t) = 1$, and when the channel is *OFF*, we have $c_i(t) = 0$. The channel state process is assumed i.i.d. over time and independent across different links, with $\mathbb{P}(c_i(t) = 1) = p_i, \forall i, t$. Let $d_i(t) \in \{0, 1\}$ be the indicator function that equals 1 when the transmission in link i during slot t is successful; otherwise, $d_i(t) = 0$. A successful transmission occurs when a link is selected and the associated channel is ON, implying that $d_i(t) = c_i(t)v_i(t), \forall i, t$.

In [131], the authors consider four low-complexity scheduling policies, namely Maximum Age First, Stationary Randomized, Max-Weight and Whittle's Index, and derive performance guarantees for each of them as a function of the network configuration.

- *Maximum Age First (MAF)*: In each slot t, select link i with highest age $\Delta_i(t)$, with ties broken arbitrarily.

It was shown, using stochastic ordering arguments, that the MAF policy minimizes the average AoI under symmetric conditions ($p_i = p \in (0, 1]$ and weights $w_i = w > 0$ for all i.) Note that MAF leverages the knowledge of $\Delta_i(t)$, but disregards the values of w_i and p_i. However, when these parameters are not symmetric, MAF age performance can be arbitrarily poor.

As an alternative to MAF, randomized scheduling policies use knowledge of w_i and p_i to achieve good performance in arbitrary settings. The randomized policy is based on a set of positive fixed values of $\{\beta_1, \ldots, \beta_N\}$.

- *Randomized*: In each slot t, select link i with probability $\beta_i / \sum_{j=1}^{N} \beta_j$.

Although this simple policy uses no information from current or past states of the network, it was shown [131] that the Randomized policy with $\beta_i = \sqrt{w_i/p_i}$ *achieves 2-optimal performance in all network configurations* (N, p_i, w_i).

Next, we describe a Max-Weight policy that leverages knowledge of w_i, p_i and $\mathbf{\Delta}(t) = [\Delta_1(t), \cdots, \Delta_N(t)]$ in making scheduling decisions. Consider the linear Lyapunov function

$$L(\mathbf{\Delta}(t)) = \frac{1}{N} \sum_{i=1}^{N} \tilde{\alpha}_i \Delta_i(t), \tag{52}$$

where $\tilde{\alpha}_i > 0$ are auxiliary parameters for performance tuning. The Max-Weight policy is defined to minimize the one-slot Lyapunov drift

$$\Phi(\mathbf{\Delta}(t)) = \mathbb{E}\left[L(\mathbf{\Delta}(t+1)) - L(\mathbf{\Delta}(t)) \middle| \mathbf{\Delta}(t) \right]. \tag{53}$$

For the linear Lyapunov function (52), this specifies the policy:

- *Max-Weight*: In each slot t, select the link i with highest value of $p_i \tilde{\alpha}_i \Delta_i(t)$, with ties broken arbitrarily.

Similar to the randomized policy, it was shown [131] that Max-Weight also achieves 2-optimal performance under all network configurations. However, in practice, the Max-Weight policy achieves much better average performance than the simple randomized policy, as shown through simulations.

The choice of a linear Lyapunov function (52) with auxiliary parameters $\tilde{\alpha}_i$ resulted in a performance guarantee of 2-optimality. Choosing a different Lyapunov function yields a different Max-Weight policy with a different performance guarantee. For example, a quadratic Lyapunov function resulted in a performance guarantee of 4-optimality [131].

Finally we consider the AoI minimization problem from a different perspective and propose an Index policy [166], also known as Whittle's Index policy. This policy is surprisingly similar to the Max-Weight policy and also yields strong performance. Whittle's Index policy is the optimal solution to a relaxation of the Restless Multi-Armed Bandit (RMAB) problem. This low-complexity heuristic policy has been extensively used in the literature [167]–[169] and is known to have a strong performance in a range of applications [170], [171]. The challenge associated with this approach is that the Index policy is only defined for problems that are *indexable*, a condition which is often difficult to establish. A detailed introduction to the relaxed RMAB problem can be found in [166], [172]. It was shown [131] that the AoI minimization problem is indeed indexable with Whittle's Index

$$C_i(\Delta_i(t)) = \frac{w_i p_i}{2} \Delta_i(t) \left[\Delta_i(t) + \frac{2}{p_i} - 1 \right]. \tag{54}$$

This specifies the policy:

- *Whittle Index*: In each slot t, select the link i with highest value of $C_i(\Delta_i(t))$, with ties being broken arbitrarily.

Notice that the Whittle's Index policy is similar to the Max-Weight policy despite being developed using entirely different approaches. Moreover, both are equivalent to the MAF policy when the network is symmetric, implying that both are AoI-optimal when $w_i = w$ and $p_i = p$. It was shown in [131] that the Whittle index policy is 8-optimal in the worst case. However, its strong performance is demonstrated through simulation.

In Fig. 10, we compare the performance of the scheduling policies in terms of the Expected Weighted Sum Age of Information. The results show that the performances of the Max-Weight and Whittle Index policies are close to optimal and generally outperform Maximum Age First and Randomized policies.

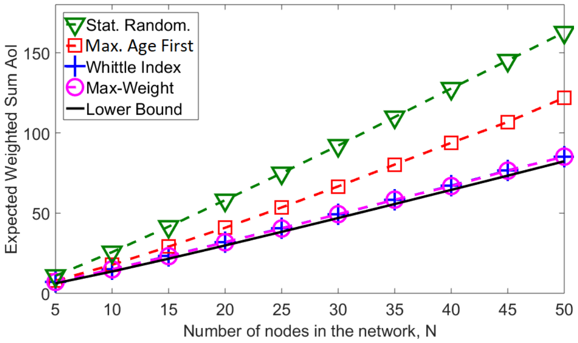


Fig. 10. Network with $T = 100,000, w_i = 1, p_i = i/N$. The simulation result for each policy and for each value of N is an average over 10 runs.

E. General Wireless Networks

In the previous section, we discussed scheduling policies for minimizing age in a broadcast network. However, many of the applications in which age is an important metric involve wireless networks with more general interference constraints. Here we examine the work [132] on scheduling policies to minimize peak and average age for a wireless network with time-varying links and general interference constraints.

The wireless network consists of a set of source-destination pairs, each connected by a wireless link. Each source generates information updates that are to be sent to its destination. The wireless network is modeled as a graph $G = (V, E)$, where V is the set of nodes and E is the set of communication links. Each link $e \in E$ is a source-destination pair in the network. Time is slotted with slot duration normalized to unity.

Wireless interference constraints limit the set of links that can be activated simultaneously. We call a set $m \subset E$ a *feasible activation set* if all links in m can be activated simultaneously without interference, and denote by $\mathcal{A}$ the collection of all feasible activation sets. We call this the *general interference model*, as it incorporates several popular interference models such as 1-hop interference, k-hop interference, and protocol interference models.

A non-interfering transmission over link e does not always succeed due to channel errors. We let $c_e(t) \in \{1, 0\}$ denote the channel error process for link e, where $c_e(t) = 1$ if a non-interfering transmission over link e succeeds and $c_e(t) = 0$ otherwise. We assume $c_e(t)$ to be independent across links, and i.i.d. across time, with known channel success probabilities $\gamma_e = \mathbb{P}[c_e(t) = 1] > 0$, for all $e \in E$.

We first consider *active sources* that can generate a new update packet at the beginning of each slot for transmission, while discarding old update packets that were not transmitted. Thus, for an active source, a transmitted packet always contains fresh information. In [132] the authors design randomized scheduling policies to minimize peak and average age. In the previous section, we saw a special case of a distributed randomized stationary policy in which a single link e was activated with probability $p_e > 0$, independent across time slots.

In the context of general interference, a randomized policy assigns a probability distribution $\mathbf{x} \in \mathbb{R}^{|\mathcal{A}|}$ over the collection of feasible activation sets, $\mathcal{A}$. Then, in each slot, activate the set $m \in \mathcal{A}$ with probability x_m, independent across time. For this policy, the activation frequency (probability) for link e is given by $f_e = \sum_{m: e \in m} x_m$.

It was shown [132] for any stationary randomized policy that the average and peak age both equal $\sum_{e \in E} w_e/(\gamma_e f_e)$. Thus, the age minimization problem can be written as

$$\begin{aligned} &\underset{\mathbf{x} \in \mathbb{R}^{|\mathcal{A}|}}{\text{Minimize}} \quad \sum_{e \in E} \frac{w_e}{\gamma_e f_e} \\ &\text{subject to} \quad \mathbf{1}^T \mathbf{x} \leq 1, \ \mathbf{x} \geq 0. \end{aligned} \tag{55}$$

Note that the optimization is over $\mathbf{x}$, the activation probabilities of feasible activation sets $m \in \mathcal{A}$. This is because the link activation frequencies f_e get completely determined by $\mathbf{x}$. The problem (55) is a convex optimization problem in standard form [173]. The solution to it is a vector $\mathbf{x} \in \mathbb{R}^{|\mathcal{A}|}$ that defines a probability distribution over link activation sets $\mathcal{A}$, and determines a centralized stationary policy that minimizes average and peak age over all randomized policies. It was also shown in [132] that this policy minimized peak age over all policies (not just randomized) and the average age for this policy is within a factor of 2 from the optimal average age over all policies.

The optimization problem (55), although convex, has a variable space that is $|\mathcal{A}|$-dimensional, and thus, its computational complexity increases exponentially in $|V|$ and $|E|$. It is, however, possible to obtain the solution efficiently in certain specific cases. For example, under single-hop interference, where links interfere with one another if they share a node, every feasible activation set is a matching on G, and therefore, $\mathcal{A}$ is a collection of all matchings in G. As a result, the constraint set in (55) is equal to the matching polytope [174]. The problem of finding an optimal schedule reduces to solving a convex optimization problem (55) over a matching polytope. This can be efficiently solved (i.e., in polynomial time) by using the Frank-Wolfe algorithm [175], and the separation oracle for matching polytope developed in [176].

Next we discuss a number of interesting extensions.

a) Buffered Sources: In some situations it is not possible for the source to generate packets on demand, and instead each source generates update packets at random. The generated packets get queued at the MAC layer FCFS queue for transmission. In this setting, the AoI minimization problem involves optimization of both the packet generation rate and the link activation frequencies. When the packets arrive according to an independent Bernoulli process, it was shown [132] that the AoI optimization problem can be decoupled; the link scheduling is done according to the same randomized policy that minimized AoI in the active source case, and the packet generation rate is given by the solution to a simple optimization problem.

b) Channel State Information: In some situations, it is possible to observe the channel in advance, and decide when to transmit based on channel conditions. Typically, the channel is modeled as either being in either the “on” or “off” state, and a transmission is successful if it takes place when the channel is on. In [160] and [161] the authors consider the impact of CSI on the age minimization problem and propose age-based scheduling policies that take CSI into account. They show that using CSI can significantly improve average

AoI. Another angle on the interplay between AoI and CSI, i.e., the age of channel state information, was considered in [177]–[179]. These studies recognized the importance of timely CSI in wireless networks and that CSI can be learned by direct channel estimation from a status update as well as indirectly through the contents of a status update. This work derived fundamental bounds and efficient schedules to minimize the age of global CSI in wireless networks with reciprocal channels.

c) Multi-Hop Networks: The multi-hop setting was first studied in the context of vehicular networks where vehicles "piggyback" status updates over multiple hops [15]. The multi-hop setting is interesting in that it typically removes the abstraction of the queue and explicitly considers the effect of the network topology and link contention in the analysis. The analysis in [15] and the subsequent studies in [26], [55], [90], [180]–[182] considered age of information in specific multi-hop network structures, e.g., line, ring, and/or two-hop networks. In [183], sources are polled for updates by a gateway that aggregates the updates and sends them to a monitor. A general multi-hop network setting where a single-source disseminates status updates through a gateway to the network was considered in [52], [56]. Another general multi-hop setting in which each node is both a source and a monitor of information was considered in [158], [165], [184]–[186]. These studies led to the formulation of fundamental limits and schedules that were shown to achieve an average age close to these limits. A practical age control protocol to improve AoI in multi-hop IP networks was proposed in [187]. The distribution of AoI for general networks with time-invariant erasure probabilities on each link was derived in [188].

VIII. Applications

In addition to queue/network focused analyses, AoI has also appeared in various application areas, including dissemination of channel state information [177]–[179], [189], timely updates via replicated servers [190], [191] including multicasting networks [192]–[196], timely source coding [197]–[203], differential encoding of temporally correlated updates [204], correlated updates from multiple cameras [205], periodic updates from correlated IoT sources [206], mobile cloud gaming [207], [208], computation-intensive updating where updates are generated after processing raw data [209]–[211], game-theoretic approaches to network resource allocation for updating sources [212]–[223], and timely updating of researchers' citations in Google Scholar [224].

In the following sections, we describe contributions to AoI analysis related to various application areas, namely, game theory, learning, caching, and network protocols. We note that these works focus on *analysis* of area-specific AoI metrics; the goodness of AoI as a performance metric is taken for granted. With respect to applications, much work needs to be done to determine when application users can perceive benefits of reduced AoI. A notable start is [208], an empirical study of quality-of-experience (QoE) in mobile cloud gaming. This work found that frame age (the average age of the most recently received game video frame) is the most important QoS metric for predicting in-game performance and QoE.

A. Age and Games

Game theory has been applied in various settings where one or more players value timeliness [212]–[223]. This includes adversarial settings where one player aims to maintain the freshness of information updates while the other player aims to prevent this [212]–[215].

In [212], the authors formulated a two-player game between a transmitter that aims to establish a connection to its receiver and an interferer that attempts to disrupt the connection. The players choose power levels as strategies and it is shown that both players have the same strategy at the Nash Equilibrium. It also shown that the Stackelberg strategy, when led by the interferer, dominates the Nash strategy.

In [213], a dynamic game between a real-time monitoring system that cares about the timeliness of status updates sent over a wireless channel and an attacker that jams the channel to delay the status updates was considered. The authors proved the existence of a unique stationary equilibrium in the game and characterized the equilibrium analytically. They showed that the attacker chooses a jamming time distribution with high variance, while the system chooses a sampling policy that results in a low variance in the time between the reception of two consecutive updates.

In [214], authors modeled the interaction between a UAV transmitter and an adversarial interferer and showed that there exists a unique Nash equilibrium but multiple Stackelberg equilibria. In [215], the authors considered a communication scheduling and remote estimation problem in the presence of an adversary and obtained a Nash equilibrium for the non-zero-sum dynamic game.

In [216], the authors formulated a two-player game to model the interaction between two transmitter-receiver pairs over an interference channel. The transmitters desire freshness of their updates at their receivers and can choose their transmit power levels. The Nash and Stackelberg strategies are derived and it is shown that the Stackelberg strategy dominates the Nash strategy.

In [217]–[219] authors studied age in the context of spectrum sharing between networks. A game-theoretic approach was proposed in [217] to study the coexistence of Dedicated Short Range Communication (DSRC) and WiFi; the DSRC network desires to minimize the average AoI and the WiFi network aims to maximize the average throughput. For the one-shot game, the Nash and Stackelberg equilibrium strategies were evaluated. The DSRC-WiFi coexistence problem from [217] was generalized in [218] to the coexistence of age and throughput optimizing networks. A repeated game approach was employed to capture the interaction of the networks over time. This line of work was further extended [219] to explore the possibility of cooperation between age and throughput optimizing networks using a randomized signaling device. It was shown that networks choose to cooperate only when they consist of a sufficiently small number of nodes, but otherwise they prefer to compete.

In [220] and [221], authors studied the coexistence of nodes that value timeliness of their information at others and provided insights into how competing nodes would coexist. In [220], authors proposed a one-shot multiple access game with nodes as players, where each node shares the spectrum using a CSMA/CA based access mechanism. Authors investigated the equilibrium strategies of nodes in each CSMA/CA slot when collision slots are shorter than successful transmissions, as well as when they are longer. They showed that when collisions are shorter, transmitting is a weakly dominant strategy. However, when collisions are longer, no weakly dominant strategy exists and a mixed strategy Nash equilibrium was derived. In [221], authors considered a distributed competition mode where each node wants to minimize a function of its age and transmission cost but information such as the number of nodes in the network and their strategies is not available. A learning strategy was proposed for each node to use its current empirical average of age and transmission cost to determine its transmit probability in each slot. They showed that for a certain set of parameters the proposed strategy converges to an equilibrium that is identified as the Nash equilibrium for a suitable virtual game.

In [222], authors proposed a Stackelberg game between an access point and a set of helpers that contribute toward charging a sensor via wireless power transfer. The access point would like to minimize a utility that includes the AoI of the sensor, the power transferred by it to charge the sensor, and the payments it makes to the helpers. The helpers benefit from the payments but bear costs of transferring power. In [223], authors designed a mobile edge computing enabled 5G health monitoring system for wireless body area networks (WBANs) in the Internet of Medical Things (IoMT). The goal was to minimize system-wide costs that depend on medical criticality, AoI, and energy consumption of health monitoring packets. The IoMT was divided into two sub-networks, namely, intra-WBANs and beyond-WBANs. For the intra-WBANs, the authors formulated a cooperative game to minimize the cost per patient. For the beyond-WBANs, where patients can choose to analyze the information either at local devices or at edge servers, the authors formulated a non-cooperative game and analyzed the Nash equilibrium.

B. Age and Learning

Data driven model-free learning techniques have been applied to optimizing age-of-information for different network assumptions and applications [113], [114], [147], [225]–[231]. In addition, age-based computational load balancing in machine learning applications with straggling servers has been studied in [232]. In [232], already aged parts of the computations are assigned to servers that are less likely to straggle in the next computation cycle.

In [113], [114], [225]–[228], the goal is to learn policies that schedule updates from one or more sources with the goal of minimizing their ages at a monitor. Authors in [113], [114] aim to schedule updates of a source over an error-prone channel under different feedback mechanisms including ARQ and Hybrid ARQ. They also proposed an on-policy SARSA algorithm [233], with the policy modeled by a softmax distribution, for when the channel error statistics are not known. The authors consider the case where a source may send updates to multiple users in [225]. In [226], the authors consider a network of devices whose updates are delivered to applications via an IoT hub. An application is updated at the end of a frame only if updates from all devices relevant to the application are scheduled for transmission during the frame. The authors use the options framework (hierarchical reinforcement learning) [234] to model scheduling as composed of two policies: the first chooses the application to be updated and the other chooses how devices corresponding to the application must be scheduled. A deep Q network based architecture is proposed to learn the scheduling policy.

In [227], the authors consider a network of sensors sending their updates to a remote monitor. The network would like to optimize a weighted sum of the average AoI and the probabilities that the ages of sensors' updates at the monitor exceed a predefined threshold. The probabilities model reliability for ultra reliable low latency communication (URLLC) systems. The actor-critic [233] algorithm A3C is used to learn a policy that chooses the sensor that must send its update. In [228], the authors empirically compare the AoI achieved using the reinforcement learning algorithms of Deep Q-Learning and Policy Gradient for scheduling packets from multiple (four) flows that arrive at a single queue single server facility. They also compare against the often analyzed maximum age first scheduling policy.

In [147] the authors consider a fully connected ad hoc network in which nodes send updates to each other. They use deep Q-learning in a multi-agent setting to learn the policy that every node must use. The policy chooses the transmit power for a node at every time slot.

Works [229]–[231] learn trajectories of one or more Unmanned Aerial Vehicles (UAVs) that collect information from sensors. In [229], mobile UAVs plan trajectories to obtain updates from IoT devices that must be delivered to a ground station. The UAV trajectories must ensure that average AoI at the ground station is below a certain threshold. However, one would like to choose trajectories such that energy efficiency is maximized. In [230], deep reinforcement learning is used to find an energy-constrained UAV trajectory (through the sensors and back to its base) that minimizes the weighted sum AoI of the sensor updates. In a similar but different problem formulation, the authors of [231] propose a deep reinforcement learning algorithm to optimize (i) the trajectory a UAV must take to obtain updates from sensors on the ground and (ii) the scheduling of updates at the sensors with the goal of minimizing the sum AoI at the UAV.

C. Caching

Gao et al. [235] were the first to introduce the concept of cache freshness in opportunistic mobile networks. They assume a single data source updating multiple caching servers (nodes), wherein each caching node may further update other nodes. The authors in [236]–[238] use AoI together with the popularity of content for cache management. In [236], the authors design a popularity weighted AoI metric based update policy for updating a local cache connected to a remote server via a limited capacity link. The authors show that the update rate of content in the local cache should be proportional to the square root of its popularity. In [238] the authors propose an update policy that minimizes the average AoI of the files with respect to a given popularity distribution in a single-server single-cache system. The authors consider AoI-dependent update durations and compare their policy with the square root policy proposed in [236].

In [237], the authors propose a caching model for requesting content based on the popularity (history of requests) and freshness (age), captured using *effective age*. They propose an optimal policy that minimizes the number of missed requests while managing the contents of the cache when the cache is full and new content arrives. In [239], the authors propose a cache-assisted lazy update and delivery (CALUD) to jointly optimize content freshness and service latency in vehicular networks. They consider multiple data sources periodically updating a single roadside unit (RSU).

In [240] the authors discuss the real-time performance of a cache enabled network with AoI as a freshness metric. They propose a random caching framework and show that it performs better than the most popular content based and uniform caching strategies with respect to minimizing the peak AoI. In [241] the authors consider AoI in serial and multi-access connected cache networks, and analyze age at intermediate caches and the end-users. Extending [241], the authors in [242] consider the trade-off between obtaining a file from the source and obtaining it from the cache. While the former could entail larger file transmission times, albeit of the freshest copy, the latter may result in an older copy obtained quickly. They consider whether a file must be stored in the cache and if so then the rate at which it must be updated.

In [243] the authors propose a freshness-aware refreshing scheme that minimizes the average service delay while guaranteeing content freshness. The cached contents are refreshed on user requests if their AoI exceeds a certain threshold (refreshing window). A small window increases service delays as the latest version is fetched more often. On the other hand, a large window increases age but keeps service delays small as the cached version is sent to the user.

In [244] the authors use Q-learning to achieve staleness control of machine learning models available at the edge for data analytics. They consider six different performance metrics including the AoI and the value of information updates.

D. Protocols

In [245] the authors measure the AoI at a remote server when using TCP/IP over a selection of access networks including WiFi, LTE, 2G/3G and Ethernet. Experiments are conducted for various application update rates, both in an emulated testbed and over the Internet. In [246], [247] authors analyzed AoI over real networks using different devices, protocols and networks (wired/wireless). The authors discuss the challenges associated in measuring AoI in real networks, including synchronization, selection of hardware, and choice of transport protocol and draw insights for AoI aware transmission protocols. In [187], [248], [249] the authors propose the Age Control Protocol, an end-to-end transport protocol that minimizes the AoI of the delivered update at the monitor in a network-transparent manner. The authors detail the protocol algorithm and empirically demonstrate its efficacy using real-world experiments that have one or more devices sending updates to a server in the cloud.

IX. Conclusion

Age of information has emerged as an end-to-end performance metric for systems that employ status update messages. In this survey, we have seen that the concept of age is so general that age-based optimization problems can be found almost everywhere, arising in all network layers and in all system components that can be viewed as communicating via updates. A source can optimize the rate at which it submits updates or the source may adopt an age-optimizing update transmission policy. A service facility that processes updates can employ scheduling policies to maximize the timeliness of delivered updates. A base station can use age to schedule downlink transmissions to its users. On the uplink, users can use age-based policies to access the channel. These policies differ substantially from rate/throughput maximization, particularly for energy harvesting sources. Moreover, in these network settings, age-based update policies are practical because senders, servers and routers understand packet time-stamps, as opposed to application-specific measurements.

This AoI introduction has described the development of new analytical models for freshness and tools for age analysis. These new methods have been used to characterize age in relatively abstract models of sensors, networks, and service facilities. Some key ideas have emerged: The updating process should be matched to the system that is handling the updates, neither underloading nor overloading the system. The system should aim to process new updates rather than old. The system should avoid processing updates that lack sufficient novelty. These observations suggest that the introduction of timeliness requirements results in systems that work smarter not harder.

We also have seen that AoI metrics are beginning to be applied in abstract models of various application domains ranging from UAV trajectory guidance to mobile gaming. However, there remain many unanswered questions, open problems, and unexplored application areas.

A promising area is edge cloud processing for low-latency edge-assisted applications. The system model is that the delivery of an update requires (timely) computation in the cloud. In this setting, mechanisms like service preemption that were introduced here in the context of queues and communication links are in fact more applicable in the context of edge-cloud processing; killing a job on a processor is a lot simpler than killing a packet in a network.

Often these edge-cloud computations will involve machine learning. The intersection of predictive machine learning and AoI could become important. Using machine learning to solve scheduling problems for reduced AoI is an obvious opportunity; but using AoI to accelerate real-time learning algorithms through prioritization of fresh data is also feasible.

With respect to networking, an interesting open problem is how to enable "age-optimality" as a network service that caters to a host of cyberphysical system applications. In such a world, a CPS application (and the many sources and monitors that underpin its functionality) would open an "age-socket." The network would then manage the end-to-end connections, including polling/sampling, and managing application queues. The networking stack of course performs some of these functions already, though not for timeliness.

We have already seen that age is applicable to various estimation problems. It should be apparent that functions of AoI can play an important role in scheduling feedback for control systems. How this should be done yields problems at the intersection of sensors, processing, and communications, where there is a tradeoff between processing (e.g., compression) and latency. More compression takes time and processing capacity, but can reduce communication latency.

Age/distortion tradeoffs represent another aspect of these types of problems, For example, suppose one can get a low resolution image with small age or a higher resolution image with increased age. In what settings would one be preferable? For remote estimation and control, a (big) open question is how to design sampling, quantization, coding, and control for multi-dimensional signals that could be correlated and non-Markovian. Of course, this will be related to age/distortion tradeoffs.

In each of these potential problem areas, there are likely to be a host of abstract but informative problem formulations. The solutions will perhaps answer the question of when or whether AoI based problem formulations will become a useful practical mechanism for optimizing the design and operation of widely deployed cyberphysical systems.

References

[1] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.

[2] P. Popovski, Č. Stefanovic, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[3] J. Sachs, L. A. A. Andersson, J. Araujo, C. Curescu, J. Lundsjo, G. Rune, E. Steinbach, and G. Wikstrom, "Adaptive 5G low-latency communication for tactile internet services," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 325–349, 2019.

[4] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, March 2012, pp. 2731–2735.

[5] X. Song and J. W.-S. Liu, "Performance of multiversion concurrency control algorithms in maintaining temporal consistency," in *Proceedings., Fourteenth Annual International Computer Software and Applications Conference*. IEEE, 1990, pp. 132–139.

[6] M. Xiong and K. Ramamritham, "Deriving deadlines and periods for real-time update transactions," in *The 20th IEEE Real-Time Systems Symposium*. IEEE, 1999, pp. 32–43.

[7] A. Karakasidis, P. Vassiliadis, and E. Pitoura, "ETL queues for active data warehousing," in *Proc. 2nd international workshop on Information quality in information systems (IQIS)*, 2005, pp. 28–39.

[8] Shigang Chen and K. Nahrstedt, "Distributed QoS routing with imprecise state information," in *Proc. 7th International Conference on Computer Communications and Networks*, 1998, pp. 614–621.

[9] H. Yu, L. Breslau, and S. Shenker, "A scalable web cache consistency architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 4, pp. 163–174, Aug. 1999.

[10] Y. C. Hu and D. B. Johnson, "Ensuring cache freshness in on-demand ad hoc network routing protocols," in *ACM international workshop on Principles of Mobile Computing (POMC)*, 2002, pp. 25–30.

[11] V. C. Giruka and M. Singhal, "Hello protocols for ad-hoc networks: overhead and accuracy tradeoffs," in *World of Wireless Mobile and Multimedia Networks, WoWMoM, Sixth IEEE International Symposium on*. IEEE, Jun. 2005, pp. 354–361.

[12] J. Cho and H. Garcia-Molina, "Effective page refresh policies for web crawlers," *ACM Transactions on Database Systems (TODS)*, vol. 28, no. 4, pp. 390–426, 2003.

[13] S. Ioannidis, A. Chaintreau, and L. Massoulie, "Optimal and scalable distribution of content updates over a mobile social network," in *Proc. IEEE INFOCOM*, 2009.

[14] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, 2011.

[15] S. K. Kaul, R. D. Yates, and M. Gruteser, "On piggybacking in vehicular networks," in *IEEE Global Telecommunications Conference, GLOBECOM 2011*, Dec. 2011.

[16] S. M. Ross, *Stochastic Processes*, 2nd ed. John Wiley & Sons, 1996.

[17] R. G. Gallager, *Stochastic processes: theory for applications*. Cambridge University Press, 2013.

[18] M. Costa, M. Codreanu, and A. Ephremides, "On the age of information in status update systems with packet management," *IEEE Trans. Info. Theory*, vol. 62, no. 4, pp. 1897–1910, April 2016.

[19] ——, "Age of information with packet management," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, June 2014, pp. 1583–1587.

[20] Y. Inoue, H. Masuyama, T. Takine, and T. Tanaka, "A general formula for the stationary distribution of the age of information and its application to single-server queues," *IEEE Trans. Info. Theory*, vol. 65, no. 12, pp. 8305–8324, 2019.

[21] J. Hespanha, "Modelling and analysis of stochastic hybrid systems," *IEE Proceedings-Control Theory and Applications*, vol. 153, no. 5, pp. 520–535, 2006.

[22] R. D. Yates and S. K. Kaul, "The age of information: Real-time status updating by multiple sources," *IEEE Trans. Info. Theory*, vol. 65, no. 3, pp. 1807–1827, March 2019.

[23] D. Vermes, "Optimal dynamic control of a useful class of randomly jumping processes," International Institute for Applied Systems Analysis, Tech. Rep. PP-80-015, 1980.

[24] M. H. A. Davis, "Piecewise-deterministic Markov processes: a general class of nondiffusion stochastic models," *J. Roy. Statist. Soc.*, vol. 46, pp. 353–388, 1984.

[25] L. DeVille, S. Dhople, A. D. Domínguez-García, and J. Zhang, "Moment closure and finite-time blowup for piecewise deterministic Markov processes," *SIAM Journal on Applied Dynamical Systems*, vol. 15, no. 1, pp. 526–556, 2016.

[26] R. D. Yates, "The age of information in networks: Moments, distributions, and sampling," *IEEE Trans. Info. Theory*, vol. 66, no. 9, pp. 5712 – 5728.

[27] T. Z. Ornee and Y. Sun, "Sampling and remote estimation for the Ornstein-Uhlenbeck process through queues: Age of information and beyond," 2021, submitted to *IEEE/ACM Trans. Netw.*, https://arxiv.org/abs/1902.03552.

[28] J. P. Champati, M. H. Mamduhi, K. H. Johansson, and J. Gross, "Performance characterization using AoI in a single-loop networked control system," in *Proc. IEEE INFOCOM Age of Information Workshop*, 2019.

[29] M. Klügel, M. H. Mamduhi, S. Hirche, and W. Kellerer, "AoI-penalty minimization for networked control systems with packet loss," in *Proc. IEEE INFOCOM Age of Information Workshop*, 2019.

[30] Y. Sun, E. Uysal-Biyikoglu, R. Yates, C. E. Koksal, and N. B. Shroff, "Update or wait: How to keep your data fresh," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7492–7508, Nov. 2017.

[31] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Age and value of information: Non-linear age case," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, June 2017, pp. 326–330.

[32] Y. Sun and B. Cyr, "Sampling for data freshness optimization: Non-linear age functions," *J. Commun. Netw.*, vol. 21, no. 3, pp. 204–219, 2019.

[33] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "The cost of delay in status updates and their value: Non-linear ageing," *IEEE Trans. Commun., in press*, 2020.

[34] C. Shapiro and H. Varian, *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business Press, 1999.

[35] A. Even and G. Shankaranarayanan, "Utility-driven assessment of data quality," *SIGMIS Database*, vol. 38, no. 2, pp. 75–93, May 2007.

[36] B. Heinrich, M. Klier, and M. Kaiser, "A procedure to develop metrics for currency and its application in CRM," *J. Data and Information Quality*, vol. 1, no. 1, pp. 5:1–5:28, 2009.

[37] E. Altman, R. El-Azouzi, D. S. Menasche, and Y. Xu, "Forever young: Aging control for smartphones in hybrid networks," 2010, https://arxiv.org/abs/1009.4733.

[38] S. Razniewski, "Optimizing update frequencies for decaying information," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 1191–1200.

[39] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *J. Commun. Netw.*, vol. 15, no. 4, pp. 338–351, Aug 2013.

[40] M. K. C. Shisher, H. Qin, L. Yang, F. Yan, and Y. Sun, "The age of correlated features in supervised learning based forecasting," in *Proc. IEEE INFOCOM Age of Information Workshops*, 2021.

[41] Y. Sun, Y. Polyanskiy, and E. Uysal, "Sampling of the Wiener process for remote estimation over a channel with random delay," *IEEE Trans. Inf. Theory*, vol. 66, no. 2, pp. 1118–1135, Feb 2020.

[42] T. Soleymani, J. S. Baras, and K. H. Johansson, "Stochastic control with stale information–part i: Fully observable systems," in *IEEE Conference on Decision and Control (CDC)*, 2019, pp. 4178–4182.

[43] Y. Sun and B. Cyr, "Information aging through queues: A mutual information perspective," in *Proc. IEEE SPAWC Workshop*, 2018.

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

[45] T. Soleymani, S. Hirche, and J. S. Baras, "Optimal self-driven sampling for estimation based on value of information," in *Proceedings of the 13th International Workshop on Discrete Event Systems (WODES)*, 2016.

[46] ——, "Maximization of information in energy-limited directed communication," in *Proc. European Control Conference (ECC)*, 2016.

[47] ——, "Optimal stationary self-triggered sampling for estimation," in *Proc. IEEE CDC*, 2016.

[48] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Minimizing the age of information through queues," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5215–5232, Aug. 2019.

[49] M. Bastopcu and S. Ulukus, "Age of information with soft updates," in *Allerton Conference*, October 2018.

[50] ——, "Minimizing age of information with soft updates," *Journal of Communications and Networks, special issue on Age of Information*, vol. 21, no. 3, pp. 233–243, July 2019.

[51] M. Shaked and J. G. Shanthikumar, *Stochastic Orders*. Springer, 2007.

[52] A. M. Bedewy, Y. Sun, and N. B. Shroff, "The age of information in multihop networks," *ACM/IEEE Trans. Netw.*, vol. 27, no. 3, pp. 1248 – 1257, Jun. 2019.

[53] Y. Sun, E. Uysal-Biyikoglu, and S. Kompella, "Age-optimal updates of multiple information flows," in *IEEE Conference on Computer Communications (INFOCOM) Workshops*, April 2018, pp. 136–141.

[54] A. Maatouk, Y. Sun, A. Ephremides, and M. Assaad, "Status updates with priorities: Lexicographic optimality," in *IEEE/IFIP WiOpt*, 2020.

[55] A. M. Bedewy, Y. Sun, and N. B. Shroff, "Optimizing data freshness, throughput, and delay in multi-server information-update systems," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, 2016, pp. 2569–2574.

[56] ——, "Age-optimal information updates in multihop networks," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, June 2017, pp. 576–580.

[57] J. P. Champati, H. Al-Zubaidy, and J. Gross, "Statistical guarantee optimization for age of information for the D/G/1 queue," in *IEEE Conference on Computer Communications (INFOCOM) Workshops*, April 2018, pp. 130–135.

[58] A. Soysal and S. Ulukus, "Age of information in G/G/1/1 systems," in *Asilomar Conference*, November 2019.

[59] ——, "Age of information in G/G/1/1 systems: Age expressions, bounds, special cases, and optimization," May 2019, available on arXiv: 1905.13743.

[60] B. Buyukates and S. Ulukus, "Age of information with Gilbert-Elliot servers and samplers," in *CISS*, March 2020.

[61] C. Kam, S. Kompella, G. D. Nguyen, J. Wieselthier, and A. Ephremides, "Age of information with a packet deadline," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, 2016, pp. 2564–2568.

[62] B. Wang, S. Feng, and J. Yang, "To skip or to switch? Minimizing age of information under link capacity constraint," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2018, pp. 1–5.

[63] K. Chen and L. Huang, "Age-of-information in the presence of error," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, 2016, pp. 2579–2584.

[64] L. Huang and L. P. Qian, "Age of information for transmissions over Markov channels," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2017.

[65] S. Kaul, R. Yates, and M. Gruteser, "Status updates through queues," in *Conf. on Information Sciences and Systems (CISS)*, Mar. 2012.

[66] R. Yates and S. Kaul, "Real-time status updating: Multiple sources," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jul. 2012.

[67] L. Huang and E. Modiano, "Optimizing age-of-information in a multi-class queueing system," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2015.

[68] E. Najm and E. Telatar, "Status updates in a multi-stream M/G/1/1 preemptive queue," in *IEEE Conference on Computer Communications (INFOCOM) Workshops*, April 2018, pp. 124–129.

[69] M. Moltafet, M. Leinonen, and M. Codreanu, "On the age of information in multi-source queueing models," *IEEE Transactions on Communications*, pp. 1–1, 2020, IEEE Early Access.

[70] E. Najm, R. Yates, and E. Soljanin, "Status updates through M/G/1/1 queues with HARQ," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2017, pp. 131–135.

[71] S. K. Kaul and R. D. Yates, "Timely updates by multiple sources: The M/M/1 queue revisited," in *54th Annual Conference on Information Sciences and Systems (CISS)*, 2020, pp. 1–6.

[72] E. Najm, R. Nasser, and E. Telatar, "Content based status updates," *IEEE Trans. Info. Theory*, vol. 66, no. 6, pp. 3846–3863, 2020.

[73] S. Kaul and R. Yates, "Age of information: Updates with priority," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2018, pp. 2644–2648.

[74] A. Maatouk, M. Assaad, and A. Ephremides, "Age of information with prioritized streams: When to buffer preempted packets?" in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, July 2019, pp. 325–329.

[75] C. Kam, S. Kompella, and A. Ephremides, "Effect of message transmission diversity on status age," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, June 2014, pp. 2411–2415.

[76] ——, "Age of information under random updates," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, 2013, pp. 66–70.

[77] C. Kam, S. Kompella, G. D. Nguyen, and A. Ephremides, "Effect of message transmission path diversity on status age," *IEEE Trans. Info. Theory*, vol. 62, no. 3, pp. 1360–1374, Mar. 2016.

[78] R. D. Yates, "Status updates through networks of parallel servers," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2018, pp. 2281–2285.

[79] V. Tripathi and S. Moharir, "Age of information in multi-source systems," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2017.

[80] B. Hajek and P. Seri, "Lex-optimal online multiclass scheduling with hard deadlines," *Mathematics of Operations Research*, vol. 30, no. 3, pp. 562–596, 2005.

[81] X. Wu, J. Yang, and J. Wu, "Optimal status update for age of information minimization with an energy harvesting source," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 193–204, March 2018.

[82] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Age-minimal transmission for energy harvesting sensors with finite batteries: Online policies," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 534–556, January 2020.

[83] S. Feng and J. Yang, "Age of information minimization for an energy harvesting source with updating erasures: Without and with feedback," available Online: arXiv:1808.05141.

[84] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Online timely status updates with erasures for energy harvesting sensors," in *Proc. Allerton*, October 2018.

[85] ——, "Using erasure feedback for online timely updating with an energy harvesting sensor," in *Proc. IEEE ISIT*, July 2019.

[86] S. Farazi, A. G. Klein, and D. R. Brown, "Average age of information for status update systems with an energy harvesting server," in *IEEE Conference on Computer Communications (INFOCOM) Workshops*, April 2018, pp. 112–117.

[87] S. Farazi, A. G. Klein, and D. R. B. III, "Age of information in energy harvesting status update systems: When to preempt in service?" in *Proc. IEEE ISIT*, June 2018.

[88] B. T. Bacinoglu, E. T. Ceran, and E. Uysal-Biyikoglu, "Age of information under energy replenishment constraints," in *Proc. ITA*, February 2015.

[89] W. Liu, X. Zhou, S. Durrani, H. Mehrpouyan, and S. D. Blostein, "Energy harvesting wireless sensor networks: Delay analysis considering energy costs of sensing and transmission," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4635–4650, July 2016.

[90] A. Arafa and S. Ulukus, "Age-minimal transmission in energy harvesting two-hop networks," in *Proc. IEEE Globecom*, December 2017.

[91] ——, "Age minimization in energy harvesting communications: Energy-controlled delays," in *Proc. Asilomar*, October 2017.

[92] B. T. Bacinoglu and E. Uysal-Biyikoglu, "Scheduling status updates to minimize age of information with an energy harvesting sensor," in *Proc. IEEE ISIT*, June 2017.

[93] R. Yates, "Lazy is timely: Status updates by an energy harvesting source," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, June 2015, pp. 3008–3012.

[94] A. Arafa and S. Ulukus, "Timely updates in energy harvesting two-hop networks: Offline and online policies," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4017–4030, August 2019.

[95] S. Feng and J. Yang, "Optimal status updating for an energy harvesting sensor with a noisy channel," in *Proc. IEEE Infocom*, April 2018.

[96] ——, "Minimizing age of information for an energy harvesting source with updating failures," in *Proc. IEEE ISIT*, June 2018.

[97] A. Baknina and S. Ulukus, "Coded status updates in an energy harvesting erasure channel," in *Proc. CISS*, March 2018.

[98] X. Zheng, S. Zhou, Z. Jiang, and Z. Niu, "Closed-form analysis of non-linear age of information in status updates with an energy harvesting transmitter," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4129–4142, Aug 2019.

[99] A. Baknina, O. Ozel, J. Yang, S. Ulukus, and A. Yener, "Sending information through status updates," in *Proc. IEEE ISIT*, June 2018.

[100] A. Arafa, J. Yang, and S. Ulukus, "Age-minimal online policies for energy harvesting sensors with random battery recharges," in *Proc. IEEE ICC*, May 2018.

[101] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Age-minimal online policies for energy harvesting sensors with incremental battery recharges," in *Proc. ITA*, February 2018.

[102] B. T. Bacinoglu, Y. Sun, E. Uysal-Biyikoglu, and V. Mutlu, "Achieving the age-energy tradeoff with a finite-battery energy harvesting source," in *Proc. IEEE ISIT*, June 2018.

[103] ——, "Optimal status updating with a finite-battery energy harvesting source," *J. Commun. Netw.*, vol. 21, no. 3, pp. 280–294, June 2019.

[104] Z. Chen, N. Pappas, E. Bjornson, and E. G. Larsson, "Age of information in a multiple access channel with heterogeneous traffic and an energy harvesting node," in *Proc. IEEE Infocom*, May 2019.

[105] G. Stamatakis, N. Pappas, and A. Traganitis, "Control of status updates for energy harvesting devices that monitor processes with alarms," available Online: arXiv:1907.03826.

[106] P. Rafiee and O. Ozel, "Active status update packet drop control in an energy harvesting node," available Online: arXiv:1911.01407.

[107] O. Ozel, "Timely status updating through intermittent sensing and transmission," available Online: arXiv:2001.01122.

[108] Y. Dong, P. Fan, and K. B. Letaief, "Energy harvesting powered sensing in IoT: Timeliness versus distortion," available Online: arXiv:1912.12427.

[109] I. Krikidis, "Average age of information in wireless powered sensor networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 628–631, April 2019.

[110] C. Tunc and S. Panwar, "Optimal transmission policies for energy harvesting age of information systems with battery recovery," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 2012–2016.

[111] S. Leng and A. Yener, "Age of information minimization for an energy harvesting cognitive radio," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 2, pp. 427–439, 2019.

[112] Y. Gu, H. Chen, Y. Zhou, Y. Li, and B. Vucetic, "Timely status update in internet of things monitoring systems: An age-energy tradeoff," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5324–5335, 2019.

[113] E. T. Ceran, D. Gündüz, and A. György, "Average age of information with hybrid ARQ under a resource constraint," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018.

[114] E. T. Ceran, D. Gündüz, and A. György, "Average age of information with hybrid ARQ under a resource constraint," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1900–1913, March 2019.

[115] M. A. Abd-Elmagid and H. S. Dhillon, "Average peak age-of-information minimization in uav-assisted iot networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 2003–2008, 2019.

[116] C. Hu and Y. Dong, "Age of information of two-way data exchanging systems with power-splitting," *Journal of Communications and Networks*, vol. 21, no. 3, pp. 295–306, 2019.

[117] E. Gindullina, L. Badia, and D. Gündüz, "Age-of-information with information source diversity in an energy harvesting system," *arXiv preprint arXiv:2004.11135*, 2020.

[118] C.-H. Tsai and C.-C. Wang, "Age-of-information revisited: Two-way delay and distribution-oblivious online algorithm," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, June 2020.

[119] A. M. Bedewy, Y. Sun, S. Kompella, and N. B. Shroff, "Age-optimal sampling and transmission scheduling in multi-source systems," in *ACM MobiHoc*, 2019, p. 121–130.

[120] ——, "Optimal sampling and scheduling for timely status updates in multi-source networks," 2020, https://arxiv.org/abs/2001.09863.

[121] V. Jog, R. J. La, and N. C. Martins, "Channels, learning, queueing and remote estimation systems with a utilization-dependent component," 2019, coRR, abs/1905.04362.

[122] C.-H. Tsai and C.-C. Wang, "Unifying AoI minimization and remote estimation — optimal sensor/controller coordination with random two-way delay," in *Proc. IEEE INFOCOM*, 2020.

[123] B. Hajek, K. Mitzel, and S. Yang, "Paging and registration in cellular networks: Jointly optimal policies and an iterative algorithm," *IEEE Trans. Inf. Theory*, vol. 54, no. 2, pp. 608–622, Feb 2008.

[124] G. M. Lipsa and N. C. Martins, "Remote state estimation with communication costs for first-order LTI systems," *IEEE Trans. Auto. Control*, vol. 56, no. 9, pp. 2013–2025, Sept. 2011.

[125] A. Nayyar, T. Başar, D. Teneketzis, and V. V. Veeravalli, "Optimal strategies for communication and remote estimation with an energy harvesting sensor," *IEEE Trans. Auto. Control*, vol. 58, no. 9, pp. 2246–2260, Sept. 2013.

[126] X. Gao, E. Akyol, and T. Başar, "Optimal communication scheduling and remote estimation over an additive noise channel," *Automatica*, vol. 88, pp. 57 – 69, 2018.

[127] J. Chakravorty and A. Mahajan, "Remote estimation over a packet-drop channel with Markovian state," *IEEE Trans. Auto. Control*, vol. 65, no. 5, pp. 2016–2031, 2020.

[128] N. Guo and V. Kostina, "Optimal causal rate-constrained sampling for a class of continuous Markov processes," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, 2020.

[129] O. Ayan, M. Vilgelm, M. Klügel, S. Hirche, and W. Kellerer, "Age-of-information vs. value-of-information scheduling for cellular networked control systems," in *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, 2019, p. 109–117.

[130] O. Ayan, M. Vilgelm, and W. Kellerer, "Optimal scheduling for discounted age penalty minimization in multi-loop networked control," in *IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2020, pp. 1–7.

[131] I. Kadota, A. Sinha, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Scheduling policies for minimizing age of information in broadcast wireless networks," *IEEE/ACM Trans. Netw.*, 2018.

[132] R. Talak, S. Karaman, and E. Modiano, "Optimizing information freshness in wireless networks under general interference constraints," in *Proc. ACM MobiHoc*, Jun. 2018.

[133] P. Parag, A. Taghavi, and J. Chamberland, "On real-time status updates over symbol erasure channels," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, March 2017.

[134] R. Yates, E. Najm, E. Soljanin, and J. Zhong, "Timely updates over an erasure channel," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2017, pp. 316–320.

[135] H. Sac, T. Bacinoglu, E. Uysal-Biyikoglu, and G. Durisi, "Age-optimal channel coding blocklength for an M/G/1 queue with HARQ," in *19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2018, pp. 486–490.

[136] Q. He, D. Yuan, and A. Ephremides, "Optimizing freshness of information: On minimum age link scheduling in wireless systems," in *Proc. IEEE/IFIP WiOpt*, May 2016, pp. 1–8.

[137] ——, "Optimal link scheduling for age minimization in wireless systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5381–5394, July 2018.

[138] R. Talak, S. Karaman, and E. Modiano, "Speed limits in autonomous vehicular networks due to communication constraints," in *Proc. IEEE CDC*, Dec 2016.

[139] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Minimizing the age of information in broadcast wireless networks," in *54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2016, pp. 844–851.

[140] Y.-P. Hsu, E. Modiano, and L. Duan, "Age of information: Design and analysis of optimal scheduling algorithms," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2017, pp. 1–5.

[141] Z. Jiang, B. Krishnamachari, X. Zheng, S. Zhou, and Z. Niu, "Decentralized status update for age-of-information optimization in wireless multiaccess channels," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, June 2018, pp. 2276–2280.

[142] Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Can decentralized status update achieve universally near-optimal age-of-information in wireless multiaccess channels?" in *2018 30th International Teletraffic Congress (ITC 30)*, vol. 01, 2018, pp. 144–152.

[143] J. Sun, Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Closed-form whittle's index-enabled random access for timely status update," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1538–1551, 2020.

[144] B. Zhou and W. Saad, "Optimal sampling and updating for minimizing age of information in the internet of things," in *Proc. IEEE GLOBECOM*, Dec 2018, pp. 1–6.

[145] A. Maatouk, M. Assaad, and A. Ephremides, "On the age of information in a CSMA environment," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 818–831, 2020.

[146] S. K. Kaul and R. Yates, "Status updates over unreliable multiaccess channels," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2017, pp. 331–335.

[147] S. Leng and A. Yener, "Age of information minimization for wireless ad hoc networks: A deep reinforcement learning approach," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[148] R. D. Yates and S. K. Kaul, "Age of information in uncoordinated unslotted updating," 2020. [Online]. Available: https://arxiv.org/abs/2002.02026

[149] X. Chen, K. Gatsis, H. Hassani, and S. S. Bidokhti, "Age of Information in Random Access Channels," 2019. [Online]. Available: http://arxiv.org/abs/1912.01473

[150] Z. Jiang, B. Krishnamachari, X. Zheng, S. Zhou, and Z. Niu, "Timely status update in wireless uplinks: Analytical solutions with asymptotic optimality," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3885–3898, 2019.

[151] H. Chen, Y. Gu, and S.-C. Liew, "Age-of-Information Dependent Random Access for Massive IoT Networks," jan 2020. [Online]. Available: http://arxiv.org/abs/2001.04780

[152] A. M. Bedewy, Y. Sun, R. Singh, and N. B. Shroff, "Optimizing information freshness using low-power status updates via sleep-wake scheduling," in *Proc. ACM MobiHoc*, 2020.

[153] N. Lu, B. Ji, and B. Li, "Age-based scheduling: Improving data freshness for wireless real-time traffic," in *Proc. ACM MobiHoc*, 2018.

[154] I. Kadota, A. Sinha, and E. Modiano, "Optimizing age of information in wireless networks with throughput constraints," in *Proc. IEEE INFOCOM*, 2018.

[155] I. Kadota and E. Modiano, "Minimizing the age of information in wireless networks with stochastic arrivals," in *Proc. ACM MobiHoc*, 2019.

[156] C. Joo and A. Eryilmaz, "Wireless scheduling for information freshness and synchrony: Drift-based design and heavy-traffic analysis," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2556–2568, Dec 2018.

[157] R. Talak, S. Karaman, and E. Modiano, "Minimizing age-of-information in multi-hop wireless networks," in *55th Annual Allerton Conference on Communication, Control, and Computing*, Oct 2017, pp. 486–493.

[158] S. Farazi, A. G. Klein, J. A. McNeill, and D. R. Brown, "On the age of information in multi-source multi-hop wireless status update networks," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2018, pp. 1–5.

[159] R. Talak, S. Karaman, and E. Modiano, "Distributed scheduling algorithms for optimizing information freshness in wireless networks," in *Proc. SPAWC (arXiv:1803.06469)*, Jun. 2018.

[160] ——, "Optimizing age of information in wireless networks with perfect channel state information," in *Proc. IEEE/IFIP WiOpt*, May 2018.

[161] R. Talak, I. Kadota, S. Karaman, and E. Modiano, "Scheduling policies for age minimization in wireless networks with unknown channel state," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2018.

[162] B. Buyukates, A. Soysal, and S. Ulukus, "Age of information scaling in large networks," in *IEEE ICC*, May 2019.

[163] ——, "Age of information scaling in large networks with hierarchical cooperation," in *IEEE Globecom*, December 2019.

[164] ——, "Scaling laws for age of information in wireless networks," *IEEE Transactions on Wireless Communications*, December 2019. Early Access.

[165] S. Farazi, A. G. Klein, and D. R. Brown III, "Fundamental bounds on the age of information in multi-hop global status update networks," *Journal of Communications and Networks*, vol. 21, no. 3, pp. 268–279, 2019.

[166] P. Whittle, "Restless bandits: Activity allocation in a changing world," *Journal of Applied Probability*, vol. 25, pp. 287–298, 1988.

[167] R. Singh, X. Guo, and P. Kumar, "Index policies for optimal mean-variance trade-off of inter-delivery times in real-time sensor networks," in *Proc. IEEE INFOCOM*, Jan. 2015, pp. 505–512.

[168] V. Raghunathan, V. Borkar, M. Cao, and P. R. Kumar, "Index policies for real-time multicast scheduling for wireless broadcast systems," in *Proc. IEEE INFOCOM*, Apr. 2008.

[169] P. Mansourifard, T. Javidi, and B. Krishnamachari, "Optimality of myopic policy for a class of monotone affine restless multi-armed bandits," in *Proc. IEEE CDC*, Dec. 2012, pp. 877–882.

[170] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, pp. 5547–5567, Nov. 2010.

[171] R. R. Weber and G. Weiss, "On an index policy for restless bandits," *Journal of Applied Probability*, vol. 27, no. 3, pp. 637–648, 1990.

[172] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed Bandit Allocation Indices*, 2nd ed. Wiley, Mar. 2011.

[173] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univerisity Press, 2004.

[174] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, 4th ed. Springer Publishing Company, Incorporated, 2007.

[175] D. Garber and E. Hazan, "A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization," *SIAM J. on Opt.*, vol. 26, no. 3, pp. 1493–1528, 2016.

[176] B. Hajek and G. Sasaki, "Link scheduling in polynomial time," *IEEE Trans. Inf. Theory*, vol. 34, no. 5, pp. 910–917, Sep. 1988.

[177] S. Farazi, A. G. Klein, and D. R. Brown, "On the average staleness of global channel state information in wireless networks with random transmit node selection," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 3621–3625.

[178] ——, "Bounds on the age of information for global channel state dissemination in fully-connected networks," in *Int'l Conference on Computer Communication and Networks (ICCCN)*, July 2017.

[179] A. G. Klein, S. Farazi, W. He, and D. R. Brown, "Staleness bounds and efficient protocols for dissemination of global channel state information," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5732–5746, Sept 2017.

[180] J. Selen, H. N. Yu, L. L. Andrew, and H. L. Vu, "The age of information in gossip networks," in *Analytical and Stochastic Modeling Techniques and Applications*. Berlin, Heidelberg: Springer, 2013, pp. 364–379.

[181] S. Farazi, D. R. Brown III, and A. G. Klein, "On global channel state estimation and dissemination in ring networks," in *Proc. Asilomar Conf. on Signals, Systems, and Computers*, Nov. 2016, pp. 1122–1127.

[182] R. D. Yates, "Age of information in a network of preemptive servers," in *Proc. IEEE Intl. Conf. on Computer Comm. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 118–123.

[183] S. Banik, S. K. Kaul, and P. B. Sujit, "Minimizing Age in Gateway Based Update Systems," in *IEEE International Symposium on Information Theory - Proceedings*, vol. 2019-July. Institute of Electrical and Electronics Engineers Inc., jul 2019, pp. 1032–1036. [Online]. Available: http://arxiv.org/abs/1903.07963

[184] S. Farazi, A. G. Klein, and D. R. Brown, "Fundamental bounds on the age of information in general multi-hop interference networks," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2019, pp. 96–101.

[185] ——, "Age of information with unreliable transmissions in multi-source multi-hop status update systems," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 2017–2021.

[186] ——, "Average age of information in multi-source self-preemptive status update systems with packet delivery errors," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 396–400.

[187] T. Shreedhar, S. K. Kaul, and R. D. Yates, "An age control transport protocol for delivering fresh updates in the internet-of-things," *20th IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks, WoWMoM 2019*, 2019.

[188] O. Ayan, H. M. Gürsu, A. Papa, and W. Kellerer, "Probability analysis of age of information in multi-hop networks," *IEEE Networking Letters*, vol. 2, no. 2, pp. 76–80, 2020.

[189] M. Costa, S. Valentin, and A. Ephremides, "On the age of channel information for a finite-state Markov model," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 4101–4106.

[190] Y. Sang, B. Li, and B. Ji, "The power of waiting for more than one response in minimizing the age-of-information," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2017.

[191] J. Zhong, R. Yates, and E. Soljanin, "Minimizing content staleness in dynamo-style replicated storage systems," in *Infocom Workshop on Age of Information*, Apr. 2018, arXiv preprint arXiv:1804.00742.

[192] ——, "Status updates through multicast networks," in *Proc. Allerton Conf. on Commun., Control and Computing*, Oct. 2017, pp. 463–469.

[193] ——, "Multicast with prioritized delivery: How fresh is your data?" in *Signal Processing Advance for Wireless Communications (SPAWC)*, Jun. 2018, pp. 476–480.

[194] B. Buyukates, A. Soysal, and S. Ulukus, "Age of information in two-hop multicast networks," in *Asilomar Conference*, October 2018.

[195] ——, "Age of information in multicast networks with multiple update streams," in *Asilomar Conference*, November 2019.

[196] ——, "Age of information in multihop multicast networks," *Journal of Communications and Networks*, vol. 21, no. 3, pp. 256–267, July 2019.

[197] J. Zhong and R. D. Yates, "Timeliness in lossless block coding," in *2016 Data Compression Conference (DCC)*, March 2016, pp. 339–348.

[198] J. Zhong, R. Yates, and E. Soljanin, "Backlog-adaptive compression: Age of information," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2017, pp. 566–570.

[199] P. Mayekar, P. Parag, and H. Tyagi, "Optimal lossless source codes for timely updates," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2018, pp. 1246–1250.

[200] M. Bastopcu, B. Buyukates, and S. Ulukus, "Optimal selective encoding for timely updates," in *CISS*, March 2020.

[201] B. Buyukates, M. Bastopcu, and S. Ulukus, "Optimal selective encoding for timely updates with empty symbol," in *IEEE ISIT*, June 2020.

[202] M. Bastopcu, B. Buyukates, and S. Ulukus, "Selective encoding policies for maximizing information freshness," April 2020, available on arXiv:2004.06091.

[203] M. Bastopcu and S. Ulukus, "Partial updates: Losing information for freshness," in *IEEE ISIT*, June 2020.

[204] S. Bhambay, S. Poojary, and P. Parag, "Differential encoding for real-time status updates," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, March 2017.

[205] Q. He, G. Dan, and V. Fodor, "Minimizing age of correlated information for wireless camera networks," in *IEEE Conference on Computer Communications (INFOCOM) Workshops*, April 2018, pp. 547–552.

[206] J. Hribar, M. Costa, N. Kaminski, and L. A. DaSilva, "Updating strategies in the internet of things by taking advantage of correlated sources," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2017.

[207] R. Yates, M. Tavan, Y. Hu, and D. Raychaudhuri, "Timely cloud gaming," in *Proc. INFOCOM*, May 2017, pp. 1–9.

[208] S. F. Lindström, M. Wetterberg, and N. Carlsson, "Cloud gaming: A QoE study of fast-paced single-player and multiplayer gaming," in *Proc. IEEE/ACM International Conference on Utility and Cloud Computing (UCC)*, Dec. 2020.

[209] M. Bastopcu and S. Ulukus, "Age of information for updates with distortion," in *IEEE ITW*, August 2019.

[210] ——, "Age of information for updates with distortion: Constant and age-dependent distortion constraints," December 2019, available on arXiv:1912.13493.

[211] B. Buyukates and S. Ulukus, "Timely distributed computation with stragglers," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5273–5282, September 2020.

[212] G. D. Nguyen, S. Kompella, C. Kam, J. E. Wieselthier, and A. Ephremides, "Impact of hostile interference on information freshness: A game approach," in *Int'l Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2017.

[213] Y. Xiao and Y. Sun, "A dynamic jamming game for real-time status updates," in *IEEE Conference on Computer Communications (INFOCOM) Workshops*, April 2018, pp. 354–360.

[214] A. Garnaev, W. Zhang, J. Zhong, and R. D. Yates, "Maintaining information freshness under jamming," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2019, pp. 90–95.

[215] X. Gac, E. Akyol, and T. Başar, "On communication scheduling and remote estimation in the presence of an adversary as a nonzero-sum game," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 2710–2715.

[216] G. D. Nguyen, S. Kompella, C. Kam, J. E. Wieselthier, and A. Ephremides, "Information freshness over an interference channel: A game theoretic view," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 908–916.

[217] S. Gopal and S. K. Kaul, "A game theoretic approach to DSRC and WiFi coexistence," in *IEEE Conference on Computer Communications (INFOCOM) Workshops*, April 2018, pp. 565–570.

[218] S. Gopal, S. K. Kaul, and R. Chaturvedi, "Coexistence of age and throughput optimizing networks: A game theoretic approach," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2019, pp. 1–6.

[219] S. Gopal, S. K. Kaul, R. Chaturvedi, and S. Roy, "Coexistence of age and throughput optimizing networks: A spectrum sharing game," *arXiv preprint arXiv:1909.02863*, 2019.

[220] ——, "A Non-Cooperative Multiple Access Game for Timely Updates," in *INFOCOM 2020 - IEEE Conference on Computer Communications Workshopss (INFOCOM WKSHPS)*, 2020.

[221] K. Saurav and R. Vaze, "Game of Ages," in *INFOCOM 2020 - IEEE Conference on Computer Communications Workshopss (INFOCOM WKSHPS)*, 2020.

[222] H. Zheng, K. Xiong, P. Fan, Z. Zhong, and K. B. Letaief, "Age-based utility maximization for wireless powered networks: A stackelberg game approach," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[223] Z. Ning, P. Dong, X. Wang, X. Hu, L. Guo, B. Hu, Y. Guo, T. Qiu, and R. Kwok, "Mobile edge computing enabled 5g health monitoring for internet of medical things: A decentralized game theoretic approach," *IEEE J. Sel. Areas Commun.*, pp. 1–16, 2020.

[224] M. Bastopcu and S. Ulukus, "Who should Google Scholar update more often?" in *IEEE Infocom*, July 2020.

[225] E. T. Ceran, D. Gündüz, and A. György, "A reinforcement learning approach to age of information in multi-user networks," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2018, pp. 1967–1971.

[226] B. Yin, S. Zhang, and Y. Cheng, "Application-oriented scheduling for optimizing the age of correlated information: A deep reinforcement learning based approach," *IEEE Internet of Things Journal*, 2020.

[227] A. Elgabli, H. Khan, M. Krouka, and M. Bennis, "Reinforcement learning based scheduling algorithm for optimizing age of information in ultra reliable low latency networks," in *2019 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2019, pp. 1–6.

[228] H. B. Beytur and E. Uysal, "Age minimization of multiple flows using reinforcement learning," in *2019 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2019, pp. 339–343.

[229] S. F. Abedin, M. Munir, N. H. Tran, Z. Han, C. S. Hong *et al.*, "Data freshness and energy-efficient UAV navigation optimization: A deep reinforcement learning approach," *arXiv preprint arXiv:2003.04816*, 2020.

[230] M. Yi, X. Wang, J. Liu, Y. Zhang, and B. Bai, "Deep reinforcement learning for fresh data collection in UAV-assisted iot networks," *arXiv preprint arXiv:2003.00391*, 2020.

[231] M. A. Abd-Elmagid, A. Ferdowsi, H. S. Dhillon, and W. Saad, "Deep reinforcement learning for minimizing age-of-information in UAV-assisted networks," *arXiv preprint arXiv:1905.02993*, 2019.

[232] E. Ozfatura, B. Buyukates, D. Gunduz, and S. Ulukus, "Age-based coded computation for bias reduction in distributed learning," in *IEEE Globecom*, December 2020.

[233] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[234] R. S. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artificial intelligence*, vol. 112, no. 1-2, pp. 181–211, 1999.

[235] W. Gao, G. Cao, M. Srivatsa, and A. Iyengar, "Distributed maintenance of cache freshness in opportunistic mobile networks," *Proceedings - International Conference on Distributed Computing Systems*, pp. 132–141, 2012.

[236] R. Yates, P. Ciblat, M. Wigger, and A. Yener, "Age-optimal constrained cache updating," in *Proc. IEEE Int'l. Symp. Info. Theory (ISIT)*, Jun. 2017, pp. 141–145.

[237] C. Kam, S. Kompella, G. D. Nguyen, J. E. Wieselthier, and A. Ephremides, "Information freshness and popularity in mobile caching," in *IEEE International Symposium on Information Theory - Proceedings*, 2017.

[238] H. Tang, P. Ciblat, J. Wang, M. Wigger, and R. Yates, "Age of Information Aware Cache Updating with File- and Age-Dependent Update Durations," pp. 1–6, 2019. [Online]. Available: http://arxiv.org/abs/1909.05930

[239] S. Zhang, J. Li, H. Luo, J. Gao, L. Zhao, and X. S. Shen, "Towards Fresh and Low-Latency Content Delivery in Vehicular Networks: An Edge Caching Aspect," *2018 10th International Conference on Wireless Communications and Signal Processing, WCSP 2018*, pp. 1–6, 2018.

[240] L. Yang, Y. Zhong, F.-C. Zheng, and S. Jin, "Edge Caching with Real-Time Guarantees," 2019. [Online]. Available: http://arxiv.org/abs/1912.11847

[241] M. Bastopcu and S. Ulukus, "Information freshness in cache updating systems," April 2020, available on arXiv:2004.09475.

[242] ——, "Maximizing information freshness in caching systems with limited cache storage capacity," in *Asilomar Conference*, November 2020.

[243] S. Zhang, L. Wang, H. Luo, X. Ma, and S. Zhou, "AoI-Delay Tradeoff in Mobile Edge Caching with Freshness-Aware Content Refreshing," vol. 100191, pp. 1–28, 2020. [Online]. Available: http://arxiv.org/abs/2002.05868

[244] A. Aral, M. Erol-Kantarci, and I. Brandić, "Staleness control for edge data analytics," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 4, no. 2, Jun. 2020.

[245] C. Sönmez, S. Baghaee, A. Ergişi, and E. Uysal-Biyikoglu, "Age-of-Information in Practice: Status Age Measured Over TCP/IP Connections Through WiFi, Ethernet and LTE," *2018 IEEE International Black Sea Conference on Communications and Networking, BlackSeaCom 2018*, pp. 1–5, 2018.

[246] H. B. Beytur, S. Baghaee, and E. Uysal, "Measuring age of information on real-life connections," *27th Signal Processing and Communications Applications Conference, SIU 2019*, 2019.

[247] ——, "Towards AoI-aware Smart IoT Systems," *2020 International Conference on Computing, Networking and Communications, ICNC 2020*, pp. 353–357, 2020.

[248] T. Shreedhar, S. K. Kaul, and R. D. Yates, "Poster: ACP: Age Control Protocol for minimizing age of information over the internet," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. ACM, 2018, p. 699–701.

[249] ——, "ACP: An end-to-end transport protocol for delivering fresh updates in the internet-of-things," *arXiv preprint arXiv:1811.03353*, 2018.