GOOD-TURING ESTIMATION OF THE NUMBER OF OPERATING SENSORS: A LARGE DEVIATIONS ANALYSIS

Cristian Budianu and Lang Tong[†]

School of Electrical and Computer Engineering,
Cornell University, Ithaca, NY, USA
{cris, ltong}@ece.cornell.edu

ABSTRACT

The paper [1] proposes an estimator for the number of operating sensors in a wireless sensor network based on the Good-Turing non-parametric estimator of the missing mass. This paper investigates the performance of this estimator using the theory of large deviations. We determine the asymptotic behavior of the large deviations exponent as the ratio n/N between the number of collected samples n and the number of operating sensors N decreases to zero. The simulations reveal that the confidence intervals obtained using the large deviations formula are upper bounds for the actual performance of the estimator. Together with the asymptotic behavior of the exponent, this suggests the surprising fact that if the scaling law $n = f(N)$ is used for the number of samples, then reliable estimation can be done if n grows at least as fast as $\sqrt{N}$. Separately, it is shown that if $\lim_{N\to\infty} \frac{n}{\sqrt{N}} = 0$ the estimator can't be used.

1. INTRODUCTION

This paper considers the problem of estimating the number of operating sensors in a large wireless sensor network. Usually, after the sensors are deployed, the number of operating sensors can vary in time due to battery consumption and/or external factors. Because the network is designed to function properly with a sufficient fraction of operating sensors, it is necessary to estimate the number of operating sensors.

We consider Sensor Network with Mobile Access (SENMA)– an architecture proposed in [2]. In SENMA, mobile access points with high processing power act as mobile base stations for the sensor nodes as shown in Fig.1. Sensors in SENMA transmit the collected data to the mobile access points. As the mobile access points collect packets from sensors, we would like to estimate the number of operating sensors in the network.

We assume that the mobile access points use a random access protocol such as ALOHA. At the end of the collection process, the mobile access point receives n packets randomly drawn from N operating sensors. The basic model considered in this paper assumes that at most one packet is collected at a time, and each received packet can come from any of the operating sensors. We consider the problem of estimating N assuming that each sensor includes its identity in its packet.

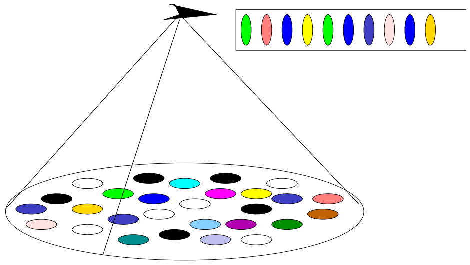

Fig. 1: The Sensor Network with Mobile Access Point.

As more packets are collected, N can be estimated with arbitrary accuracy. Given a performance requirement, the objective is to collect as few packets as possible. While this is a classical problem of estimating deterministic parameters, the standard maximum likelihood estimator leads to a numerical search, possibly in high dimensions when there are multiple categories of sensors.

In [1], an estimator based on the Good-Turing algorithm [3] was proposed. The Good-Turing algorithm is a nonparametric technique that estimates the missing mass based on those samples that appear only once. This simple estimator has a remarkable asymptotic performance, close to that of the ML estimator [4].

In this paper, we characterize the asymptotic behavior of the estimator proposed in [1] using the large deviation theory for the urn model, [5]. Such an analysis characterizes the decay rate of the probability that the estimation error exceeds a certain level. In other words, this analysis provides a guideline for determining approximately how many packets need to be collected. We determine the asymptotic behavior of the large deviations exponent as the ratio n/N decreases to zero. Numerical examples coupled with the exponent analysis revealed that the proposed estimator has the surprising property that any fixed confidence interval for its relative error can be achieved asymptotically by collecting only $C\sqrt{N}$ samples, where C is a constant that depends on the confidence interval selected. Moreover, if the scaling law $n = f(N)$ used for the number of samples is such that $\lim_{N\to\infty} \frac{n}{\sqrt{N}} = \infty$, any confidence interval for the performance of the estimator can be achieved. Separately, we show that the required number of packets n should grow at least as fast as $\sqrt{N}$.

2. THE GOOD-TURING ESTIMATOR

Consider a finite or countable set $\mathcal{N}$, a probability distribution P on this set, and a sample $\mathbf{X} = (X_1, \ldots, X_n)$, where $X_i \in \mathcal{N}$ are

[†] This work was supported in part by the Multidisciplinary University Research Initiative (MURI) under the Office of Naval Research Contract N00014-00-1-0564, and Army Research Laboratory CTA on Communication and Networks under Grant DAAD19-01-2-0011.

i.i.d. random variables with distribution P. For $x \in \mathcal{N}$, denote $p_x \triangleq \mathbb{P}\{X_i = x\}$ the probability of class x. Note that a uniform distribution is not required, besides the fact that $\mathcal{N}$ can be an infinite set.

For the observed sample $\mathbf{X}$, define the function $t : \mathcal{N} \to \mathbb{N}$, where $t(x)$ gives the number of samples in $\mathbf{X}$ equal to x. Using the multiplicity function t, we group the classes that appear the same number of times into sets :

$$\mathcal{S}_k \triangleq \{x \in \mathcal{N} : t(x) = k\} .$$

Note that the function t and the sets $\mathcal{S}_k$ are function of the observed sample $\mathbf{X}$, thus they are random variables. We use the notation $S_k \triangleq |\mathcal{S}_k|$. Now we define P_k to be the probability that a new sample, drawn (i.i.d.) with distribution P, belongs to set $\mathcal{S}_k$, $P_k \triangleq \sum_{x \in \mathcal{S}_k} p_x$. For $k = 0$, P_0 is the probability that if a new item is observed, it belongs to a new class. The probability P_0 is called the missing mass and $1 - P_0$ is called the coverage of sample $\mathbf{X}$. The probabilities P_k depend on the sample $\mathbf{X}$, and thus are random variables.

The following estimator for the missing mass, known as Good-Turing estimator was proposed in [3]

$$\hat{P}_0 = \frac{S_1}{n}.$$

This estimator estimated the missing mass using the number of classes that appear in the sample exactly once.

3. ESTIMATION OF THE NUMBER OF OPERATING SENSORS

Consider a sensor network having N operating sensors, each of them being identified by an ID that is an element of set $\mathcal{N}$, with $|\mathcal{N}| = N$. In this paper N is assumed constant during the collection time, but unknown. In each time slot the received packet can belong to any of the sensors with equal probability,*i.e.,*

$$\forall\, x \in \mathcal{N} \ : \ p_x \triangleq \mathbb{P}\{X_i = x\} = \frac{1}{N}.$$

This model identical to an urn model with replacement.

Denote $S \triangleq \sum_{k=1}^{n} S_k$ the total number of (different) sensors that appear in the sample; $S_0 \triangleq N - S$ represents the number of operating sensors that do not appear in the current sample. The problem is to estimate N using the received sample $\mathbf{X}$. Since S is observed, this is equivalent to estimating S_0, the number of sensors that are hidden to the operator.

The following estimation method for the number of operating sensors in a sensor network was proposed in [1]. First, use the Good-Turing formula to obtain $\hat{P}_0$. Then, using the assumption of equally likely classes, the missing mass is given by $P_0 = 1 - \frac{S}{N}$. Using the estimated value of P_0, we have the following estimator for N:

$$\hat{N} = \frac{S}{1 - \hat{P}_0} = \frac{S}{1 - \frac{S_1}{n}}. \tag{1}$$

4. PERFORMANCE ANALYSIS

4.1. On the Minimum Number of Samples

The estimated value $\hat{N}$ can be infinity if the number of collected samples is lower than N. On the other hand, the simulations showed that one can obtain accurate estimates using less than N samples. Thus, we want to see, for large N, how many samples should one collect such that we have a "small" probability of having an infinite estimator ? We have $\hat{N} = \infty$ if and only if $S_1 = n$, more exactly if all the elements of the vector sample $\mathbf{X}$ are different. We have

$$P(N) \triangleq \mathbb{P}\{\hat{N} = \infty\} = \mathbb{P}\{S_1 = n\} = \frac{N!}{(N-n)!N^n}$$

When $N \to \infty$ and $n = f(N) < N$, we want to determine those functions f that satisfy $\lim_{N\to\infty} P(N) = 0$. The answer is given by the following proposition (which is a classical result).

Proposition 1 *Consider* $n = f(N) < N$ *such that* $\lim_{N\to\infty} \frac{n}{N} = 0$. *Denoting* $L \triangleq \lim_{N\to\infty} \frac{n^2}{N}$, *we have*

$$\lim_{N\to\infty} \mathbb{P}\{\hat{N} = \infty\} = \begin{cases} 0 & \text{if } L = \infty \\ \exp\left(-\frac{1}{2}L\right) & \text{if } L \in (0, \infty) \\ 1 & \text{if } L = 0 \end{cases} .$$

Proof: Use Stirling approximation of factorial (Robbins' sharpening of Stirling's formula, [6], p. 39):

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n+1}} \le n! \le \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}},$$

and elementary limit calculation. □

The proposition says that one needs a sample size that increases with N faster than $\sqrt{N}$ in order to guarantee that the probability of having an infinite estimated value is asymptotically 0.

4.2. Confidence Intervals Using the Large Deviations Approach

Introduce a constant β, and consider that $N,\ n \to \infty$ such that $n = \lfloor \beta N \rfloor$. Thus, β is the fraction of sensors that were seen by the operator. For $i = 0, \ldots, N$ denote

$$\gamma_i^N \triangleq \frac{S_i}{N},$$

the fraction of the sensors that appear i times in the current sample. With these notations, the performance of the estimator proposed can be written as

$$\frac{\hat{N}}{N} = \frac{S}{N} \frac{1}{1 - \frac{S_1}{N}\frac{N}{n}} = \frac{1 - \gamma_0^N}{1 - \gamma_1^N \frac{N}{n}}$$

We are interested in the probability of the following two events when N is very large

$$\left\{ \frac{\hat{N}}{N} > c > 1 \right\}, \quad \left\{ \frac{\hat{N}}{N} < c < 1 \right\}.$$

The tool used to evaluate the probability of the events given above is the large deviations theory for occupancy problems developed by Dupuis, Nuzman and Whiting in [5]. In the rest of the paper only the probability of $\left\{ \frac{\hat{N}}{N} > c > 1 \right\}$ is analyzed; the other event is just discussed briefly.

This theory applied to our problem provides the following asymptotic result

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P} \left\{ \frac{\hat{N}}{N} > c \right\} = -J(\beta, c).$$

The large deviations exponent $J(\beta, c)$ is given by the following constraint minimization problem :

$$J(\beta, c) = \min_{\Gamma \in F(\beta,c)} \{D(\Gamma || \mathcal{P}_\beta)\}.$$

The optimization domain $F(\beta, c)$ is formed by all discrete distributions $\Gamma = [\gamma_0, \gamma_1, \dots]$ which satisfy

$$\sum_{i=0}^{\infty} \gamma_i = 1, \quad \sum_{i=0}^{\infty} i\gamma_i = \beta, \quad \frac{1-\gamma_0}{1-\gamma_1 \frac{1}{\beta}} \geq c.$$

The expression $D(P||Q)$ is the Kullback-Leibler distance between two distributions, *i.e.,*

$$D(P||Q) = \begin{cases} \sum_i P_i \log \frac{P_i}{Q_i} & \text{if } P \gg Q \\ \infty & \text{otherwise} \end{cases},$$

and $\mathcal{P}_\beta$ is the Poisson distribution with parameter β, *i.e.,* $\mathcal{P}_\beta(i) = \exp(-\beta)\frac{\beta^i}{i!}$.

The solution to this optimization problem can be found using Lagrange multipliers [7]. As usual, the Lagrange multipliers approach implies an optimization that can be solved only numerically. Except that it can be used for testing, this solution provides little insight into the problem.

Since usually we are interested in estimating the number of nodes using as few samples as possible, we consider the asymptotic behavior of the exponent $J(\beta, c)$ as the ratio $\beta \to 0$. This behavior will be investigated by deriving upper and lower bounds for the exponent that are tight for small values of parameter β. Also, the simulations reveal that the lower bound derived is tight enough to derive confidence intervals for the estimator performance.

For $\Gamma \in F(\beta, c)$, the solution of the optimization problem given above satisfies the last inequality from the definition of $F(\beta, c)$ with equality. For these distributions we have the following bounds for γ_0 and γ_1

$$\gamma_{0L} \triangleq 1 - \beta + \beta^2 \frac{1}{2c+\beta} \leq \gamma_0 \leq 1 - \beta + \beta^2 \frac{1}{c+\beta},$$

$$\gamma_{1L} \triangleq \beta - \frac{\beta^2}{c} + \frac{\beta^3}{c(2c+\beta)} \leq \gamma_1 \leq \beta - \frac{\beta^2}{c} + \frac{\beta^3}{c(c+\beta)}.$$

For $c > 1$, the optimization region for γ_0 and γ_1 and the bounds given above are represented in Fig. 2.

For the case $c > 1$, we derive upper and lower bounds on the large deviations exponent. Denote $\bar{\gamma}_{01L} \triangleq 1 - \gamma_{0L} - \gamma_{1L}$, $\bar{\gamma}_{1L} \triangleq 1 - \gamma_{1L}$, $\bar{\mathcal{P}}_{\beta,01} = 1 - \mathcal{P}_\beta(0) - \mathcal{P}_\beta(1)$, and $\bar{\mathcal{P}}_{\beta,1} = 1 - \mathcal{P}_\beta(1)$.

The upper bound is obtained by considering a particular value $\Gamma^* \in F$ and computing $D(\Gamma||\mathcal{P}_\beta)$. The value chosen is the left corner of the optimization region :

$$\Gamma^* = (\gamma_{0L}, \gamma_{1L}, 1 - \gamma_{0L} - \gamma_{1L}, 0, \dots),$$

which gives

$$D^*(\beta, c) = \gamma_{0L} \log \frac{\gamma_{0L}}{\mathcal{P}_\beta(0)} + \gamma_{1L} \log \frac{\gamma_{1L}}{\mathcal{P}_\beta(1)} + \bar{\gamma}_{01L} \log \frac{\bar{\gamma}_{01L}}{\mathcal{P}_\beta(2)}.$$

The lower bound is obtained by finding a convex domain F_*, $F \subset F_*$, such that the optimization problem over F_* can be obtained in closed form. The choice made is

$$F_* = \left\{\Gamma : \sum \gamma_i = 1, \gamma_0 \geq \gamma_{0L}, \gamma_1 \geq \gamma_{1L}\right\}.$$

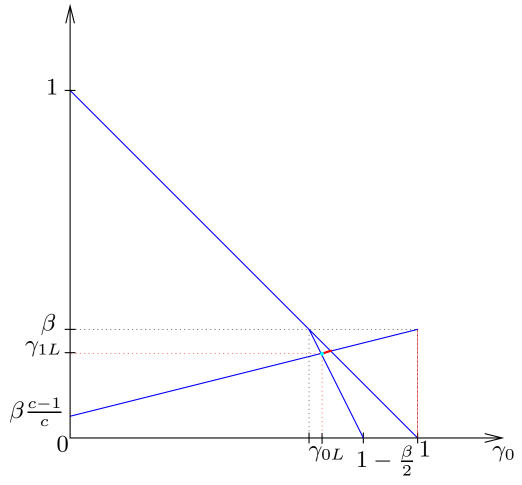

Fig. 2: The optimization region for γ_0 and γ_1 and the lower bounds γ_{0L} and γ_{1L}.

The solution $\Gamma_* = \arg\min_{\gamma \in F_*} D(\Gamma||\mathcal{P}_\beta)$ provides the lower bound $D_*(\beta, c)$. Note that this lower bound is nontrivial only if $\mathcal{P}_\beta \notin F_*$. The bound is given in the following proposition.

Proposition 2 *For any* $\beta \in [0, 1)$ *and* $c > \frac{\beta(1+\exp(-\beta))}{2(1-\exp(-\beta))}$*, we have the following lower bound on the exponent*

$$D_*(\beta, c) = \begin{cases} \gamma_{0L} \log \frac{\gamma_{0L}}{\mathcal{P}_\beta(0)} + \gamma_{1L} \log \frac{\gamma_{1L}}{\mathcal{P}_\beta(1)} + \bar{\gamma}_{01L} \log \frac{\bar{\gamma}_{01L}}{\bar{\mathcal{P}}_{\beta,01}} \\ \quad if \;\; \mathcal{P}_\beta(0) \frac{1-\gamma_{1L}}{1-\mathcal{P}_\beta 1} < \gamma_{0L} \\ \gamma_{1L} \log \frac{\gamma_{1L}}{\mathcal{P}_\beta(1)} + \bar{\gamma}_{1L} \log \frac{\bar{\gamma}_{1L}}{\bar{\mathcal{P}}_{\beta,1}}, \quad otherwise \end{cases}.$$

In particular, the bound holds for all $c > 1.0821$.

The conditions imposed depend on β and c; for each c, it can be shown that for β in an interval $[0, \beta_{max}(c))$, the condition $\mathcal{P}_\beta(0)\frac{1-\gamma_{1L}}{1-\mathcal{P}_\beta 1} < \gamma_{0L}$ is true. Thus for evaluation of the asymptotic behavior of $D_*(\beta, c)$ we consider the first equation.

The upper and lower bounds are in closed form and their limits when $\beta \to 0$ can be calculated. For the choices made, the two limits are equal, which proves the following theorem.

Theorem 1 *We have the following asymptotic behavior of the large deviations exponent :*

$$\lim_{\beta \to 0} \frac{J(\beta, c)}{\beta^2} = \frac{c - 1 - \log(c)}{2c} \triangleq B.$$

□

Thus, for any fixed, small β there is a large N such that we have

$$\frac{1}{N} \log \mathbb{P}\left\{\frac{\hat{N}}{N} > c\right\} \approx -J(\beta, c) \approx -\beta^2 B.$$

Although the approach of the proof presented is valid only for $c > 1$, the theorem holds identically for the large deviations exponent associated with $\left\{\frac{\hat{N}}{N} < c < 1\right\}$. In this case, however, the proof is done using a different approach, mainly because a tight lower bound can't be found by the method used above.

5. SIMULATIONS AND NUMERICAL RESULTS

In Fig. 3 the confidence region for the relative error of the estimator is represented for $\varepsilon = 0.001$. More exactly, for $N = 16000$ and $\varepsilon = 0.001$, the y-axis gives the levels c such that $\tilde{\mathbb{P}}\left\{\frac{\hat{N}}{N} > c > 1\right\} = \varepsilon$ (upper bound) and $\tilde{\mathbb{P}}\left\{\frac{\hat{N}}{N} < c < 1\right\} = \varepsilon$ (lower bound), where we denoted by $\tilde{\mathbb{P}}$ the observed empirical probability of an event. For the upper bound, the same quantity c is derived using the formula $\frac{1}{N}\log\mathbb{P}\left\{\frac{\hat{N}}{N} > c > 1\right\} = -J(\beta, c)$ and two approximations given by the upper and lower bounds on the exponent $D^*(\beta, c)$ and $D_*(\beta, c)$. One might note that the curve obtained using D_* is very close to the curve obtained using the true exponent function. A complete discussion on the bounds on the exponent is given in [8].

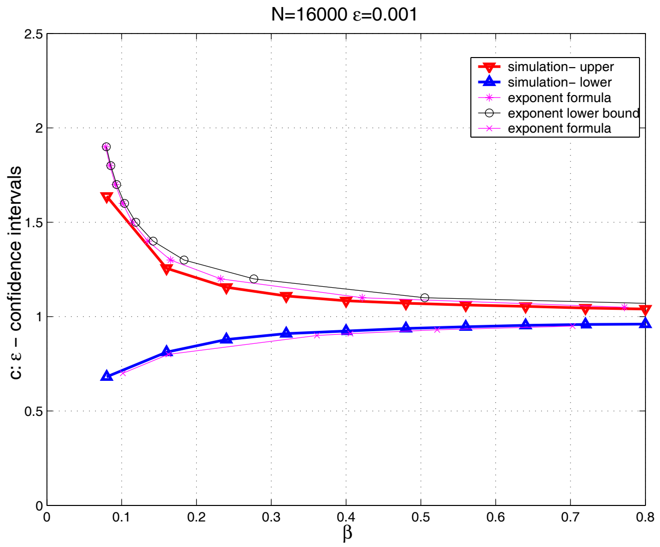

Fig. 3: Confidence intervals for the performance of proposed estimator. The way they were obtained is given in the legend.

Similar to the plot above, the simulations done for different parameters $(N, \beta, c, \varepsilon)$, revealed the following remarkable fact

$$\tilde{\mathbb{P}}\left\{\frac{\hat{N}}{N} > c\right\} < \exp\left(-NJ(\beta, c)\right).$$

This suggests that the right hand side expression that uses the large deviations exponent is an upper bound for the true probability $\mathbb{P}\left\{\frac{\hat{N}}{N} > c\right\}$. Using Theorem 1, it follows that for any fixed confidence interval (c, ε), when N becomes large, one needs to acquire only

$$\bar{n}(c, \varepsilon) = g(c, \varepsilon)\sqrt{N}$$

samples to achieve the desired performance. Moreover, using a scaling law $n = f(N)$ such that $\lim_{N\to\infty}\frac{n}{\sqrt{N}} = \infty$, any confidence interval for the performance of the estimator can be achieved. The simulations confirm this statement. For example, if $N \in [2000, 64000]$, $c = 1.12$ and $\varepsilon = 1.001$, the ratio $\frac{\bar{n}}{\sqrt{N}} \in (37, 40.5)$. The constant obtained is smaller than the one implied by the result of Theorem 1 because of the upper bound effect. The same discussion holds for the event $\left\{\frac{\hat{N}}{N} < c < 1\right\}$.

6. CONCLUSIONS

The performance of the estimator of the number of operating sensors based on the Good-Turing estimator is analyzed. First, we showed that one needs the number of samples n to grow faster than $\sqrt{N}$ in order to be able to use this estimator. Then, using the theory of large deviations we derived the behavior of the exponent for small ratios n/N. Although suggested by this asymptotic behavior, the simulations revealed the surprising fact that if the law of the number of samples satisfies $\lim_{N\to\infty}\frac{n}{\sqrt{N}} = \infty$ then arbitrary performance can be achieved asymptotically. Thus, the performance of the estimator under study exhibits a phase transition for the scaling laws of the number of samples $n = f(N)$ that satisfy $\lim_{N\to\infty}\frac{n}{\sqrt{N}} \in (0, \infty)$.

Acknowledgment: The authors acknowledge useful discussions with Prof. Shai Ben-David. The first author acknowledges useful discussions with Gökhan Mergen.

7. REFERENCES

[1] C. Budianu and L. Tong, "Estimation of the Number of Operating Sensors in a Sensor Network," in *Proc of 2003 Asilomar Conference on Signals, Systems and Computers*, (Monterey, CA), Nov. 2003. `http://acsp.ece.cornell.edu/pubC.html/`.

[2] L. Tong, Q. Zhao, and S. Adireddy, "Sensor Networks with Mobile Agents," in *Proc. 2003 Military Communications Intl 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] W. Esty, "The Efficiency of Good's Nonparametric Coverage Estimator," *The Annals of Statistics*, vol. 14, pp. 1257–1260, Sept. 1986.

[5] P. Dupuis, C. Nuzman, and P. Whiting, "Large Deviations Asymptotics for Occupancy Problems." `http://cm.bell-labs.com/cm/ms/who/nuzman/`, 2003.

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York, NY: Academic Press, 1981.

[7] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[8] C. Budianu, S. Ben-David, and L. Tong, "Estimation of the Number of Operating Sensors in a Sensor Network." to be submitted to IEEE Trans. on Signal Processing, 2004.