Estimating Sensor Population via Probabilistic Sequential Polling

Amir Leshem, *Member, IEEE,* and Lang Tong, *Fellow, IEEE*

***Abstract*—A probabilistic sequential polling protocol (PSPP) is presented for the estimation of the sensor population in a large-scale sensor network with a mobile access point. It is shown that PSPP requires $O(\log_2 N)$ sensor transmissions and a total of $O\left((\log_2 N)^2\right)$ polls to achieve an arbitrarily predetermined level of accuracy.**

***Index Terms*—Estimation, polling, probabilistic algorithms, sensor network with mobile access (SENMA), sensor networks.**

I. Introduction

WE CONSIDER the problem of estimating the number of sensors that belong to a certain category. For example, to maintain a large-scale network, one may be interested in how many sensors are still operating. A related problem is to estimate the number of sensors whose measurements have certain attributes.

The specific network architecture considered in this letter is sensor network with mobile access (SENMA) [2], in which a mobile access point is capable of interrogating sensors. Such an operation can be viewed, mathematically, as one of sampling the sensor field. From a communication system design point of view, on the other hand, the process of extracting information from the sensor network has two separate phases: a down-link broadcast from the mobile access point and an uplink transmission from sensors, both of which are energy consuming.

The primary design objective for large-scale sensor network is energy efficiency. To this end, we are interested in procedures that minimize the number of required transmissions. The brute-force scheme that schedules [e.g., via time division multiple access (TDMA)] transmissions of all sensors and counts the number of receptions at the access point requires the number of uplink transmissions at the order of $O(N)$, where N is the number of sensors in the category of interest. If a random sampling strategy is used and the number of sensors N is estimated by counting the number of distinct replies, the number of required transmissions is $O(N \log_2 N)$ in order to have diminishing probability of estimation error. In [4]–[6], the Good–Turing estimator [3] for missing mass is adapted for the problem of estimating the number of operating sensors, which requires, roughly, $O(\sqrt{N})$ number of transmissions [6], [7].

The strategies described above do not require down-link transmission, and the energy consumption occurs mostly at the uplink. In this letter, we present a polling strategy that reduces the number of transmissions to the level of $O(\log_2 N)$ for sensors and within $O\left((\log_2 N)^2\right)$ down-link polls. What makes this possible is the use of a sequential polling strategy coupled with a stopping rule that guarantees any predetermined accuracy.

II. Protocol Description

Consider a sensor network with N sensors, where N is unknown, and each sensor has an identity number from 1 to N. This might be the case when polling an uncooperative sensor network. We will focus on the case when each sensor belongs to one of the two classes, e.g., either operating or nonoperating. Alternatively, some sensors detected a given event while others have not. We assume that the probability of observing an operating sensor is larger than an unknown p_0. Note that at the first stage, we do not require accurate estimation of the probability of a sensor to operate but only need an upper bound. The way to obtain such an overestimate using polling with arbitrary accuracy will be discussed at a later stage. However, in some cases, we can obtain this probability based on *a priori* survivability analysis of the sensors (e.g., expected battery-life time of a sensor). The number of operating sensors that is unknown will be denoted by N_0 . We would like to estimate N_0. Since we do not know the total sensor population, this is not an ordinary sampling problem, where only the percentage of operating sensors needs to be estimated. We also need to estimate the total population size. To that end, we propose a probabilistic sequential polling protocol (PSPP). The protocol is probabilistic in the sense that it has a design parameter ε that is decided *a priori*, independent of the actual number of sensors, and the probability of success is above $1 - \varepsilon$. This approach is widely used, e.g., in the Rabin and Solovay–Strassen primality test algorithms, where we set up the probability of failure low enough that we are certain that the failure will actually not happen.

An important fact regarding the polling is that when the mobile agent polls sensor number ℓ, there can be the following three cases.

1) Sensor number ℓ is operating. In this case, a positive answer is obtained.
2) Sensor number ℓ is not operating. In this case, the sensor does not answer.

Manuscript received August 8, 2004; revised November 30, 2004. This work was supported in part by he Multidisciplinary University Research Initiative (MURI) by the Army Research Laboratory CTA on Communication and Networks under Grant DAAD19-01-2-0011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Erchin Serpedin.

A. Leshem is with School of Engineering, Bar-Ilan University, Ramat-Gan, 52900 Israel.

L.Tong is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA.

Digital Object Identifier 10.1109/LSP.2005.845595

3) ℓ is an index of a nonexisting sensor. In this case, no answer is obtained.

Based on the protocol, we cannot infer whether we are in case two or three. However, we assume that each sensor has the same probability to operate; therefore, when polling existing sensors, we will not see long runs of sensors not providing an answer. However, if $\ell > N$, then for every k, $\ell + k$ is an index of a nonexisting sensor. Therefore, we will see long runs of negative answers. This is the only possible way to bound the total population size. Our analysis in the second section will show that for each ℓ, we can choose $k(l)$ such that we maintain the probability of obtaining a run of no answers when starting to poll sensor ℓ below ε/ℓ^2. Using a union bound, we will obtain the desired probability of failure of the algorithm.

The PSPP is composed of three steps.

1) Estimate the probability p_0 that a sensor belongs to the population of operating sensors by sampling of M sensors with identity 1 to M.
2) Estimate $N_{\max}$—the maximum index of the operating sensor. This is done by sequential polling. Whenever we poll an operating sensor (with identity number ℓ) and obtain a positive answer, we continue to poll sensor number 2ℓ. This ensures that, at most, $O(\log_2(N))$ operating sensors are polled before we obtain an index of a nonexisting sensor. If a negative answer is received, sensor $\ell + 1$ is polled until $k(l)$ consecutive sensors fail to reply, where $k(l)$ is determined based on ϵ. See Table I.
3) Enhancing the accuracy of the estimator $\hat{p}$ of p_0 to a required degree, and estimate the sensor population N_0 by $\hat{p}\hat{N}_{\max}$.

The purpose of the first step is to find a lower bound p_0 on the probability of a sensor to operate. This is done through an ordinary random sampling technique by sampling an initial segment of the sensors, i.e., the first M sensors for a fixed M. The second step, which is the main contribution of the PSPP protocol, consists of the estimation of the total size of the population. We show that this can be done in $O\left((\log_2 N)^2\right)$ down-link transmissions and $2\log_2 N$ up-link transmissions. The last step is a refinement of the random sampling phase to obtain a more accurate estimate of the population size.

Table I provides the detail of the sequential polling strategy to estimate $N_{\max}$. Assume that p_0 and ε are given (where $0 < \varepsilon$ is any positive number). The protocol first ensures that the correct estimate is found with probability greater than $1-\varepsilon$. The second phase is described in Table I.

The second phase is composed of two main components. Steps (1)–(6) find identity of an operating sensor N_{low} such that there are no operating sensors above $N_{\text{high}} = 2N_{\text{low}}$. The decision that there are no operating sensors above sensor l is achieved if $k(l)$ consecutive sensors above it are not operating. Steps (7)–(11) estimate the identity of the maximal operating sensor using a binary search with a similar decision rule based on $k(l)$ consecutive nonoperating sensors. In Section III, we will show that the probability that there is an operating sensor above $N_{\max}$ is less than $\varepsilon/{N_{\max}}^2$. Moreover, the total probability of underestimating $N_{\max}$ is less than ε.

TABLE I
SECOND PHASE OF THE SEQUENTIAL POLLING PROTOCOL

```
 (1) For each $l = 1, \ldots, N$ define $k(l) = a + b\log_2(l)$ where
         $a = \frac{\log_2\left(\frac{6\varepsilon}{\pi^2}\right)}{\log_2(1-p_0)}$
         $b = \frac{-2\log_2(l)}{\log_2(1-p_0)}$                          (11)
 (2) $\ell = 1$.
 (3) Poll sensor number $\ell$.
 (4) If sensor $\ell$ is active set $\ell \leftarrow 2\ell$ and goto step (3).
 (5) If no answer is received set $u_\ell \leftarrow \ell + k(\ell)$.
 (6) For $k = \ell + 1, \ldots, u_\ell$
             if      $k = u_\ell + 1$
                     set $N_{\text{low}} \leftarrow \frac{1}{2}\ell$, $N_{\text{high}} \leftarrow \ell$
             else
                     Poll sensor $k$
                     if sensor $k$ is operating
                     (answered received) set $\ell \leftarrow 2k$ and goto (3)
             end
       end
 (7) If $N_{\text{high}} - N_{\text{low}} < k(N_{\text{low}})$ goto (11).
 (8) Set $\ell \leftarrow \frac{1}{2}(N_{\text{low}} + N_{\text{high}})$.
 (9) Set $u_\ell \leftarrow \ell + k(\ell)$.
(10) For $k = \ell, \ldots, u_\ell + 1$
         if    $k = u_\ell + 1$
               set $N_{\text{high}} \leftarrow \ell$, goto (7)
         if    sensor $k$ is operating
               set $N_{\text{low}} \leftarrow k$, goto (7)
       end
(11) For $k = N_{\text{low}}, \ldots, N_{\text{high}}$
     if sensor $k$ is operating $N_{\text{low}} \leftarrow k$
     end
```

III. Analysis of PSPP

A. Performance of Sequential Polling

In this section, we prove the following theorem.

Theorem 3.1: Assume that we have an unknown number of sensors N and that we know *a priori* that the probability of a sensor to operate properly is greater than p_0. Then, for every ε, there is a probabilistic protocol such that with probability greater than $1 - \varepsilon$, it estimates the number of operating sensors using $I_{\text{dl}}(N) = O\left((\log_2 N)^2\right)$ pollings and at most $I_{\text{ul}}(N) = 2\log_2 N$ up-link transmissions. Furthermore, the total number of pollings grows linearly with $\log_2 \varepsilon$ and $\log_2(1 - p_0)$.

Proof: Let p_0 be a lower bound on the probability of a sensor to operate properly. Let $\varepsilon' = 6\varepsilon/\pi^2$ be given. For each l, let $k(l)$ be large enough so that

$$(1 - p_0)^{k(l)} = \frac{\varepsilon'}{l^2}. \tag{1}$$

$k(l)$ grows linearly with $\log_2 l$ since

$$k(l)\log_2(1 - p_0) < \log_2 \varepsilon' - \log_2(l^2). \tag{2}$$

Hence, $k(l) = \lceil(\log_2 \varepsilon' - 2\log_2(l))/\log_2(1 - p_0)\rceil$ is sufficient so that $k(l) = a + b\log_2(l)$. Note also that $k(l)$ depends linearly on $\log_2 \varepsilon$ and $1/\log_2(1 - p_0)$, where

$$a = \frac{\log_2 \varepsilon'}{\log_2(1 - p_0)} \qquad b = \frac{-2}{\log_2(1 - p_0)}.$$

Claim 3.1: The probability of underestimating the identity of the operating sensor with the largest identity number is less than $\varepsilon = \pi^2\varepsilon'/6$.

Proof: We use a union bound. The probability that we have a sequence of $k(l)$ failed sensors beginning at sensor l is $(1-p_0)^{k(l)} = \varepsilon'/l^2$, by the choice of $k(l)$. Hence, the probability of obtaining a negative answer at any step is bounded by

$$P_{UE} \le \sum_{l<N} \frac{\varepsilon'}{l^2} < \sum_{l=1}^{\infty} \frac{\varepsilon'}{l^2} = \frac{\pi^2 \varepsilon'}{6} = \varepsilon. \tag{3}$$

Note also that for every l satisfying $l/2 < N < l$, we certainly get $k(l)$ consecutive negative answers when beginning polling sensor l, so the stopping of the algorithm is ensured.

We had at most $\log_2 N$ doubling steps (where sensor ℓ responded and we continued with polling sensor 2ℓ) before polling a nonexisting sensor. After a jump occurred ($\ell \rightarrow 2\ell$) we had at most $k(2\ell)$ pollings of consecutive sensors, which result in no answer before a positive answer occurred (since otherwise, the algorithm terminates prematurely. As shown in claim 3.1, this event occurs with a probability of less than ε). Since $k(\ell)$ is monotonically increasing and for every $m < \log_2 N$ we had at most one positive answer between 2^m and 2^{m+1} (a positive answer implies an index doubling), the total number of sensor polling during the first part of this phase ${I_{\mathrm{dl}}}^{(1)}(N)$ is bounded by

$${I_{\mathrm{dl}}}^{(1)}(N) \le \sum_{m=1}^{\log_2 N} k(2^m). \tag{4}$$

Since $k(2^m) = a + bm$, we obtain

$${I_{\mathrm{dl}}}^{(1)}(N) \le \left(a - \frac{1}{2}b\right) \log_2 N + \frac{1}{2} b \log_2^2 N. \tag{5}$$

The number of sensor responses is bounded by ${I_{\mathrm{ul}}}^{(1)}(N) \le \log_2 N$. Now, we start a binary search for the maximum by testing $3l/4, \ldots, 3l/4 + k(3l/4)$ in a similar way. If all these failed, we know that the maximum is between $3l/4$ and l; otherwise, we know that the maximum is between $l/2$ and $3l/4$. This process continues for $\log_2 N$ steps until the maximum is exactly localized. The total number of pollings in this stage is bounded by ${I_{\mathrm{dl}}}^{(2)}(N) \le (1/2) k(N) \log_2 N$; hence, the overall number of polling $I_{\mathrm{dl}}(N)$ is given by

$$\begin{aligned} I_{\mathrm{dl}}(N) &= {I_{\mathrm{dl}}}^{(1)}(N) + {I_{\mathrm{dl}}}^{(2)}(N) < k(N) \log_2(N) \\ &= 2a \log_2(N) + b \left(\log_2 N\right)^2. \end{aligned}$$

The total number of replies is bounded by

$$I_{\mathrm{ul}}(N) \le 2 \log_2 N$$

since we have another $\log_2(N)$ intervals with a single positive answer at each interval until the maximum $N_{\max}$ is found.

B. Estimating p_0

We now estimate the number of pollings that is necessary for estimating p_0. To estimate the number of operations needed to estimate p_0, we use the following application of the Chernoff bound [1].

Theorem 3.2: Let $p_1, \ldots, p_M$ be given. $p = (1/M) \sum_{i=1}^{M} p_i$. Assume that each X_i is a random variable satisfying

$$p\left(X_i = 1 - p_i\right) = p_i, \qquad p\left(X_i = -p_i\right) = 1 - p_i \tag{6}$$

and let $X = \sum_{i=1}^{M} X_i$. Then

$$Pr\left(X > a\right) < e^{-2a^2/M}. \tag{7}$$

Note that each X_i is a centering of a Bernoulli random variable with probability of success p_i. In our context, we have $p_i =$ p for all i. The X_i's are obtained by subtracting p from the replies (positive reply is 1, and no reply is 0), to obtain zero mean random variables. Now, we can estimate p_o using a sequential sampling of the first M sensors. Using the above theorem, we will evaluate how large M needs to be. The estimate of p_o is given by $\hat{p} = (1/M)Y$, where $Y = \sum_{i=1}^{M} Y_i$, $Y_i = X_i + p$ is the result of polling sensor i, and $Y = X + pM$ is the total number of positive replies. We want to estimate the probability $Pr(\hat{p} > p + \delta)$. To that end, note that

$$\begin{aligned} Pr(\hat{p} > p + \delta p) &= Pr\left(\frac{X}{M} + p > p + \delta p\right) \\ &= Pr(X > \delta p M) < e^{-2(\delta p)^2 M}. \end{aligned} \tag{8}$$

This is true for any value of δ, and the probability of overestimating p by δp is exponentially decaying. To ensure that our p_o used for computing $k(l)$ is indeed an underestimate, we can define $\hat{p}_o$ by $(1-\beta)\hat{p}$. In this case, we obtain

$$\begin{aligned} Pr((1-\beta)\hat{p} > p) &= Pr\left((1-\beta)\frac{X}{M} + (1-\beta)p > p\right) \\ &= Pr(X > \beta p M) < e^{-2(\beta p)^2 M}. \end{aligned} \tag{9}$$

Hence, we make the probability of overestimating p_o arbitrarily small. Choosing $M = K/p^2$ and $\beta = 1/\sqrt{2}$ will ensure that

$$Pr(\hat{p}_o > p) < e^{-K}.$$

Since we need $e^{-K} < \varepsilon$, the proper choice of K is given by

$$K = -\log_2 \varepsilon. \tag{10}$$

This K is independent of N and, hence, constitutes a constant overhead in the algorithm.

After $N_{\max}$ has been obtained, we estimate the number of sensors by $\hat{\boldsymbol{N}} = \hat{p} N_{\max}$. We use $N\mathrm{max}$ as an estimate of N. This is an underestimate, but the probability of underestimating by k decays exponentially with k, i.e., there might truly be k nonoperating sensors above $N_{\max}$, but this occurs with probability $P(N > N_{\max+k}) < (1-p_0)^k/(1-p_0)$. The error in the estimation of the number of operating sensors now consists of the error in estimating $\hat{p}$. This decays with M the number of samples used to estimate p. If we choose $M = f(N)$, we obtain

$$Pr\left(\left|\frac{\hat{\boldsymbol{N}}}{N_o} - 1\right| > \delta\right) = 2e^{-(\delta p)^2 f(N)}.$$

More specifically, if we choose $M = r(\log_2 N/\delta^2 p^2)$, we obtain

$$Pr\left(\left|\frac{\hat{\boldsymbol{N}}}{N_o} - 1\right| > \delta\right) = \frac{2}{N^r}.$$

If we want to make convergence more rapid, we can choose $M = (\log_2 N)^r/(p\delta^2)$ and obtain

$$Pr\left(\left|\frac{\hat{\boldsymbol{N}}}{N_o} - 1\right| > \delta\right) = \frac{2}{N^{(\log_2 N)^{r-1}}}.$$

This is a price paid for reducing the number of samples. We obtain a rapidly converging estimator with a small number of total transmissions; however, the convergence rate is subexponential.

IV. Conclusions

In this letter, we have presented a probabilistic algorithm for estimating the number of operating sensors in a sensor network. The probability of failure of the algorithm ε affects the complexity linearly with $\log_2 \varepsilon$, so the probability of failure can be made arbitrarily small at a low computational cost. The algorithm requires $O(\log_2 N)$ up-link transmissions and $O\left((\log_2 N)^2\right)$ down-link transmissions (polling). We have also demonstrated the tradeoffs between number of transmissions and accuracy of the estimator.

References

[1] N. Alon, *The Probabilistic Method*, 2nd ed. New York: Wiley, 2000.

[2] L. Tong, Q. Zhao, and S. Adireddy, "Sensor networks with mobile agents," in *Proc. Military Commun. Int. Symp.*, Boston, MA, Oct. 2003.

[3] I. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, pp. 237–264, 1953.

[4] C. Budianu and L. Tong, "Estimation of the number of operating sensors in a sensor network," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Monterey, CA, Nov. 2003.

[5] ——, "Good–Turing estimation of the number of operating sensors: A large deviations analysis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2004, pp. 1029–1032.

[6] C. Budianu, S. Ben-David, and L. Tong, "Estimation of the number of operating sensors in large-scale sensor network with mobile access," *IEEE Trans. Signal Process.*, submitted for publication.

[7] C. Budianu, "Estimation in wireless communication systems," Ph.D. dissertation, Cornell Univ., Ithaca, NY, 2004.