On the Minimum Average Age of Information in IRSA for Grant-free mMTC

Subham Saha, Vineeth Bala Sukumaran and Chandra R. Murthy

***Abstract*—We consider the optimal design of the frame-based irregular repetition slotted ALOHA (IRSA) protocol for minimizing the average age of information (AAoI) in grant-free massive machine-type communications (mMTC). To this end, we first characterize the AAoI as a function of the number of UEs, the frame duration, and the repetition distribution of IRSA. We present this characterization for IRSA schemes with packet recovery at the end of frame and packet generation either at the beginning of the frame or just in time before first transmission in a frame. We also propose and characterize the AAoI of a novel early packet recovery method which further reduces the average age of information. In all cases, the analysis reveals that, as a function of normalized channel traffic (defined as the ratio of number of UEs to frame duration), the AAoI first decreases linearly due to more frequent updates received from the UEs, and increases sharply beyond a critical point due to packet recovery failures caused by collisions. We then consider the problem of minimizing AAoI by optimizing over the normalized channel traffic and repetition distribution for all the proposed sampling and recovery schemes. The optimization problem is challenging since the objective function is semi-analytical and can only be completely characterized using simulations. In an asymptotic regime where the number of UEs as well as the frame size is large, we characterize the AAoI using upper and lower bounds. We also obtain a locally optimal normalized channel traffic and repetition distribution using differential evolution. Based on the insights obtained from the asymptotic analysis, we also propose a pragmatic approach to obtain a normalized channel traffic and repetition distribution for AAoI reduction in the non-asymptotic case. Finally, we empirically show that our AAoI minimizing schemes outperform conventional throughput optimal schemes.**

I. Introduction

The recent growth in Internet of Things (IoT), fueled by a variety of sensing and monitoring applications such as smart cities and smart industries, has spurred research interest in massive machine-type communication (mMTC) systems. These systems have a massive number of miniature, resource constrained, and low cost user elements (UEs) [1] deployed for sensing and monitoring purposes. These UEs sporadically communicate to a central base station (BS), for which, low-complexity grant-free random access schemes such as irregular repetition slotted ALOHA (IRSA) are used [2], [3].

Typically, the UEs sample and measure a physical process, e.g., pollution levels, and report the measurements to the BS. The BS uses the measurements to remotely estimate the physical process. The performance of such sampling, measurement, and remote estimation systems critically depends on the freshness of the reported measurements [4]. The age of information (AoI), introduced in [5], is a relevant metric to measure freshness of measurements. The AoI is a function of time, computed at the BS, and measures the age of the last successfully delivered measurement. The average age of information (AAoI) is the time-average of the AoI. In this paper, we consider the design of grant-free IRSA schemes for minimizing the AAoI for mMTC scenarios.

A key challenge in mMTC systems is to provide reliable connectivity to large number of UEs. While grant-based access schemes are reliable, the control overhead cost is prohibitively high when the number of UEs is large. In contrast, grant-free random access (RA) schemes remove the requirement of connection establishment, thus reducing the number of interactions between UEs and the BS. However, as the channel access is random, collisions can occur, leading to a loss in reliability as the number of UEs is increased [2]. In the past decade, several grant-free RA schemes have been introduced and their performances analyzed [3], [6]–[12]. It was shown in [13] that improved throughput is possible in grant-free repetition based RA schemes such as IRSA using a successive interference cancellation (SIC) decoder. An SIC decoder allows a colliding packet can be recovered if the interference caused by other packets can be cancelled. The throughput and packet decoding probability of frame based IRSA and its variants are analyzed in [2], [3], [9]–[11], typically by considering the asymptotic regime where the number of UEs and slots in each frame goes to infinity. An important design parameter of IRSA is the repetition distribution, which can be optimized to maximize the achievable throughput. Narayanan et al. [3] showed that a throughput arbitrarily close to one packet per slot, which the maximum possible throughput under a packet collision model, is in fact asymptotically achievable by choosing the repetition distribution as a truncated soliton distribution. In [14], [15] a close relation of the SIC process to the decoding of low-density parity-check code over binary erasure channels was exploited to obtain the throughput and packet decoding success probability for IRSA in the non-asymptotic regime.

We note that optimization of delay or throughput is not equivalent to the optimization of AAoI (e.g., [16]). The AAoI characterization and optimization problem has been considered, e.g., in [4], [17]–[20], but these studies do not consider IRSA as the underlying communication scheme.

The work presented in this paper is closest to [21], where an analysis of AAoI for frame-based IRSA was carried out for a model where UEs are active or inactive according to

Subham Saha and Vineeth Bala Sukumaran are with the Dept. of Avionics, Indian Institute of Space Science and Technology, Thiruvananthapuram 695547, India. (e-mails: subhamsaha0216@gmail.com, vineethbs@iist.ac.in). Chandra R. Murthy is with the Dept. of ECE, Indian Institute of Science, Bangalore 560012, India (e-mail: cmurthy@iisc.ac.in).

The work of C. R. Murthy was supported by a young faculty research grant from the Ministry of Electronics & Technology, Govt. of India.

a stochastic process. The AAoI was characterized only for an IRSA scheme in which a UE samples its packet at the beginning of a frame and the SIC process is applied at the end. In contrast, we characterize the AAoI for IRSA schemes that samples packets just-in-time and or recovers packets early in a frame. We also consider the minimization of AAoI for frame-based IRSA by optimizing over various parameters of IRSA, including the repetition scheme, which is not available in prior work such as [21].

Contributions:

1) We first consider the AAoI for the basic IRSA scheme in which packets are generated at the start of a frame and packet recovery is performed at the end of the frame. We characterize the AAoI of a UE as a function of system parameters such as the frame length and IRSA repetition scheme. This is presented in Section IV.
2) We propose the following modifications to the basic IRSA scheme. We consider an IRSA scheme in which packets are generated at a UE just-in-time before the first transmission of that packet. We characterize the AAoI in this case and show that the AAoI reduces from the basic scheme by an amount proportional to the frame duration. The proportionality factor depends on the repetition distribution and is larger for distributions with a smaller number of maximum repetitions. We also consider an IRSA scheme with early-packet recovery. The SIC process is applied in every slot in order to recover UE packets early and thus reduce AAoI even further. A complexity comparison of the two recovery methods (i.e., end-of-frame and early recovery methods) shows that the complexity associated with the early recovery method is identical to the end-of-frame recovery method. Furthermore, in the early recovery method, computations are distributed throughout the frame, and it therefore incurs lower packet decoding delay compared to end-of-frame recovery. We note that our characterization of the AAoI for all these schemes is semi-analytical since the derived AAoI expressions contain terms which are obtained from numerical or simulation based methods. In particular, the packet success probability can be characterized using either a simulation or an iterative density evolution method [9]. The density evolution method uses an asymptotic assumption which does not always match with the packet success probability in the non-asymptotic regime. These modifications to the basic IRSA protocol and the insights on AAoI minimization revealed by our analysis are not available in prior work (e.g., [21]). We discuss this in Section IV.
3) We consider the problem of minimizing the AAoI under the IRSA protocol, and for each sampling and packet recovery method, by finding the optimal normalized channel traffic (ratio of number of UEs to frame duration) and repetition distribution. The limitations in the AAoI expressions alluded to above make this optimization problem challenging. However, in the asymptotic regime of large number of UEs, the density evolution based characterization of the packet success probability can be used to obtain a tractable AAoI optimization problem. We characterize the optimal AAoI (an appropriately normalized AAoI) using upper and lower bounds. We also obtain locally optimal normalized channel traffic and repetition distributions using differential evolution, a numerical search method. We compare the asymptotic scaling of the AAoI as a function of number of users with that of an equivalent fully centralized time division multiple access (TDMA) scheme. We show that, for the basic IRSA scheme, the AAoI scaling is *the same as* that of the equivalent centralized TDMA scheme. This is presented in Section V.
4) In the non-asymptotic regime, where the AAoI optimization problem is even more challenging, we observe that the use of the normalized channel traffic and repetition distribution from the asymptotic solution is sub-optimal. We then propose a pragmatic approach in Section V which modifies the asymptotic solution to obtain normalized channel traffic and repetition distributions which are empirically shown to reduce the non-asymptotic AAoI. The AAoI performance obtained is as close as 12% and 13% of the asymptotic values, for the basic IRSA scheme and the scheme with just-in-time sampling, respectively. The use of early recovery along with just-in-time sampling is shown to further reduce the AAoI. For example, our proposed scheme with early recovery achieves an AAoI which is 54% less than that of slotted ALOHA and 53% less than CRDSA [9].

Notation: We denote the set of all non-negative integers by $\mathbb{Z}_+$ and the set of all positive integers by $\mathbb{Z}_{++}$. The expectation of a random variable X is denoted as $\mathbb{E}[X]$. We summarize the key notations in Table I.

II. Irregular repetition slotted ALOHA

In this section, we present a brief overview of IRSA and describe related notation. Our discussion is based on [9]. We consider a single hop wireless network scenario with M UEs transmitting data packets to a BS over non-fading, noiseless links. Time is divided into slots, and the transmissions are organized into frames, where each frame consists of N slots. All UEs and the BS are slot and frame synchronized with each other. A UE has a single data packet to be transmitted to the BS in a given frame. In the frame, the UE repeats the transmission of the data packet in multiple slots. A UE first selects the number of repetitions $D \in \{2, 3, \ldots, N\}$ by sampling D from a probability mass function $f_D[d], d \in \{2, 3, \ldots, N\}$. The D repetition slots in which the UE transmits the data packet are then chosen uniformly without replacement from the slot indices $\{1, 2, \ldots, N\}$. The transmitted data packet contains a header with the list of D repetition slots. Every UE transmits data packets as described above using independent samples of D as well as repetition slots. IRSA is thus a fully distributed transmission protocol.

In each slot, the BS receives a signal which is the superposition of all data packets transmitted in that slot by the UEs. In the basic IRSA scheme, the BS stores the received signal in every slot, and performs the SIC procedure at the

TABLE I: Notation

M	Number of UEs in a frame of IRSA.
N	Number of slots in a frame of IRSA.
G	Normalized channel traffic (M/N).
G_τ	Maximum value of G with vanishing probability of failure in the asymptotic regime.
P_s	Probability of successful decoding of a UE at the end of a frame.
P_s^{de}	Estimate of P_s using density evolution, in the asymptotic regime.
P_s^{sim}	Estimate of P_s using simulation with finite M and N.
$T_u[i]$ or T	Slot at which UE u samples and generates a packet in frame i.
$R_u[i]$ or R	Slot at which the BS recovers the data packet from UE u in frame i.
U	$N - R$.
n	The number of SIC iterations in every slot other than the last slot.
n_e	The number of SIC iterations in the last slot of a frame.
$\overline{A_u}$ or $\overline{A}$	Average age of information (AAoI) for UE u.
$\overline{A}_{norm}$	Normalized AAoI ($\overline{A}/M$).
A^*	Minimum value of the AAoI.
A^*_{norm}	Minimum value of the normalized AAoI.
f_D^*	Optimal distribution f_D for Fixed-M.
G^*	Optimal operating throughput for Fixed-M.
$A^*_{norm,df}$	Minimum value of the normalized AAoI for Fixed-M, with f_D having finite support.
$f^*_{D,df}$	Optimal distribution f_D having finite support for Fixed-M.
G^*_{df}	Optimal value of G for Fixed-M with f_D having finite support.

end of the frame on the signals thus stored to recover the data packets. A *singleton slot* is a slot with transmission by a single UE. The BS can recover the data packet transmitted by a UE in any singleton slot. When multiple UEs transmit their data packets in a slot (i.e., a collision occurs), the BS is not able to directly recover any data packet from the signal received in that slot. The SIC process starts with singleton decoding where packets in the singleton slots are recovered. The headers of these recovered packets contain the slot indices in which these packets were repeated. These recovered packets are then cancelled from the received signals in the slots where the packet was repeated. This procedure is used to successively cancel packet interference and recover packets until no further packets can be recovered in the frame.

A. Probability of successful decoding

We denote the probability of successful decoding of a UE's packet using SIC at the end of the frame as P_s. Also, we define normalized channel traffic $G = M/N$. Then, the system throughput is GP_s packets per slot. We note that P_s is a function of M, N, and f_D. It was shown in [9] that in the asymptotic regime where M and $N \to \infty$, P_s can be characterized through an iterative computational procedure referred to as density evolution, by drawing on the similarity between the SIC process and the process of decoding low density parity check (LDPC) codes for binary erasure channels. Here, we briefly present this computation. We define the polynomial $\Lambda(x) = \sum_{d=1}^{N} x^d f_D[d]$ and its derivative $\Lambda'(x) = \sum_{d=1}^{N} dx^{d-1} f_D[d]$. Then, from [9], the probability that a transmission is not recovered at the l^{th} iteration is $p_l = 1 - e^{-G\Lambda'(p_{l-1})}$, with the initialization $p_1 = 1$. The asymptotic probability of success at the end of the SIC decoding operation is given by $P_s = \lim_{l\to\infty} 1 - \Lambda(p_l)$.[1] We denote this characterization of P_s as P_s^{de}; note that it is implicitly a function of G and f_D. It is also known that there exists a threshold value of the normalized channel traffic, G_τ, such that $P_s^{de} \approx 1$ for $G < G_\tau$, where G_τ depends on f_D. For $G > G_\tau$, the SIC process performs poorly, and P_s^{de} is small but bounded away from zero [22].

III. System model and problem statement

We consider the wireless network setup described in Section II, where, in every frame, M UEs sample a process of interest and generate a data packet to be communicated to the BS. Since our objective is to minimize the age of information, the UEs send fresh information to the BS by sampling and generating a data packet once every frame.[2] The sampling, measurement, data packet generation and transmission time is considered to be negligible compared to the slot duration. We index slots and frames by $t \in \mathbb{Z}_{++}$ and $i \in \mathbb{Z}_{++}$, respectively. In the ith frame, UE u samples and generates a packet at the beginning of slot $T_u[i]$, $u \in \{1, 2, \ldots, M\}$. We note that $T_u[i] \in \{1, 2, ..., N\}$. In this paper, we consider the following choices for $T_u[i]$ for a UE:

G1: $T_u[i] = 1, \forall i$, i.e., each UE samples the process at the beginning of every frame.

GT: $T_u[i] = s_1$, where s_1 is the first transmission slot for the UE under IRSA. In this case, the UE samples just-in-time before the first transmission.

G1 models scenarios where the sampling process is periodic. We also consider GT since it is intuitive that a UE u can choose $T_u[i] = s_1$ to reduce the age.

We recall that in IRSA the BS stores the signals received in each slot of a frame and applies the SIC procedure at the end of the frame in order to recover the packets. We denote this recovery method as **REND**. In REND, we assume that a maximum of $n_e \in \mathbb{Z}_+$ iterations of SIC is applied at the end of a frame. For example, REND with $n_e = 0$ will recover the packets from the singleton slots but will not proceed to interference cancellation (IC), while REND with $n_e = \infty$ will continue SIC process as long as an IC can be performed.

[1] The sequence $\Lambda(p_l)$ converges to a fixed point. For practical computation, we take $l = 1000$.

[2] We assume that there is no channel state feedback since we are considering grant-free communication systems.

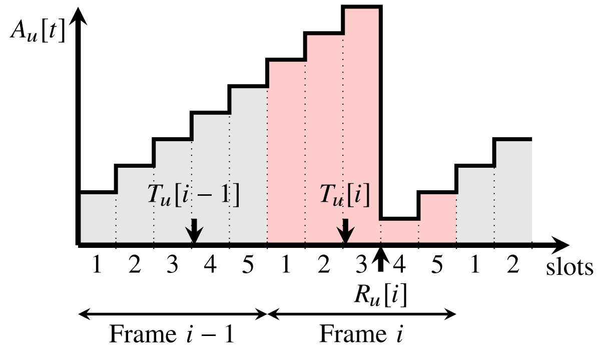


Fig. 1: Illustration of the evolution of AoI $A_u[t]$ for a UE u. The IRSA frame length $N = 5$. The packet is generated at the start of slot $T_u[i]$ in frame i. In this example, UE fails to update in frame $i-1$ and AoI increases linearly. In the next frame the UE is recovered successfully at slot $R_u[i] = 3$. Since $T_u[i] = 3$, the AoI is 1 at the beginning of the slot 4.

Intuitively, the AoI can be reduced by applying SIC at the end of every slot, instead of waiting till the end of the frame. We denote this packet recovery method as **REARLY-n**. In REARLY-n, we assume that SIC is applied for a maximum of $n \in \mathbb{Z}_+$ iterations in every slot other than the last slot of a frame. In the last slot, a maximum of n_e SIC iterations are performed. For example, in REARLY-0, packets received in singleton slots are recovered immediately at the end of the slot, while packets received in slots with collisions are recovered via SIC at the end of the frame.

Suppose the packet transmitted by user u in frame i is successfully recovered. Then, we denote the slot in which the data packet is recovered as $R_u[i] \in \{1, 2, \ldots, N\}$. We assume that the packets are recovered at the end of a slot. For REND, we have that $R_u[i] = N$, while in REARLY-n, $R_u[i] \le N$. We also note that $R_u[i] \ge T_u[i]$ for any frame i and UE u.

A. Average age of information (AAoI)

The age of information for UE u at the start of slot t is denoted as $A_u[t]$. We assume that $A_u[0] = a_{u,0}$ for $a_{u,0} \in \mathbb{Z}_+, \forall u \in \{1, 2, \ldots, M\}$. Within a frame i, $A_u[t]$ increases linearly with each slot till UE u's packet is recovered using REND or REARLY-n. Suppose the packet is recovered at $t_{rec} = (i-1)N + R_u[i]$. Then, the age drops to $R_u[i] - T_u[i] + 1$ at the start of slot $t_{rec} + 1$, where $T_u[i]$ depends on whether sampling is done using GT or G1. Thus, the evolution of $A_u[t]$ can be described as

$$A_u[t+1] \quad = \quad \begin{cases} A_u[t] + 1, & \text{if } u \text{ is not recovered at } t, \\ R_u[i] - T_u[i] + 1, & \text{if } u \text{ is recovered at } t. \end{cases} \tag{1}$$

Here, $i = \lceil \frac{t}{N} \rceil$. We illustrate the above evolution by an example in Figure 1. We note that in G1 and REND, the process $A_u[t]$ is random since the successful decoding of the packet depends on whether the SIC process over the random slot selections by the UEs is successful. In other cases, the randomness in the slot selection process causes the sampling time $T_u[i]$ as well as the decoding time $R_u[i]$ to be random, which introduces additional randomness in $A_u[t]$.

For each user, we are interested in the average age of information (AAoI), defined as

$$\overline{A}_u = \limsup_{K \to \infty} \frac{1}{K} \sum_{t=0}^{K-1} A_u[t].$$

We note that both AoI and AAoI have units of slots. For brevity of notation, the index u is dropped in further discussion since the UEs are identical in every respect (e.g., the average age is represented as $\overline{A}$ for any UE).

B. Problem statement

Our objective is to obtain an IRSA design so as to minimize the AAoI. In the mMTC scenario, we assume that there is a large but fixed number M of UEs that the system design has to support. The frame length N, the repetition distribution f_D to be used by each UE, the sampling method, and the recovery method have to be chosen so as to minimize the AAoI. We recall that the choice of sampling method is either G1 or GT, and the choice of recovery method is either REND or REARLY-n. Formally, the problem is to

$$\underset{\{\text{G1, GT}\}, \{\text{REND, REARLY-n}\}, N, f_D}{\text{minimize}} \quad \overline{A} \tag{2}$$

We denote this IRSA design problem as **Fixed-M**. An approach to solve Fixed-M is to minimize the AAoI over N and f_D for a given combination of sampling (G1 or GT) and recovery method (REND or REARLY-n). For each combination (e.g., G1 and REND), the minimum value of Fixed-M is denoted as A^*, while the optimal value of N and optimal distribution f_D are denoted as N^* and f_D^*, respectively. In the next section, we characterize the AAoI $\overline{A}$ as a function of N, M, and f_D for each combination of sampling and recovery methods. Following this, we optimize the AAoI over N and f_D in Section V.

IV. Average age of information for IRSA

In this section, we characterize $\overline{A}$ for the system discussed in Section III as a function of N, M, f_D, the choice of sampling method (G1 or GT), and the choice of recovery method (REND or REARLY-n). Since the packet transmission and recovery process are independent from frame to frame, we drop the frame index i in the notations below. We consider REND first.

Theorem IV.1. *For G1 and REND,*

$$\overline{A} = \frac{N}{2} + \frac{N}{P_s} - \frac{1}{2}. \tag{3}$$

Theorem IV.1 is proved in Appendix A by identifying a renewal process for each UE with the renewal cycle being the interval between two successful updates. For every renewal cycle, the age is obtained as a reward which depends on the duration of the renewal cycle. The AAoI is computed using the renewal reward theorem. We observe that $\overline{A}$ is inversely proportional to P_s.

In Figure 2, we plot (3) with $P_s = P_s^{de}$. We note that $P_s = P_s^{de}$ is appropriate for mMTC scenario with $M \longrightarrow \infty$.

In this illustration, we choose the repetition distribution f_D with $\Lambda(x) = x^3$. We show $\overline{A}_{norm} = \overline{A}/M$ in the plot since in the mMTC scenario $\overline{A} \to \infty$ as $M \to \infty$. Furthermore, this normalization by M clearly brings out the dependence of $\overline{A}$ on the normalized channel traffic G and P_s for the asymptotic and non-asymptotic cases over a wide range of values of G. We note that G is varied by varying N, keeping M fixed. We observe that $\overline{A}_{norm}$ initially decreases and then increases rapidly as G increases. The initial decrease occurs due to the decrease in N with increasing G (as $N = \lceil M/G \rceil$). But for larger G, the recovery probability P_s rapidly transitions from 1 to a small non-zero value and $\overline{A}_{norm}$ increases. This transition in $\overline{A}_{norm}$ occurs for G in the neighbourhood of G_τ (for this example of $\Lambda(x) = x^3$, $G_\tau = 0.818$).

We also plot the AAoI obtained from simulation for the non-asymptotic cases of $M = 100$ and 1000, to validate whether (3) with P_s^{de} matches with simulated AAoI values for non-asymptotic M and N. For each M, SIC is performed with $n_e = 100$ and the simulations are run for 10000 frames. Two quantities are estimated from the simulations: (a) the SIC success probability P_s^{sim} and (b) the AAoI $\overline{A}^{sim}$. For all values of G, we observe that the value of (3) with $P_s = P_s^{sim}$ matches with $\overline{A}^{sim}$. For both $G < G_\tau$ and $G > G_\tau$, we observe that (3) with $P_s = P_s^{de}$ closely matches (3) with $P_s = P_s^{sim}$ as well as $\overline{A}^{sim}$ for non-asymptotic M and N. In a small interval around G_τ, there is a mismatch[3] due to P_s^{de} being different from P_s^{sim}. This shows that, in the non-asymptotic regime with fixed M, the accuracy of (3) is limited by the accuracy of the characterization of P_s via the density evolution analysis, for $G \approx G_\tau$; the result in Theorem IV.1 itself is accurate. We also observe that for a fixed f_D, the G at which minimum AAoI occurs is different at finite values of M compared to the asymptotic M case, for the same reason.

We now discuss the case of GT and REND. We let $\alpha(f_D) \triangleq \sum_{d=2}^{N} \frac{f_D[d]}{d+1}$.

Theorem IV.2. *For GT and REND,*

$$\overline{A} = \frac{N}{2} + \frac{N}{P_s} - (N+1)\alpha(f_D) + \frac{1}{2}. \tag{4}$$

To prove Theorem IV.2, we identify a renewal process in the evolution of the age for every UE, with the renewal cycle being the duration between two successful updates. But, due to the just-in-time sampling, the reward depends not only on the duration of the renewal cycle but also on the sampling time T, which is random due to the randomness in the slot selection process. We derive the distribution of T in Appendix B and formally prove the theorem in Appendix C.

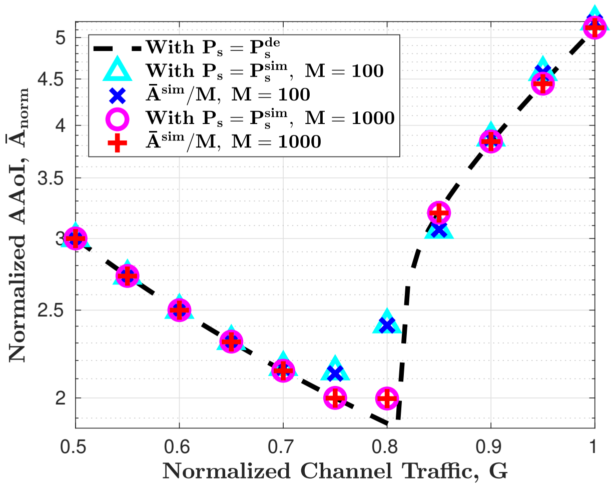


Fig. 2: Plot of $\overline{A}_{norm}$ for G1 and REND using (3) with $P_s = P_s^{de}$ and $P_s = P_s^{sim}$ (for $M = 100$ and 1000) compared with $\overline{A}^{sim}/M$ (for $M = 100$ and 1000). The repetition distribution has $\Lambda(x) = x^3$. The simulations are done for 10000 frames with $n_e = 100$ and $N = \lceil M/G \rceil$. We observe that (3) with $P_s = P_s^{sim}$ matches $\overline{A}^{sim}$ for all G and approaches (3) with $P_s = P_s^{de}$ as M increases.

We note that since $\alpha(f_D) \geq \frac{1}{N+1}$,[4] compared to the AAoI for G1 and REND given in Theorem IV.1, the AAoI in this scheme is reduced by $(N+1)\alpha(f_D) - 1$. This reduction in the age is a function of f_D as well as N. We analytically compute $\overline{A}_{norm}$ by substituting $P_s = P_s^{de}$. Simulations show similar behaviour in the case of GT and REND as in G1 and REND, but with the reduced AAoI as mentioned above. We omit the detailed discussion here to avoid repetition.

We have the following characterization of $\overline{A}$ for G1 and REARLY-n.

Theorem IV.3. *For G1 and REARLY-n,*

$$\overline{A} = \frac{N}{2} + \frac{N}{P_s} - \mathbb{E}[U|\mathcal{S}] - \frac{1}{2}, \tag{5}$$

where $U = N - R$.

The proof is given in Appendix D. Compared to G1 and REND, we observe a reduction of $\mathbb{E}[U|\mathcal{S}]$ in $\overline{A}$ due to early recovery. We do not have an analytical characterization of $\mathbb{E}[U|\mathcal{S}]$ for a general n in REARLY-n. However, for REARLY-0, we obtain the distribution of R for finite N in Appendix E from which we derive $\mathbb{E}[U|\mathcal{S}]$. Simulations of REARLY-n are used to compute $\overline{A}^{sim}$ as before, for finite M and N. In Figure 3, we compare $\overline{A}_{norm}$ for G1-REND, G1-REARLY-0 and G1-REARLY-5. Comparing cases (b) and (d) in Figure 3, we see that even REARLY-0 (where only singleton packets are decoded immediately and no SIC is performed till the end of the frame) leads to a nearly 19% reduction in AAoI at $G = 0.5$ compared to G1 and REND. Also, comparing (b), (d) and (f) in Figure 3, REARLY-5 leads to a 34% reduction in AAoI compared to G1 and REND, and nearly 18% reduction in AAoI compared to G1 and REARLY-0, at $G = 0.5$.

[3] The SIC process stops at a stopping set (a set of UEs which have transmitted in slots where every slot has at least two transmissions from the same set). The size of these stopping sets, i.e., the number of UEs in the stopping set, grows with the normalized channel traffic. While density evolution is able to capture the effect of large stopping sets (at high G), the contribution of moderate to small stopping sets (which are observed at finite M) are not captured by the analysis. However, at low G, this effect is not very prominent since the size of the stopping set is small (and hence $P_s \approx 1$). In the transition region (G near G_τ), we observe a small mismatch between P_s^{de} and P_s^{sim} [22].

[4] $\alpha(f_D) = \sum_{d=2}^{N} \frac{f_D[d]}{d+1} \geq \frac{\sum_{d=2}^{N} f_D[d]}{N+1} = \frac{1}{N+1}$

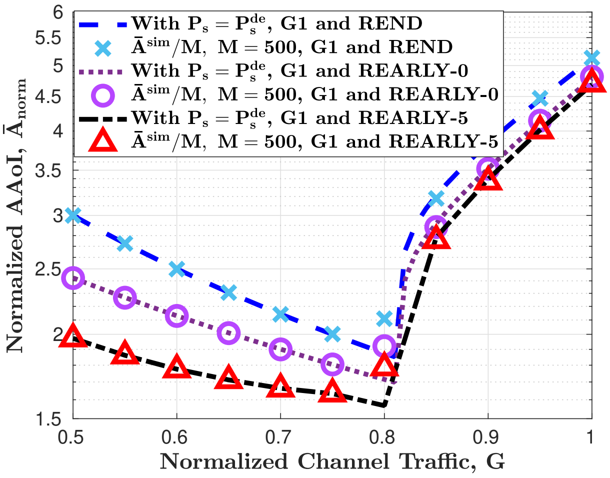


Fig. 3: Plot of $\overline{A}_{norm}$ for G1 from: (a) (3) with $P_s = P_s^{de}$, (b) $\overline{A}^{sim}$ for REND, (c) (5) with $P_s = P_s^{de}$, (d) $\overline{A}^{sim}$ for REARLY-0, (e) (5) with $P_s = P_s^{de}$ and (f) $\overline{A}^{sim}$ for REARLY-5. For (c) $\mathbb{E}[U|\mathcal{S}]$ in (5) is obtained from Appendix E while for (e) $\mathbb{E}[U|\mathcal{S}]$ is estimated from the simulations. The repetition distribution has $\Lambda(x) = x^3$. The simulations are done for 1000 frames with $n_e = 100$ and $N = \lceil M/G \rceil$.

We now consider the sampling method GT and REARLY.

Theorem IV.4. *For GT and REARLY-n,*

$$\begin{aligned}\overline{A} \quad = \quad & \frac{N}{2} + \frac{N}{P_s} - [(N+1)\alpha(f_D) - 1]\left[1 - \frac{P_s}{N}\mathbb{E}[U|\mathcal{S}]\right] \\ & -\mathbb{E}[U|\mathcal{S}] - \frac{P_s}{N}\mathbb{E}[(T-1)U|\mathcal{S}] - \frac{1}{2}, \qquad (6)\end{aligned}$$

where $\alpha(f_D) \triangleq \sum_{d=1}^{N} \frac{f_D[d]}{d+1}$ *and* T *is the slot index of first transmission for the UE.*

The proof is given in Appendix F. Compared to G1 and REND, we note that there is a reduction in $\overline{A}$ caused due to both sampling just before transmission as well as early recovery of the packets.[5] However, we do not have analytical characterizations of $\mathbb{E}[U|\mathcal{S}]$ or $\mathbb{E}[TU|\mathcal{S}]$ for a general n. These quantities need to be estimated from simulations.

A. Complexity of the recovery methods

In a practical implementation of an SIC receiver, the two computationally expensive operations are (a) decoding a data packet from a singleton slot and (b) cancelling its interference in other slots. In order to analyze the complexity of the recovery methods, we consider the case $n = \infty$ and $n_e = \infty$. The SIC process in both the recovery methods (REND and REARLY-∞) stops if no singleton slot is found. The stopping set is defined as the set of UEs which cannot be recovered. These UEs have transmissions in slots where each one of those slots have at least two transmissions from UEs in the stopping set. This renders the slots not recoverable by IC. We note that, for a given frame, the stopping set of the SIC process is unique (which can be an empty set) [22]. Thus, the UEs in the stopping set as well as the UEs which are recovered are the same for REND and REARLY-∞. Furthermore, the number of decoding operations is equal to the number of recovered UEs, while the total number of decoding operations and ICs is equal to the total number of transmissions from the recovered UEs. Hence, we conclude that the complexity of the recovery process is the same for both REND and REARLY-∞.

In REND, the decoding process happens at one shot, at the end of the frame, while in REARLY-n, the decoding process is performed whenever a singleton packet is received. As a consequence, the computations in REARLY-n are distributed across the frame. In this work, we do not consider the time involved in executing the decoding and IC operations, as this is dependent on the specific algorithm and architecture used for data decoding. Nonetheless, it is clear that since REARLY-n distributes the computations over the duration of the frame, in addition to the advantage obtained in age by early decoding, REARLY-n can achieve a lower packet decoding delay compared to REND, where all the computations are executed at the end of the frame.

B. Discussion

We now summarize the results obtained in this section. We have characterized the AAoI of four schemes: G1-REND, GT-REND, G1-REARLY-n, and GT-REARLY-n up to expressions containing P_s, $\mathbb{E}[U|\mathcal{S}]$, and $\mathbb{E}[TU|\mathcal{S}]$. GT strictly improves the AAoI compared to G1, and REARLY-n strictly improves the AAoI compared to REND. Hence, for any given repetition distribution f_D, the least AAoI is achieved by GT-REARLY-n, and the largest AAoI of these schemes is achieved by G1-REND. Quantitative performance comparison of the different schemes can be done via simulation. We illustrate this in Figure 4. We see that, as the normalized channel traffic G is increased, the AAoI for all the schemes decreases up to around $G = 0.75$, and increases thereafter. In the $G < G_\tau$ regime, which is the preferred regime of operation for achieving low AAoI, GT-REARLY-5 lowers the AAoI by a factor of 2 compared to G1-REND. The other schemes offer a performance between these two extremes, with G1-REARLY-5 outperforming GT-REND at low values of G. In the next section, we use the AAoI characterizations developed above to obtain an normalized channel traffic and repetition distribution for IRSA that minimizes the AAoI.

V. Design of IRSA for minimum AAoI

In this section, we consider the optimal choice of the frame duration N and repetition distribution $f_D[d]$ for minimizing AAoI of IRSA. We consider different combinations of G1, GT and REND, REARLY-n with fixed M. The optimization of AAoI for these combinations in the non-asymptotic regime of finite M and N requires a simulation based approach and is tedious. However, for REND with both G1 and GT, we show that it is possible to formulate a tractable optimization problem in the asymptotic regime where $M, N \to \infty$ using the density evolution based approximation P_s^{de} for P_s. Therefore, we first consider the optimization problems in this asymptotic regime, and use insights gained from its solution to obtain

[5] We note that in (6) the expectations in the third and fifth terms are non-negative, $\mathbb{E}[U|\mathcal{S}] \leq N$, and $(N+1)\alpha(f_D) \geq 1$ from the footnote on page 5.

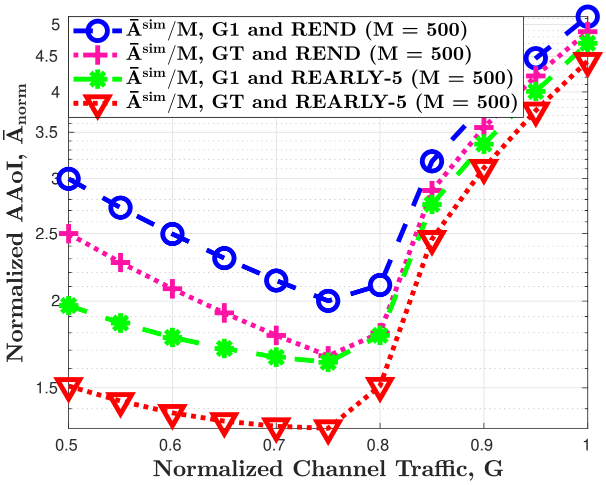


Fig. 4: A plot comparing $\overline{A}^{sim}/M$ for (a) G1 and REND, (b) GT and REND, (c) G1 and REARLY-5 and (d) GT and REARLY-5. The repetition distribution has $\Lambda(x) = x^3$. The simulations are run for 1000 frames with $M = 500$ and $n_e = 100$. GT and REARLY-5 offers the lowest AAoI compared to other combinations of sampling and recovery methods.

a judicious choice of N and $f_D[d]$ for non-asymptotic cases for all combinations. We note that the solution to GT and REARLY-n also gives us the solution to Fixed-M, since GT and REARLY-n offers the lowest AAoI among all the schemes considered. We start with the optimal design of IRSA for G1 and REND.

A. IRSA design for G1 and REND

We consider the optimization of AAoI in the asymptotic regime where $M, N \longrightarrow \infty$. In this regime, we have that $\overline{A} \longrightarrow \infty$. Therefore, we consider $\overline{A}_{norm} = \overline{A}/M$ for minimization. Using $\overline{A}$ from (3) and replacing P_s with P_s^{de}, we have the AAoI minimization problem

$$\underset{G, f_D}{\text{minimize}} \quad \frac{1}{2G} + \frac{1}{GP_s^{de}}, \tag{7}$$

where the optimization is carried out over $f_D[d]$ (which can have infinite support) and $G = M/N$ for a fixed M. The optimal solution is denoted as A^*_{norm}. We recall that P_s^{de} is a function of G and f_D, and the throughput in IRSA is given by GP_s. The solution of (7) achieves a trade-off between throughput and the normalized channel traffic to achieve minimum $\overline{A}_{norm}$ and is therefore different from the problem of maximizing the throughput GP_s. For example, in case of contention resolution diversity slotted ALOHA (CRDSA) [9], the maximum throughput of 0.541 and a normalized AAoI of 2.69 is achieved in the asymptotic regime at $G = 0.597$. However, a marginally lower minimum normalized AAoI of 2.64 is achieved at $G = 0.668$ which nonetheless shows that the throughput optimal solution is different from the AAoI optimal solution. We have the following result for A^*_{norm}.

Theorem V.1. *For G1 and REND in the asymptotic regime of $M, N \longrightarrow \infty$,*

1) *A^*_{norm} is bounded below by $\frac{3}{2}$.*
2) *A sequence of truncated soliton distributions with maximum degree d_m achieves this lower bound in the limit as $d_m \longrightarrow \infty$. A truncated soliton distribution with maximum degree d_m is defined by the probability mass function $f_D[d] = \frac{d_m}{d(d-1)(d_m-1)}$ for $d \in \{2, ..., d_m\}$.*

Proof. We note that a lower bound to the optimal solution of this problem is obtained by using the maximum value of 1 for both G and P_s. So $A^*_{norm} \geq \frac{3}{2}$. To show that this lower bound is indeed achievable, we evaluate $\overline{A}_{norm}$ for a class of truncated soliton distributions. From [3], we have that a truncated soliton distribution with maximum degree d_m achieves $G_\tau = 1 - 1/d_m$. Therefore, with $G = G_\tau$ we have that $\overline{A}_{norm} = \frac{3}{2(1-1/d_m)}$ with $P_s \longrightarrow 1$ as the number of SIC iterations $n_e \longrightarrow \infty$. Therefore, this lower bound is achievable in the limit as $d_m \longrightarrow \infty$. □

We also consider the following constrained[6] form of (7) where the distribution $f_D[d]$ has a maximum number of repetitions Λ_{max}.

$$\begin{aligned} \underset{G, f_D}{\text{minimize}} \quad & \frac{1}{2G} + \frac{1}{GP_s^{de}} \\ \text{subject to} \quad & \sum_{d=2}^{\Lambda_{max}} f_D[d] = 1, \\ & 0 \leq f_D[d] \leq 1, \forall d \geq 2. \end{aligned} \tag{8}$$

Solving the above optimization problem analytically is challenging as P_s^{de} is not available in closed form and the optimization problem is not convex. However, locally optimal solutions[7] to this constrained optimization problem (with P_s^{de} evaluated using the iterative density evolution method outlined earlier) can be obtained using differential evolution (DE) [23]. The locally optimal value of (8) obtained using DE is denoted as $A^*_{norm,df}$ and the corresponding choices G and f_D as G^*_{df} and $f^*_{D,df}$ respectively. In Table II, for $\Lambda_{max} \in \{4, 5, 6, 7, 8\}$ we present G^*_{df}, $f^*_{D,df}$ and $A^*_{norm,df}$ obtained using DE.

Discussion: From Theorem V.1 we see that, for large M, the AAoI scales linearly as $3M/2$. We compare the AAoI scaling with an equivalent centralized time division multiple access (TDMA) method (see Appendix G for a discussion.). In TDMA, we assume that transmissions from users are scheduled in a frame of length M slots, with the user scheduled in distinct slots. The users sample packets at the beginning of the TDMA frame with M slots. The users are decoded only at the end of the TDMA frame. From Theorem G.1, we have that the TDMA scheme has an AAoI of $3M/2 - 1/2$. This shows that the IRSA based distributed access scheme has the same AAoI scaling with M, namely, $3M/2$, as a fully centralized scheme (a comparison of IRSA with other distributed random access schemes is given in the sequel.) Thus, for G1 and REND, IRSA achieves not only the same throughput as TDMA but also the same AAoI, in the asymptotic regime with $M, N \longrightarrow \infty$.

[6] The constraints ensure that $f_D[d]$ is a valid repetition distribution with at least two repetitions.

[7] Such solutions are useful since they are amenable to implementation in the non-asymptotic setting where $\Lambda_{max} \leq N$.

TABLE II: Solutions $f^*_{D,df}, G^*_{df}$, and the local optimum $A^*_{norm,df}$ for (8) obtained using differential evolution. The obtained $f^*_{D,df}$ are given labels O4-O8 for easy reference.

Λ_{max}	Label	$f^*_{D,df}$	G^*_{df}	$A^*_{norm,df}$
4	O4	$0.531x^2 + 0.469x^4$	0.868	1.7281
5	O5	$0.57x^2 + 0.034x^3 + 0.396x^5$	0.899	1.6708
6	O6	$0.546x^2 + 0.165x^3 + 0.289x^6$	0.915	1.6397
7	O7	$0.528x^2 + 0.232x^3 + 0.001x^4 + 0.239x^7$	0.929	1.6146
8	O8	$0.5118x^2 + 0.258x^3 + 0.0101x^4 + 0.0042x^5$ $+0.0006x^6 + 0.0045x^7 + 0.2108x^8$	0.939	1.5974

Theorem V.1 indicates that truncated soliton distributions are good candidates for $f_D[d]$ for large M, N; they might work well even for the non-asymptotic AAoI optimization problem. The locally optimal finite degree $f^*_{D,df}$ obtained for (8) are another set of candidates for $f_D[d]$ in the non-asymptotic regime. In order to evaluate the usefulness of these candidates for the repetition distribution in the non-asymptotic case, we compare the $\overline{A}_{norm}$ of the truncated soliton distributions (with different maximum repetition degrees d_m) and $f^*_{D,df}$ from Table II using simulations. In Table III, we present the minimum $\overline{A}_{norm}$ for the distribution O8 (since O8 has the minimum $\overline{A}_{norm}$ of O4-O8) and truncated soliton distributions with $d_m \in \{8, 20, 50, 100, 200\}$. From Table III we observe that at $M = 200$, the truncated soliton distribution with $d_m = 20$ offers the least $\overline{A}_{norm}$, while at $M = 500$, $d_m = 50$ outperforms $d_m = 20$. This shows that, among the truncated soliton distributions, using a larger value of d_m is helpful as M gets large (which matches with the achievability result in Theorem V.1.) However, for non-asymptotic M, we observe that distribution O8 and a truncated soliton distribution with d_m either being 20 or 50 have lower $\overline{A}_{norm}$ compared with truncated soliton distributions with $d_m = 100$ or 200. Even though Theorem V.1 suggests that a truncated soliton distribution achieves an AAoI close to 3/2 as $M \to \infty$ with $d_m = M$, choosing $d_m = M$ for finite M is sub-optimal. For non-asymptotic M, having a larger number of repetitions leads to an increase in the number of collisions, which causes the SIC process to fail.

We observe that for $f^*_{D,df}$ from Table II, $A^*_{norm,df}$ decreases as Λ_{max} increases. We conjecture that $A^*_{norm,df} \to 1.5$, which is the asymptotic lower bound, as $\Lambda_{max} \to \infty$. We also expect that $G^*_{df} \to 1$ in this case. An important property of $f^*_{D,df}$ is that they are irregular distributions. We do not expect such convergence results for regular $f_D[d]$ with $\Lambda(x) = x^n, n \in \{2, 3, \cdots\}$, which use a fixed number of repetitions. We explain this by drawing a parallel between SIC in IRSA and LDPC decoding [3] and using the observations in Luby et al. [24] for LDPC codes. The irregularity leads to a better balance of competing requirements of high repetition degree for UEs (message-nodes in the LDPC terminology) and low degree (or collisions) for slots (check-nodes in LDPC terminology), which is needed for improved SIC performance.

Non-asymptotic M and N: For non-asymptotic M and N we consider the AAoI minimization problem (7), with P_s^{de} replaced by the actual packet success probability P_s. The minimization is carried out over the choice of $f_D[d]$ and N for a fixed M. The optimization problem is not tractable since at finite M, P_s can only be obtained from simulations. We propose the following two-step approach to obtain a good choice of $f_D[d]$ and N: (a) we solve the asymptotic problem with $P_s \approx P_s^{de}$ for finite degree distributions using differential evolution; the solution yields a candidate distribution $f^*_{D.df}$ as well as an operating point G^*_{df}, (b) we then do a one-dimensional grid search over the value of G (or equivalently N for a fixed M) starting with G^*_{df} in order to minimize the simulated value of AAoI further. In this minimization, the AAoI is evaluated using simulations. The solution is then $f^*_{D,df}$ obtained in (a), with the operating point G (or equivalently $N = M/G$) obtained via grid search in (b).

We illustrate the performance of our pragmatic approach with an example. We consider O8 for this example. Table II suggests that a good choice of G for O8 is $G^*_{df} = 0.939$ for minimum $\overline{A}_{norm}$. That is, for a given number of UEs M, the frame length N can be chosen as $\lceil M/0.939 \rceil$. In Figure 5, we illustrate that such a choice results in a larger $\overline{A}$ (by comparing $\overline{A}$ from (3) with P_s replaced by P_s^{de} and $N = \left\lceil M/G^*_{df} \right\rceil$ against $\overline{A}^{sim}$ for $N = \left\lceil M/G^*_{df} \right\rceil$). Thus, the use of $N = \left\lceil M/G^*_{df} \right\rceil$ is clearly sub-optimal for finite M. We note that P_s is less than 1 for G close to G^*_{df} for finite M, while in the asymptotic case, $P_s^{de} \lnapprox 1$ for all $G \leq G^*_{df}$. As a consequence, operating at $G = G^*_{df}$ leads to a low value of P_s and the corresponding $\overline{A}^{sim}$ is significantly higher that obtained in the asymptotic case. Therefore, step (b) reduces the AAoI by choosing $G < G^*_{df}$ using a grid search. In Figure 5, we refer to this as the minimum $\overline{A}^{sim}$. We observe that this minimum $\overline{A}^{sim}$ is 6% away from the asymptotic minimum. This difference is found to decrease further, as M is increased. We also observe that by backing off from G^*_{df}, we obtain $\overline{A}$ close to that expected from asymptotic analysis. We compare this $\overline{A}$ with conventional CRDSA (which samples and recovers packets at the beginning and end of the frame respectively) with two repetitions [9]. The pragmatic approach that we have proposed improves the AoI by almost 32% compared to CRDSA based access. We also note that our approach achieves

TABLE III: Minimum $\overline{A}_{norm}$ in non-asymptotic regime for truncated soliton distribution with $d_m \in \{8, 20, 50, 100, 200\}$ and O8 from Table II. The minimum $\overline{A}_{norm}$ is obtained for $M \in \{200, 300, 500\}$ from simulation with $n_e = 100$. Each system is simulated for 10000 frames.

M	Minimum $\overline{A}_{norm}$					
	O8	Truncated soliton distribution with				
		$d_m = 8$	$d_m = 20$	$d_m = 50$	$d_m = 100$	$d_m = 200$
200	1.889	1.987	1.883	1.890	1.949	2.068
300	1.840	1.956	1.833	1.826	1.873	1.958
500	1.790	1.911	1.785	1.764	1.797	1.852

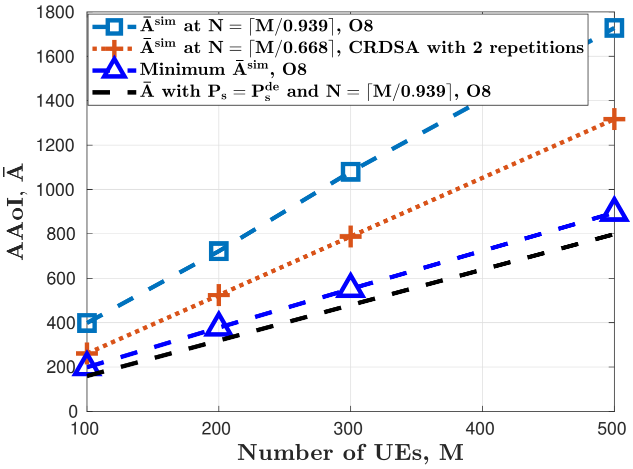


Fig. 5: For G1 and REND, a plot of (a) $\overline{A}^{sim}$ for O8 from Table II at $N = \lceil M/0.939 \rceil$ (b) $\overline{A}^{sim}$ for CRDSA with two repetitions at $N = \lceil M/0.668 \rceil$, (c) $\overline{A}^{sim}$ for O8 at a G chosen via grid search (with a step size of 0.01) to minimize $\overline{A}^{sim}$ and (d) $\overline{A}$ from (3) with P_s^{de} and $N = \lceil M/0.939 \rceil$, from top to bottom. The CRDSA scheme with two repetitions has $\Lambda(x) = x^2$ and $A^*_{norm,df} = 2.64$ at $G^*_{df} = 0.668$. The candidate solution (O8) from P_s^{de} at $N = \left\lceil M/G^*_{df} \right\rceil$ is not optimal in the non-asymptotic regime.

an AAoI which is within 20% of the fully centralized TDMA scheme (in Appendix G) at $M = 500$.

B. IRSA design for GT and REND

We now consider the case of GT and REND. We consider $\overline{A}_{norm}$ from (4) with P_s replacing P_s^{de} to obtain the following optimization problem

$$\underset{G, f_D}{\text{minimize}} \quad \frac{1}{2G} + \frac{1}{GP_s^{de}} - \frac{\alpha(f_D)}{G}. \tag{9}$$

In contrast to the optimization problem in (7) for G1-REND, the optimal distribution and the choice of G here should trade-off the term $-\frac{\alpha(f_D)}{G}$ with $\frac{1}{2G} + \frac{1}{GP_s^{de}}$. Therefore, the truncated soliton distribution from Theorem V.1 or even a throughput maximizing $f_D[d]$ may not be optimal here. For example, a soliton distribution (which was optimal for G1 and REND) has $\alpha(f_D) = \frac{1}{4}$, but the CRDSA scheme with two repetitions has the maximum possible[8] $\alpha(f_D)$ which is $\frac{1}{3}$. We reuse the notation A^*_{norm} to denote the optimal solution in GT and REND. We have the following result.

Theorem V.2. *For GT and REND in the asymptotic regime of $M, N \to \infty$,*

1) *A^*_{norm} is bounded below by $\frac{7}{6}$.*
2) *A sequence of truncated soliton distributions with maximum degree d_m achieves $\overline{A}_{norm}$ of $\frac{5}{4}$ as $d_m \to \infty$.*

Proof. We consider the objective function $\frac{1}{2G} + \frac{1}{GP_s^{de}} - \frac{\alpha(f_D)}{G}$. From Theorem V.1, we have that $\frac{1}{2G} + \frac{1}{GP_s^{de}} \geq 3/2$. From our discussion above, we have that CRDSA with two repetitions attains $\alpha(f_D) = 1/3$, which is the maximum value among all distributions with at least two repetitions in a frame. Furthermore, $G \leq 1$, so that $-\frac{\alpha(f_D)}{G} \geq -1/3$. Therefore, $A^*_{norm} \geq 7/6$. We now consider a sequence of truncated soliton distributions for obtaining a characterization of the achievable $\overline{A}_{norm}$. For a truncated soliton distribution $f_D[d] = \frac{d_m}{d(d-1)(d_m-1)}$ with d_m being the maximum number of repetitions, from [3], we have that $G_\tau = 1 - 1/d_m$ and for any $G \leq G_\tau$, $P_s^{de} = 1$. Then, at $G = G_\tau$, we have that $\frac{1}{2G} + \frac{1}{GP_s^{de}} = \frac{3}{2(1-1/d_m)}$, which has a limit of 3/2 as $d_m \to \infty$. Also, as $d_m \to \infty$, we have that $\alpha(f_D)/G \to 1/4$, so that $\overline{A}_{norm} \to 5/4$ at $G = G_\tau$ as $d_m \to \infty$. □

As for G1 and REND, we consider the following constrained optimization problem:

$$\begin{aligned} \underset{G, f_D}{\text{minimize}} \quad & \tfrac{1}{2G} + \tfrac{1}{GP_s} - \tfrac{\alpha(f_D)}{G} \\ \text{subject to} \quad & \textstyle\sum_{d=2}^{\Lambda_{max}} f_D[d] = 1, \\ & 0 \leq f_D[d] \leq 1, \forall d \geq 2, \end{aligned} \tag{10}$$

where the repetition distribution is constrained to have a maximum degree of Λ_{max}. We obtain locally optimal solutions using DE for $\Lambda_{max} \in \{4, 5, 6, 7, 8\}$. The results are shown in Table IV. We note that the notations are similar to that used for G1 and REND.

Discussion: For large M, from Theorems V.1 and V.2, we observe that GT achieves an AAoI that is $M/4$ less than that of G1 (or 16.67% less than that of G1). Comparing $A^*_{norm,df}$ from Table II and IV, we observe that sampling just-in-time

[8] CRDSA achieves the maximum value of $\alpha(f_D)$ among all distributions that at least transmit twice in a frame. We note that IRSA assumes at least two transmissions. Since $\alpha(f_D) = \sum_{d=2}^{\infty} \frac{f_D[d]}{d+1}$, we have that the maximum is when $f_D[2] = 1$ and $f_D[d] = 0$ for $d > 2$.

TABLE IV: Solutions $f^*_{D,df}$,G^*_{df}, and local optimum $A^*_{norm,df}$ for the problem (10) obtained using differential evolution. The obtained $f^*_{D,df}$ are given labels OT4-8 for easy reference.

Λ_{max}	Label	$f^*_{D,df}$	G^*_{df}	$A^*_{norm,df}$
4	OT4	$0.5624x^2 + 0.4376x^4$	0.8660	1.4146
5	OT5	$0.5624x^2 + 0.0891x^3 + 0.3485x^5$	0.8964	1.3744
6	OT6	$0.5446x^2 + 0.1899x^3 + 0.2654x^6$	0.9139	1.3493
7	OT7	$0.5276x^2 + 0.2321x^3 + 0.0007x^4$ $0.0002x^5 + 0.0002x^6 + 0.2392x^7$	0.929	1.3305
8	OT8	$0.5020x^2 + 0.2825x^3 + 0.0123x^4 + 0.2033x^8$	0.938	1.31878

offers about 17% lower AAoI compared to sampling at the beginning of the frame using the best among the finite degree distributions considered. Also, with the same Λ_{max} of 8, the optimal distribution O8 is able to achieve a value of $A^*_{norm,df}$ which is 6% away from the asymptotic lower bound for G1, while the optimal distribution OT8 is 13% away from the lower bound for GT. Since the first sampling time reduces as a function of the number of repetitions, one would expect the optimal repetition distribution to have lower maximum degree, but we observe from Table IV that the increase in the age due to P_s has a much larger effect, so that higher degrees are still preferred. Further, as in the case of G1 and REND, minimizing the AAoI requires the repetition distributions to be irregular. We also note that compared to an equivalent TDMA scheme[9] which achieves an asymptotic scaling of M (from Theorem G.1), the best possible scaling is $7M/6$. The difference represents the loss in AAoI performance due to the distributed nature of IRSA which necessitates the use of repeated transmissions.[10]

Non-asymptotic M and N: Similar to the case of G1, we propose a pragmatic approach to solve the problem in the non-asymptotic case using: (a) a solution (consisting of a finite degree distribution $f_D[d]$ and an operating point G) of the asymptotic problem with $P_s = P_s^{de}$ using differential evolution, and (b) a local search over the value of G (or equivalently over N, for a fixed M) in order to further reduce the AAoI. We illustrate the performance of this approach below. Similar to the case of G1, we use the distribution OT8 and $N = \left\lceil M/G^*_{df} \right\rceil$ in the non-asymptotic case (from Table IV). We show the simulation results for 10000 frames and $n_e = 100$ in Figure 6. The use of $N = \left\lceil M/G^*_{df} \right\rceil$ results in sub-optimal AoI performance. A local search over $G < G^*_{df}$ leads to an improved AAoI performance that is 13% away from the asymptotic value for $M = 500$. We also found that this difference decreases further as M increases. We also compare the AoI with that in a slotted ALOHA system and CRDSA with two repetitions. For slotted ALOHA, the probability of transmission is chosen to be $1/M$ to minimize $\bar{A}$ [4] (plotted as Figure 6(b)). For CRDSA, we compute the AAoI with just-in-time packet sampling. At $M = 500$, the solution based on OT8 and local search for G achieves an AAoI which is 29% and 44% lower than that of CRDSA and slotted ALOHA, respectively. At $M = 500$, the AAoI is 50% higher than that of the equivalent TDMA scheme in Appendix G.

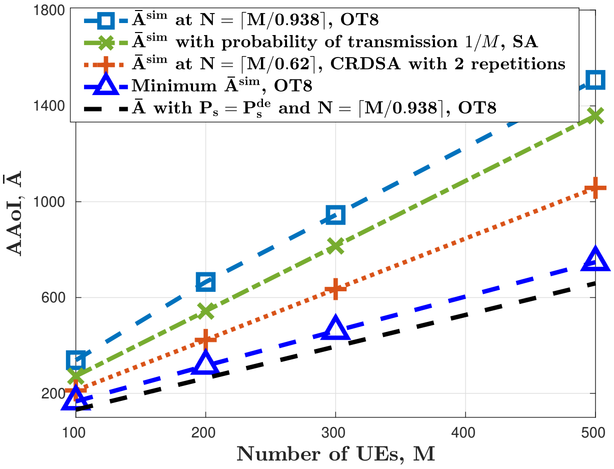


Fig. 6: For GT and REND, a plot of (a) $\overline{A}^{sim}$ for OT8 from Table IV at $N = \lceil M/0.938 \rceil$ (since $G^*_{df} = 0.938$ for OT8), (b) $\overline{A}^{sim}$ for slotted ALOHA with probability of transmission $1/M$, (c) $\overline{A}^{sim}$ for CRDSA with two repetitions at $N = \lceil M/0.62 \rceil$ (for CRDSA with two repetitions, $\overline{A}^*_{norm,df} = 2.1224$ and $G^*_{df} = 0.62$), (d) $\overline{A}^{sim}$ for OT8 at a G chosen via grid search (with step size of 0.01) to minimize $\overline{A}^{sim}$, and (e) $\overline{A}$ for OT8 from (4) with P_s^{de} and $N = \lceil M/0.938 \rceil$, from top to bottom.

C. IRSA design for REARLY-n

For REARLY-n (with G1 or GT) we recall that in addition to P_s, $\mathbb{E}[U|S]$ and $\mathbb{E}[TU|S]$ (which appear along with P_s in the objective function) are characterized using simulations, so that the optimization problem can only be solved using a simulation based approach. We use the following

[9] For this comparison, we consider a TDMA scheme whose behavior is equivalent to GT and REND, where each user is scheduled in a unique slot, and the frame consists of M slots. Like GT, each UE samples its packet at the start of its scheduled slot, and like REND, the decoding of all packets is done at the end of the TDMA frame.

[10] We note that the asymptotic lower bound of $7M/6$ is obtained with the assumption that at least two repetitions are needed in an IRSA frame. However, if we also allow for the case that a user can have only one transmission, then we note that the asymptotic lower bound is M (since $\alpha(f_D)$ would have a maximum value of $1/2$ under this assumption.)

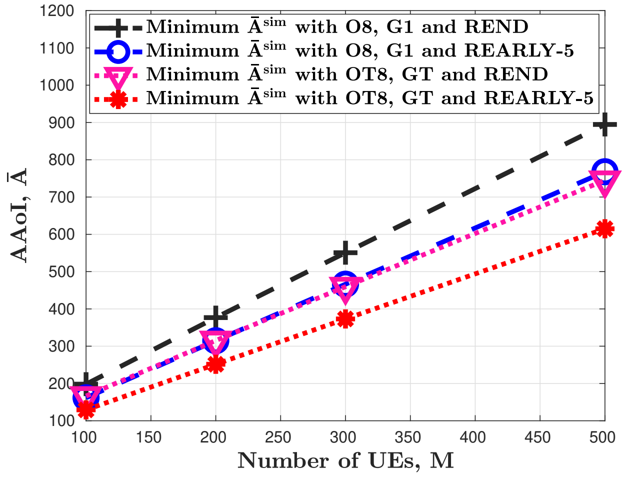


Fig. 7: A plot comparing minimum $\overline{A}^{sim}$ for (a) G1 and REND, (b) G1 and REARLY-5, (c) GT and REND, and (d) GT and REARLY-5. In G1, the repetition distribution is O8 from Table II and in GT the repetition distribution is OT8 from Table IV. The simulations are run for 1000 frames with $n_e = 100$ and $M = 100, 200, 300$ and 500. The minimum is obtained using grid search over G with step size of 0.01. GT and REARLY-5 achieves a significantly better AAoI performance compared to all the other schemes. The minimum $\overline{A}^{sim}$ in case of G1 and REARLY-5 is similar to GT and REND.

insights and observations obtained in the previous sections to obtain an IRSA design for Fixed-M in this case. These are: (a) the methodology of using a repetition distribution $f_D[d]$ obtained from the asymptotic analysis with a local search based modification to the asymptotic operating point substantially reduces the AAoI in the non-asymptotic case, and (b) the AAoI reduction obtained from the SIC iterations in every slot for REARLY is marginal beyond 5 iterations. Therefore, we suggest the following solution: (a) obtain a repetition distribution $f_D[d]$ and an operating point G from the asymptotic analysis with REND and (b) obtain a local minimum of the simulated AAoI over $G = M/N$ with the simulation carried out for REARLY-5.

Discussion: The performance of our approach is shown in Figure 7, where we compare GT-REARLY-5 with G1-REND, GT-REND, and G1-REARLY-5. We see that GT-REARLY-5 achieves an AAoI reduction of approximately $3M/10$ in comparison with G1-REARLY-5 or GT-REND, and a significant reduction in AAoI compared to CRDSA and slotted ALOHA (see Figure 6). Also, the AAoI is similar for G1-REARLY-n and GT-REND (which is supported by a similar observation in Theorem G.1 for TDMA).

We compare the REARLY-n schemes with corresponding centralized TDMA schemes[11] discussed in Appendix G. From Theorem G.1, we have that for large M, the AAoI performance of TDMA corresponding to REARLY-n with GT and G1 are $M/2$ and M, respectively. However, we expect that these lower bounds are optimistic for the REARLY-n schemes as packets are not always decoded in every slot in the case of IRSA. For example, from Figure 7, we observe that the AAoI of GT and G1 (with REARLY-n) scale approximately as $6M/5$ and $3M/2$, respectively. This deviation from the TDMA performance is due to a larger fraction of packets getting decoded at the end of the frame for REARLY-n when G is large (which is usually the case when AAoI is minimized) and is the price to pay for having a decentralized transmission scheme. With just-in-time sampling and early recovery of packets, our proposed scheme reduces the AAoI by 54% compared to SA, and 53% compared to conventional CRDSA. In comparison with an equivalent CRDSA scheme (in which we use the CRDSA repetition distribution from [9] but with just-in-time sampling and early recovery) we observe a reduction of 16% for our proposed scheme.

VI. Conclusions and future Work

In this paper, we considered the problem of minimizing the AAoI for an mMTC system which uses frame-based IRSA. We first analyzed a basic IRSA scheme where packets are generated by the users at the beginning of a frame and are recovered at the BS using SIC at the end of a frame. Then we proposed and analyzed just-in-time sampling and early recovery schemes which improve the AAoI performance of IRSA. For each combination of sampling and recovery schemes, we obtained semi-analytical AAoI expressions as a function of the number of UEs, the frame duration, and the IRSA repetition distribution. In all cases, for a given IRSA repetition distribution, the AAoI initially decreases and then sharply increases, as a function of the normalized channel traffic. We also considered the challenging problem of optimizing the IRSA repetition distribution and the normalized channel traffic for minimizing the AAoI. For the case where packet recovery is done at the end of the frame, we showed that tractable AAoI minimization problems can be formulated in the asymptotic regime where the number of UEs is large.

We obtained upper and lower bounds on the minimum AAoI: they scale linearly with the number of users. We compared the minimum AAoI achieved by IRSA with a equivalent centralized TDMA scheme. In the case where the sampling is done at the beginning of a frame, the asymptotic scaling of AAoI with number of UEs for IRSA is the same as that of TDMA. In the case where sampling is just-in-time, we showed that there is a fundamental difference in asymptotic scaling of AAoI for IRSA in comparison to TDMA due to the distributed nature of IRSA. Using differential evolution, we also obtained locally optimal normalized channel traffic and repetition distribution for IRSA. In the non-asymptotic case, we judicially modified the above solutions via a one-dimensional search procedure, and evaluated them empirically via simulations. The IRSA scheme which samples just in time and does early decoding of packets achieves a significantly better AAoI scaling with the number of users M compared to CRDSA, slotted ALOHA, and basic IRSA.

Future work could aim towards deriving bounds on the AAoI and obtain better insight into the behaviour of AAoI, and extend this work to fading wireless channels.

[11] We note that a TDMA scheme corresponding to REARLY-n would be one in which users are decoded at the end of their transmission slots. The sampling times are either all at the beginning of a TDMA frame (for the comparison with G1) or at the beginning of their respective transmission slots (for the comparison with GT).

Appendix A
Proof of Theorem IV.1

For a particular UE in the system, let $F_k \in \{1, 2, 3, \dots\}$ denote the number of frames between $(k-1)^{\text{th}}$ and k^{th} successful packet recovery. For G1 and REND, the sampling instant is fixed ($T = 1$) and collisions are resolved at the end of the frame. We note that P_s is dependent on the repetition distribution and the SIC decoding process; it is independent of the frame index. We have that $\mathbb{E}[F_k] = \frac{1}{P_s}$ and $\mathbb{E}[F_k^2] = \frac{2-P_s}{P_s^2}$, for any k. Let $J_k = \sum_{j=1}^{k} F_j$ denote the number of frames elapsed before the kth successful update. Then $X_i = \sup\{k : J_k \leq i\}$, the number of successful updates up to frame i, is a *renewal process*.

In the k^{th} epoch, the AoI starts from N (since $T = 1$ and $R = N$ at J_{k-1} due to successful recovery at the end of the frame) and increases linearly up to NF_k slots. After the successful update at frame J_k, the AoI drops to N. The cumulative AoI for the kth epoch, $W_k = \sum_{j=0}^{NF_k-1} N+j = N^2 F_k + \frac{NF_k}{2}(NF_k - 1)$. As F_ks are *iid*, $\{W_1, W_2, W_3, \dots\}$ is a also sequence of positive *iid* random variables with

$$\begin{aligned}\mathbb{E}[W_k] &= N^2\mathbb{E}[F_k] + \frac{N^2}{2}\mathbb{E}[F_k^2] - \frac{N}{2}\mathbb{E}[F_k] \\ &= \frac{N^2}{P_s} + \frac{N^2}{2}\frac{2-P_s}{P_s^2} - \frac{N}{2P_s} < \infty.\end{aligned} \tag{11}$$

W_k can be considered as *rewards* of the renewal process X_i. Let $Y_i = \sum_{k=1}^{X_i} W_k$. Then, from the *renewal reward theorem* [25] the AAoI is

$$\overline{A} = \lim_{I\to\infty} \frac{Y_I}{NI} = \frac{\mathbb{E}[W_k]}{N\mathbb{E}[F_k]} = \frac{N}{P_s} + \frac{N}{2} - \frac{1}{2}. \tag{12}$$

Appendix B
Distribution of sampling instant in GT

In GT, a UE samples just-in-time before the first transmission in a frame. The sampling instant is random, independent of the frame index, and is determined by the repetition distribution. We recall that the number of repetitions is denoted by $D \in \{2, \cdots, N\}$ (which follows the distribution f_D) and the sampling instant (i.e., the slot of first transmission by the UE) is denoted by the random variable T. In a frame, for a UE with $D = d$, the first slot can only be selected from 1 to $N - d + 1$ slots. As the d slots are chosen uniformly,

$$\mathbb{P}(T = j | D = d) = \begin{cases} \frac{\binom{j-1}{0}\binom{N-j}{d-1}}{\binom{N}{d}}, & \text{for } j \in \{1, 2, \dots, N-d+1\} \\ 0, & j \in \{N-d+2, \dots, N\} \end{cases}$$

and

$$\begin{aligned}\mathbb{P}(T = j) &= \sum_d \mathbb{P}(T = j | D = d)\mathbb{P}(D = d) \\ &= \sum_{d=1}^{N-j+1} \frac{\binom{N-j}{d-1}}{\binom{N}{d}} f_D[d].\end{aligned}$$

We compute the expectation of T as

$$\mathbb{E}[T] = \sum_{j=1}^{N} j \sum_{d=2}^{N-j+1} \frac{\binom{N-j}{d-1}}{\binom{N}{d}} f_D[d]. \tag{13}$$

Exchanging the limits of summation

$$\mathbb{E}[T] = \sum_{d=2}^{N} \left\{ \sum_{j=1}^{N-d+1} j \frac{\binom{N-j}{d-1}}{\binom{N}{d}} \right\} f_D[d] = \sum_{d=1}^{N} g_d f_D[d] \tag{14}$$

where $g_d = \sum_{j=1}^{N-d+1} j \frac{\binom{N-j}{d-1}}{\binom{N}{d}}$. In order to simplify g_d, we consider $\beta_1(x) = \sum_{j=d}^{N}(1+x)^j$ and $\beta_1'(x) = \sum_{j=d}^{N} j(1+x)^{j-1}$. From the expansion of $\beta_1'(x)$ we obtain that the coefficient of the term where exponent of x is $(d-1)$ is given by $\sum_{j=1}^{N-d+1} j \frac{\binom{N-j}{d-1}}{\binom{N}{d}}$. Since $\beta_1(x) = \frac{(1+x)^{N+1} - (1+x)^d}{x}$ and

$$\beta_1'(x) = \frac{(N+1)(1+x)^N - d(1+x)^{d-1}}{x} - \frac{(1+x)^{N+1} - (1+x)^d}{x^2},$$

in the compact form of $\beta_1'(x)$, the coefficient of the term where x's exponent is $(d-1)$ is $(N+1)\binom{N}{d} - \binom{N+1}{d+1}$. Equating the coefficients, we get

$$\sum_{j=1}^{N-d+1} j \frac{\binom{N-j}{d-1}}{\binom{N}{d}} = (N+1)\binom{N}{d} - \binom{N+1}{d+1}. \tag{15}$$

Similarly, we consider $\beta_2(x) = \sum_{j=d}^{N-1}(1+x)^j$. In the expansion of $\beta_2(x)$, the coefficient of the term where x has exponent $(d-1)$ is given as $\sum_{j=1}^{N-d+1}\binom{N-j}{d-1}$. In compact form, $\beta_2(x) = \frac{(1+x)^N - (1+x)^{d-1}}{x}$ the coefficient of the term where x's exponent is $(d-1)$ is $\binom{N}{d}$. Hence,

$$\sum_{j=1}^{N-d+1} \binom{N-j}{d-1} = \binom{N}{d}. \tag{16}$$

From (15) and (16),

$$\begin{aligned} g_d &= \frac{\sum_{j=1}^{N-d+1} j\binom{N-j}{d-1}}{\binom{N}{d}} \\ &= \frac{\sum_{j=1}^{N-d+1}(N+1)\binom{N-j}{d-1} - (N-j+1)\binom{N-j}{d-1}}{\binom{N}{d}} \\ &= \frac{\binom{N+1}{d+1}}{\binom{N}{d}} = \frac{N+1}{d+1}, \end{aligned}$$

and from (14), we arrive at

$$\mathbb{E}[T] = (N+1)\sum_{d=2}^{N} \frac{f_D[d]}{d+1}. \tag{17}$$

Appendix C
Proof of Theorem IV.2

We follow the notation of Appendix A. The samples are obtained just before transmission and the distribution of sampling instant T is as given in Appendix B. We identify a renewal process where the renewal epochs are the times at which packets are successfully recovered. We now obtain the cumulative reward in every renewal cycle. For a UE, at the end of $(k-1)^{\text{th}}$ successful packet recovery, the AoI drops to $N - T[J_{k-1}] + 1$, where we recall that $T[i]$ is the sampling slot in the i^{th} frame. Thus, the AoI in k^{th} epoch starts from $N - T[J_{k-1}] + 1$ and increases linearly up to NF_k slots. The cumulative AoI for k^{th} epoch $W_k = \sum_{j=0}^{NF_k-1} N - T[J_{k-1}] +$

$1 + j = NF_k(N - T[J_{k-1}] + 1) + \frac{NF_k}{2}(NF_k - 1)$. We note that F_k and $T[J_{k-1}]$ are mutually independent and *iid* random variables. Therefore, the rewards associated with successive renewal cycles are also *iid* random variables.[12] Hence,

$$\begin{aligned}\mathbb{E}[W_k] &= \frac{N}{P_S}(N - \mathbb{E}\left[T[J_k - 1]\right] + 1) + N^2\frac{2-P_s}{2P_s^2} - \frac{N}{2P_s}\\ &= \frac{N^2}{P_s^2} + \frac{N^2}{2P_s} - \frac{N(N+1)}{P_s}\alpha(f_D) + \frac{N}{2P_s}\end{aligned}$$

where $\mathbb{E}[T]$ is obtained from (17) and $\alpha(f_D) = \sum_{d=2}^{N}\frac{f_D[d]}{d+1}$. Similar to Appendix A, W_k can be considered as rewards of the renewal process X_i. Then for $Y_i = \sum_{k=1}^{X_i} W_k$, from the *renewal reward theorem* [25], the AAoI is obtained as

$$\overline{A} = \lim_{I\to\infty}\frac{Y_I}{NI} = \frac{\mathbb{E}[W_k]}{N\mathbb{E}[F_k]} = \frac{N}{2} + \frac{N}{P_s} - (N+1)\alpha(f_d) + \frac{1}{2}. \quad (18)$$

Appendix D
Proof of Theorem IV.3

The notation is the same as in Appendix A. Samples are taken in the beginning of every frame and $T = 1$. We identify a renewal process in the evolution where a renewal cycle corresponds to the time between the $(k-1)^{\text{th}}$ and k^{th} packet recovery for a UE. As the SIC process is used in every slot, the recovery instant R is not always the end of the frame. At the end of $(k-1)$th epoch, the AoI of the UE is N and increases linearly for $(N-1)F_k + R[J_k]$ slots in kth epoch. At J_k the UE is recovered at slot $R[J_k]$ and the AoI drops to $R[J_k]$. Afterwards the AoI increases linearly for $N - R[J_k]$ slots till the end of frame J_k. The cumulative age in F_k, i.e.,

$$\begin{aligned}W_k &= \sum_{j=0}^{(N-1)F_k+R[J_k]-1} N + j + \sum_{j=0}^{N-R[J_k]-1} R[J_k] + j\\ &= \frac{N^2}{2}F_k^2 + NF_kR[J_k] - \frac{NF_k}{2}, \text{ and}\\ \mathbb{E}[W_k] &= \frac{N^2}{2}\frac{2-P_s}{P_s^2} + N\mathbb{E}[F_kR[J_k]] - \frac{N}{2}\frac{1}{P_s}.\end{aligned}$$

We note that $R[J_k]$ depends on the frame J_k or more generally on event of successful update $\mathcal{S}$. Thus, $\mathbb{E}[F_kR[J_k]] = \mathbb{E}[F_k]\mathbb{E}[R[J_k]|F_k] = \mathbb{E}[F_k]\mathbb{E}[R[J_k]|\mathcal{S}]$. Hence, for $Y_i = \sum_{k=1}^{X_i} W_k$ and $U = N - R$, applying the *renewal reward theorem* [25], we obtain the AAoI as

$$\begin{aligned}\overline{A} = \lim_{I\to\infty}\frac{Y_I}{NI} = \frac{\mathbb{E}[W_k]}{N\mathbb{E}[F_k]} &= N\frac{2-P_s}{2P_s} + \mathbb{E}[R|\mathcal{S}] - \frac{1}{2}\\ &= \frac{N}{2} + \frac{N}{P_s} - \mathbb{E}[U|\mathcal{S}] - \frac{1}{2}.\end{aligned} \quad (19)$$

Appendix E
Distribution of R in REARLY-0

We denote the M UEs by $u_1, u_2, ..., u_M$. In REARLY-0, the singleton slots in $\{1, 2, ..., N-1\}$ are recovered in the same slot where they are received. The recovery instant $R_{u_m}[i]$ for UE u_m in frame i is a random variable which can take values in $\{1, 2, ..., N\}$ if the UE is successful in frame i. Since the recovery method does not depend on the frame index and the UEs are identical, without loss of generality, we drop the frame index and derive the distribution of the sampling instant for u_1. We denote the event of successful recovery by $\mathcal{S}$. In REARLY-0, the UE can be recovered successfully either from a singleton slot or via the SIC process carried out at the end of the frame. Since all singleton slots are recovered in first iteration using SIC, the event $\mathcal{S}$ is the same as the event that the packet is successfully recovered. Therefore, the probability of successful recovery $\mathbb{P}\{\mathcal{S}\} = P_s$, the probability of success in the SIC process. Furthermore, $\mathbb{P}\left\{R_{u_1} = j|\mathcal{S}\right\} = \frac{\mathbb{P}\{R_{u_1}=j\cap\mathcal{S}\}}{\mathbb{P}\{\mathcal{S}\}} = \frac{\mathbb{P}\{R_{u_1}=j\}}{\mathbb{P}\{\mathcal{S}\}}$ for $j \in \{1, 2, ..., N\}$ (as the recovery instant is defined only when the UE is successful in delivering a packet).

Let the random variables $D_1, D_2, ..., D_M$ denote the number of repetitions in a frame by $u_1, u_2, ..., u_M$ respectively, which are independent of the frame index. In a particular frame, let $D_1 = d_1, D_2 = d_2, ..., D_M = d_M$. For UE u_1, we denote the d_1 selected slots as $c_1 < c_2 \cdots < c_{d_1}$. Then for $j \in \{1, 2, ..., N-1\}$, $\mathbb{P}\left\{R_{u_1} = j|D_1 = d_1, ..., D_M = d_M\right\}$ is given by (20) at the top of the next page, and

$$\begin{aligned}&\mathbb{P}\left\{R_{u_1} = j|\mathcal{S}\right\} = \frac{1}{P_s}\Bigg(\sum_{d_1,...,d_M}\\ &\mathbb{P}\left\{R_{u_1} = j|D_1 = d_1, ..., D_M = d_M\right\} f_D(d_1)\ldots f_D(d_M)\Bigg).\end{aligned}$$

At the last slot, u_1 can either be recovered from a singleton slot or via the SIC process. Thus, at $j = N$, since $\sum_{j=1}^{N}\mathbb{P}\left\{R_{u_1} = j|\mathcal{S}\right\} = 1$, $\mathbb{P}\left\{R_{u_1} = N|\mathcal{S}\right\} = 1 - \sum_{j=1}^{N-1}\mathbb{P}\left\{R_{u_1} = j|\mathcal{S}\right\}$.

Appendix F
Proof of Theorem IV.4

We use the same notation as in Appendix A. The samples are taken just-in-time before transmission and the sampling instant T depends on the repetition distribution as described in Appendix B. In every slot, the SIC process is employed to recover the UEs' packets. The recovery instant $R[J_k]$ is a random variable which depends on the slot selection and SIC process. We proceed as in the case of GT and REND by considering the cumulative age in a duration between the $(k-1)^{\text{th}}$ and k^{th} successful packet recovery times. In the k^{th} epoch, the AoI of the UE starts from $N - T[J_{k-1}] + 1$ and increases linearly for $(N-1)F_k + R[J_k]$ slots. In frame J_k, packet recovery occurs at slot $R[J_k]$, and the AoI drops to $R[J_k] - T[J_k] + 1$. Thereafter, the AoI increases linearly to $N - R[J_k]$ at the end of frame J_k. The cumulative AoI W_k is

[12] We note that for the k^{th} renewal cycle, the duration of the cycle is determined by F_k. However, the reward is partly determined by $T[J_{k-1}]$ which is sampled in the previous renewal cycle. We associate $T[J_{k-1}]$ to the reward in the k^{th} renewal cycle. There is no dependence across renewal cycles since $T[J_{k-1}]$ is independent of $T[J_{k-2}]$ as well as F_k.

$$\mathbb{P}\left\{R_{u_1} = j | D_1 = d_1, ..., D_M = d_M\right\} = \mathbb{P}\left\{c_1 = j; j \text{ is singleton}\right\} + \sum_{c_1<j} \mathbb{P}\left\{c_2 = j; j \text{ is singleton but } c_1 \text{ is selected by others}\right\}$$
$$+ \cdots + \sum_{c_1<\ldots<c_{k-1}<j} \mathbb{P}\left\{c_k = j; j \text{ is singleton but } c_1, \ldots, c_{k-1} \text{ are selected by others}\right\}$$
$$+ \cdots + \sum_{c_1<\ldots<c_{d_1-1}<j} \mathbb{P}\left\{c_{d_{u_1}} = j; j \text{ is singleton but } c_1, ..., c_{d_1-1} \text{ are selected by others}\right\}$$
$$= \frac{\binom{N-j}{d_1-1}}{\binom{N}{d_1}} \prod_{n=2}^{M} \frac{\binom{N-1}{d_n}}{\binom{N}{d_n}} + \frac{\binom{N-j}{d_1-2}\binom{j-1}{1}}{\binom{N}{d_1}} \left[\prod_{n=2}^{M} \frac{\binom{N-1}{d_n}}{\binom{N}{d_n}} - \prod_{n=2}^{M} \frac{\binom{N-2}{d_n}}{\binom{N}{d_n}}\right] + ... + \frac{\binom{N-j}{d_1-k}\binom{j-1}{k-1}}{\binom{N}{d_1}} \left[\prod_{n=2}^{M} \frac{\binom{N-1}{d_n}}{\binom{N}{d_n}} - \binom{k-1}{1} \prod_{n=2}^{M} \frac{\binom{N-2}{d_n}}{\binom{N}{d_n}} + \right.$$
$$\left. \binom{k-1}{2} \prod_{n=2}^{M} \frac{\binom{N-3}{d_n}}{\binom{N}{d_n}} - ... + \prod_{n=2}^{M} \frac{\binom{N-k}{d_n}}{\binom{N}{d_n}}\right] + \cdots + \frac{\binom{j-1}{d_1-1}}{\binom{N}{d_1}} \left[\prod_{n=2}^{M} \frac{\binom{N-1}{d_n}}{\binom{N}{d_n}} - \binom{d_1-1}{1} \prod_{n=2}^{M} \frac{\binom{N-2}{d_n}}{\binom{N}{d_n}} + \binom{d_1-1}{2} \prod_{n=2}^{M} \frac{\binom{N-3}{d_n}}{\binom{N}{d_n}} - ... + \prod_{n=2}^{M} \frac{\binom{N-d_1}{d_n}}{\binom{N}{d_n}}\right] \tag{20}$$

obtained as

$$W_k = \sum_{m=0}^{(N-1)F_k+R[J_k]-1} N - T[J_{k-1}] + 1 + m \quad +$$
$$\sum_{n=0}^{N-R[J_k]-1} R[J_k] - T[J_k] + 1 + n = \frac{N^2F_k^2}{2} + NR[J_k]F_k - \frac{NF_k}{2}$$
$$- (T[J_{k-1}]-1)(NF_k - N + R[J_k]) - (T[J_k]-1)(N - R[J_k]).$$

We note that F_k,$T[J_k]$, and $R[J_k]$ are dependent on each other (their statistics are determined by the repetition distribution). However, all of them are independent of $T[J_{k-1}]$.

In contrast to the proof of GT and REND, we observe that W_k now has dependency across the identified cycles since $T[J_{k-1}]$ as well as $T[J_k]$ determine W_k. However, if we consider the evolution of $T[J_{k-1}]$ as that of a Markov chain (even though it is *iid*), then we have that $(T[J_{k-1}], F_k)$ is a *Markov Renewal process* [25]. Then, W_k is the reward in a Markov Renewal process which is dependent on the state $T[J_{k-1}]$ of the Markov chain. Let $U[J_k] = N - R[J_k]$. Similar to Appendix D, $\mathbb{E}[F_kU[J_k]] = \mathbb{E}[F_k]\mathbb{E}[U[J_k]]|\mathcal{S}]$ and

$$\mathbb{E}[W_k] = N^2\frac{2-P_s}{2P_s^2} + \frac{N}{P_s}(N - \mathbb{E}[U[J_k]|\mathcal{S}])$$
$$-(\mathbb{E}[T[J_{k-1}]] - 1)(\frac{N}{P_s} - E[U[J_k]|\mathcal{S}])$$
$$-\mathbb{E}[(T[J_k] - 1)U[J_k]|\mathcal{S}] - \frac{N}{2P_s}.$$

Since W_k is the reward in the Markov renewal process, with $Y_i = \sum_{k=1}^{X_i} W_k$, from Appendix B and by applying the *renewal reward theorem* [25], the AAoI is given by

$$\overline{A} = \lim_{I\to\infty} \frac{Y_I}{NI} = \frac{\mathbb{E}[W_k]}{N\mathbb{E}[F_k]} = \frac{N}{2} + \frac{N}{P_s} - \mathbb{E}[U|\mathcal{S}] - ((N+1)$$
$$\alpha(f_D) - 1)(1 - \frac{P_s}{N}E[U|\mathcal{S}]) - \frac{P_s}{N}\mathbb{E}[(T-1)U|\mathcal{S}] - \frac{1}{2}.$$

Appendix G
Average age of information for time division multiple access (TDMA)

In this section, we analyze the AAoI for the fully centralized TDMA scheme, under packet sampling and decoding time assumptions which correspond to that of G1, GT and REND, REARLY-n. The AAoI performance of TDMA forms a baseline for comparison of the performance of IRSA schemes discussed in this paper. We note that, in TDMA, a user is allotted a slot by the base station to transmit its packet without any collisions. Thus, M users transmit one packet each in M non-overlapping slots. We define a TDMA frame as consisting of these M slots (i.e., the frame size $N = M$). We note that G1 corresponds to the case where the users sample packets at the beginning of the frame and GT corresponds to the users sampling packets at the beginning of their assigned TDMA slots. Similarly, REND corresponds to the users being decoded at the end of the TDMA frame, while REARLY-n corresponds to each user being decoded at the end of its own TDMA slot. The AAoI for TDMA is given by the following theorem.

Theorem G.1. *The AAoI for TDMA with G1-REND is $\frac{3M}{2} - \frac{1}{2}$, GT-REND or G1-REARLY is M, and GT-REARLY is $\frac{M}{2} + \frac{1}{2}$.*

Proof. For G1 and REND, we note that the age process $A[t]$ of any user is periodic with period of $M = N$. Within each period, the cumulative age can be shown to be $M^2 + \sum_{m=0}^{M-1} m = M^2 + M(M-1)/2$. Thus, the AAoI is $3M/2 - 1/2$. In GT and REND, the i^{th} user samples a packet at the start of the i^{th} slot and the packet is decoded only at the end of a TDMA frame. Then, $A_u[t]$ for user u is a periodic function with a period of M and the total age over the period is $uM + \sum_{m=0}^{M-1} m = uM + M(M-1)/2$. So the AAoI for UE u is $u + (M-1)/2$ and the AAoI averaged over all users is M. The proof for G1 and REARLY is similar and is skipped. For GT-REARLY, from [4, Lemma 1], the AAoI is $M/2 + 1/2$. □

A. Discussion

We note that the AAoI expressions derived above for TDMA holds in both asymptotic and non-asymptotic regimes. Thus, they serve as lower bounds for comparing the performance of the distributed IRSA schemes in asymptotic and non-asymptotic regimes. However, since transmitted packets are always successfully received in the case of TDMA, the age expressions can be expected to be optimistic compared to that derived for IRSA. We note that the performance of sampling just-in-time but decoding the packets at the end of the frame is equivalent to sampling at the beginning of the frame but decoding the packet of a user in the slot of transmission in the case of TDMA. This is observed to be true in the case of IRSA as well. For large M, we find that the AAoI decreases by $M/2$ when we move from G1-REND to GT-REND or G1-REARLY and then reduces by an additional $M/2$ when we move to GT-REARLY. Even though a reduction of $M/2$ is not observed in the case of IRSA, there is a reduction of $\approx 3M/10$ when we move from G1-REND to GT-REND or G1-REARLY and then a further reduction of again $\approx 3M/10$ when we move to GT-REARLY. These gains are approximately the same and are additive.

References

[1] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, pp. 10–17, Sep. 2015.

[2] S. Moon and J. Lee, "Performance study of repetition-based grant-free schemes in the mMTC scenario," in *International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pp. 1–2, Jun. 2019.

[3] K. R. Narayanan and H. D. Pfister, "Iterative collision resolution for slotted ALOHA: An optimal uncoordinated transmission policy," in *International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, pp. 136–139, Aug. 2012.

[4] R. D. Yates and S. K. Kaul, "Status updates over unreliable multiaccess channels," in *Proc. IEEE Int. Symp. Inf. Theory*, pp. 331–335, Jun. 2017.

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

[6] Z. Yang, M. Chen, Y. Wang, and Y. Pan, "Compressive sensing based multiuser detection for asynchronous machine-to-machine systems," in *International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, Oct. 2017.

[7] G. Ma, B. Ai, F. Wang, X. Chen, Z. Zhong, Z. Zhao, and H. Guan, "Coded tandem spreading multiple access for massive machine-type communications," *IEEE Wireless Communications*, vol. 25, pp. 75–81, Apr. 2018.

[8] E. Paolini, G. Liva, and M. Chiani, "Coded slotted ALOHA: A graph-based method for uncoordinated multiple access," *IEEE Trans. Inf. Theory*, vol. 61, pp. 6815–6832, Dec. 2015.

[9] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA," *IEEE Trans. Commun.*, vol. 59, pp. 477–487, Feb. 2011.

[10] E. Paolini, G. Liva, and A. Graell i Amat, "A structured irregular repetition slotted ALOHA scheme with low error floors," in *Proc. ICC*, pp. 1–6, May 2017.

[11] C. R. Srivatsa and C. R. Murthy, "Throughput analysis of PDMA/IRSA under practical channel estimation," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, Jul. 2019.

[12] E. Casini, R. De Gaudenzi, and O. Del Rio Herrero, "Contention resolution diversity slotted aloha (CRDSA): An enhanced random access scheme for satellite access packet networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1408–1419, 2007.

[13] S. Sen, N. Santhapuri, R. R. Choudhury, and S. Nelakuditi, "Successive interference cancellation: A back-of-the-envelope perspective," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, Hotnets-IX, (New York, NY, USA), 2010.

[14] A. G. i Amat and G. Liva, "Finite-length analysis of Irregular Repetition Slotted ALOHA in the waterfall region," *IEEE Commun. Lett.*, vol. 22, pp. 886–889, 2018.

[15] E. Paolini, "Finite length analysis of irregular repetition slotted ALOHA (IRSA) access protocols," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, pp. 2115–2120, Jun. 2015.

[16] Y. Sun, I. Kadota, R. Talak, and E. Modiano, "Age of information: A new metric for information freshness," *Synthesis Lectures on Communication Networks*, vol. 12, no. 2, pp. 1–224, 2019.

[17] I. Kadota, A. Sinha, and E. Modiano, "Scheduling algorithms for optimizing age of information in wireless networks with throughput constraints," *IEEE/ACM Trans. Netw.*, vol. 27, pp. 1359–1372, Aug. 2019.

[18] X. Chen, K. Gatsis, H. Hassani, and S. S. Bidokhti, "Age of information in random access channels," *ArXiv*, vol. abs/1912.01473, 2019.

[19] H. bin Chen, Y. Gu, and S.-C. Liew, "Age-of-information dependent random access for massive IoT networks," *ArXiv*, vol. abs/2001.04780, Feb. 2020.

[20] 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, pp. 1118–1135, Feb. 2020.

[21] A. Munari and A. A. Frolov, "Average age of information of Irregular Repetition Slotted ALOHA," *ArXiv*, vol. abs/2004.01998, May 2020.

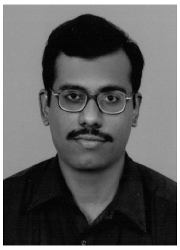
[22] T. Richardson and R. Urbanke, *Modern Coding Theory*. USA: Cambridge University Press, 2008.

[23] R. Storn and K. Price, "Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces," *Journal of Global Optimization*, 1997.

[24] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, "Improved low-density parity-check codes using irregular graphs," *IEEE Trans. Inf. Theory*, vol. 47, pp. 585–598, Feb. 2001.

[25] A. Kumar, D. Manjunath, and J. Kuri, *Wireless Networking*. Elsevier Science, 2008.

Subham Saha received his B.Tech. degree in Electronics and Communication Engineering from Indian Institute of Space Science and Technology, Trivandrum, India, in 2020. His research interest is in the area of multiple access protocols for massive machine-type communications applications. Since 2021, he is working as Scientist/Engineer SC at U. R. Rao Satellite Center, Indian Space Research Organization, Bangalore, India.



Vineeth Bala Sukumaran received the B. Tech degree in Electronics and Communication Engineering from College of Engineering, Thiruvananthapuram, and Ph. D. degree in Electrical Communication Engineering from Indian Institute of Science, Bangalore, India. Currently, he is an Assistant Professor in the Department of Avionics at the Indian Institute of Space Science and Technology, Thiruvananthapuram. His research interests are in the area of communication networks, stochastic modelling and performance analysis, and reinforcement learning.

Chandra R. Murthy (S'03–M'06–SM'11) received the B. Tech. degree in Electrical Engineering from the Indian Institute of Technology, Madras, the M. S. and Ph. D. degrees in Electrical and Computer Engineering from Purdue University and the University of California, San Diego. Currently, he is a Professor in the Department of Electrical Communication Engineering at the Indian Institute of Science, Bangalore, India. His research interests are in the areas of energy harvesting communications, 5G/6G technologies and compressed sensing. He is currently serving as a senior area editor for the IEEE Transactions on Signal Processing, and as an associate editor for the IEEE Transactions on Communications and the IEEE Transactions on Information Theory.