Average Counting via Approximate Histograms - Preliminary Report

Jacek Cichoń and Karol Gotfryd
Department of Computer Science
Faculty of Fundamental Problems of Technology
Wrocław University of Technology
Poland
{jacek.cichon, karol.gotfryd}@pwr.edu.pl

***Abstract*—In this paper we propose a novel method of solving the averaging problem for distributed Wireless Sensors Networks. Our method uses a set of probabilistic counters and allows to find the approximation of the average of a set of measures done by sensor network with arbitrary, controlled by two parameters, precision. The exchange of information is based on broadcasting method exploiting extreme propagation technique. Our method require O(D) rounds, where D is the diameter of the network.**

***Index Terms*—message propagation, distributed algorithm, extreme propagation, average, probabilistic counter, exponential distribution, Erlang distribution, Delta method**

I. Introduction

The problem of averaging in distributed Wireless Sensors Networks (WSN) have been widely studied in a series of paper (see e.g. [1], [2]). Recent versions of this algorithms (see [3]–[5]) take benefit of the broadcast nature of the wireless communication channels. But in all these algorithms the convergence speed to the true average is large and highly exceed the diameter of the network, which is the obvious lower bound on the number of rounds needed to compute the exact average.

The idea of using probabilistic counters for estimation of aggregates in networks was introduced in [6], [7]. In this paper we propose a novel method of estimation of average in the distributed enviroment. Our method is based on the extreme propagation technique popularized by C. Baquero, P. S. Almeida, and R. Menezes in 2009 in [8] and on the notion of probabilistic counters invented by 1977 by Robert Morris in [9]. It runs in $\mathrm{O}(D)$ steps, where D is the diameter of network. Its precision is controlled by two parameters. In our method the approximation of the average is build from approximated histograms built using probabilistic counters.

Let us mention, that we are not assuming any knowledge on the size of the network. We assume only that we know some reasonable upper bound on the network diameter.

A. Mathematical Notation and Background

We denote by $|A|$ the cardinality of a set A. By $\Gamma(x)$ we denote the standard generalization of the factorial function. Notice that $\Gamma(n) = (n-1)!$ for any integer $n \geq 1$.

We denote by $\mathbf{E}[X]$ and $\mathbf{var}[X]$ the expected value and the variance of the random variable X, respectively. We denote by $\xrightarrow{d}$ the convergence in distribution of random variables. We will use the following property of this convergence: if $X_n \xrightarrow{d} X$ and g is a continuous function, then $\lim_n \mathbf{E}[g(X_n)] = \mathbf{E}[g(X)]$.

Let us recall that a random variable X has the exponential distribution with parameter λ ($X \sim \mathrm{Exp}(\lambda)$) if its density function f_X is given by the formula $f_X(x) = \lambda \exp(-\lambda x)$. If $X_1, \ldots, X_n \sim \mathrm{Exp}(\lambda)$ and are independent and $Y = \min\{X_1, \ldots, X_n\}$ then $Y \sim \mathrm{Exp}(\lambda)$. If $X_1, \ldots, X_L$ are independent random variables with a common $\mathrm{Exp}(\mu)$ distribution, then the sum $S = X_1 + \ldots + X_L$ have the Erlang distribution with parameters L and μ ($S \sim \mathrm{Erl}(L, \mu)$), i.e. its density function is given by formula

$$f_{L,\mu}(x) = \frac{\mu^L x^{L-1} e^{-\mu x}}{(L-1)!} \ . \tag{1}$$

II. Histograms

We assume that network is modeled by a connected graph with relatively small diameter D. The edges of this graph corresponds with bidirectional communication links. Suppose that the network consist on n nodes numbered by $\{1, \ldots, n\}$ and that each node stores a value T_k. Let $\vec{T} = (T_i)_{i=1,\ldots,n}$. Our goal is to estimate the mean

$$\mathrm{avg}(\vec{T}) = \frac{1}{n} \sum_{k=1}^{n} T_k$$

in an efficient and easy way.

Using the extreme propagation technique in its basic form we may assume that each node knows the values $m = \min\{T_i : i = 1, \ldots, n\}$ and $M = \max\{T_i : i = 1, \ldots, n\}$. If $m = M$ then the average value of the sequence (T_i) is known. Suppose hence that $m < M$ and let $\Delta = M - m$.

We fix a parameter K and we split the interval $[m, M]$ into K intervals of equal length: we put $I_i = [m + \frac{\Delta}{K}(i-1), m + \frac{\Delta}{K} i)$ for $i = 1, \ldots K-1$ and $I_K = [m + \frac{\Delta}{K}(K-1), M]$. Let w_i denotes the middle point of the interval I_i, i.e. we put $w_i = m + \frac{\Delta}{K}(i - \frac{1}{2})$.

Let $H_i = |\{k : T_k \in I_i\}|$, for $i \in \{0, \ldots, K-1\}$. We call the vector $\vec{H} = (H_i)_{i=1 \ldots K}$ a *histogram* of the data $(T_i)_{i=1,\ldots n}$. We are going to approximate the average

value of observed data $(T_i)_{i=1,\ldots,n}$. For arbitrary vector $\vec{k} = (k_1, \ldots, k_K)$ we define a function

$$\mathrm{am}_{\vec{k}}(x_1, \ldots, x_K) = \frac{\sum_{i=1}^{K} k_i x_i}{\sum_{i=1}^{K} x_i}$$

We approximate the average value of observed data $(T_i)_{i=1,\ldots,n}$ be the value $\mathrm{am}(\vec{H}) = \mathrm{am}_{\vec{w}}(\vec{H})$, where $\vec{w}$ is the sequence of middle poins of histograms intervals, i.e. we define

$$\mathrm{am}(\vec{H}) = \frac{\sum_{i=1}^{K} w_i H_i}{\sum_{i=1}^{K} H_i} \tag{2}$$

In this approach each observed value is approximated by the nearest element from the set of middle points $(w_i)_{i=1,\ldots,K}$, so some errors in this method is unavoidable. We call this error a *discretization error*. This error is controlled by the number K of sub-intervals into which we divide the range of observed data and by the spread of observed data:

Theorem 1 (Discretization error)**.** *For arbitrary vector $\vec{T}$ of observed data we have*

$$\frac{\left|\mathrm{am}(\vec{H}) - \mathrm{avg}(\vec{T})\right|}{M - m} \leq \frac{1}{2K} \ ,$$

where $m = \min\{T_i : i = 1, \ldots, n\}$ and $M = \max\{T_i : i = 1, \ldots, n\}$.

A. Approximate Counters

Probabilistic counters where intensively investigated in last years. They were invented in 1977 by Robert Morris (see [9]). This version was carefully analyzed in the early 1980s by Philippe Flajolet (see [10]), who coined the name Approximate Counting In a more recent investigations some other methods where proposed for estimation of a cardinality of a stream of data. Some of them are well suited for counting the size of distributed network (see e.g. [11], [12]).

In this paper we use a method based on exponential distributions. It uses the following property of this distribution: if $X_1, \ldots, X_n$ are independent random variables with the common distribution $\mathrm{Exp}(1)$, then the random variable $Y = \min\{X_1, \ldots, X_n\}$ has the distribution $\mathrm{Exp}(n)$. One random variable with $\mathrm{Exp}(n)$ is not sufficient for the estimation of the the parameter n. However, if we have a sequence $Y_1, \ldots, Y_L$ of independent random variables with $\mathrm{Exp}(n)$ distribution where $L > 2$, then the random variable $Z = Y_1 + \ldots + Y_L$ has the Erlang distribution $\mathrm{Erl}(L, n)$. Moreover, from Eq. 1 we easily deduce that $\mathbf{E}\left[\frac{L-1}{Z}\right] = n$ and $\mathbf{var}\left[\frac{L-1}{Z}\right] = \frac{n^2}{L-2}$. Therefore, the random variable $C = \frac{L-1}{Z}$ is an unbiased estimator of the number n and its precision is controlled by the parameter L. We will use this approach in this paper.

B. Approximated Histograms

Let $\vec{T} = (T_i)_{i=1,\ldots,n}$ be the sequence of observed values. We split the interval $[\min(\vec{T}), \max(\vec{T})]$ into K intervals $(I_i)_{i=1,\ldots,I_K}$ of equal lengths. We associate with each interval I_i a approximate counter $C_{L,i}$ counting the number $H_i = |\{k : T_k \in I_i\}|$ based on the Erlang distribution $\mathrm{Erl}(L, H_i)$.

We call the vector $\vec{C}_L = (C_{L,i})_{i=1\ldots K}$ *an approximate histogram* of the data $(T_i)_{i=1,\ldots n}$. Let $\vec{H}$ be the histogram obtained from $\vec{T}$. We will prove the number $\mathrm{am}(\vec{C}_L)$ is an asymptotically unbiased estimator of the number $\mathrm{am}(\vec{H})$.

Theorem 2. *Let $\vec{H} \in \mathbb{R}^K$ be a vector of non-negative numbers such that $C = \sum_{i=1}^{K} H_i > 0$. Then*

$$\sqrt{L}\big(\mathrm{am}(\vec{C}_L) - \mathrm{am}(\vec{H})\big) \xrightarrow{d} \mathcal{N}(0, s^2) \ ,$$

where $s^2 = \sum_{i=1}^{K} \left(\sum_{j=1}^{K} (j-i) H_i H_j\right)^2 \cdot C^{-4}$.

Proof. Let us fix i such that $H_i \geq 1$. Then $C_{L,i} = \frac{L-1}{X}$, where $X \sim \mathrm{Erl}(L, H_i)$. From Lemma 1 proved in Section VI we deduce that the sequence $\sqrt{L}(C_{L,i} - H_i)$ converges (if L grows to infinity) in distribution to the normal distribution $\mathcal{N}(0, H_i^2)$. Notice that if $H_i = 0$, then $C_{L+1,i} = 0$, so $\sqrt{L}(C_{L,i} - H_i) = 0$, hence also in this case we have a convergence to $\mathcal{N}(0, 0)$, interpreted as the Dirac's delta function. Observe also that random variables $C_{L,1}, \ldots, C_{L,K}$ are independent. Therefore

$$\sqrt{L}(C_{L,1} - H_1, \ldots, C_{L,K} - H_K) \xrightarrow{d} \mathcal{N}(0, \Sigma) \ ,$$

where $\Sigma = \mathrm{diag}(H_1^2, \ldots, H_K^2)$ is the square diagonal matrix with elements $(H_1^2, \ldots, H_K^2)$ on the main diagonal.

We are going to apply the Multivariate Delta Method to the function $\mathrm{am}()$. Notice that

$$\frac{d}{dx_i}\mathrm{am}() = \frac{d}{dx_i}\frac{\sum_{j=1}^{K} w_j x_j}{\sum_{j=1}^{K} x_j} = \frac{\sum_{j=1}^{K} (w_i - w_j) x_j}{(\sum_{j=1}^{K} x_j)^2}$$

Let $\nabla_{\vec{H}}$ be the gradient $(\frac{d}{dx_1}\mathrm{am}(), \ldots, \frac{d}{dx_K}\mathrm{am}())$ evaluated at the point $\vec{H} = (H_1, \ldots, H_K)$. From the Multivariate Delta Method we get

$$\sqrt{L}(\mathrm{am}(\vec{C}_L) - \mathrm{am}(\vec{H})) \xrightarrow{d} \mathcal{N}(0, \nabla_{\vec{H}}^T \Sigma \nabla_{\vec{H}}) \ ,$$

hence

$$\sqrt{L}(\mathrm{am}(\vec{C}_L) - \mathrm{am}(\vec{H})) \xrightarrow{d} \mathcal{N}(0, s^2) \ ,$$

where

$$s^2 = \sum_{i=1}^{K} \left(\frac{\sum\limits_{j=1}^{K} (w_j - w_i) H_j}{(\sum\limits_{j=1}^{K} H_j)^2} \right)^2 H_i^2 = \frac{\sum_{i=1}^{K} \left(\sum\limits_{j=1}^{K} (w_j - w_i) H_i H_j \right)^2}{\left(\sum\limits_{i=1}^{K} H_i \right)^4} \ .$$

Hence theorem is proved. □

Corollary 1. $\lim_{L\to\infty} \mathbf{E}\left[\mathrm{am}(\vec{C}_L)\right] = \mathrm{am}(\vec{H})$

Corollary 2. *Let* $C = \sum_{i=1}^{K} H_i$. *If* $C > 0$ *then*

$$\mathbf{var}\left[\mathrm{am}(\vec{C}_L)\right] = \frac{1}{L}\cdot\sum_{i}^{K}\left(\sum_{j=1}^{K}(w_j - w_i)H_iH_j\right)^2 C^{-4} + \mathrm{o}\left(\frac{1}{L}\right) .$$

Proof. If $X_n \xrightarrow{d} Z$ then for every continuous function g we have $\lim_n E[g(X_n)] = E[g(Z)]$. If we apply this property for function $g(x) = x^2$ to conclusion of Theorem 2 then we get

$$\lim_{L\to\infty} L\cdot\mathbf{var}\left[\mathrm{am}(\vec{C}_L)\right] = \mathbf{var}\left[\mathcal{N}(0,s^2)\right] = s^2 ,$$

so the Corollary is proved. □

III. Precision of Approximated Histograms

Our goal is to compare the number $\mathrm{am}(\vec{H})$ (see Formula 2) with $\mathrm{am}(\vec{C}_L)$. In section V we shall discuss series of experimental result. For a proper interpretation of obtained result we will use the following measure of error of the estimate $\mathrm{am}(\vec{C}_L)$:

$$\mathrm{err}(\mathrm{am}(\vec{H}),\mathrm{am}(\vec{C}_L)) = \frac{|\mathrm{am}(\vec{H}) - \mathrm{am}(\vec{C}_L)|}{M-m} .$$

Notice that $0 \le \mathrm{err}(\mathrm{am}(\vec{H}),\mathrm{am}(\vec{C}_L)) \le 1$.

Theorem 3. *Let* $\vec{b} = (b_1,\ldots,b_k)$,$\alpha,\beta \in \mathbb{R}$ *and* $\alpha > 0$. *Let* $\vec{v} = (\alpha b_1+\beta,\ldots,\alpha b_k+\beta)$. *Then for arbitrary* $\vec{x},\vec{y} \in \mathbb{R}^k$ *we have*

$$\mathrm{err}(\mathrm{am}_{\vec{b}}(\vec{x}),\mathrm{am}_{\vec{b}}(\vec{y})) = \mathrm{err}(\mathrm{am}_{\vec{v}}(\vec{x}),\mathrm{am}_{\vec{v}}(\vec{y})) .$$

Proof. Notice that the distance $M-n$ may be calculated from coefficients wK and x_1, namely $M - m = (w_K - w_1)\frac{K+1}{K}$. For arbitrary $\vec{z} \in \mathbb{R}^k$ we have

$$\mathrm{am}_{\vec{v}}(\vec{z}) = \frac{\sum_{i=1}^{k} v_i z_i}{\sum_{i=1}^{k} z_i} = \frac{\sum_{i=1}^{k}(\alpha b_i+\beta)z_i}{\sum_{i=1}^{k} z_i} = \frac{\alpha\sum_{i=1}^{k} b_i z_i + \beta\sum_{i=1}^{k} z_i}{\sum_{i=1}^{k} z_i} = \alpha\frac{\sum_{i=1}^{k} b_i z_i}{\sum_{i=1}^{k} z_i} + \beta = \alpha\cdot\mathrm{am}_{\vec{b}}(\vec{z}) + \beta .$$

Therefore

$$\mathrm{err}(\mathrm{am}_{\vec{v}}(\vec{x}),\mathrm{am}_{\vec{v}}(\vec{y})) = \frac{|\mathrm{am}_{\vec{v}}(\vec{x}) - \mathrm{am}_{\vec{v}}(\vec{y})|}{(v_k - v_1)\frac{K+1}{K}} = \frac{\alpha|\mathrm{am}_{\vec{b}}(\vec{x}) - \mathrm{am}_{\vec{b}}(\vec{y})|}{\alpha(v_k - v_1)\frac{K+1}{K}} = \mathrm{err}(\mathrm{am}_{\vec{b}}(\vec{x}),\mathrm{am}_{\vec{b}}(\vec{y})) .$$

□

From this theorem we deduce that investigation of errors of estimator of average values based on probabilistic counters may be reduced to such data, where the middle points $(w_i)_{i=1,\ldots,K}$ are fixed and are equal to $\vec{b} = (1,2,\ldots,K)$. In this case we have

$$\mathrm{am}_{\vec{b}}(\vec{x}) = \frac{\sum_{i=1}^{K} i\cdot x_i}{\sum_{i=1}^{K} x_i}$$

and (see Corollary 2) $\mathbf{var}\left[\mathrm{err}(\mathrm{am}(\vec{H}),\mathrm{am}(\vec{C}_L))\right] \approx h(H_1,\ldots,H_K)$ where

$$h(x_1,\ldots,x_K) = \frac{1}{L\cdot(K+1)^2}\frac{\left(\sum_{j=1}^{K}\left(\sum_{i=1}^{K}(j-i)x_ix_j\right)\right)^2}{\left(\sum_{i=1}^{K} x_i\right)^4}$$

Theorem 4 implies that when $\sum_{i=1}^{K} H_i = C$ is fixed, then the function h attached its maximum value at point $\vec{c} = (\frac{C}{2},0,\ldots,0,\frac{C}{2})$. In this case we have $h(\vec{c}) = \frac{1}{8L}\frac{(K-1)^2}{(K+1)^2}$. This case of highly concentrated data at two extremal values will be carefully discussed is Section V where we present results of numerical experiments.

In the case when $H_i = a$ for each $i = 1,\ldots,K$ we have $h(a,a,\ldots,a) = \frac{1}{12L}\frac{K^2-1}{K(K+1)^2} \le \frac{1}{12\cdot L\cdot K}$.

IV. Algorithm

In this section we show a pseudo-code of a a discussed in this paper algorithm. This algorithm is used by every node in the network. We assume that the communication in the network is divided into rounds and that in each round each pair of connected nodes can exchange information in both directions.

The input of this algorithm are:

1) D: upper approximation of a diameter of a network
2) m: minimal value of observed data
3) M: maximal value of observed data
4) K: number of sub-intervals dividing the range $[m, M]$
5) L: number of exponential random variables connected with each subintervals

We assume the in an initial phase, before running this algorithm an another algorithm calculate numbers m and M. W also assume that the number D is known. In fact, it is sufficient to know some upper bound on the network diameter, because this algorithm stabilize (no new messages is send) after D^* where D^* is the precise network diameter. Both numbers K and L have influence on the precision of obtained estimator of the average of observable data.

```
1: function CountAvgMean(D,m,M,K,L)
2:     // Initialization
3:     T = observed value
4:     for a=1 ... K do
5:         for j=1 ... L do
6:             X[a][j] = +∞;
7:         end for
```

```
    end for
    Δ = M − m
    find a such that T ∈ [m + (Δ/K)(a − 1), m + (Δ/K)a]
    for j=1 ... L do
        X[a][j] = RandomExp(1)
    end for
    send pair (a,X[a]) to all neighbors
    // broadcasting loop
    for I=1 ... D do
        C = X;
        for all received (a,Y) do
            for j=1 ... L do
                C[a][j] = min(C[a][j],Y[j])
            end for
        end for
        for a=1 ... K do
            if C[a] ≠ X[a] then
                X[a] = C[a]
                send pair (a,X[a]) to all neighbors
            end if
        end for
    end for
    // final calculations
    for a=1 ... K do
        S[a] = 0;
        for j=1 ... L do
            S[a] = S[a] + X[a][j]
        end for
        H[a] = (L-1)/S[a]
    end for
    S1 = Σ_{i=1}^{K} (m + (Δ/K)(i − 1/2)) H[i]
    S2 = Σ_{i=1}^{K} H[i]
    return S1/S2
end function
```

V. Experiments

At the end of Sec. III we showed that we should check the precision of proposed estimator on a symmetric distribution concentrated on two points. This case will be discussed in Sec. V-A. In the next section we will show how our estimator behaves on randomly distributed data.

Let us notice that in out experiments we take into account both kinds of errors. The first one is due to the discretization error (see Thm. 1) and is controlled by the number K of sub-intervals representing data. The second one is due to probabilistic nature of probabilistic counters and it is controlled by the number L of of probabilistic counters attached to every sub-interval.

A. Worst case

Fig. 1 depicts the outcomes of the experiments of the worst case data for different network sizes n varying from 50 to 10000 with step 10. For each n we performed 100 independent experiments where $n/2$ nodes have the value 0 and the remaining $n/2$ the value 1. The parameters were set to $K = 4$ and $L = 50$. We can observe that in all experiments our algorithm counts the average with 20% precision.

Fig. 2 shows the maximal and average errors for these experiments as a function of the network size. We can observe that regardless of the number of nodes in almost all experiments the average is counted with 20% precision and the mean error is about 5%.

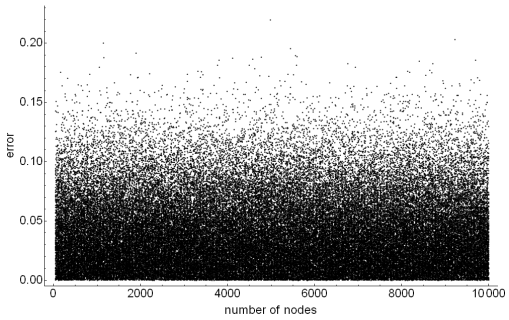

Fig. 1. Errors of algorithm CountAvgMean with parameters $K = 4$, $L = 50$ for data concentrated on end points with respect to the number of nodes in the network.

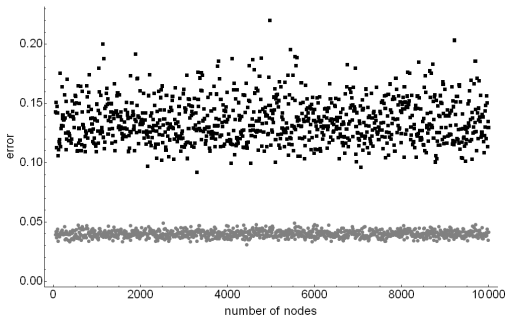

Fig. 2. Maximal and average errors of algorithm CountAvgMean with parameters $K = 4$, $L = 50$ for data concentrated on end points with respect to the number n of nodes in the network. For each n we run 100 experiments.

Finally, Fig. 3 and 4 show the results of experiments performed for data concentrated at two extremal points, where a fraction p of n nodes have value 0 and $(1 - p)$ value 1 for $n = 100$, 1000 and 10000 and for p from the set $\{0.05i\colon 1 \leq i < 20\}$. As previously we chose $K = 4$ and $L = 50$. For each n and p 1000 independent experiments were performed. We can observe that both mean and maximal error of the proposed estimator don't depend on the network size and decrease as the distribution of the values becomes more skewed.

B. Uniform and Normal Distribution

Fig. 5 presents the outcomes of the experiments for different network sizes n for the case where the randomly generated data are distributed *uniformly* over the unit interval. We performed 100 independent experiments for each n in the range from 50 to 5000 with step 10. In each experiment the

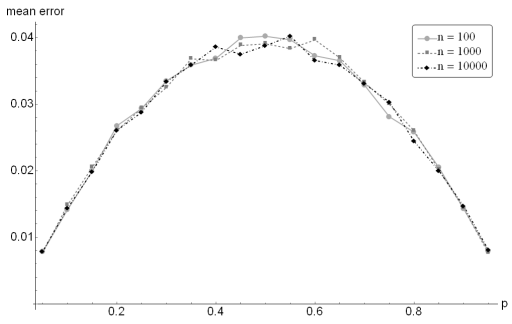

Fig. 3. Mean errors of algorithm COUNTAVGMEAN with parameters $K = 4$, $L = 50$ for data concentrated on end points with respect to the fraction p of nodes with minimal value. Experiments were repeated independently 1000 times for networks of size 100, 1000 and 10000.

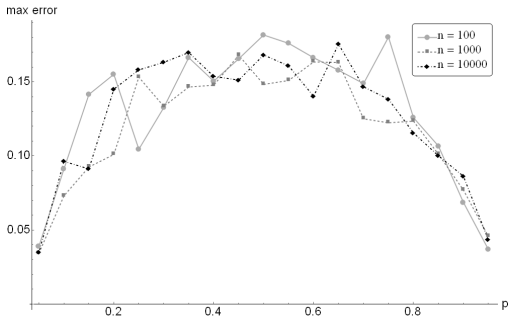

Fig. 4. Maximal errors of algorithm COUNTAVGMEAN with parameters $K = 4$, $L = 50$ for data concentrated on end points with respect to the fraction p of nodes with minimal value. Experiments were repeated independently 1000 times for networks of size 100, 1000 and 10000.

interval between the minimal and maximal value was split into $K = 20$ equal sub-intervals and $L = 20$ probabilistic counters were used. The maximal and average errors as a function of the network size are shown in Fig. 6. We can see that for each n the mean error of our estimator is below 2% and in all experiments the approximation error doesn't exceed 8%.

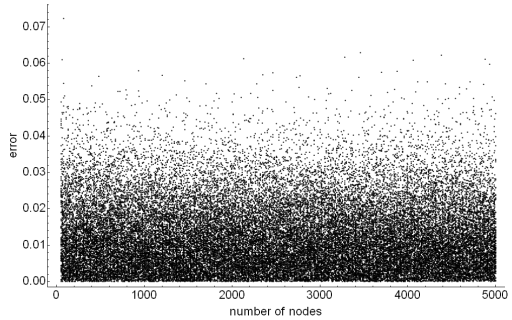

Fig. 5. Errors of algorithm COUNTAVGMEAN with parameters $K = 20$, $L = 20$ for randomly generated data from uniform distribution over $[0, 1]$. with respect to the number of nodes in the network.

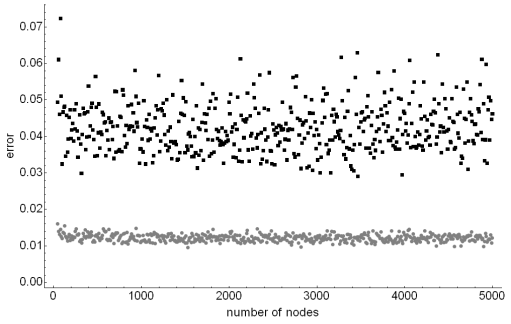

Fig. 6. Maximal and average errors of algorithm COUNTAVGMEAN with parameters $K = 20$, $L = 20$ for randomly generated data from uniform distribution over $[0, 1]$. with respect to the number n of nodes in the network. For each n we run 100 experiments.

We performed similar experiments to the previous one for random data following the *normal distribution* with mean 1000 and the variance equals to 100. As before, for each network size n between 50 and 5000 (with step 10) we ran 100 independent simulations with the same choice of parameters (i.e. $K = L = 20$). Fig. 7 and 8 depict the errors of the individual experiments and the maximal and average errors for each n, respectively. Observe that in this case the average is estimated with 5% precision.

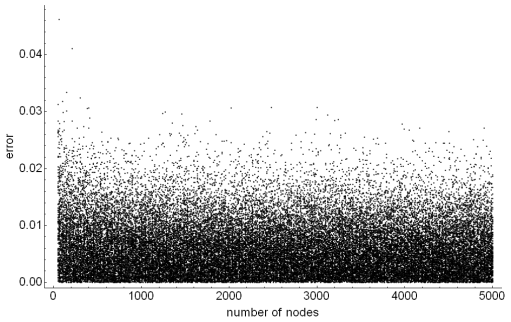

Fig. 7. Errors of algorithm COUNTAVGMEAN with parameters $K = 20$, $L = 20$ for random data following normal distribution with mean equal to 1000 and variance 100 with respect to the number of nodes in the network.

VI. PROOFS

Lemma 1. *Suppose that* $X_L \sim \mathrm{Erl}(L, m)$*, where* $L > 2$*. Let* $Y_L = \frac{L-1}{X}$*. Then* $\mathbf{E}[Y_L] = m$*,* $\mathbf{var}[Y_L] = \frac{m^2}{L-2}$ *and the sequence* $\sqrt{L}(Y_L - m)$ *converges in distribution to the normal distribution* $\mathcal{N}(0, m^2)$*.*

Proof of this lemma is skipped and due to restrictions on the length of the article.

Theorem 4. *Let* $c > 0$*,* $k \geq 2$*,* $\Sigma_{c,k} = \{\vec{x} \in \mathbb{R}^k : \sum_{i=1}^k x_i = c \wedge \bigwedge_{i=1}^k (x_i \geq 0)\}$ *and*

$$f(x_1, \ldots, x_k) = \sum_{j=1}^{k} x_j^2 \left(\sum_{i=1}^{k} (j - i) x_i \right)^2 .$$

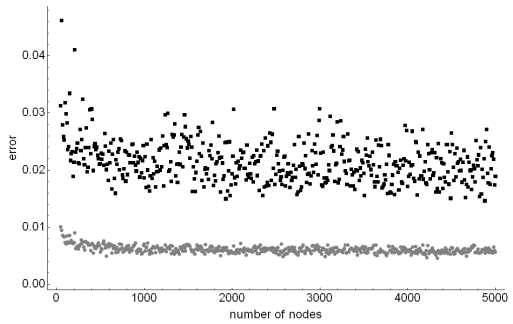

Fig. 8. Maximal and average errors of algorithm COUNTAVGMEAN with parameters $K = 20$, $L = 20$ for random data following normal distribution with mean equal to 1000 and variance 100 with respect to the number n of nodes in the network. For each n we run 100 experiments.

Let $\vec{b} = (\frac{c}{2}, 0, \ldots, 0, \frac{c}{2}) \in \Sigma_{c,k}$. *Then* $f(\vec{b}) = \sup\{f(\vec{x}) : \vec{x} \in \Sigma_{c,k}\}$ *and* $f(\vec{b}) = \frac{(k-1)^2 c^4}{8}$.

Proof. Notice that $\Sigma_{c,k}$ is a compact subset of $\mathbb{R}^k$ and that f is a continuous function on $\Sigma_{c,k}$. Therefore there exists a point $\vec{b} \in \Sigma_{c,k}$ such that

$$f(\vec{b}) = \sup\{f(\vec{x}) : \vec{x} \in \Sigma_{c,k}\} \ .$$

We shall prove that $f(\vec{b}) = f((\frac{c}{2}, 0, \ldots, 0, \frac{c}{2}))$.

Lemma 2. *Suppose that* $\vec{x} = (x_1, \ldots, x_k) \in \Sigma_{c,k}$, $1 < l < k$ *and* $x_l > 0$. *Let*

$$\vec{x'} = \left(x_1 + \frac{k-l}{k-1} x_l, x_2, \ldots, x_{l-1}, 0, x_{l+1}, \ldots, x_{k-1}, x_k + \frac{l-1}{k-1} x_l\right) \ .$$

Then $f(\vec{x}) \leq f(\vec{x'})$.

We omit the proof of this lemma. We show only main hint: namely if we define $I_j(y_1, \ldots, y_k) = \sum_{i=1}^{k} (j-i) y_i$ then we have $I_j(\vec{x'}) = I_j(\vec{x})$ for each $j = 1, \ldots, k$.

From Lemma 2 we deduce that the maximal value of the function f on the set $\Sigma_{c,n}$ is attached on the subset $\{(a, 0, \ldots, 0, c-a) : 0 \leq a \leq c\}$. Let us observe that

$$f(\alpha, 0, \ldots, 0, c-\alpha) = 2(k-1)^2 \alpha^2 (c-\alpha)^2.$$

Therefore the function $g(\alpha) = f(\alpha, 0, \ldots, 0, c-\alpha)$ reaches its maximum on the interval $[0, c]$ at point $\alpha = \frac{c}{2}$ and $g(\frac{c}{2}) = \frac{(k-1)^2 c^4}{8}$. Hence the theorem is proved. □

VII. CONCLUSIONS AND FUTURE WORKS

The proposed in paper method of counting the average value in distributed environment may be summarize as follows: represent data by a histogram of K bins and use a sequence of L independent probabilistic counters connected with each bin to count approximately the number of balls in each bin. This can be done in an efficient way using broadcasting and the extreme propagation technique. The worst case for the proposed method are symmetric data concentrated at two points. Using 200 probabilistic counters we get a precision of order 20%. However, at the end of algorithm each node has at its disposal an approximated histogram, so it can recognize this phenomena may take appropriate action.

In the further work we are planning to to investigate the best strategy for choosing optimal parameters K and L given the constrain $K \cdot L = C$ and to extend our method to other aggregates than the mean value.

ACKNOWLEDGMENTS

This paper was supported by Polish NCN grant nr 2013/09/B/ST6/02258.

REFERENCES

[1] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE/ACM Trans. Netw.*, vol. 14, no. SI, pp. 2508–2530, Jun. 2006. [Online]. Available: http://dx.doi.org/10.1109/TIT.2006.874516

[2] A. G. Dimakis, A. D. Sarwate, and M. J. Wainwright, "Geographic gossip: Efficient averaging for sensor networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 3, pp. 1205–1216, 2008. [Online]. Available: http://dx.doi.org/10.1109/TSP.2007.908946

[3] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione, "Broadcast gossip algorithms for consensus," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2748–2761, July 2009.

[4] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *44th Symposium on Foundations of Computer Science (FOCS 2003), 11-14 October 2003, Cambridge, MA, USA, Proceedings*. IEEE Computer Society, 2003, pp. 482–491. [Online]. Available: http://dx.doi.org/10.1109/SFCS.2003.1238221

[5] F. Iutzeler, P. Ciblat, W. Hachem, and J. Jakubowicz, "New broadcast based distributed averaging algorithm over wireless sensor networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 3117–3120.

[6] J. Considine, F. Li, G. Kollios, and J. Byers, "Approximate aggregation techniques for sensor databases," in *Data Engineering, 2004. Proceedings. 20th International Conference on*, March 2004, pp. 449–460.

[7] S. Nath, P. B. Gibbons, S. Seshan, and Z. R. Anderson, "Synopsis diffusion for robust aggregation in sensor networks," in *Proceedings of the 2Nd International Conference on Embedded Networked Sensor Systems*, ser. SenSys '04. New York, NY, USA: ACM, 2004, pp. 250–262. [Online]. Available: http://doi.acm.org/10.1145/1031495.1031525

[8] C. Baquero, P. S. Almeida, and R. Menezes, "Fast estimation of aggregates in unstructured networks," in *ICAS*, R. Calinescu, F. Liberal, M. Marín, L. P. Herrero, C. Turro, and M. Popescu, Eds. IEEE Computer Society, 2009, pp. 88–93.

[9] R. Morris, "Counting large numbers of events in small registers," *Commun. ACM*, vol. 21, no. 10, pp. 840–842, October 1978. [Online]. Available: http://dx.doi.org/10.1145/359619.359627

[10] P. Flajolet, "Approximate counting: A detailed analysis." *BIT*, vol. 25, no. 1, pp. 113–134, 1985.

[11] J. Cichoń, J. Lemiesz, and M. Zawada, "On cardinality estimation protocols for wireless sensor networks," in *ADHOC-NOW*, ser. Lecture Notes in Computer Science, H. Frey, X. Li, and S. Rührup, Eds., vol. 6811. Springer, 2011, pp. 322–331.

[12] J. Cichoń, J. Lemiesz, W. Szpankowski, and M. Zawada, "Two-phase cardinality estimation protocols for sensor networks with provable precision," in *2012 IEEE Wireless Communications and Networking Conference, WCNC 2012, Paris, France, April 1-4, 2012*. IEEE, 2012, pp. 2009–2013. [Online]. Available: http://dx.doi.org/10.1109/WCNC.2012.6214120