Stochastic analysis of algorithms for collecting longitudinal data

Frédérique Robin, Bruno Sericola, Emmanuelle Anceaume

Stochastic analysis of algorithms for collecting longitudinal data

Frédérique Robin[1] Bruno Sericola[2]
Emmanuelle Anceaume[3]
[1] Inria, Univ. Rennes, CNRS, IRISA Inria,
Campus de Beaulieu, 35042 Rennes Cedex, France
frederique.robin@inria.fr
[2] Inria, Univ. Rennes, CNRS, IRISA Inria,
Campus de Beaulieu, 35042 Rennes Cedex, France
bruno.sericola@inria.fr
[3] CNRS, Univ. Rennes, Inria, IRISA
Campus de Beaulieu, 35042 Rennes Cedex, France
emmanuelle.anceaume@irisa.fr

May 3, 2021

Abstract This paper proposes and analyses the performance and the vulnerability to attacks of three algorithms for collecting longitudinal data in a large scale system. A monitoring device is in charge of continuously collecting measurements from end-devices. The communication graph is connected but not necessarily complete. For scalability reasons, at each collect, a single end-device is randomly selected among all the end-devices to send the content of its local buffer of data to the monitoring device. Once sent, the end-device resets its buffer, and resumes its measurement process. Two of the three algorithms are randomized algorithms while the third one is deterministic. The difference between the randomized algorithms stems from the random choice policy: in the first algorithm, choice is uniform while in the second one the random choice is weighted by the current amount of measurements at end-devices. The third algorithm is deterministic. End-devices are successively chosen in a round robin way. We study the transient and stationary maximum load distribution at end-devices when collects are made using the first and third algorithm, and by providing bounds via a coupling argument when the second algorithm is used. While the third algorithm provides the best performance, it is highly vulnerable to attacks.

keywords Collecting longitudinal data, coupling technique, balls and urn models.

1 Introduction

The objective of this paper is to conduct a thorough analysis of three distributed algorithms (two rely on randomness and one is deterministic) dedicated to the collect of data through longitudinal measurements. A longitudinal study consists, through repeated observations over a long period of time, in collecting information of a cohort of individuals in order to observe the evolution of some measurable variables of interest. Longitudinal studies are highly relevant in epidemiological investigations (including the physiological changes in pregnancy described by [3, 5], or the follow-up of diabetics of a given age group studied by [8]), the early childhood educational programs addressed by [11], but also to measure and compare various business and branding initiatives, product feedback, or customer satisfaction. In longitudinal surveys, measurements (e.g., temperature, CO_2 level, physiological indicators, etc.) are collected over time and are interpreted. Interpretation is conducted once outliers are detected and possibly corrected (see [2]).

Internet-of-Things (IoT) technology, which is a growing network of interconnected objects, facilitates developments of such monitoring systems. As described by [4], in most of the IoT architectures, end-devices

measure data and route them to one or more processing centers through a network. Hence, in this paper, we consider a processing center that we call the monitoring device, and K end-devices. End-devices do not necessarily know each other, while the monitoring device knows and communicates with each of the K end-devices.

The load balancing problem is close to the data collect problem since the former one consists in distributing tasks among multiple resources (see [1] for more details). The main difference between both is that for longitudinal data collection, all end-devices perform their measurements at the same time when required by the monitoring device. In contrast, in the load balancing problem, only few tasks are allocated at the same time to end-devices and the processing time usually takes a non negligible amount of time. Note that the Age of Information, initially introduced by [7], is a measure determining whether a set of information is up-to-date for decision making purpose in communication system.

Providing fine-grain measurements in the data collect problem may require to frequently query end-devices to cope with characteristics that fluctuate very quickly. However, when the cohort of end-devices under study is densely populated, i.e. K is in thousands or millions, very frequent collects of all the end-devices may give rise to vast amount of data that transit towards the monitoring device. This may rapidly overload this device, preventing it accordingly from correctly handling the input stream of information. To cope with such critical issue, the monitoring device periodically samples a small subset of end-devices to collect their measurements instead of collecting data from all the K end-devices. When an end-device is selected, it sends back the content of its buffer of measurements to the monitoring device, empties its buffer and resumes its measurement process. In the following, during each collect, a single end-device is queried.

There are several ways to select end-devices, and the objective of this paper is to study in depth three algorithms that perform this task. The two first algorithms rely on randomization to select end-devices, while the third one is deterministic. In the first algorithm, called Algorithm A in the following, choice is random and uniform. In the second one, called Algorithm B, the choice is random but weighted by the current amount of measurements at end-devices. Finally, in the third algorithm, called Algorithm C, the choice is deterministic and follows a round-robin strategy: each end-device is successively chosen.

In this paper we are interested in analysing the effect of sampling on the size of the local buffer at each end-device at the end of each collect. Indeed, since during a collect a single end-device is queried for sending back its buffer content at the monitoring device, buffers at the other $K-1$ end-devices continue to grow. To evaluate the performance of these algorithms we study the transient and stationary maximum buffer size distribution at end-devices at the end of the n-th collect, for $n \geq 1$. We show that with Algorithm A, the limiting distribution of the buffer size at any end-device k is geometric with parameter $1/K$, the stationary distribution of the maximal buffer size of the K end-devices is upper bounded by $\ell \geq K$ with probability $K!S(\ell, K)/K^{\ell}$ where $S(\ell, K)$ are the Stirling numbers of the second kind, and the stationary average maximal buffer size is $\Theta(K \ln K)$. For Algorithm B, the impact of the random weighted sampling policy makes the moments of the buffer size at any end-devices not easy to obtain analytically. Hence by using a coupling argument, we provide bounds on the maximal and total buffer size. These bounds show that Algorithm B performs better than Algorithm A. Finally, by an easy argument, we show that Algorithm C provides better performance than Algorithm B, and thus better than Algorithm A.

Since we are considering large scale distributed architectures, the presence of malicious entities that devise adversarial strategies to prevent the correct functioning of algorithms is unavoidable (see [6]). In this work we consider an omniscient entity, called the adversary, that has full knowledge of the code run by the different devices, and in particular the one of the monitoring device. We focus on Deny-of-Services (DoS) attacks. A Denial of Service attack tries to take down an Internet resource by flooding this resource with more requests than it is capable of handling. In this work, we assume that the monitoring device includes advanced intrusion prevention and threat management systems, which combine firewalls, VPN, anti-spam, content filtering, load balancing, and other layers of DoS defense techniques. Together they enable constant and consistent network protection to prevent a DoS attack from happening. This includes everything from identifying possible traffic inconsistencies with the highest level of precision in blocking the attack. On the other hand, given the characteristics of end-devices, theses devices are not protected by such mechanisms, and thus are vulnerable to DoS attacks. One way for the adversary to target some victim is through sniffing.

Sniffing corresponds to theft or interception of data by capturing the network traffic using a sniffer (an application aimed at capturing network packets). When data is transmitted across networks, if the data packets are not encrypted, the data within the network packet can be read using a sniffer. By a simple argument, we will show that Algorithm C cannot protect end-devices from DoS attack, while randomization makes Algorithm A less vulnerable than Algorithm B.

The remaining of the paper is organized as follows. Section 2 describes the model of the system in terms of participating entities, and communication capability, as well as the main principles of the three collect algorithms. Section 3 describes the dynamic of the stochastic (resp. deterministic) process as an urn and balls model. Sections 4, 5, and 6 provide an in depth analysis of the algorithms. The trade-off between their performance and their vulnerability to deny-of-service attacks launched by an omniscient adversary is discussed in Section 7. Finally, Section 8 concludes.

2 Assumptions on the system and data collect algorithm

We consider a set of $K \geq 1$ end-devices, $1, 2, \ldots K$ with both measurement and communication capabilities. These K end-devices are monitored by another device, called monitoring device, whose role is to periodically collect measurements from these end-devices. The time needed for each end-device to make a measurement is negligible with respect to the collect periodicity. Those $K+1$ devices are interconnected through a connected but not necessarily complete communication network.

The monitoring device communicates with the end-devices by invoking a reliable broadcast primitive. Such a primitive guarantees that if the monitoring device broadcasts some message then all the end-devices will eventually receive this message. The end-devices communicate with the monitoring device by invoking a reliable send primitive. This primitive ensures that the sent message is eventually received by the monitoring device. Note that from an implementation point of view, since the communication graph is not necessarily complete, messages may need to be forwarded by intermediate end-devices before being received by the monitoring device.

All the devices have access to public key cryptography (PKC). PKC employs two keys, the private and public keys, that are mathematically related although knowledge of one key does not allow someone to easily determine the other key. Public keys (pk) are publicly known, while private keys (sk) are kept private by their owner. One key is used to encrypt the plaintext into something that appears to be random and meaningless (the ciphertext) and the other key is used to decrypt the ciphertext back to the plaintext. Let sk_k and pk_k be respectively the secret and public keys of device k. We rely on encryption to keep some pieces of information secret during transmission between the monitoring device and end-devices (this will be detailed later). In the following, encryption of message m is done with the public key of the recipient k of data m and is denoted by $\mathsf{enc}_{\mathsf{pk}_k}(m)$, while decryption at k is done with k's secret key and is denoted by $\mathsf{dec}_{\mathsf{sk}_k}(m)$.

2.1 Main principles of the collect algorithms

We propose three algorithms to collect measurements from a set of $K \geq 1$ end-devices. Only new measurements are collected, that is, once an end-device k has sent its buffer of measurements to the monitoring device, then k resets its buffer, and continues its data measurement process. So the next time end-device k will be queried by the monitoring device, k will (only) provide measurements it has performed since the last time it was queried.

2.2 Algorithm A

Algorithm A consists of a potentially infinite sequence of collects. Each collect is triggered by the monitoring device. At the n-th collect, $n \geq 1$, the monitoring device draws at random and uniformly an integer k in $\{1, \ldots, K\}$. Integer k represents the end-device that will provide its buffer of measurements to the monitoring device during this n-th collect. To prevent the adversary from discovering that end-device k is the one whose data will be collected, the monitoring device encrypts k's identifier using k's public key, i.e $\mathsf{enc}_{\mathsf{pk}_k}(k)$, and

broadcasts a collect query parameterized with $\mathsf{enc}_{\mathsf{pk}_k}(k)$ to all the end-devices. Upon receipt of this n-th collect query, all the K end-devices perform d new data measurements and store them in their buffer. However, prior to do that, end-device k that successively discovered that it is the one that must send its buffer of measurements to the monitoring device, i.e., $\mathsf{dec}_{\mathsf{sk}_k}(\mathsf{enc}_{\mathsf{pk}_k}(k))= k$, sends back its buffer to the monitoring device, and locally resets its. Thus at the end of collect n, the buffer of end-device k contains exactly d data, while for all the other end-devices, the size of their buffer is equal to the one after collect $n-1$ plus d. Without loss of generality, we assume in the following that $d = 1$.

2.3 Algorithm B

Algorithm B differs from Algorithm A only in the way the monitoring device randomly chooses the end-device from which it collects measurements. Specifically, at the n-th collect, $n \geq 1$, the monitoring device draws at random an integer k in $\{1, \ldots, K\}$, with a probability proportional to the size of the buffer of end-device k. Thus an end-device whose buffer of measurements has not been collected for a long time will be more likely to be chosen during this collect.

2.4 Algorithm C

Algorithm C queries end-devices according to the round-robin schema: at the n-th collect, $n \geq 1$, the chosen end-device is end-device $k = n \bmod (K) + K_{\{n \bmod (K)=0\}}$, where $K_{\{E\}}$ is the function equal to K if E is true and 0 otherwise. Thus each end-device is chosen in circular order without priority, starting by end-device $k = 1$ when $n = 1$.

3 Modeling Algorithms A, B, and C

We model those three algorithms to determine, at any collect $n \geq 1$, the amount of data measurements at any end-device k, i.e., the size of its buffer, denoted by buffer k, and the total amount of measurements over all the K buffers.

3.1 Modeling Algorithm A

Let $\{A(n),\ n \geq 0\}$ be a discrete time stochastic process, with $A(n) = (A_1(n), \ldots, A_K(n))$, where $A_k(n)$ represents the total amount of data measurements in the buffer of end-device k at the n-th collect. The dynamic of this stochastic process can be seen as an urn and balls problem. There are K urns, each one modeling the local buffer of end-devices. Balls represent the number of measurements in the local buffer of each end-device, i.e., one ball represents one measurement (recall that $d = 1$). We suppose that initially, i.e., at $n = 0$, all the local buffers contain one data measurement and the collect process start at $n \geq 1$. Thus at $n = 0$ there is one ball in each of the K urns.

At the n-th collect, end-device k is chosen randomly and uniformly, i.e. with probability $1/K$. All the balls of urn k are withdrawn from it, and one ball is added to each of the K urns (see Section 2.2). For $n \geq 1$, $A_k(n)$ is equal to the number of balls in urn k at collect n (i.e., is equal to the amount of data measurements in k's buffer at collect n).

We denote by $S_A(n)$ the total number of balls in the K urns at collect n, which thus corresponds to the total amount of measurements over all the K buffers at the end of the n-th collect. This quantity is defined by

$$S_A(n) = \sum_{k=1}^{K} A_k(n). \tag{1}$$

Let $(U_n)_{n\geq 0}$ be a sequence of random variables independent and identically uniformly distributed on interval $[0, 1]$. This sequence is used in Relation (2) to determine the urn which is selected at each instant. We suppose

that, for all $k = 1, \ldots, K$, random variables $A_k(n)$ and U_n are independent. For all $n \geq 0$, we define the sequence of random variables $J_A(U_n)$ by

$$J_A(U_n) = \sum_{k=1}^{K} k \mathbb{1}_{\{(k-1)/K \leq U_n < k/K\}}, \tag{2}$$

which gives the urn selected by the monitoring device, using Algorithm A, to collect its balls at time n. The evolution of process $\{A(n),\ n \geq 0\}$ is thus given, for all $k = 1, \ldots, K$, by $A_k(0) = 1$ and, for all $n \geq 1$, by

$$A_k(n) = \begin{cases} A_k(n-1) + 1 & \text{if} \quad J_A(U_{n-1}) \neq k \\ 1 & \text{if} \quad J_A(U_{n-1}) = k. \end{cases} \tag{3}$$

It is easily checked from this definition that for each $n \geq 1$, there exists a unique urn u such that $A_u(n) = 1$ and $A_k(n) \geq 2$, for $k = 1, \ldots, K$, with $k \neq u$.

To analyse the maximum amount of uncollected measurements at any end-device k, we introduce the process $\{A'(n),\ n \geq 0\}$, where $A'(n) = (A'_1(n), \ldots, A'_K(n))$, which is obtained by reordering the entries of vector $A(n)$ in the ascending order, that is with $A'_1(n) = 1 < A'_2(n) \leq \cdots \leq A'_K(n)$. More precisely, for all $k = 1, \ldots, K$, we define the $A'_k(n)$ by $A'_k(0) = 1$ and, for $n \geq 1$, by $A'_1(n) = 1$ and for $k = 2, \ldots, n$, by

$$A'_k(n) = \begin{cases} A'_{k-1}(n-1) + 1 & \text{if } J_A(U_{n-1}) \geq k \\ A'_k(n-1) + 1 & \text{if } J_A(U_{n-1}) \leq k-1. \end{cases}$$

Note that for the definition of process $\{A'(n),\ n \geq 0\}$, we keep the notation $J_A(U_n)$ since it does not depend on the number of balls in each urn but only on urn k. Note in particular that $S_{A'}(n) \stackrel{\mathcal{D}}{=} S_A(n)$, which means that $S_A(n)$ and $S_{A'}(n)$ have the same distribution. This relation can also be written as

$$A'_k(n) = A'_{k-1}(n-1)1_{\{J_A(U_{n-1}) \geq k\}} + A'_k(n-1)1_{\{J_A(U_{n-1}) \leq k-1\}} + 1. \tag{4}$$

3.2 Modeling Algorithm B

Let $\{B(n),\ n \geq 0\}$ be a discrete time stochastic process, with $B(n) = (B_1(n), \ldots, B_K(n))$, where $B_k(n)$ represents the total amount of uncollected measurements in the buffer of end-device k at the n-th collect. Similarly to Algorithm A, the dynamic of the stochastic process $\{B(n),\ n \geq 0\}$ can be represented by an urn and balls model. There are K urns, each one modeling the local buffer of end-devices. Balls represent the number of measurements in the local buffer of each end-device. Similarly to Algorithm A, we suppose that initially, i.e., at $n = 0$, all the local buffers contain d data measurements and the collect process start at $n \geq 1$. Recall that wlog we assume that $d = 1$, thus at $n = 0$ there is one ball in each of the K urns.

We denote by $S_B(n)$ the total number of balls in the K urns at collect n, which corresponds to the total number of uncollected measurements at the n-th collect. This quantity is defined by

$$S_B(n) = \sum_{k=1}^{K} B_k(n). \tag{5}$$

At the n-th collect, a node k is chosen with probability $B_k(n)/S_B(n)$, and all the balls of urn k are withdrawn from that urn, and one ball is added in each of the K urns. For $n \geq 1$, $B_k(n)$ is equal to the number of balls in urn k at collect n.

Let $(V_n)_{n \geq 0}$ be a sequence of random variables independent and identically uniformly distributed on interval $[0, 1]$. This sequence is used in Relation (7) to determine the urn which is selected at each instant. We suppose that, for all $k = 1, \ldots, K$, random variables $B_k(n)$ and V_n are independent. The evolution of process B is thus given by $B(0) = A(0) = (1, \ldots, 1)$ and, for $n \geq 1$ and $k = 1, \ldots, K$, by

$$B_k(n) = \begin{cases} B_k(n-1) + 1 & \text{if} \quad J_B(V_{n-1}) \neq k \\ 1 & \text{if} \quad J_B(V_{n-1}) = k, \end{cases} \tag{6}$$

where $J_B(V_n)$ is defined, for all $n \geq 0$, by

$$J_B(V_n) = \sum_{k=1}^{K} k \mathbb{1}_{\{s_B(n,k-1) \leq V_n < s_B(n,k)\}}, \text{ with } s_B(n,k) = \frac{1}{S_B(n)} \sum_{j=1}^{k} B_j(n). \tag{7}$$

Note that, as for Algorithm A, $J_B(V_n)$ is the urn selected by the monitoring device, using Algorithm B, to collect its balls at time n. As we did for stochastic process $\{A(n),\ n \geq 0\}$, we reorder the entries of random vector $(B_1(n), \ldots, B_K(n))$ in the ascending order. Thus, we introduce the process $\{B'(n),\ n \geq 0\}$, where $B'(n) = (B'_1(n), \ldots, B'_K(n))$, with $B'_k(0) = 1$ and, for $n \geq 1$, $B'_1(n) = 1 < B'_2(n) \leq \cdots \leq B'_K(n)$ is defined by

$$B'_1(n) = 1 \text{ and } B'_k(n) = \begin{cases} B'_{k-1}(n-1) + 1 & \text{for } k = 2, \ldots, J_B(V_{n-1}) \\ B'_k(n-1) + 1 & \text{for } k = J_B(V_{n-1}) + 1, \ldots, K. \end{cases} \tag{8}$$

In the same way the distribution of the total number of balls $S_{B'}(n)$ at collect n remains unchanged, that is $S_B(n) \stackrel{\mathcal{D}}{=} S_{B'}(n)$.

3.3 Modeling Algorithm C

Let $\{C(n),\ n \geq 0\}$ be a discrete time deterministic process, with $C(n) = (C_1(n), \ldots, C_K(n))$, where $C_k(n)$ represents the total amount of uncollected measurements in the buffer of end-device k at the n-th collect. Since Algorithm C queries end-devices according to the round-robin schema, at the n-th collect, $n \geq 1$, the chosen end-device (or the chosen urn) is end-device (or urn) $k = n \bmod (K) + K_{\{n \bmod (K)=0\}}$. For $n \geq 1$, $C_k(n)$ is equal to the number of balls in urn k at collect n.

The sequence $C(n)$ is thus given by $C(0) = (1, \ldots, 1)$, for $n = 1, \ldots K-1$ by

$$C_k(n) = \begin{cases} n & \text{if} \quad k = 1 \\ C_{k-1}(n-1) & \text{if} \quad k = 2, \ldots, n-1 \\ 1 & \text{if} \quad k = n \\ C_{k-1}(n-1) + 1 & \text{if} \quad k = n+1, \ldots, K, \end{cases}$$

and, for all $n \geq K$ and $k = 1, \ldots, K$, by

$$C_k(n) = \begin{cases} n \bmod (K) + K_{\{n \bmod (K)=0\}} & \text{if} \quad k = 1 \\ C_{k-1}(n-1) & \text{if} \quad k = 2, \ldots, n-1. \end{cases}$$

For instance, for $K = 5$, we obtain

$$\begin{array}{lll} C(0) = (1,1,1,1,1), & C(5) = (5,4,3,2,1), & C(10) = (5,4,3,2,1), \quad \cdots \\ C(1) = (1,2,2,2,2), & C(6) = (1,5,4,3,2), & C(11) = (1,5,4,3,2), \quad \cdots \\ C(2) = (2,1,3,3,3), & C(7) = (2,1,5,4,3), & C(12) = (2,1,5,4,3), \quad \cdots \\ C(3) = (3,2,1,4,4), & C(8) = (3,2,1,5,4), & C(13) = (3,2,1,5,4), \quad \cdots \\ C(4) = (4,3,2,1,5), & C(9) = (4,3,2,1,5), & C(14) = (4,3,2,1,5), \quad \cdots \end{array}$$

We denote by $S_C(n)$ the total number of balls in the K urns at collect n, which corresponds to the total number of uncollected measurements at the n-th collect. This quantity is defined by

$$S_C(n) = \sum_{k=1}^{K} C_k(n). \tag{9}$$

As we did for stochastic processes $\{A(n),\ n \geq 0\}$ and $\{B(n),\ n \geq 0\}$, we reorder the entries of random vector $(C_1(n), \ldots, C_K(n))$ in the ascending order. Thus, we introduce the process $\{C'(n),\ n \geq 0\}$, where $C'(n) = (C'_1(n), \ldots, C'_K(n))$, with $C'(0) = (1, \ldots, 1)$ and, for $n \geq 1$,

$$C'_1(n) = 1 \text{ and } C'_k(n) = C'_k(n-1) + 1, \text{ for } k = 2, \ldots, K-2.$$

For $K = 5$, we easily get $C'(0) = (1,1,1,1,1)$, $C'(1) = (1,2,2,2,2)$, $C'(2) = (1,2,3,3,3)$, $C'(3) = (1,2,3,4,4)$ and $C'(n) = (1,2,3,4,5)$, for all $n \geq 4$. In the same way the total number of balls $S_{C'}(n)$ at collect n remains unchanged, that is $S_C(n) = S_{C'}(n)$. This quantity is easily obtained by

$$S_C(n) = S_{C'}(n) = (n+1)\left(K - \frac{n}{2}\right) 1_{\{0 \leq n \leq K-2\}} + \frac{K(K+1)}{2} 1_{\{n \geq K-1\}}. \tag{10}$$

4 Analysis of Algorithm A

We study in this section the stochastic processes $\{A(n),\ n \geq 0\}$ and $\{A'(n),\ n \geq 0\}$.

4.1 Distribution of $A_k(n)$

We start by analyzing the amount of uncollected measurements at any end-device k at any collect $n \geq 1$.

Theorem 4.1 *For all $n \geq 0$, $k = 1, \ldots, K$ and $\ell = 1, \ldots, n+1$, we have*

$$\mathbb{P}\{A_k(n) = \ell\} = \left(1 - \frac{1}{K}\right)^{\ell-1} \left[\frac{1}{K} 1_{\{\ell \leq n\}} + 1_{\{\ell = n+1\}}\right].$$

Proof. Proof. From the definition of $A_k(n)$ in Relation (3), we have, for all $k = 1, \ldots, K$, $A_k(0) = 1$ and for all $n \geq 1$,

$$A_k(n) = (A_k(n-1)+1) 1_{\{J_A(U_{n-1}) \neq k\}} + 1_{\{J_A(U_{n-1}) = k\}} = A_k(n-1) 1_{\{J_A(U_{n-1}) \neq k\}} + 1.$$

Note that this relation implies that $A_k(n) \leq A_k(n-1) + 1$, which means that the random variable $A_k(n)$ takes its values in the set $\{1, \ldots, n+1\}$.

By conditioning on $A_k(n-1)$ and using the fact that $A_k(n-1)$ and U_{n-1} are independent, we obtain, for all $\ell, j \geq 1$,

$$\begin{aligned}
\mathbb{P}\{A_k(n) = \ell \mid A_k(n-1) = j\} &= \mathbb{P}\{A_k(n-1) 1_{\{J_A(U_{n-1}) \neq k\}} + 1 = \ell \mid A_k(n-1) = j\} \\
&= \mathbb{P}\{1_{\{J_A(U_{n-1}) \neq k\}} = (\ell-1)/j \mid A_k(n-1) = j\} \\
&= \mathbb{P}\{1_{\{J_A(U_{n-1}) \neq k\}} = (\ell-1)/j\} \\
&= \begin{cases} \mathbb{P}\{J_A(U_{n-1}) = k\} & \text{if } \ell = 1 \\ \mathbb{P}\{J_A(U_{n-1}) \neq k\} & \text{if } \ell \geq 2 \text{ and } j = \ell - 1 \\ 0 & \text{otherwise.} \end{cases} \\
&= \begin{cases} 1/K & \text{if } \ell = 1 \\ 1 - 1/K & \text{if } j = \ell - 1 \text{ and } \ell \geq 2 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Unconditioning, we obtain $\mathbb{P}\{A_k(n) = 1\} = 1/K$ and, for $\ell \geq 2$,

$$\mathbb{P}\{A_k(n) = \ell\} = \left(1 - \frac{1}{K}\right) \mathbb{P}\{A_k(n-1) = \ell - 1\}.$$

This simple recurrence relation leads to

$$\mathbb{P}\{A_k(n) = \ell\} = \begin{cases} \dfrac{1}{K}\left(1 - \dfrac{1}{K}\right)^{\ell-1} & \text{if } 1 \leq \ell \leq n \\ \left(1 - \dfrac{1}{K}\right)^{n} & \text{if } \ell = n+1, \end{cases}$$

which completes the proof. ■

Observe that, at each collect instant n, the amount of measurements $A_k(n)$, collected at any device k has the same distribution for $k = 1, \ldots, K$. This is due to the fact that the U_n are uniformly distributed. Note that $A_1(n), \ldots, A_K(n)$ are not independent. We get from Theorem 4.1 that, as expected, the uncollected measurements of any end-device k follows a kind of geometric law with parameter $1/K$.

Corollary 4.2 highlights that the expected amount of uncollected measurements in any buffer at collect n is not exactly K but increases towards K when the number n of collects goes to infinity. It also shows that the expected total amount of uncollected measurements at collect n, $\mathbb{E}(S_A(n))$, converges increasingly to K^2 when the number of collects n goes to infinity.

Corollary 4.2 *For all $k = 1, \ldots, K$, we have*

$$\mathbb{E}(A_k(n)) = K\left(1 - \left(1 - \frac{1}{K}\right)^{n+1}\right) \quad \textit{and} \quad \mathbb{E}(S_A(n)) = K^2\left(1 - \left(1 - \frac{1}{K}\right)^{n+1}\right).$$

Moreover, $\lim_{n\longrightarrow\infty} \mathbb{E}(A_k(n)) = K$ *and* $\lim_{n\longrightarrow\infty} \mathbb{E}(S_A(n)) = K^2$.

Proof. Proof. From Theorem 4.1, we have, for all $\ell = 0, \ldots, n$,

$$\mathbb{P}\{A_k(n) > \ell\} = \sum_{j=\ell+1}^{n+1} \mathbb{P}\{A_k(n) = j\} = \frac{1}{K}\sum_{j=\ell+1}^{n}\left(1 - \frac{1}{K}\right)^{j-1} + \left(1 - \frac{1}{K}\right)^{n} = \left(1 - \frac{1}{K}\right)^{\ell}.$$

We thus get

$$\mathbb{E}(A_k(n)) = \sum_{\ell=0}^{n} \mathbb{P}\{A_k(n) > \ell\} = \sum_{\ell=0}^{n}\left(1 - \frac{1}{K}\right)^{\ell} = K\left(1 - \left(1 - \frac{1}{K}\right)^{n+1}\right).$$

The second equality follows immediately from Relation (1) and the limits are trivial. ■

Corollary 4.3 shows that the limiting distribution of the uncollected measurements of end-device k is geometric with parameter $1/K$.

Corollary 4.3 *For all $k = 1, \ldots, K$ and $\ell \geq 1$, we have*

$$\lim_{n\longrightarrow\infty} \mathbb{P}\{A_k(n) = \ell\} = \left(1 - \frac{1}{K}\right)^{\ell-1}\frac{1}{K}.$$

Proof. Proof. The proof is straightforward from Theorem 4.1 and Corollary 4.2. ■

4.2 Distribution of $A'_k(n)$

In this section, we study the maximum amount of uncollected measurements distribution of end-devices, that is the distribution of $A'_K(n)$. Recall that $A'_1(n) = 1$ for all $n \geq 0$ and that $A'_2(n) \leq \ldots \leq A'_K(n)$, with $A'_k(0) = 1$ for all $k = 1, \ldots, K$ and $2 \leq A'_k(n) \leq n + 1$ for all $n \geq 1$. The distribution of $A'_2(n)$ is given in Theorem 4.4.

Theorem 4.4 $A'_2(0) = 1$ *and for all $n \geq 1$ and $\ell = 2, \ldots, n + 1$, we have*

$$\mathbb{P}\{A'_2(n) = \ell\} = \left(\frac{1}{K}\right)^{\ell-2}\left[\left(1 - \frac{1}{K}\right)1_{\{\ell\leq n\}} + 1_{\{\ell=n+1\}}\right].$$

Proof. Proof. By definition of $A'_k(n)$, see Relation (4), we have $A'_2(0) = 1$ and for all $n \geq 1$,

$$A'_2(n) = 1_{\{J_A(U_{n-1})\geq 2\}} + A'_2(n-1)1_{\{J_A(U_{n-1})=1\}} + 1.$$

By conditioning on $A'_2(n-1)$ and using the fact that $A'_k(n-1)$ and U_{n-1} are independent, we obtain, for all $\ell \geq 1$ and $j \geq 2$,

$$\begin{aligned}
\mathbb{P}\{A'_2(n) = \ell \mid A'_2(n-1) = j\} &= \mathbb{P}\{1_{\{J_A(U_{n-1})\geq 2\}} + j1_{\{J_A(U_{n-1})=1\}} + 1 = \ell\} \\
&= \begin{cases} \mathbb{P}\{J_A(U_{n-1}) \geq 2\} & \text{if } \ell = 2 \\ \mathbb{P}\{J_A(U_{n-1}) = 1\} & \text{if } \ell \geq 3 \text{ and } j = \ell - 1 \\ 0 & \text{otherwise.} \end{cases} \\
&= \begin{cases} 1 - 1/K & \text{if } \ell = 2 \\ 1/K & \text{if } \ell \geq 3 \text{ and } j = \ell - 1 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Unconditioning, we obtain $\mathbb{P}\{A'_2(n) = 2\} = 1 - 1/K$ and, for $\ell \geq 3$,

$$\mathbb{P}\{A'_2(n) = \ell\} = \frac{1}{K}\mathbb{P}\{A'_2(n-1) = \ell - 1\}.$$

This simple recurrence relation leads, for $\ell = 2, \ldots, n+1$, to

$$\mathbb{P}\{A'_2(n) = \ell\} = \left(\frac{1}{K}\right)^{\ell-2}\left[\left(1 - \frac{1}{K}\right)1_{\{\ell\leq n\}} + 1_{\{\ell=n+1\}}\right],$$

which completes the proof. ■

The next result gives the expected value of $A'_2(n)$.

Corollary 4.5 *For all $n \geq 0$, we have*

$$\mathbb{E}(A'_2(n)) = 1 + \frac{K\left(1 - (1/K)^n\right)}{K-1}.$$

Proof. Proof. The relation is true for $n = 0$. For $n \geq 1$, From Theorem 4.4, we have, for all $\ell = 1, \ldots, n$,

$$\mathbb{P}\{A'_2(n) > \ell\} = \sum_{j=\ell+1}^{n+1} \mathbb{P}\{A'_2(n) = j\} = \left(1 - \frac{1}{K}\right)\sum_{j=\ell+1}^{n}\left(\frac{1}{K}\right)^{j-2} + \left(\frac{1}{K}\right)^{n-1} = \left(\frac{1}{K}\right)^{\ell-1}.$$

We thus get

$$\mathbb{E}(A'_2(n)) = \sum_{\ell=0}^{n}\mathbb{P}\{A'_2(n) > \ell\} = 1 + \sum_{\ell=1}^{n}\left(\frac{1}{K}\right)^{\ell-1} = 1 + \frac{K\left(1 - (1/K)^n\right)}{K-1},$$

which completes the proof. ■

The distribution of $A'_k(n)$ for all k is given in Theorem 4.6.

Theorem 4.6 *For all $k \geq 2$, we have $A'_k(1) = 2$. For all $k \geq 3$, $n \geq 2$ and $\ell = 3, \ldots, n+1$, we have*

$$\mathbb{P}\{A'_k(n) = \ell\} = \left(1 - \frac{k-1}{K}\right)\mathbb{P}\{A'_{k-1}(n-1) = \ell - 1\} + \frac{k-1}{K}\mathbb{P}\{A'_k(n-1) = \ell - 1\}.$$

Proof. Proof. By definition of $A'_k(n)$ given in Relation (4), we have $A'_k(0) = 1$, for all $k \geq 1$, anf for all $k \geq 2$ and $n \geq 1$,

$$A'_k(n) = A'_{k-1}(n-1)1_{\{J_A(U_{n-1})\geq k\}} + A'_k(n-1)1_{\{J_A(U_{n-1})\leq k-1\}} + 1.$$

Since $A'_{k-1}(0) = 1$ and $A'_k(0) = 1$, we obtain $A'_k(1) = 2$.

Now, for $k \geq 3$, by conditioning on $A'_{k-1}(n-1)$ and $A'_k(n-1)$ and using the fact that $A'_k(n-1)$ and U_{n-1} are independent, we obtain, for all $n \geq 2$ and $\ell, i, j \geq 2$,

$$\begin{aligned}\mathbb{P}\{A'_k(n) = \ell \mid A'_{k-1}(n-1) = i, A'_k(n-1) = j\} &= \mathbb{P}\{i1_{\{J_A(U_{n-1}) \geq k\}} + j1_{\{J_A(U_{n-1}) \leq k-1\}} = \ell - 1\} \\ &= \begin{cases} 0 & \text{if } \ell = 2 \\ 1 - (k-1)/K & \text{if } \ell \geq 3 \text{ and } i = \ell - 1 \\ (k-1)/K & \text{if } \ell \geq 3 \text{ and } j = \ell - 1 \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

Unconditioning, we obtain for all $n \geq 2$ and $\ell \geq 3$,

$$\mathbb{P}\{A'_k(n) = \ell\} = \left(1 - \frac{k-1}{K}\right) \mathbb{P}\{A'_{k-1}(n-1) = \ell - 1\} + \frac{k-1}{K} \mathbb{P}\{A'_k(n-1) = \ell - 1\}.$$

which completes the proof. ■

The expected value of $A'_k(n)$ is obtained by recurrence in the following result.

Corollary 4.7 *For all $n \geq 0$, we have $\mathbb{E}(A'_1(n)) = 1$. For all $k = 1, \ldots, K$, we have $\mathbb{E}(A'_k(0)) = 1$. For all $n \geq 1$ and $k = 2, \ldots, K$, we have*

$$\mathbb{E}(A'_k(n)) = \left(1 - \frac{k-1}{K}\right) \mathbb{E}(A'_{k-1}(n-1)) + \frac{k-1}{K} \mathbb{E}(A'_k(n-1)) + 1.$$

Proof. Proof. Let $n \geq 1$ and $k \geq 2$. By taking the expectation in Relation (4) and using the fact that $A'_k(n-1)$ and U_{n-1} are independent, we obtain

$$\begin{aligned}\mathbb{E}(A'_k(n)) &= \mathbb{E}(A'_{k-1}(n-1))\mathbb{P}\{J_A(U_{n-1}) \geq k\} + \mathbb{E}(A'_k(n-1))\mathbb{P}\{J_A(U_{n-1}) \leq k-1\} + 1 \\ &= \left(1 - \frac{k-1}{K}\right) \mathbb{E}(A'_{k-1}(n-1)) + \frac{k-1}{K} \mathbb{E}(A'_k(n-1)) + 1,\end{aligned}$$

which completes the proof. ■

Using Theorem 4.6, we compute in Figure 1, $\mathbb{P}\{A'_K(n) \geq \ell\}$ for different values of K and n.

We analyze in the following theorem the limiting behavior of the distribution of the $A'_k(n)$. We denote by $\pi_k(\ell)$, for all $k = 1, \ldots, K$, this stationnary distribution, that is

$$\pi_k(\ell) = \lim_{n \longrightarrow \infty} \mathbb{P}\{A'_k(n) = \ell\}.$$

The existence of this limit follows by recurrence as shown in Theorem 4.8.

Theorem 4.8 *We have $\pi_1(\ell) = 1_{\{\ell=1\}}$ and, for $k = 2, \ldots, K$,*

$$\pi_k(\ell) = \left[\prod_{r=1}^{k-1} \left(1 - \frac{r}{K}\right)\right] \sum_{r=1}^{k-1} a_r(k) \left(\frac{r}{K}\right)^{\ell-k} 1_{\{\ell \geq k\}}, \tag{11}$$

where the coefficients $a_r(k)$ are given by $a_1(2) = 1$ and, for $k \geq 3$,

$$\begin{cases} a_r(k) & = -\dfrac{r}{k-r-1} a_r(k-1) \quad \text{for } r = 1, \ldots, k-2 \\ a_{k-1}(k) & = 1 - \displaystyle\sum_{r=1}^{k-2} a_r(k). \end{cases}$$

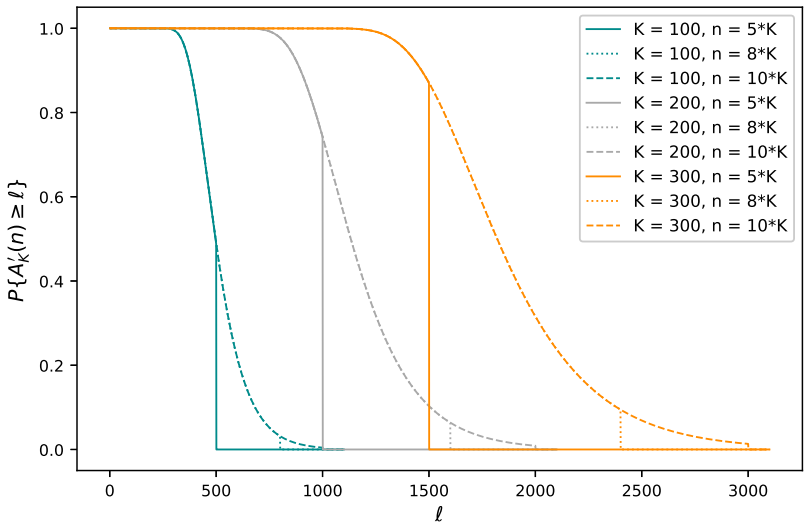

Figure 1: Using Theorem 4.6, we compute $\mathbb{P}\{A'_K(n) \geq \ell\}$ for different values of K ($K = 100, 200$ and 300) and at different collects n ($n = 5K, 8K$ and $10K$). The plain lines represent $\mathbb{P}\{A'_{5K}(n) \geq \ell\}$, the dotted lines represent $\mathbb{P}\{A'_{8K}(n) \geq \ell\}$ and the dashed lines represent $\mathbb{P}\{A'_{10K}(n) \geq \ell\}$.

Proof. Proof. For $k = 1$, the result is trivial since $A'_1(n) = 1$, for all $n \geq 0$. For $k = 2$, by taking the limit when n tends to infinity in Theorem 4.4, we easily get

$$\pi_2(\ell) = \left(1 - \frac{1}{K}\right)\left(\frac{1}{K}\right)^{\ell-2} 1_{\{\ell \geq 2\}}, \tag{12}$$

which is exactly Relation (11) when $k = 2$.

For $k \geq 3$, we take the limit when n tends to infinity in Theorem 4.6. We obtain

$$\pi_k(\ell) = \frac{k-1}{K}\pi_k(\ell - 1) + \left(1 - \frac{k-1}{K}\right)\pi_{k-1}(\ell - 1). \tag{13}$$

It is easily checked by recurrence, using Relation (12), that $\pi_k(\ell) = 0$, for $\ell \leq k - 1$. Using this property in Relation (13) with $\ell = k$, we get

$$\pi_k(k) = \frac{k-1}{K}\pi_k(k-1) + \left(1 - \frac{k-1}{K}\right)\pi_{k-1}(k-1) = \left(1 - \frac{k-1}{K}\right)\pi_{k-1}(k-1),$$

which leads to

$$\pi_k(k) = \left(1 - \frac{k-1}{K}\right)\cdots\left(1 - \frac{2}{K}\right)\pi_2(2).$$

Since, from Relation (12), we have $\pi_2(2) = 1 - 1/K$, it follows that

$$\pi_k(k) = \prod_{r=1}^{k-1}\left(1 - \frac{r}{K}\right).$$

We now verify that the right hand side of Relation (11) satisfies Relation (13). In order to do that, we denote by $p_k(\ell)$ the right-hand side of (11), that is

$$p_k(\ell) = \left[\prod_{r=1}^{k-1}\left(1 - \frac{r}{K}\right)\right]\sum_{r=1}^{k-1} a_r(k)\left(\frac{r}{K}\right)^{\ell-k} 1_{\{\ell \geq k\}},$$

where $a_1(2) = 1$ and, for $k \geq 3$,

$$a_r(k) = -\frac{r}{k-r-1} a_r(k-1) \qquad \text{for } r = 1, \ldots, k-2 \tag{14}$$

$$a_{k-1}(k) = 1 - \sum_{r=1}^{k-2} a_r(k), \tag{15}$$

For $\ell = k$, we have, using Relation (15),

$$p_k(k) = \left[\prod_{r=1}^{k-1} \left(1 - \frac{r}{K}\right)\right] \sum_{r=1}^{k-1} a_r(k) = \prod_{r=1}^{k-1} \left(1 - \frac{r}{K}\right) = \pi_k(k).$$

For $\ell \geq k+1$, we have

$$\begin{aligned}
\frac{k-1}{K} p_k(\ell-1) &= \pi_k(k) \sum_{r=1}^{k-1} a_r(k) \frac{k-1}{K} \left(\frac{r}{K}\right)^{\ell-1-k} \\
&= \pi_k(k) \sum_{r=1}^{k-1} a_r(k) \frac{k-1}{r} \left(\frac{r}{K}\right)^{\ell-k} \\
&= \pi_k(k) \left[\sum_{r=1}^{k-2} a_r(k) \frac{k-1}{r} \left(\frac{r}{K}\right)^{\ell-k} + a_{k-1}(k) \left(\frac{k-1}{K}\right)^{\ell-k}\right].
\end{aligned}$$

In the same way, we have, for $\ell \geq k+1$,

$$\left(1 - \frac{k-1}{K}\right) p_{k-1}(\ell-1) = \pi_k(k) \sum_{r=1}^{k-2} a_r(k-1) \left(\frac{r}{K}\right)^{\ell-k}.$$

By adding these two expressions, we obtain

$$\begin{aligned}
&\frac{k-1}{K} p_k(\ell-1) + \left(1 - \frac{k-1}{K}\right) p_{k-1}(\ell-1) = \\
&\pi_k(k) \left[\sum_{r=1}^{k-2} \left(a_r(k) \frac{k-1}{r} + a_r(k-1)\right) \left(\frac{r}{K}\right)^{\ell-k} + a_{k-1}(k) \left(\frac{k-1}{K}\right)^{\ell-k}\right].
\end{aligned}$$

Observing that Relation (14) can be written, for $r = 1, \ldots, k-2$, as

$$a_r(k) = a_r(k) \frac{k-1}{r} + a_r(k-1),$$

we get

$$\frac{k-1}{K} p_k(\ell-1) + \left(1 - \frac{k-1}{K}\right) p_{k-1}(\ell-1) = p_k(\ell),$$

which proves that $\pi_k(\ell) = p_k(\ell)$. ■

Explicit expressions of coefficients $a_r(k)$ and properties of these coefficients are given in the following corollary. Let us first recall that the Stirling numbers of the second kind are defined by $S(\ell, 1) = S(\ell, \ell) = 1$ for $\ell \geq 1$ and by $S(\ell, k) = kS(\ell-1, k) + S(\ell-1, k-1)$, for $\ell \geq k$.

Corollary 4.9 shows that the distribution of all the order statistics of the $A_k(n)$ are easily obtained using the Stirling numbers of the second kind.

Corollary 4.9 *For all $k = 2, \ldots, K$ and $r = 1, \ldots, k-1$, we have*

$$a_r(k) = \frac{(-1)^{k-1-r} r^{k-1}}{r!(k-1-r)!}. \tag{16}$$

For all $k = 2, \ldots, K$ and $\ell \geq k$, we have

$$\lim_{n \longrightarrow \infty} \mathbb{P}\{A'_k(n) = \ell\} = \pi_k(\ell) = \frac{(K-1)! S(\ell-1, k-1)}{(K-k)! K^{\ell-1}}, \tag{17}$$

and

$$\lim_{n \longrightarrow \infty} \mathbb{P}\{A'_k(n) \geq \ell\} = \sum_{m=\ell}^{\infty} \pi_k(m) = \frac{(K-1)!}{(K-k)! K^{\ell-2}} \sum_{r=1}^{k-1} \frac{(-1)^{k-1-r} r^{\ell-1}}{r!(k-1-r)!(K-r)}. \tag{18}$$

Proof. Proof. It is easily checked that Relation (16) satisfies $a_1(2) = 1$ and the recursive definition of the $a_r(k)$ given in Theorem 4.8, for $k \geq 3$ and $r = 1, \ldots, k-2$. It remains to study the case $r = k-1$. In that case, using the Euler's finite difference theorem (see for instance Relation (6.19) of [13]), which tells us that

$$\sum_{r=1}^{k-1} \frac{(-1)^{k-1-r} r^{k-1}}{r!(k-1-r)!} = 1, \quad \text{for} \quad k \geq 2$$

we get

$$a_{k-1}(k) = 1 - \sum_{r=1}^{k-2} a_r(k) = \frac{(k-1)^{k-1}}{(k-1)!},$$

which completes the proof of Relation (16).
In order to prove Relation (17), observe first that

$$\prod_{r=1}^{k-1} \Big(1 - \frac{r}{K}\Big) = \prod_{r=1}^{k-1} \Big(\frac{K-r}{K}\Big) = \frac{(K-1)!}{(K-k)! K^{k-1}}.$$

Using this relation and the expression of the $a_r(k)$ obtained in Relation (16) that we insert in Relation (11), we obtain, for all $k = 2, \ldots, K$ and $\ell \geq k$,

$$\begin{aligned} \pi_k(\ell) &= \frac{(K-1)!}{(K-k)! K^{k-1}} \sum_{r=1}^{k-1} \frac{(-1)^{k-1-r} r^{k-1}}{r!(k-1-r)!} \Big(\frac{r}{K}\Big)^{\ell-k} \\ &= \frac{(K-1)!}{(K-k)! K^{\ell-1}} \sum_{r=1}^{k-1} \frac{(-1)^{k-1-r}}{r!(k-1-r)!} r^{\ell-1}. \end{aligned} \tag{19}$$

The Euler's finite difference theorem (see for instance Relation (6.20) of [13]) also tells us that for $\ell \geq k$, we have

$$\sum_{r=1}^{k-1} \frac{(-1)^{k-1-r} r^{\ell-1}}{r!(k-1-r)!} = S(\ell-1, k-1), \tag{20}$$

where $S(\ell, k)$ are the Stirling numbers of the second kind defined above. Using this relation, we get

$$\pi_k(\ell) = \frac{(K-1)! S(\ell-1, k-1)}{(K-k)! K^{\ell-1}}.$$

To prove Relation (18), we use Relation (19) which leads to

$$\begin{aligned}\sum_{m=\ell}^{\infty}\pi_k(m) &= \frac{(K-1)!}{(K-k)!}\sum_{r=1}^{k-1}\frac{(-1)^{k-1-r}}{r!(k-1-r)!}\sum_{m=\ell}^{\infty}\Big(\frac{r}{K}\Big)^{m-1}\\ &= \frac{(K-1)!}{(K-k)!}\sum_{r=1}^{k-1}\frac{(-1)^{k-1-r}}{r!(k-1-r)!}\Big(\frac{r}{K}\Big)^{\ell-1}\frac{K}{K-r}\\ &= \frac{(K-1)!}{(K-k)!K^{\ell-2}}\sum_{r=1}^{k-1}\frac{(-1)^{k-1-r}r^{\ell-1}}{r!(k-1-r)!(K-r)},\end{aligned}$$

which completes the proof. ■

Theorem 4.10 *The stationary distribution of the maximal amount of uncollected data over the K end-devices is upper bounded by $\ell \geq K$ with probability $K!S(\ell,K)/K^{\ell}$. More precisely, we have*

$$\lim_{n\longrightarrow\infty}\mathbb{P}\{A'_K(n)\leq\ell\} = \sum_{m=K}^{\ell}\pi_K(m) = \frac{K!S(\ell,K)}{K^{\ell}},$$

Proof. Proof. By definition of the $A'_k(n)$, the maximal amount of uncollected data is obtained by end-device number K. Thus, applying Relation (18) of Corollary 4.9 with $k=K$ and $\ell+1$ instead of ℓ, we obtain for all $\ell \geq K$,

$$\sum_{m=\ell+1}^{\infty}\pi_K(m) = \frac{(K-1)!}{K^{\ell-1}}\sum_{r=1}^{K-1}\frac{(-1)^{K-1-r}r^{\ell}}{r!(K-r)!} = \frac{(K-1)!}{K^{\ell-1}}\left[-\sum_{r=1}^{K}\frac{(-1)^{K-r}r^{\ell}}{r!(K-r)!}+\frac{K^{\ell}}{K!}\right].$$

Using Relation (20), we get

$$\sum_{m=\ell+1}^{\infty}\pi_K(m) = 1-\frac{(K-1)!S(\ell,K)}{K^{\ell-1}} = 1-\frac{K!S(\ell,K)}{K^{\ell}},$$

which completes the proof. ■

Figure 2 illustrates, for Algorithm A, the stationary maximal amount of uncollected data over the K end-devices (see Theorem 4.10). Specifically, let $T(\ell,K)$ be defined, for all $\ell \geq K$, by

$$T(\ell,K) = \frac{K!S(\ell,K)}{K^{\ell}},$$

and let $\ell(K,\varepsilon)$ be the smallest value of ℓ for which $T(\ell,K)$ is greater than or equal to $1-\varepsilon$, for $\varepsilon \in (0,1)$, that is

$$\ell(K,\varepsilon) = \inf\{j \mid T(j,K) \geq 1-\varepsilon\}.$$

Observe that $T(\ell,K)$ can be computed, for $\ell \geq K$, as

$$T(\ell,K) = T(\ell-1,K) + \left(1-\frac{1}{K}\right)^{\ell-1}T(\ell-1,K-1)$$

with $T(\ell,2) = 1-(1/2)^{\ell-1}$, for $\ell \geq 2$. For instance, Figure 2 indicates that when $K=3$ and $\varepsilon = 10^{-3}$, the amount of accumulated measurements at any end-device does not exceed 11 with probability greater than or equal to 0.999. Table 1 gives the values of $\ell(K,\varepsilon=10^{-3})$ for some large values of K.

To complete our analysis, we give in Corollary 4.11 the expected value of the stationary maximal amount of measurements collected at any end-device.

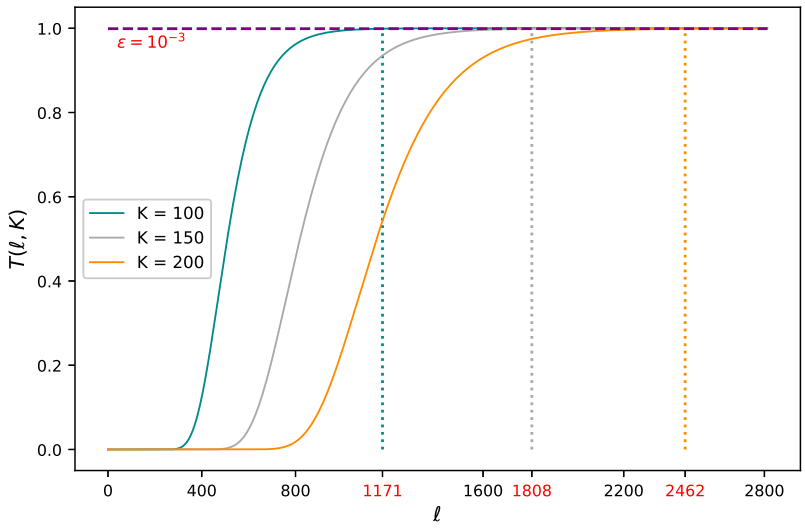

Figure 2: Illustration of the stationary maximal amount of uncollected data over the K end-devices. For $K = 100, 150$ and 200 and for each $\ell = K, \ldots, 2,800$, we compute $T(\ell, K)$ and thresholds $\ell(K, \varepsilon)$ for $\varepsilon = 10^{-3}$ (i.e., $\ell(100K, \varepsilon) = 1171$, $\ell(150K, \varepsilon) = 1808$, and $\ell(200K, \varepsilon) = 2462$). The purple line represents $1-\varepsilon = 0.999$.

K	250	500	750	1,000	10,000
$\ell(K, 10^{-3})$	3,101	6,555	10,139	13,809	161,168

Table 1: Thresholds $\ell(K, \varepsilon)$ of the stationary maximal amount of measurements collected at any end-device for large values of K when $\varepsilon = 10^{-3}$.

Corollary 4.11 *For all $k = 1, \ldots, K$, we have*

$$\lim_{n \longrightarrow \infty} \mathbb{E}(A'_k(n)) = K \sum_{\ell=K-k+1}^{K} \frac{1}{\ell}.$$

Proof. Proof. In order to simplify the writing we introduce the notation $x_k = \lim_{n \longrightarrow \infty} \mathbb{E}(A'_k(n))$. By taking the limit when n tends to infinity in Corollary 4.7, we get $x_1 = 1$ and, for all $k = 2, \ldots, K$,

$$x_k = \left(1 - \frac{k-1}{K}\right) x_{k-1} + \frac{k-1}{K} x_k + 1,$$

which can be written as $x_k = x_{k-1} + K$, that is

$$x_k = K \sum_{\ell=K-k+1}^{K} \frac{1}{\ell},$$

which completes the proof. ■

Taking $k = K$, we get from Corollary 4.11 that the average of the maximum amount of uncollected measurement $A'_K(n)$ converges to $K \sum_{\ell=1}^{K} 1/\ell$ when n goes to infinity. This indicates that the stationary average maximum uncollected measurements is $\Theta(K \ln K)$.

5 Analysis of Algorithm B

This section is devoted to the study of the impact of the "random weighted sampling policy" on the distribution and the moments of the uncollected measurements of any end-devices. Actually both of them are not easy to obtain analytically, nevertheless we show in this section that stochastic process $\{A(n),\ n \geq 0\}$ is, in terms of distribution, an upper bound of stochastic process $\{B(n),\ n \geq 0\}$.

To derive an upper bound of stochastic process $\{B(n),\ n \geq 0\}$ from process $\{A(n),\ n \geq 0\}$, we apply a coupling technique consisting in using the same sequences of independent and uniformly distributed random variables for both processes. In this subsection, we thus assume that the two sequences of independent and uniformly distributed random variables U_n and V_n are equal. We denote by $\widetilde{A}$ and $\widetilde{B}$ the two coupled stochastic processes constructed from processes $\{A(n),\ n \geq 0\}$ and $\{B(n),\ n \geq 0\}$ respectively.
We first need two lemmas which allow us to compare the quantities $s_{\widetilde{B}'}(n,k)$ and k/K as well as variables $J_{\widetilde{B}}(U_n)$ and $J_{\widetilde{A}}(U_n)$. As we did for processes $\{A(n),\ n \geq 0\}$ and $\{B(n),\ n \geq 0\}$, we introduce the process $\widetilde{A}'$ and $\widetilde{B}'$ which are obtained by reordering the entries of vectors $\widetilde{A}(n)$ and $\widetilde{B}(n)$ in the ascending order.

Lemma 5.1 *For all $n \geq 0$ and for all $k = 1, \ldots, K$, we have $s_{\widetilde{B}'}(n,k) \leq k/K$.*

Proof. Proof. Recall that $s_{\widetilde{B}'}(n,k)$ is defined by

$$s_{\widetilde{B}'}(n,k) = \frac{1}{S_{\widetilde{B}'}(n)} \sum_{j=1}^{k} \widetilde{B}'_j(n) \text{ where } S_{\widetilde{B}'}(n) = \sum_{k=1}^{K} \widetilde{B}'_k(n).$$

If, for all $k = 1, \ldots, K$, $\widetilde{B}'_k(n)/S_{\widetilde{B}'}(n) = 1/K$ then the result is true since in that case, we have $s_{\widetilde{B}'}(n,k) = k/K$.

Otherwise, the sequence $\widetilde{B}'_k(n)/S_{\widetilde{B}'}(n)$ being non-decreasing with k and since we have $s_{\widetilde{B}'}(n,K) = 1$, there exists a unique index $i \in \{1, \ldots, K-1\}$ such that

$$\frac{\widetilde{B}'_1(n)}{S_{\widetilde{B}'}(n)} \leq \cdots \leq \frac{\widetilde{B}'_i(n)}{S_{\widetilde{B}'}(n)} < 1/K \leq \frac{\widetilde{B}'_{i+1}(n)}{S_{\widetilde{B}'}(n)} \leq \cdots \leq \frac{\widetilde{B}'_K(n)}{S_{\widetilde{B}'}(n)}.$$

It follows that, for all $k = 1, \ldots, i$, we have $s_{\widetilde{B}'}(n,k) \leq k/K$. For $k = i+1, \ldots, K$, we have

$$s_{\widetilde{B}'}(n,k) = \sum_{j=1}^{k} \frac{\widetilde{B}'_j(n)}{S_{\widetilde{B}'}(n)} = 1 - \sum_{j=k+1}^{K} \frac{\widetilde{B}'_j(n)}{S_{\widetilde{B}'}(n)} \leq 1 - \frac{K-k}{K} = \frac{k}{K},$$

which completes the proof. ■

Lemma 5.2 *For all $n \geq 0$, we have $J_{\widetilde{A}}(U_n) \leq J_{\widetilde{B}'}(U_n)$.*

Proof. Proof. The definition of $J_{\widetilde{B}'}(U_n)$ given in Relation (7), with B' instead of B, is

$$J_{\widetilde{B}'}(U_n) = \sum_{k=1}^{K} k 1_{\{s_{\widetilde{B}'}(n,k-1) \leq U_n < s_{\widetilde{B}'}(n,k)\}},$$

in which we set $s_{\widetilde{B}'}(n,0) = 0$, for all $n \geq 0$. This relation and the fact that the $s_{\widetilde{B}'}(n,k)$ are non decreasing in k imply that

$$s_{\widetilde{B}'}(n, J_{\widetilde{B}'}(U_n) - 1) \leq U_n < s_{\widetilde{B}'}(n, J_{\widetilde{B}'}(U_n)) < s_{\widetilde{B}'}(n, J_{\widetilde{B}'}(U_n) + 1). \tag{21}$$

Note that $J_{\widetilde{A}}(U_n) = J_A(U_n)$. Similarly, from the definition of $J_A(U_n)$ in Relation (2), we get that

$$\frac{J_{\widetilde{A}}(U_n) - 1}{K} \leq U_n < \frac{J_{\widetilde{A}}(U_n)}{K} < \frac{J_{\widetilde{A}}(U_n) + 1}{K}. \tag{22}$$

We prove this Lemma by contradiction. Suppose that there exists an index n such that $J_{\widetilde{B}'}(U_n) < J_{\widetilde{A}}(U_n)$. Using this assumption and applying Lemma 5.1, we have from Relation (21) that

$$U_n < s_{\widetilde{B}'}(n, J_{\widetilde{B}'}(U_n)) \leq \frac{J_{\widetilde{B}'}(U_n)}{K} < \frac{J_{\widetilde{A}}(U_n)}{K}.$$

Using again the assumption $J_{\widetilde{B}'}(U_n) < J_{\widetilde{A}}(U_n)$ or equivalently $J_{\widetilde{B}'}(U_n) \leq J_{\widetilde{A}}(U_n) - 1$, we obtain from Relation (22)

$$\frac{J_{\widetilde{B}'}(U_n)}{K} \leq \frac{J_{\widetilde{A}}(U_n) - 1}{K} \leq U_n < \frac{J_{\widetilde{A}}(U_n)}{K}.$$

We thus have both $U_n < J_{\widetilde{B}'}(U_n)/K$ and $U_n \geq J_{\widetilde{B}'}(U_n)/K$, which is a contradiction. We thus have that for all $n \geq 0$, we have $J_{\widetilde{A}}(U_n) \leq J_{\widetilde{B}'}(U_n)$. ■

Lemma 5.3 *For all $n \geq 0$ and $k = 1, \ldots, K$, we have $\widetilde{B}'_k(n) \leq \widetilde{A}'_k(n)$ and $S_{\widetilde{B}'}(n) \leq S_{\widetilde{A}'}(n)$.*

Proof. Proof. We prove the first result by recurrence on n. When $n = 0$, we trivially have $\widetilde{A}'(0) = A(0) = B(0) = \widetilde{B}'(0)$. Suppose that for a index n, we have $\widetilde{B}'_k(n) \leq \widetilde{A}'_k(n)$, for all $k = 1, \ldots, K$. From Lemma 5.2, we have $J_{\widetilde{A}}(U_n) \leq J_{\widetilde{B}'}(U_n)$. We thus distinguish the following four disjoints cases: $k = 1$, $k \in \{2, \ldots, J_{\widetilde{A}}(U_n)\}$, $k \in \{J_{\widetilde{A}}(U_n) + 1, \ldots, J_{\widetilde{B}'}(U_n)\}$ and $k \in \{J_{\widetilde{B}'}(U_n) + 1, \ldots, K\}$.

Note that if $J_{\widetilde{B}'}(U_n) = K$ then only the first three cases have to be considered and if $J_{\widetilde{A}}(U_n) = K$ then only the first two cases have to be considered.

For each case, we use Relations (8) and (4) and the recurrence hypothesis.

- For $k = 1$, we have $\widetilde{B}'_1(n+1) = \widetilde{A}'_1(n+1) = 1$.
- For $k \in \{2, \ldots, J_{\widetilde{A}}(U_n)\}$, we have

$$\widetilde{B}'_k(n+1) = \widetilde{B}'_{k-1}(n) + 1 \leq \widetilde{A}'_{k-1}(n) + 1 = \widetilde{A}'_k(n+1).$$

- For $k \in \{J_{\widetilde{A}}(U_n) + 1, \ldots, J_{\widetilde{B}'}(U_n)\}$, we have

$$\widetilde{B}'_k(n+1) = \widetilde{B}'_{k-1}(n) + 1 \leq \widetilde{A}'_{k-1}(n) + 1 \leq \widetilde{A}'_k(n) + 1 = \widetilde{A}'_k(n+1).$$

- For $k \in \{J_{\widetilde{B}'}(U_n) + 1, \ldots, K\}$, we have

$$\widetilde{B}'_k(n+1) = \widetilde{B}'_k(n) + 1 \leq \widetilde{A}'_k(n) + 1 = \widetilde{A}'_k(n+1).$$

This completes the proof of the first inequality. The second one is now trivial since

$$S_{\widetilde{B}'}(n) = \sum_{k=1}^{K} \widetilde{B}'_k(n) \leq \sum_{k=1}^{K} \widetilde{A}'_k(n) = S_{\widetilde{A}'}(n),$$

which completes the proof. ■

We are now able to prove the main result of this section, that is that Algorithm B performs better than Algorithm A. We first recall the following lemma, see [9].

Lemma 5.4 *[Strassen's theorem] The real random variable X stochastically dominates Y if and only if there exists a coupling $(\widetilde{X}, \widetilde{Y})$ of X and Y such that $\mathbb{P}[\widetilde{X} \geq \widetilde{Y}] = 1$.*

Theorem 5.5 *For all $n \geq 0$ and $k = 1, \ldots, K$, we have*

$$B'_k(n) \overset{s.t.}{\preceq} A'_k(n) \text{ and } S_B(n) \overset{s.t.}{\preceq} S_A(n).$$

Proof. Proof. Combining Lemma 5.4 and Lemma 5.3, we directly deduce that for all $n \geq 0$ and $k = 1, \ldots, K$, we have $B'_k(n) \overset{s.t.}{\preceq} A'_k(n)$ and $S_{B'}(n) \overset{s.t.}{\preceq} S_{A'}(n)$. Then, using the fact that $S_{A'}(n) \overset{\mathcal{D}}{=} S_A(n)$ and $S_{B'}(n) \overset{\mathcal{D}}{=} S_B(n)$, we conclude that $S_B(n) \overset{s.t.}{\preceq} S_A(n)$. ■

Theorem 5.5 shows that the amount of accumulated measurements at any end-device is lower for algorithm B than for Algorithm A, i.e. for any $\ell \geq 0$

$$\mathbb{P}\{B'_K(n) \geq \ell\} \leq \mathbb{P}\{A'_K(n) \geq \ell\}.$$

In other words, the probability that the maximum uncollected measurements exceeds ℓ with Algorithm B is greater than or equal to the probability that the maximum uncollected measurements exceeds ℓ with Algorithm A. Then, we also get that the total sum of uncollected measurements over all end-devices is also lower for Algorithm B than for Algorithm A (in the sense that $S_B(n)$ is stochastically dominated by $S_A(n)$).

6 Analysis of Algorithm C

We show in this section that Algorithm C performs better than algorithm B. More precisely we show that deterministic process $\{C(n),\ n \geq 0\}$ is a lower bound of stochastic process $\{B(n),\ n \geq 0\}$

Theorem 6.1 *For any $n \geq 0$ and for all $k = 1, \ldots, K$, we have*

$$C'_k(n) \leq B'_k(n) \text{ and } S_C(n) = S_{C'}(n) \leq S_{B'}(n).$$

Proof. Proof. We prove the first inequality by recurrence. The result is clearly true for $n = 0$, since $C'_k(0) = B'_k(0) = 1$. Suppose that for a fixed $n \geq 1$, we have $C'_k(n-1) \leq B'_k(n-1)$, for all $k = 1, \ldots, K$. For $k = 1$, we have $C'_1(n) = B'_1(n) = 1$. From Relation (8) and using the fact that the sequence $B'_k(n)$ is non decreasing in k we obtain

$$\begin{aligned} B'_k(n) &= \left(B'_{k-1}(n-1)+1\right) 1_{\{k=2,\ldots,J_{B'}(V_{n-1})\}} + \left(B'_k(n-1)+1\right) 1_{\{k=J_{B'}(V_{n-1})+1,\ldots,K\}} \\ &\geq \left(B'_{k-1}(n-1)+1\right) 1_{\{k=2,\ldots,J_{B'}(V_{n-1})\}} + \left(B'_{k-1}(n-1)+1\right) 1_{\{k=J_{B'}(V_{n-1})+1,\ldots,K\}} \\ &= B'_{k-1}(n-1)+1 \geq C'_{k-1}(n-1)+1 = C'_k(n). \end{aligned}$$

The second inequality is then immediate by definition of $S_{C'}(n)$ and $S_{B'}(n)$. ■

7 Discussion : performance and vulnerability to attacks

We have shown in the previous section that Algorithm C performs better than Algorithm B which in turn performs better than Algorithm A and we have obtained several performance measures for both Algorithms A and C. This means that from a pure performance criterion it is better to use Algorithm C.

Now let us investigate the capacity of these three algorithms to tolerate the presence of deny-of-service attacks against end-devices. We model such an attack by an omniscient entity that has a full knowledge of the code run by both the monitoring devices and end-devices. In particular the adversary knows the distribution of the random variable used by the monitoring device to select the next end-device to be queried, but not the values of the random variable.

Let us first consider the randomized algorithms A and B. The random variable used by the monitoring device to select the end-device from which data is collected at collect n has been denoted by U_n, for Algorithm A and by V_n, for Algorithm B. The distribution of U_n is uniform over $\{1, \ldots, K\}$ and the distribution of V_n is proportional to the number of balls in each urn at collect n and thus depends on n. The adversary knows the distributions of U_n and V_n and tries to determine using these distributions which urn (i.e., end-user) is selected by the monitoring device at each collect n. We denote by Y the random variable used by the adversary at each collect n to mimic the distribution of U_n and V_n

Consider first Algorithm A. By definition of the adversary, random variables Y and U_n are independent and identically uniformly distributed. The adversary will succeed in determining the end-user to be queried by the monitoring device at collect n with probability $\mathbb{P}\{Y = U_n\}$. We thus define the vulnerability Vu(A) of algorithm A by the probability $\mathbb{P}\{Y = U_n\}$. It is given by

$$\mathrm{Vu(A)} = \mathbb{P}\{\mathrm{Y} = \mathrm{U_n}\} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathbb{P}\{\mathrm{Y} = \mathrm{k}, \mathrm{U_n} = \mathrm{k}\} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \left(\mathbb{P}\{\mathrm{U_n} = \mathrm{k}\}\right)^2 = \frac{1}{\mathrm{K}}.$$

Concerning Algorithm B, again by definition of the adversary, the random variables Y and V_n are independent and identically distributed, but non uniform. The adversary will succeed in determining the end-user to be queried by the monitoring device at collect n with probability $\mathbb{P}\{Y = V_n\}$. The vulnerability Vu(B) of algorithm B at collect n is then given by

$$\mathrm{Vu(B)} = \mathbb{P}\{\mathrm{Y} = \mathrm{V_n}\} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \left(\mathbb{P}\{\mathrm{V_n} = \mathrm{k}\}\right)^2.$$

Clearly, Vu(B) is not easy to obtain because the random variables V_n depends on n, but we have the following bound. The function $f(x) = x^2$ being strictly convex, the Jensen inequality gives easily

$$\frac{1}{K}\sum_{k=1}^{K} \left(\mathbb{P}\{V_n = k\}\right)^2 > \frac{1}{K^2}\left(\sum_{k=1}^{K} \mathbb{P}\{V_n = k\}\right)^2 = \frac{1}{K^2}.$$

It follows that $\mathrm{Vu(A)} = \mathbb{P}\{\mathrm{Y} = \mathrm{U_n}\} < \mathbb{P}\{\mathrm{Y} = \mathrm{V_n}\} = \mathrm{Vu(B)}$, which means that Algorithm A is strictly less vulnerable than Algorithm B. Observe that since Algorithm C is deterministic, we clearly have Vu(C) = 1. Finally, the global result of this analysis shows a trade-off between the performance and vulnerability aspects of these algorithms.
Algorithm C *performs better than* Algorithm B, which *performs better than* Algorithm A
Algorithm A *is less vulnerable than* Algorithm B, which *is less vulnerable than* Algorithm C to deny-of-service attacks.

8 Conclusion

In this paper, we have proposed and analyzed the performance of three algorithms for collecting longitudinal data in a large scale system. A monitoring device is in charge of continuously collecting measurements from K end-devices. We have studied the transient and stationary distributions of the uncollected data at end-devices as a function of the collect policy implemented by the monitoring device. We have also compared their capability to be resilient to deny-of-service attacks launched by an omniscient adversary and have shown a trade-off between vulnerability and performance.

It would be relevant to explore other collect policies, starting with those proposed to solve the load balancing problem such as selecting uniformly a end-device among the d end-devices with the most uncollected measurements (multiple-choice paradigm [1]), the two choice paradigm [10] or the $(1 + \beta)$-choice [12].

The algorithms studied in this paper are based on the hypothesis that at each collect, the selected end-device sends its measurements in time. It would be interesting to challenge these algorithms in an environment where some end-device may not respond due to transient partitions. By transient partitions, we mean that among the K end-devices, some of them can be temporarily partitioned from the remaining of the devices, in the sense that communications between these end-devices and the remaining of the system are temporarily delayed.

References

[1] Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal. “Balanced allocations (extended abstract)”. In: *Proceedings of the ACM Symposium on Theory of Computing (STOC)* (1994).

[2] I. Azimi, O. Oti, S. Labbaf, H. Niela-Viln, A. Axelin, N. Dutt, P. Liljeberg, and A. M. Rahmani. "Personalized Maternal Sleep Quality Assessment: An Objective IoT-based Longitudinal Study". In: *IEEE Access* 7 (2019), pp. 93433–93447.

[3] I. Azimi, T. Pahikkala, A. M. Rahmani, H. Niela-Vilén, A. Axelin, and P. Liljeberg. "Missing data resilient decision-making for health care IoT through personalization: A case study on maternal health". In: *Future Generation Computer Systems* 96 (2019), pp. 297–308.

[4] O. Dieng, B. Diop, O. Thiare, and C. Duc Pham. "A Study on IoT Solutions for Preventing Cattle Rustling in African Context". In: 2017.

[5] Feig DS et al. "Continuous glucose monitoring in pregnant women with type 1 diabetes (concept): a multicentre international randomised controlled trial". In: *The Lancet* 390.10110 (2017), pp. 2347–2359.

[6] Michael J. Fischer and Michael Merritt. "Appraising two decades of distributed computing theory research". In: *Distributed Computing* 16.2–3 (2003), pp. 239–247.

[7] S. Kaul, M. Gruteser, V. Rai, and J. Kenney. "Minimizing age of information in vehicular networks". In: *2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*. 2011, pp. 350–358.

[8] P. Lange, J. Parner, P. Schnohr, and G. Jensen. "Copenhagen City Heart Study: longitudinal analysis of ventilatory capacity in diabetic and nondiabetic adults". In: *European Respiratory Journal* 20 (2002), pp. 1406–1412.

[9] T. Lindvall. *Lectures on the Coupling Method*. Dover Publications Inc., New York, 2002.

[10] Y. Mocquard, B. Sericola, and E. Anceaume. "Balanced allocations and global clock in population protocols: An accurate analysis". In: *SIROCCO 2018 : 25th Colloquium on Structural Information and Communication Complexity* (2018).

[11] G. T. Moor and S. Takemi. "The Childrens Physical Environment Rating Scale (CPERS): Reliability and Validity for Assessing the Physical Environment of Early Childhood Educational Facilities". In: 17.4 (1999), pp. 24–53.

[12] Y. Peres, K. Talware, and U. Wieder. "The $(1 + \beta)$-Choice Process and Weighted Balls-into-Bins". In: *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2010).

[13] J. Quaintance and H. W. Gould. *Combinatorial Identities for Stirling Numbers. The unplublished notes of H. W. Gould*. World Scientific Publishing, 2006.