On Distributed Cardinality Estimation: Random Arcs Recycled*

Marcin Kardas† Mirosław Kutyłowski† Jakub Lemiesz†

Abstract

We introduce and analyze a distributed cardinality estimation algorithm for a network consisted of not synchronized nodes. Our solution can be regarded as a generalization of the classic approximate counting algorithm based on the balls and bins model and is connected to the well studied process of covering the circle with random arcs. Although the algorithm is presented in the context of a radio network, the basic idea is applicable to any system in which many uncoordinated nodes communicate over a shared medium. In the paper we prove the correctness of the algorithm and by the methods of complex analysis we carefully examine the accuracy and precision of the estimator we have proposed. We also show that the construction of the proposed algorithm is a backbone for simple distributed summation.

1 Introduction

Designing practical or at least realistic and yet provably correct algorithms for wireless distributed systems is usually a challenging task. Communication models that take into account a lot of physical details make the design and analysis of the algorithms difficult. On the other hand, by accepting many simplifying assumptions on the environment and the communication capabilities of nodes, one can cause that the practical implementation of an algorithm will be difficult or impossible. This motivates the search for possibly simple but realistic models, in which nodes would have modest yet reliable communication capabilities. Such model, based on the carrier sensing mechanism has been recently considered in several papers (e.g. [1, 2, 10, 15, 22]). Algorithms in this model, which is known as the beeping communication model, exchange information not by passing messages but by sending and detecting a jamming signal (called a beep). Namely, each device is capable of identifying the channel status as either busy (at least one device is sending signal) or idle (no device is sending signal). Such way of communication is typically much more reliable and requires significantly less energy than transmitting and receiving actual messages, see e.g. [1]. Although the beeping model makes almost no assumptions besides availability of the reliable carrier sensing mechanism, it has allowed to solve efficiently a few complex problems (such as vertex coloring or computing a maximal independent set, see Section 2).

Inspired by these results, we introduce a new cardinality estimation algorithm for unstructured wireless networks. The knowledge of the network cardinality or its good estimation is essential for setting the values of parameters in most of the protocols, particularly in a dynamic and loaded network environment. Similarly, in the context of multi-hop topologies it is essential for a node to know at least the size of its neighborhood. Several distributed size estimation algorithms dedicated to ad hoc wireless systems have been proposed recently, e.g. [3, 5, 19, 20, 21]. Most of them could be also easily adapted to estimate the size of a node's neighborhood. However, a common trait of these algorithms is the assumption that a communication framework is established, typically that there exists a permanent network-wide slot synchronization (TDMA).

Precisely establishing the slot boundaries requires global synchronization of clocks. If a network consists of cheap, battery-powered devices not equipped with atomic or GPS-based clocks some synchronization procedure is required. However, even in the single-hop setting the existing procedures are non-trivial and the synchronization entails a considerable expenditure of time and energy (see [15, 24, 28]). Moreover, in many scenarios the composition of a network may change dynamically and some clocks may drift. Then, the obtained synchronization is ephemeral and needs frequent updates.

Obviously, if the communication graph is not complete, the synchronization problem is even more complex. Thus, in most deployed multi-hop networks the access to the channel is based on CSMA protocols, which typically don't require slot synchronization (see [23, 27]). However, the main difficulty in CSMA protocols is adjusting back-off values that are used to

[*] The research was financed by the Polish National Science Centre, based on the decision DEC-2013/08/M/ST6/00928.

[†] Department of Computer Science at the Faculty of Fundamental Problems of Technology, Wrocław University of Technology, Poland.

schedule retransmissions after collisions. For a given node the effectiveness of such adjusting strongly depends on the knowledge of the cardinality of its one-hop neighborhood. Therefore, in such a setting a procedure for the estimation of a neighborhood size that does not require the synchronization may be crucial (cf. [8, 17]).

The algorithm presented in this paper can be applied to estimate the size of a single-hop network as well as to estimate the size of a node's neighborhood in a multi-hop network. The solution complies with the beeping model, therefore is lightweight, reliable and effective. The main advantage of the algorithm is that it can be executed concurrently on each node in the network without the need for the previous synchronization or additional communication arrangements. We also show that the construction of the algorithm is quite universal and can serve as a backbone for the approximate summation and averaging of distributed values. We would like to note that although the algorithm is presented in the context of a radio channel, the basic idea is applicable to any system in which many uncoordinated nodes communicate over a shared medium and the carrier sensing is available.

2 Related Work

In the beeping communication model nodes rely entirely on carrier sensing. At any particular time a node can be in beeping or listening mode and cannot distinguish between a single beep and a collision of two or more beeps (no collision detection). Although one could define a coding scheme to encode bit messages with beeps[1], this would require additional overhead and be prone to collisions, thus the focus is on different techniques. Suitable techniques have recently been considered in a number of papers. In [15] Flury et al. show how the carrier sensing can be used as an elegant way for coordination in practice. In [1] Afek et al. consider the problem of computing a maximal independent set under the beeping communication model. The results of [9, 2] show that this model is also interesting from the biological point of view, e.g., in the analysis of processes, in which cells communicate by secreting certain proteins that are sensed by neighboring cells. In [22] the authors study the decentralized interval coloring problem, a variant of vertex coloring specially suited for the beeping model that is directly connected to the task of establishing reliable transmission scheduling. The authors of [10] improve the result of [22] and draw attention to the fact that in [22] prior to establishing communication nodes need to determine the size of their neighborhoods (potentially a chicken-and-egg problem). They also show how the size of a neighborhood can be roughly bounded by the number of detected beeps.

The main problem in the network size estimation is to avoid overestimating by processing the same message several times and underestimating by not distinguishing different massages. Therefore, the task is closely related to the problem of estimating the number of distinct elements in a data set, which is well studied in the literature and for which many excellent algorithms have been proposed (e.g., [4, 11, 12, 16, 29]). Most of them are based on discrete random variables: e.g. maximum of geometric random variables is used in Probabilistic Counting [12] and HyperLogLog [11], the balls-and-bins model in Linear Counting [29] and Bernoulli trials in Two-Phase Algorithm [5]. A few algorithms based on continuous random variables have been proposed as well: Minima and Optimal Counting [16, 4] and COMP [21] use order statistics of the uniform and exponential distribution.

Building on these solutions several distributed size estimation protocols dedicated to wireless systems have been proposed recently (e.g., [3, 5, 6, 7, 19, 20]). However, these protocols implicitly assume that a communication framework is already established and hide the complexity of the underlying implementation. In contrast, the algorithm presented in this paper is straightforward to implement and does not require any prior communication arrangements. The random variable employed in the algorithm resembles the continuous version of the variable used in Linear Counting [29], a classic solution that can be easily adapted to the beeping model and thus can be a good reference point.

Our Results

In Section 3 we introduce the random process that underlies the idea of our solution, present a new size estimation algorithm and prove its correctness. In Section 4 we analytically investigate the accuracy and precision of an applied estimator. The results prove that our algorithm performs better than Linear Counting – its closest discrete analogue, the fact that we further investigate in Section 5 by numerical simulations.

3 Size Estimation Algorithm

In this Section we introduce a size estimation algorithm that is based on a random process of throwing arcs on a circle. The algorithm does not require a precise synchronization of devices but merely requires the carrier sensing ability. We prove that the algorithm is valid and we show that with minor changes it can be used for distributed computing of other statistics, in particular for distributed approximate summation of real numbers.

[1] Note that a beep carries less information than a bit. In case of a bit there are in fact three states: zero, one and no signal.

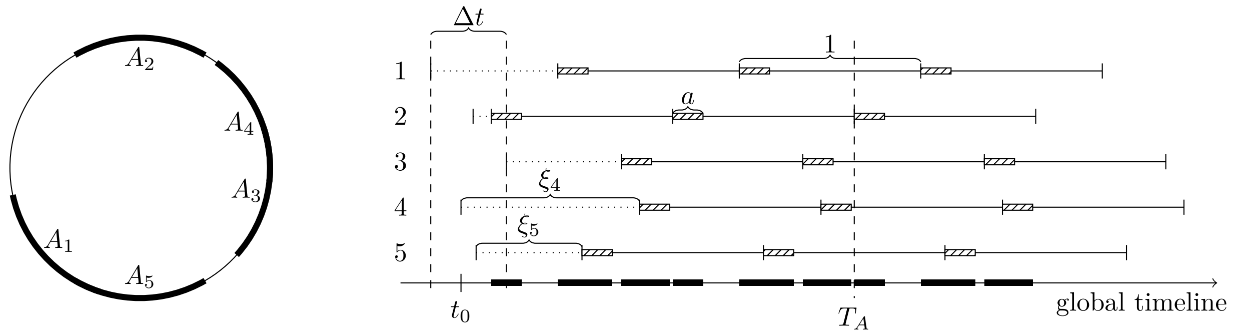


Figure 1: An instance of Siegel's process with 5 arcs of length $a = 1/6$ each (left) and an execution of RAR algorithm performed by $n = 5$ nodes (right). The maximal relative time difference Δt is smaller than 1, hence $k = 3$ cycles are sufficient. Nodes $i = 1, \ldots, 5$, start at time t_0 (according to local clock) and wait for a period $\xi_i \in [0, 1)$ (dotted line). A cycle of each node begins with a transmission of duration a, followed by the channel sensing for a period of duration $1 - a$. By connecting endpoints of the cycle in which T_A appears (usually different for distinct devices), each device gets similar pattern of arcs on the circle.

3.1 Random Arcs on the Circle The starting point of our algorithm is the following random process: n arcs $A_1, A_2, \ldots, A_n$, each of length $a < 1$, are placed at random over the circumference K of a circle of length 1 (see Figure 1). The arcs are placed independently with counter-clockwise endpoints distributed uniformly. The covered part of the circumference A is simply the union taken over all arcs, $A = \bigcup_{i=1}^{n} A_i$. For μ denoting the Lebesgue measure on K we define a random variable $S = 1 - \mu(A)$ denoting the total length of an unoccupied part of the circle. The mathematical properties of S have been well studied. Siegel [26] has derived formulas for the moments and distribution of the random variable S (see Lemma 4.1). In fact, the moments of S can be obtained from the Robbins' theorem (see [25]). For example, by this theorem we have

$$\mathbb{E}(S) = \int_K \Pr(x \notin A)\, d\mu(x),$$

where $\Pr(x \notin A)$ is the probability that a given point x will not be covered by any of n arcs. Because the positions of arcs are chosen independently we have $\Pr(x \notin A) = (\Pr(x \notin A_1))^n = (1-a)^n$. Thus, for the circle of unit circumference $\mathbb{E}(S) = (1-a)^n$.

Our key observation is that if the number of arcs n is initially unknown, one can infer about n from the value taken by S. For example, knowing the formula of $\mathbb{E}(S)$ one can instantly derive an estimator of n by the method of moments:

$$\hat{n} = \ln(S)/\ln(1-a)\ . \tag{3.1}$$

The underlying idea of the estimator (3.1) seems to be quite universal, and indeed it can be easily generalized to estimate quantities other than cardinality. For instance, let us derive an estimator of a sum

$$x = x_1 + x_2 + \ldots + x_n,$$

where

$$x_1, x_2, \ldots, x_n \in (1-\varepsilon, 1+\varepsilon) \text{ and } 0 < \varepsilon \leq 1.$$

Assume that arcs are randomly placed on the circumference of a circle of unit length and that the ith arc has length

$$a_i = 1 - e^{cx_i},$$

where $c = \ln(1-a)$ and $a < 1$. Notice that we have $a_i = \Theta(a)$ as $a \to 0$ for fixed ε. Let now T denote the random proportion of the circumference that is not contained in any arc. From the Robbins' theorem we get

$$\mathbb{E}(T) = (1-a_1)(1-a_2)\ldots(1-a_n).$$

Finally, since

$$\ln(\mathbb{E}(T)) = c(x_1 + x_2 \ldots x_n)$$

by the method of moments we obtain the estimator

$$\hat{x} = \ln(T)/\ln(1-a)$$

of the sum x. What is noteworthy is that the application of the estimator $\hat{x}$ does not require any knowledge of n. If n or its estimation is known we obtain the estimation of the average value x/n as well. Although in what follows we will present the formal analysis concerning only the size estimation algorithm, some illustrative numerical results concerning the summation process are included in Section 5.

Algorithm 1 RAR(t_0, a, k)

```
At time $t_0$ according to my clock:
$\xi \leftarrow$ random($[0, 1)$)
wait for time of length $\xi$
for $i = 1, 2, \ldots, k$ do
    send a signal of length $a$ (beep)
    sense the channel for time of length $1 - a$
    $S_i \leftarrow$ the total length of sensed "silence"
$S \leftarrow \min\{S_1, S_2, \ldots, S_k\}$
return: $\hat{n} \leftarrow \ln(S) \;/\; \ln(1 - a)$
```

3.2 Algorithm Description Based on the above, we design a distributed RAR (Random Arcs Recycling) algorithm in which nodes simulate the random process of placing arcs on a circle. The circle is represented by the shared channel and arcs are represented by beeps of duration a. The number of participating nodes is inferred from the length of an unoccupied part S of the circle which corresponds to the time when the channel is idle. As we prove in Section 4, smaller values of a result in a more accurate and precise estimation. On the other hand, a transmission should be long enough to be sensed by nodes participating in the algorithm. Therefore, we leave a as an input parameter dependent on additional technical aspects.

In the baseline scenario RAR is executed once, at roughly the same time by each device. Namely, we assume that each device has an internal clock and executes the algorithm at time t_0, but clocks may differ in their time indication. We assume[2] that the length of a cycle is 1 and the length of a beep is $a > 0$. Additionally, we shall assume that a is sufficiently short, namely $a < 1/n$, where $n \geq 2$ is the number of devices.

RAR is executed independently by each device. Initially, a device picks uniformly at random $\xi \in [0, 1)$ and then at the time t_0 waits for the time of duration ξ. Next, a device executes $k \geq 3$ cycles of the following procedure. In its ith round a device sends a signal of duration a, senses the channel for the time of duration $1 - a$ and returns the total time S_i when the channel was idle. Finally, the algorithm returns the estimation based on the minimal S_i. RAR aims to simulate Siegel's process (see Section 2) in a periodic manner such that one cycle flows smoothly into the next one and a transmission pattern stabilizes (see Figure 1.) The validity of RAR follows from Theorem 3.1, which shows that it actually simulate the random process of placing arcs on a circle and justifies the use of estimator $\hat{n}$.

THEOREM 3.1. *Assume that* RAR *is executed with parameter $k \geq 3$ and let $\Delta t \leq k - 2$ be the maximal difference between clocks' indications over all pairs of devices. Then, for an arbitrary chosen device D there exists at least one $i \in \{1, 2, \ldots, k\}$ such that each device contributes a beep of duration a in the ith cycle of device D and $S_i = \min\{S_1, S_2, \ldots S_k\}$.*

Proof. Since devices transmit at intervals equal to the length of the cycle, any device contributes to any cycle of device D a beep of the total length at most a. We will show that D always has a cycle in which each device contributes a beep of the total length a. Let A be a device which has begun its first transmission at the earliest time and let T_A be the starting point of its kth transmission (absolute time). Then,

1) since $\Delta t \leq k-2$ and $\xi \in [0, 1)$, the time lag between the first transmissions of any two devices is less than $k - 1$, thus each device starts at least one transmission before time T_A,

2) each device starts one transmission in the interval $[T_A - 1, T_A)$ and one in the interval $[T_A, T_A + 1)$ and these moments are at distance 1,

3) a cycle of D in which the moment T_A appears is the cycle in which each device contributes a beep of length a. Note that if some beep is not contained entirely in the cycle's boundary, then its projecting part overlaps with the position of a signal of D and is accounted anyway.

Let us finally remark that in each cycle of device D a given beep appears in the same position or does not appear at all. Thus, the cycle mentioned in step 3) is the cycle with the minimal value of the total silence S.

4 Analysis of the Estimator

In Theorem 3.1 we have established a direct connection between RAR and the original process considered by Siegel. To conclusively confirm the usefulness of RAR we have to analyze the properties of estimator (3.1). As a starting point we will use the following result:

LEMMA 4.1. (SIEGEL [26]) *Let S be a random variable denoting the total length of unoccupied space of circle circumference after throwing at random n arcs of length a each. Then the CDF is given by $F(t) = \Pr(S \leq t) =$*

$$1+\sum_{l=1}^{n}\sum_{k=0}^{n-1}(-1)^{k+l}\binom{n}{l}\binom{l-1}{k}\binom{n-1}{k}t^k(1-la-t)_+^{n-k-1},$$

where $(x)_+ = \max\{0, x\}$ with the convention $(x)_+^0 = 0$ for $x \leq 0$.

[2] Such assumption will not influence the properties of the estimator as only the ratio of the length of a single transmission to the length of one cycle matters.

In the analysis we repeatedly encounter a somewhat cumbersome issue of evaluating the nth finite difference of a function $f(x)$ for $x = 0$, which can be written as (cf. [13, 18])

$$D_n[f] = \sum_{k=0}^{n} (-1)^k \binom{n}{k} f(k) \ . \tag{4.2}$$

Such alternate binomial sums gained the interest of many researchers after Flajolet and Sedgewick emphasized their role in the average-case analysis of algorithms (see [13]). Despite the binomial coefficients get close to 2^n, for many explicitly given functions such sums tend to be polynomially bounded or even decreasing in n. For example, it is apparent in the following facts (see e.g. [18]) that we use extensively in the analysis:

FACT 4.1. *If f is a polynomial of degree at most $m-1$ then $D_m[f] = 0$.*

FACT 4.2. *$\sum_{k=1}^{m} (-1)^k \binom{m}{k} \frac{1}{k} = -H_m$, where H_m is the mth harmonic number, $H_m = 1 + 1/2 + \ldots + 1/m$.*

Because of this phenomenon of *exponential cancellation* inherent in most sums of this type, they resist elementary techniques of analysis based on the asymptotic evaluation of individual terms. The general approach that has proved successful in the asymptotic analysis of high order differences is based on the Cauchy's residue theorem and its consequences, for example Rice's method (see [13]). Such approach is often unavoidable − the rapid growth of binomial coefficients causes that it could be extremely hard to evaluate the formula (4.2) numerically. This approach facilitates the proof of the following theorem that determines the bias of the estimator (3.1):

THEOREM 4.1. *If $a < 1/n$ then the expectation of estimator $\hat{n} = \ln(S)/\ln(1-a)$ is*

$$\mathbb{E}(\hat{n}) = (n-1) - \frac{1}{2}an + \frac{e^{an}-1}{an} + \mathcal{O}\left(n^{-1}\right).$$

Proof. Throughout the proof we assume $a < 1/n$. We regroup terms of the cumulative distribution function $F(t)$ from Lemma 4.1 and split it into:

$$P_1(t) = \sum_{l=1}^{n} \sum_{k=1}^{n-1} (-1)^{k+l} \binom{n}{l} \binom{l-1}{k} \binom{n-1}{k} \cdot \\ \cdot t^k \left((1-la-t)_+^{n-k-1} - (1-t)^{n-k-1} \right)$$

and

$$P_2(t) = \sum_{l=0}^{n} (-1)^l \binom{n}{l} (1-la-t)_+^{n-1},$$

so

$$\mathbb{E}(\ln S) = \int_0^1 \ln t \, dF(t) = -\int_0^1 P_1(t) \frac{dt}{t} - \int_0^1 P_2(t) \frac{dt}{t}.$$

In order to integrate $P_1(t)/t$ term-wise we have to deal with integrals

$$\begin{aligned} I(l,k) &= \int_0^{1-la} t^{k-1} (1-la-t)^{n-k-1} dt \\ &= (1-la)^{n-1} B(k, n-k) = \frac{(1-la)^{n-1}}{k\binom{n-1}{k}} \ , \end{aligned}$$

where we have disposed of the symbol $(\cdot)_+$ by changing the upper limit of the integration and by using the substitution $u = t/(1-la)$ we obtain an integral defining the Eulerian Beta function $B(k, n-k)$ (see [14]). Hence by Fact 4.2

$$\begin{aligned} \int_0^1 P_1(t) \frac{dt}{t} &= -\sum_{l=1}^{n} (-1)^l \binom{n}{l} \left((1-la)^{n-1} - 1 \right) H_{l-1} \\ &= -D_n[b_1], \end{aligned}$$

with $b_1(k) = \left((1-ak)^{n-1} - 1\right) H_{k-1}$ for $k \in \mathbb{N}_+$ and $b_1(0) = 0$.

Recall that $H_n = \psi(n+1) + \gamma$, where $\psi(z)$ is the digamma function (see [14]) and γ is the Euler's constant. Let

$$f(z) = \left((1-az)^{n-1} - 1\right)\left(\psi(z) + \gamma\right)$$

and let $\mathcal{C}$ be positively oriented closed contour encircling points $\{1, \ldots, n\}$. Then by the Rice integral we obtain

$$D_n[b_1] = \frac{(-1)^n}{2\pi i} \int_{\mathcal{C}} \varphi(z) \ dz \ ,$$

where

$$\varphi(z) = \frac{n! f(z)}{z(z-1)\cdots(z-n)} \ .$$

Let us consider a family of positively oriented concentric circles $\mathcal{C}_R$ centered at 0 with radii $R + \frac{1}{2}$, $R \in \mathbb{N}$. The radii are chosen so that the contours cut negative part of the real line exactly in between of consecutive singularities of $\psi(z)$. Thus $\psi(z) = \mathcal{O}(\ln z)$ and $\varphi(z) = \mathcal{O}\left(\frac{\ln z}{z^2}\right)$ on $\mathcal{C}_R$ as $R \to \infty$. Hence

$$\lim_{R\to\infty} \int_{\mathcal{C}_R} \varphi(z) \ dz = 0$$

and by the Cauchy integral formula we get

$$D_n[b_1] = (-1)^{n-1} \sum_{z_k} \operatorname{Res}(\varphi(z))_{z=z_k},$$

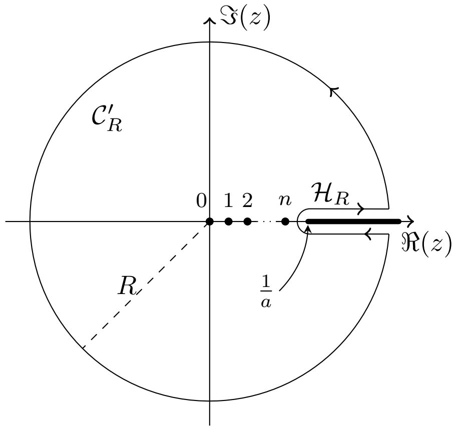


Figure 2: Contour for Rice integral.

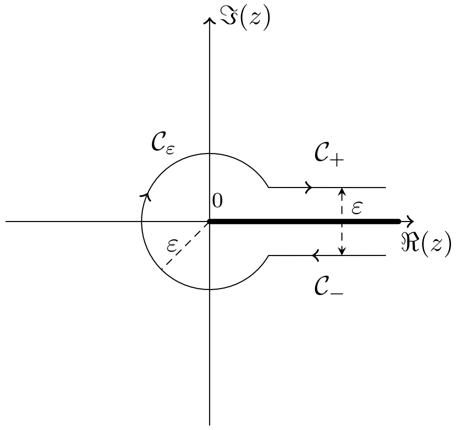


Figure 3: Hankel contour.

where the sum is taken over all singularity points z_k of $\varphi(z)$ that are not encircled by $\mathcal{C}$. The only singularities we have to consider are simple poles located at points $z = 0, -1, -2, \ldots$, induced by the singularities of $\psi(z)$. Using asymptotic series for $\psi(z)$ we obtain

$$\operatorname{Res}(\varphi(z))_{z=0} = (-1)^n a(n-1)$$

and

$$\operatorname{Res}(\varphi(z))_{z=-k} = (-1)^n \frac{(1+ak)^{n-1}-1}{k\binom{n+k}{k}} \ ,$$

for $k \in \mathbb{N}_+$. It suffices to consider the first two residues to finally get

$$\begin{aligned}\int_0^1 P_1(t)\frac{dt}{t} &= a(n-1) + \frac{(1+a)^{n-1}-1}{n+1} + \mathcal{O}\left(\frac{a}{n}\right)\\ &= a(n-1) + \frac{e^{an}-1}{n} + \mathcal{O}\left(\frac{a}{n}\right).\end{aligned}$$

With similar methods we show that the integral of $P_2(t)/t$ can be represented as an alternating binomial sum:

$$\int_0^1 P_2(t)\frac{dt}{t} = \sum_{l=0}^{n}(-1)^l\binom{n}{l}(1-la)^{n-1}\ln(1-la) \ .$$

Once again we use the Rice method. To address a branch cut along the semi-line $[\frac{1}{a}, \infty)$ induced by the logarithm we consider a contour of integration as the one depicted in Figure 2. It is easy to see that as radius $R \to \infty$ the integral over $\mathcal{C}'_R$ tends to 0 as before. By substitution $z = (t+1)/a$ we transform $\mathcal{H}_R$ into the Hankel contour $\mathcal{H}$ that starts at ∞ in the lower half-plane, encircles the origin clockwise and returns to ∞ in the upper half-plane (c.f. [14] and Figure 3). We prove in Lemma 4.2 that the contour integral can be evaluated by ordinary improper Riemann integral. With a little more effort we show that for $a < 1/n$ the integral of $P_2(t)/t$ is negligible:

$$\int_0^1 P_2(t)\frac{dt}{t} = \mathcal{O}\left(a^{n-1}n!\right).$$

Therefore with the formula for $\mathbb{E}(\ln S)$, we can use asymptotic expansion of $\ln(1-a)$ to obtain the final result.

Lemma 4.2. *Let $h(z) = g(z)(\operatorname{Log}(-z))^k$, where $k \in \mathbb{N}_+$ is constant and $g(z)$ is meromorphic and has no poles on $[0, \infty)$. Let $\mathcal{H} = \mathcal{C}_\varepsilon \cup \mathcal{C}_- \cup \mathcal{C}_+$ be the contour defined in Figure 3. If the integral of $h(z)$ over $\mathcal{C}_\varepsilon$ tends to zero as $\varepsilon \to 0$ then*

$$\lim_{\varepsilon\to 0}\int_{\mathcal{H}} h(z)dz = \int_0^\infty g(x)\left((\ln(x) - i\pi)^k - (\ln(x) + i\pi)^k\right)dx.$$

Proof. Let for $x \in \mathbb{R}$

$$(h(x))^+ = \lim_{y\to 0^+} h(x+iy)$$

and

$$(h(x))^- = \lim_{y\to 0^-} h(x+iy) \ .$$

By this notation for $x > 0$ we have

$$(h(x))^+ = g(x)(\ln(x) - i\pi)^k$$

and

$$(h(x))^- = g(x)(\ln(x) + i\pi)^k.$$

Then

$$\lim_{\varepsilon\to 0}\int_{\mathcal{C}_-\cup\mathcal{C}_+} h(z)\,dz = \int_0^\infty (h(x))^+ - (h(x))^- \ dx \ .$$

The general idea of the proof of Theorem 4.2, which determines the variance of the estimator (3.1), resembles the proof of Theorem 4.1. However, in case of Theorem 4.2 the computations are somewhat more convoluted and require greater precision. Below we shall present only the final result and some general conclusions.

THEOREM 4.2. *If $a < 1/n$ then the variance of estimator $\hat{n} = \ln(S)/\ln(1-a)$ is*

$$\mathbb{V}\mathrm{ar}\left(\hat{n}\right) = \left(2\sigma_3(an)\right) an^2 \; + \; \left(1 - 3\sigma_1(an)\right) an \\ + \; \left(4\sigma_1^2(an) + 4\sigma_2(an) - 6\sigma_1(an)\right) \; + \; o(an) \, ,$$

where $\sigma_1(x) = (e^x - 1)/x$, $\sigma_2(x) = (e^x - 1 - x)/x^2$ and $\sigma_3(x) = (e^x - 1 - x - \frac{1}{2}x^2)/x^3$.

Note that for $0 \leq x \leq 1$ and $j = 1, 2, 3$, we have $\sigma_j(x) = 1/j! + \mathcal{O}\left(x\right)$. Therefore, as $a < 1/n$ we get $\mathbb{V}\mathrm{ar}\left(\hat{n}\right) = \mathcal{O}\left(an^2\right)$. In fact Theorem 4.2 determines $\mathbb{V}\mathrm{ar}\left(\hat{n}\right)$ up to a constant. Contributions of individual terms depend on a, however the dominant term of the variance is always at most of order n. Consequently, for $a < 1/n$ the estimator is asymptotically highly concentrated (cf. Figure 4a).

5 Simulations

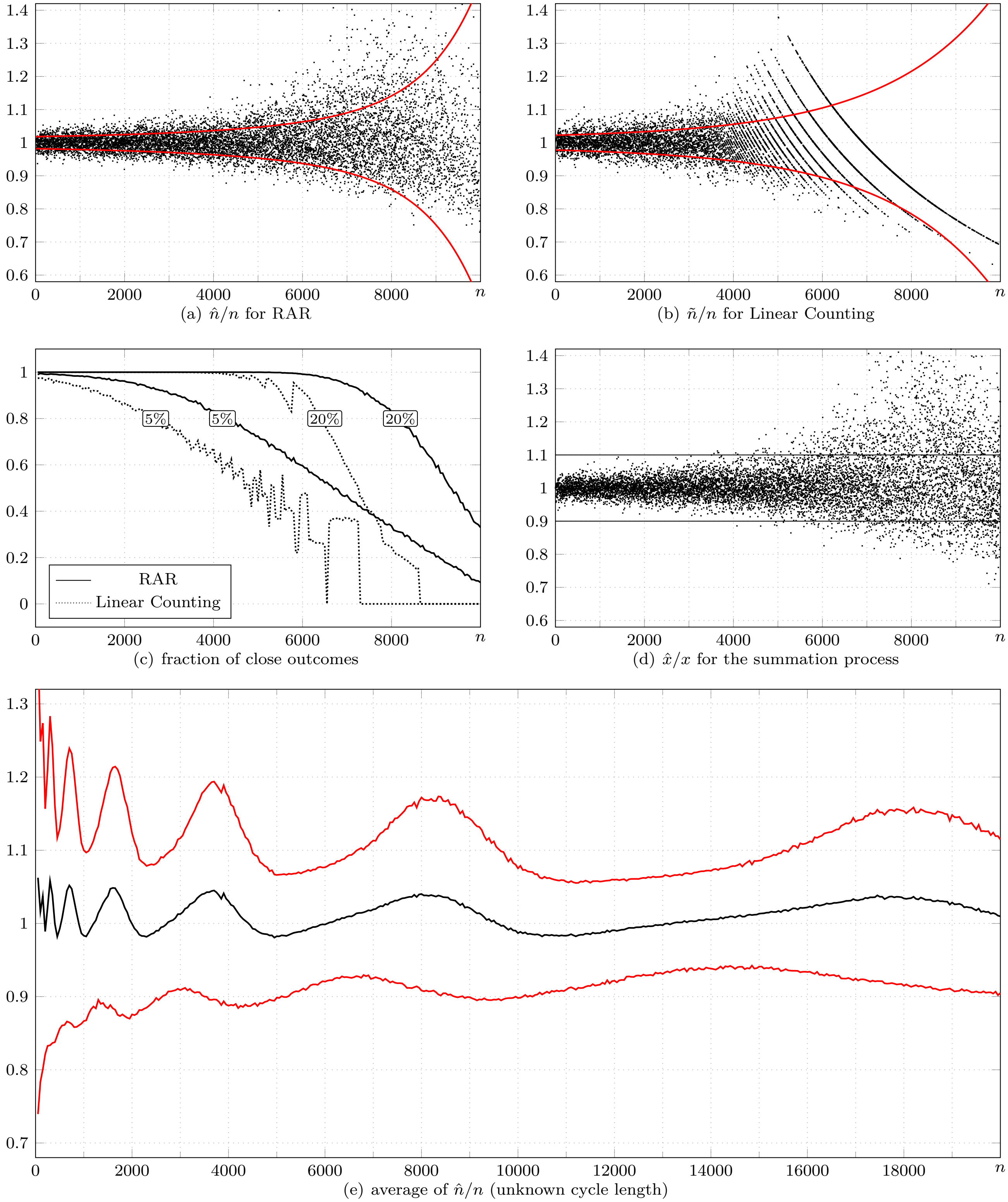
The fact that RAR can be executed asynchronously follows from the continuous nature of the underlying process. Interestingly, our analysis as well as computer simulations have shown that the continuous nature brings some additional advantages. These advantages are evident when RAR is compared with Linear Counting by Whang et al. [29], which can be considered as a discrete analogue of our solution. In Linear Counting each node places a ball (beeps in our scenario) randomly in one out of m bins (m time slots). The size estimation is based on the fraction V of empty bins (idle slots), $\tilde{n} = -m\ln(V)$. To compare the estimators we assume $m = 1/a$, so that a single pass of both algorithms ends within one time unit. By the notation of Theorem 4.2 we have $\mathbb{E}\left(\hat{n}\right) = n + \sigma_3(an)a^2n^2 + \mathcal{O}\left(n^{-1}\right)$ and $\mathbb{V}\mathrm{ar}\left(\hat{n}\right) = 2\sigma_3(an)an^2 + \mathcal{O}\left(an\right)$. From [29] we get $\mathbb{E}\left(\tilde{n}\right) = n + \sigma_2(an)a^2n^2/2$ and $\mathbb{V}\mathrm{ar}\left(\tilde{n}\right) = \sigma_2(an)an^2$. Because $0 < \sigma_3(x) < \sigma_2(x)/2$ our estimator is less biased and more concentrated. Moreover, unlike in RAR, implementation of balls-and-bins model would required synchronization mechanism and wouldn't allow for summation of fractional values.

In Figures 4a and 4b we present a typical results of a simulation of RAR (with a beep length $a = 1/1000$) and Linear Counting (with $m = 1000$ slots), respectively. For each network size $n \in \{1, \ldots, 10^4\}$ a dot represents a value taken by the corresponding estimator normalized by n. Lines represent, accordingly, theoretical values of $1 \pm \sqrt{\mathbb{V}\mathrm{ar}\left(\hat{n}/n\right)}$ and $1 \pm \sqrt{\mathbb{V}\mathrm{ar}\left(\tilde{n}/n\right)}$. One can see that results of Linear Counting are more dispersed and have artifacts related to its discrete nature. The superiority of RAR is even more apparent in Figure 4c, where for each network size n we calculated the fraction of simulations for which the relative error was smaller than 5% and 20% (10^4 simulations per data point). The results in Figures 4a and 4c confirm that $\hat{n}$ is almost unbiased and highly concentrated for the wide range of network sizes n. In fact, it is clear that the scope of its validity is much broader than implied by the formal analysis, which works only up to $n < 1/a$. Naturally, both algorithms cease to work when the channel is totally occupied. However, Figure 4c suggests that the continuous process is more robust in this regard. The expected number of arcs N_a of length a needed to cover the whole circle is asymptotically $\mathbb{E}\left(N_a\right) \sim a^{-1}\log a^{-1}$ as $a \searrow 0$ (see [26]). When the channel is totally occupied RAR returns $+\infty$. One could consider a modified algorithm in which circle length is enlarged until finite estimation is obtained. That is, we start with $a = 1/2$ and a circle of unit length. If the channel is totally occupied we double the length of circle and rerun RAR. Figure 4e presents empirical values of $\mathbb{E}\left(\hat{n}/n\right)$ (black) and standard deviation (red) for the modified algorithm (results averaged over 10^4 simulations for each network size).

In Figure 4d we plotted (similarly as in Figure 4a) results for the estimator $\hat{x}$ used in the summation algorithm sketched in Subsection 3.1. In this case we have fixed $\varepsilon = 1/2$ and for each n the numbers $x_1, x_2, \ldots, x_n$ have been chosen independently and uniformly at random. The results suggest that the estimator $\hat{x}$ is reasonable and seems to have properties similar to properties of the estimator $\hat{n}$.

6 Final Remarks

The beeping model has proved that it allows to design interesting algorithms with strong theoretical guarantees and yet implementable in realistic scenarios. Our solution shows that the model establishes a good framework for continuous processes, which in practice requires weaker assumptions about the underlying communication mechanism. More interestingly, our algorithm behaves better than its discrete (and synchronous) counterpart. It would be interesting to find other instances of this phenomenon. Let us also remark that the extension of the scope of the formal analysis for $a \geq 1/n$ and its generalization to the estimator of the sum seems to be attainable by the technique we have proposed.



(a) $\hat{n}/n$ for RAR

(b) $\tilde{n}/n$ for Linear Counting

(c) fraction of close outcomes

(d) $\hat{x}/x$ for the summation process

(e) average of $\hat{n}/n$ (unknown cycle length)

Figure 4: Results of computer simulations of RAR and Linear Counting.

References

[1] Yehuda Afek, Noga Alon, Ziv Bar-Joseph, Alejandro Cornejo, Bernhard Haeupler, and Fabian Kuhn. Beeping a maximal independent set. *Distributed Computing*, 26(4):195–208, 2013.

[2] Yehuda Afek, Noga Alon, Omer Barad, Eran Hornstein, Naama Barkai, and Ziv Bar-Joseph. A biological solution to a fundamental distributed computing problem. *science*, 331(6014):183–185, 2011.

[3] Carlos Baquero, Paulo S. Almeida, Raquel Menezes, and Paulo Jesus. Extrema propagation: Fast distributed estimation of sums and network sizes. *IEEE Transactions on Parallel and Distributed Systems*, 23(4):668–675, 2012.

[4] Philippe Chassaing and Lucas Gerin. Efficient estimation of the cardinality of large data sets. In *4th Colloquium on Mathematics and Computer Science*, pages 419–422. DMTCS Proceedings, 2006.

[5] Jacek Cichoń, Jakub Lemiesz, Wojciech Szpankowski, and Marcin Zawada. Two-phase cardinality estimation protocols for sensor networks with provable precision. In *Proceedings of WCNC'12*, Paris, France, 2012. IEEE.

[6] Jacek Cichoń, Jakub Lemiesz, and Marcin Zawada. On cardinality estimation protocols for wireless sensor networks. In *ADHOC-NOW*, volume 6811 of *Lecture Notes in Computer Science*, Paderborn, Germany, July 2011. Springer.

[7] Jacek Cichoń, Jakub Lemiesz, and Marcin Zawada. On size estimation protocols for sensor networks. In *CDC*, Proceedings of 51st Annual Conference on Decision and Control, pages 5234–5239. IEEE, 2012.

[8] Reuven Cohen and Boris Kapchits. Continuous neighbor discovery in asynchronous sensor networks. *Networking, IEEE/ACM Transactions on Networking*, 19(1):69–79, Feb 2011.

[9] Joanne R. Collier, Nicholas A. M. Monk, Philip K. Maini, and Julian H. Lewis. Pattern formation by lateral inhibition with feedback: a mathematical model of delta-notch intercellular signalling. *Journal of Theoretical Biology*, 183(4):429–446, 1996.

[10] Alejandro Cornejo and Fabian Kuhn. Deploying wireless networks with beeps. In Nancy A. Lynch and Alexander A. Shvartsman, editors, *DISC*, volume 6343 of *Lecture Notes in Computer Science*, pages 148–162. Springer, 2010.

[11] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *Proceedings of the Conference on Analysis of Algorithms (AofA'07)*, pages 127–146, 2007.

[12] Philippe Flajolet and G. Nigel Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31(2):182–209, 1985.

[13] Philippe Flajolet and Robert Sedgewick. Mellin transforms and asymptotics: finite differences and Rice's integrals. *Theoretical Computer Science*, 144(1–2):101–124, 1995.

[14] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[15] Roland Flury and Roger Wattenhofer. Slotted programming for sensor networks. In Tarek F. Abdelzaher, Thiemo Voigt, and Adam Wolisz, editors, *IPSN*, pages 24–34. ACM, 2010.

[16] Frédéric Giroire. Order statistics and estimating cardinalities of massive data sets. *Discrete Applied Mathematics*, 157(2):406–427, 2009.

[17] Venkat Iyer, Andrei Pruteanu, and Stefan Dulman. Netdetect: Neighborhood discovery in wireless networks using adaptive beacons. In *Self-Adaptive and Self-Organizing Systems (SASO), 2011 Fifth IEEE International Conference on*, pages 31–40, Oct 2011.

[18] Károly Jordán. *Calculus of Finite Differences*. AMS Chelsea Publishing Series, 1965.

[19] Tomasz Jurdziński, Mirosław Kutyłowski, and Jan Zatopiański. Energy-efficient size approximation of radio networks with no collision detection. In *Proceedings of COCOON '02*, pages 279–289. Springer-Verlag, 2002.

[20] Marcin Kardas, Marek Klonowski, Piotr Syga, and Szymon Wilczek. Obfuscated counting in single-hop radio network. In *Proceedings of ICPADS'12*, Singapore, 2012. IEEE.

[21] Damon Mosk-Aoyama and Devavrat Shah. Computing separable functions via gossip. In *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, PODC '06, pages 113–122, 2006.

[22] Arik Motskin, Tim Roughgarden, Primoz Skraba, and Leonidas J. Guibas. Lightweight coloring and desynchronization for networks. In *INFOCOM*, pages 2383–2391, 2009.

[23] Kyung-Joon Park, Jihyuk Choi, Jennifer C. Hou, Yih-Chun Hu, and Hyuk Lim. Optimal physical carrier sense in wireless networks. *Ad Hoc Networks*, 9(1):16–27, 2011.

[24] Ill-Keun Rhee, Jaehan Lee, Jangsub Kim, Erchin Serpedin, and Yik-Chung Wu. Clock synchronization in wireless sensor networks: An overview. *Sensors*, 9(1):56–85, 2009.

[25] Herbert E. Robbins. On the measure of a random set. *The Annals of Mathematical Statistics*, 15(1), 1944.

[26] Andrew F. Siegel. Random arcs on the circle. *Journal of Applied Probability*, pages 774–789, 1978.

[27] Andrew S. Tanenbaum and David Wetherall. *Computer Networks*. Pearson Prentice Hall, 5th ed, 2011.

[28] Jana van Greunen and Jan Rabaey. Lightweight time synchronization for sensor networks. In *Proceedings of WSNA'03*, pages 11–19, San Diego, USA, 2003. ACM.

[29] Kyu-Young Whang, Brad T. Vander Zanden, and Howard M. Taylor. A linear-time probabilistic counting algorithm for database applications. *ACM Transactions on Database Systems*, 15(2):208–229, 1990.