Consensus based estimation of anonymous networks size using Bernoulli trials

Damiano Varagnolo, Gianluigi Pillonetto, Luca Schenato

***Abstract*— To maintain and organize distributed systems it is necessary to have a certain degree of knowledge of their status like the number of cooperating agents. The estimation of this number, usually referred as the network size, can pose challenging questions when agents' identification information cannot be disclosed, since the exchanged information cannot be associated to who originated it. In this paper we propose a totally distributed network size estimation strategy based on statistical inference concepts that can be applied under anonymity constraints. The scheme is based on the following paradigm: agents locally generate some Bernoulli trials, then distributedly compute averages of these generated data, finally locally compute the Maximum Likelihood estimate of the network size exploiting its probabilistic dependencies with the previously computed averages. In this work we study the statistical properties of this estimation strategy, and show how the probability of returning a wrong evaluation decreases exponentially in the number of locally generated trials. Finally, we discuss how practical implementation issues may affect the estimator, and show that there exists a neat phase transition between insensitivity to numerical errors and uselessness of the results.**

***Index Terms*— anonymous networks, size estimation, number of agents, number of nodes, sensor networks, consensus, distributed estimation, distributed identification.**

I. Introduction

Increasing scalability, robustness and flexibility requests drawn the attention of the scientific communities to networked systems, where a multitude of agents collaborate in order to achieve a common task. Brilliant examples of such collective entities are the Internet, the mobile telephony, the peer-to-peer networks.

Despite their appealing qualities like the robustness to failures, networked systems come with peculiar problematics posing interesting yet difficult questions such as the privacy - cooperation dichotomy. Indeed, collaboration may be obstructed when arising privacy concerns push the agents to be anonymous: a significative example is about buildings heating/cooling control, where the need for the knowledge of the rooms occupancies conflicts with people's need of being not gauged.

In this work we develop a *general purely distributed networks size estimation strategy*. In the proposed scheme each node executes the same operations, moreover the following constraints are explicitly taken into account: anonymity, small computational/memory/communication capabilities, no existence of leaders or overlay structures, very limited knowledge of the network topology as in ad-hoc and mobile networks. Our labor is thus in the framework of the so-called *anonymous networks* [1], [2], where agents' identification information (ID) is or not unique or not exploitable. We notice that this framework is often used to obtain computability proofs for distributed algorithms, see e.g. [3], [4]. Our contribution originates from [5], and addresses the same problem of distributed size estimation under bounded computational, memory and bandwidth resources constraints. The estimation strategy can be summarized as follows:

1) each agent $i = 1, \dots, S$ locally generates a vector of $M \in \mathbb{N}_+$ i.i.d. Bernoulli trials $y_{i,m} \in \{0, 1\}$;
2) starting from the values $y_{i,m}$, agents distributedly compute the M averages of the locally generated data, i.e. the various $f_m = \frac{1}{S} \sum_{i=1}^{S} y_{i,m}$;
3) since the joint probability of the values f_m's depends on S, agents locally obtain an estimate $\widehat{S}$ of S via statistical inference.

(We notice that we consider S, the network size, as a deterministic but unknown parameter. $\widehat{S}$ is nonetheless a random variable.) This strategy is similar to the one in [5]: here we consider Bernoulli trials rather than general continuous random variables. But even if this seems a rather small change, the resulting estimator have completely different structure and properties. Interestingly, we show how these properties are strongly connected with the Newton-Pepys problem [6] and the distribution of the totatives of the network size [7] (more details in the following sections).

Literature review: we briefly review the relevant literature on the estimation of network sizes, starting from recalling that in general the estimation of the size of a group using statistical inference is an old problem that can be traced back at least to the *German tank problem* during World War II [8].

We notice that most of the proposed strategies sample a subset of the whole available information because of querying all the agents may be too costly. E.g., network traffic flow estimation is usually performed from data packets that are randomly sampled from the main stream [9].

Several sampling methods are based on *random walk strategies* [10], [11] [12], relying on passing a token through the network to collect information each time it visits an agent. There are mainly two different approaches [13]: the *return-time* one, exploiting the number of steps made by

All the authors are with the Department of Information Engineering, University of Padova, via Gradenigo 6/b, 35131 Padova, Italy. Emails: { varagnolo | giapi | schenato } @dei.unipd.it.

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement n°257462 HYCON2 Network of excellence and n°223866 FeedNetBack, by Progetto di Ateneo CPDA090135/09 funded by the University of Padova, and by the Italian PRIN Project "New Methods and Algorithms for Identification and Adaptive Control of Technological Systems".

the random walk of the token to return to the sender, and the *time-to-vanish* one. In this method the token contains a counter that is decreased every time the token is passed. The decrease is stochastically proportional to the number of neighbors of the receiving node. The network size is then inferred exploiting the statistical properties of the return-time and time-to-vanish.

We notice that sampling methods have been proposed also in conjunction to *randomly generated IDs* [14], [15], [16]. In general it is possible to map agents' random IDs into $[0,1]$, and then query the network for who have its ID belonging to a certain given interval $I \subset [0,1]$. From the answer and the size of I it is then possible to infer the network size. Obviously the size of I dominates the (stochastic) amount of information to be exchanged and the performances of the algorithm.

Capture-recapture strategies, sometimes referred also as *Lincoln-Petersen methods* [17], can also be used for size estimation purposes. Here a master disseminates a certain number of "seeds" (messages) that are then passed among agents. The strategy is then to ask to a certain set of agents whether they hold a seed or not. From the number of positive answers it is possible to estimate the size of the network [18]. We notice that these methods are strongly connected to the *sampling of finite populations* [19]: each of S individuals have its weight y_i and contributes to the population total, denoted by $\tau := \sum_{i=1}^{S} y_i$. A classical estimator for τ is the so-called *Hansen-Hurwitz estimator* [20]: select k individuals with selection probabilities p_i, and then estimate τ by means of $\widehat{\tau} := \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p_i}$. In the network size estimation problem $y_i = 1$ and p_i is the inverse of the quantity to be estimated. A similar approach is the Horvitz-Thompson estimator [21], where $\widehat{\tau} := \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ with π_i an opportune modification of the previous p_i's, see e.g. [22]. These methods are related to the so-called *inverted birthday paradox*. The direct paradox, called also the *birthday problem*, indicates the fact that the probability that two persons share the birthday in a group of 23 people is approximatively $\frac{1}{2}$. The inverse paradox indicates the fact that the number of how many people share the birthday date gives statistical information on the size of the group [23], [24].

Contributions: the strategy we propose is similar in spirit to the ones proposed in [25], [26], and belongs to the category of estimators that apply distributed consensus algorithms to randomly generated numbers, see, e.g., [27], [28]. In particular we show that exploiting average consensus on Bernoulli trials it is possible to obtain a Maximum Likelihood (ML) estimator of the network size which probability of error decays exponentially with the amount of locally generated data. (Remarkably, none of the strategies mentioned in the literature survey achieve an exponentially decaying error rate.) We also analyze the robustness of the offered strategy to numerical errors, and show that there is a neat phase transition between insensitiveness to errors and loss of meaningfulness of the final results.

The structure of the paper is the following. In Sec. II we develop the main algorithm starting from the single generation case to arrive to the general multiple generations case. In Sec. III we consider the sensitivity of the strategy to numerical errors, and then in Sec. IV we draw some concluding remarks[1].

Remark 1. The estimator proposed in this work is coherent with the following impossibility result from [29], see also [30]:

Theorem 2 (Thm. 9 [29])**.** There exists no algorithm that is able to compute the number of nodes in an anonymous network that terminates with the correct result for every finite execution with probability one and that has a bounded average bit complexity (i.e. s.t. the average number of bits used by the algorithm is bounded).

Indeed, we propose an estimator whose probability of error can be made arbitrarily *close* to zero.

II. The estimation algorithm

In this section the network will satisfy the following simplificative assumptions:

Assumption 3. There are no quantization effects, i.e. numbers are represented by an unlimited number of bits. Consensus algorithms are performed using an infinite number of consensus steps. Communication among agents is reliable, i.e. there is no packet loss. Finally, there exists an upper bound on the number of agents that actually constitutes the network, i.e. $S_{\max} \in \mathbb{N}_+$ s.t. $S \leq S_{\max}$ is known.

We will always assume that the network is not-time varying and capable to compute averages by means of average consensus algorithms [31], [32].

For ease of comprehensibility, we start considering the case where each agent generates only one scalar.

A. The single generation case: $M = 1$

As introduced in Sec. I, if each agent i locally generates

$$y_i \sim \mathcal{B}(p) \qquad \text{i.i.d.} \tag{1}$$

with $\mathcal{B}(p)$ the Bernoulli distribution with success probability p, then

$$\sum_{i=1}^{S} y_i \sim \mathrm{Bin}(S,p) \ , \tag{2}$$

with $\mathrm{Bin}(S,p)$ the binomial distribution of S experiments with success probability p. Assume that, by means of an average-consensus on the various y_i ran under the previous simplificative assumptions, agents compute

$$f := \frac{1}{S}\sum_{i=1}^{S} y_i \ . \tag{3}$$

Since $\left(\sum_{i=1}^{S} y_i\right) \in \{0,\dots,S\}$, it follows that it must be $fS \in \{0,\dots,S\}$. If S is unknown, it follows that f must

[1]The proofs of the proposed propositions can be found in the homonymous technical report available on the authors' webpages.

belong to the finite set

$$\mathbb{F}_{S_{\max}} := \left\{ f = \frac{k}{S} \text{ s.t. } k = 0, \ldots, S, \text{ and } S = 1, \ldots, S_{\max} \right\} . \tag{4}$$

It is straightforward that if S is fixed then f corresponds to the exact fraction of agents generating ones. It also follows that the probability mass function for f is

$$\mathbb{P}\left[f \ ; \ S, p\right] = \begin{cases} \binom{S}{fS} p^{fS} \left(1-p\right)^{S-fS} & \text{if } \begin{cases} fS \in \mathbb{N}_+ \\ f \in [0,1] \end{cases} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

We notice that $\mathbb{P}\left[f \ ; \ S, p\right] \neq 0$ if and only if there exists a network of S agents s.t. exactly fS of those generated $y_i = 1$ while the rest generated $y_i = 0$.

If f is observed, p is known, and S is a guess on the actual number of agents, then (5) represents the likelihood of the hypothesis S. A remarkable fact is that, if f is observed and p is known, the set $\mathcal{I}_f$ of the hypotheses S having strictly positive likelihood is finite, and given by

$$\begin{aligned} \mathcal{I}_f &:= \left\{ S \leq S_{\max} \mid \mathbb{P}\left[f \ ; \ S, p\right] > 0 \right\} \\ &= \left\{ S \leq S_{\max} \mid fS \in \mathbb{N}_+ \right\} \\ &= \left\{ S = \ell \bar{S} \leq S_{\max} \mid \begin{cases} \ell \in \mathbb{N}_+ \\ f\bar{S} = \bar{k} \\ \left(\bar{k}, \bar{S}\right) \text{ are coprime} \end{cases} \right\} . \end{aligned} \tag{6}$$

The variable $\bar{S} = \bar{S}(f)$ introduced by definition (6) plays a key role in our algorithm. First of all, it is unique and it depends only on f, not on p. Moreover $S \in \mathcal{I}_f$, therefore the true hypothesis S must be a multiple of $\bar{S}$. Finally, $\bar{S}$ is the ML estimator for S independently of the particular realizations of the y_i, as shown in Fig. 1 and formally stated in proposition 4.

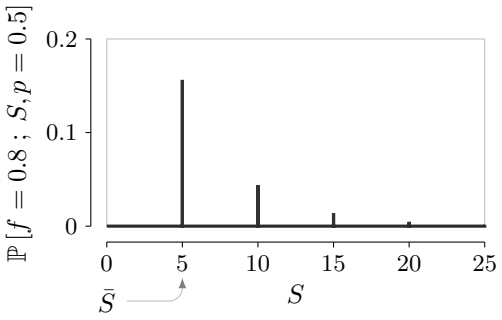

Fig. 1. Graphical representation of the likelihood given in (5) for $f = 0.8$, $p = 0.5$. In this case, $\bar{S} = 5$.

Proposition 4. Given the likelihood in (5), the ML estimator is given by

$$\widehat{S}(f;p) := \arg\max_{S \in \mathcal{I}_f} \mathbb{P}\left[f \ ; \ S, p\right] = \min \mathcal{I}_f = \bar{S}(f) \tag{7}$$

for every $p \in [0,1]$. This strategy cannot overestimate the true S and is also biased, i.e.

$$\widehat{S}(f;p) \leq S, \qquad \mathbb{E}_f\left[\widehat{S}(f;p)\right] < S \text{ for } S \geq 2 ,$$

where the first inequality should be intended as no outcomes of $\widehat{S}$ can be strictly bigger than S.

(7) has the following Occam's razor interpretation: if f is the measured fraction of agents generating ones, then $\widehat{S}$ is the smallest ("simplest", hence most probable) network that could have generated that fraction f. We notice that this motivates the biasedness of $\widehat{S}$, since this property is connected to the fact that the choice always selects the simplest hypothesis among the plausible ones.

Remarkably, using the same techniques of Prop. 4 we can derive the following corollary 5, that can be used to generalize the so called *Newton-Pepys problem*, an old question about if it is more probable to have at least one six when throwing six dice or to have at least two six when throwing twelve dice [6].

Corollary 5.

$$\mathbb{P}\left[f \ ; \ \nu\, \widehat{S}, p\right] \geq \mathbb{P}\left[f \ ; \ \kappa\, \nu\, \widehat{S}, p\right] \tag{8}$$

$\forall \kappa, \nu \in \mathbb{N}, \ p \in [0,1]$.

As it can be seen in Fig. 2, the map $\widehat{S} = \bar{S}(f)$ is extremely nonlinear. Without bounds on the maximal number of agents, it is defined over the positive rational numbers in $[0,1]$. If the upper bound $S_{\max}$ is known, $\widehat{S}$ is defined over the set $\mathbb{F}_{S_{\max}}$ defined in (4).

B. Performances of the estimation algorithm

Although $\widehat{S}$ does not explicitly depend on p, its performance does, and we are thus interested in optimizing this parameter. Consider then as the performance index the estimation error probability

$$\alpha(p, S) := \mathbb{P}\left[\widehat{S} \neq S \ ; \ S, p\right] . \tag{9}$$

Considering that $\widehat{S} = \bar{S}(f) = S$ if and only if $fS = k$ and the pair (k, S) is coprime, defining

$$\mathcal{F}_S := \left\{ f \mid fS = k \text{ with } (k, S) \text{ coprime} \right\} , \tag{10}$$

it follows that

$$\alpha(p, S) := \mathbb{P}\left[\widehat{S} \neq S \ ; \ S, p\right] = \mathbb{P}\left[f \notin \mathcal{F}_S \ ; \ S, p\right] . \tag{11}$$

We have thus a numerical procedure to compute the performance index (i.e. the estimator error probability) $\alpha(p, S)$: a) compute the set $\mathcal{F}_S$ which does not depend on p, b) compute $\mathbb{P}\left[f \notin \mathcal{F}_S \ ; \ S, p\right]$ exploiting (5).

We notice now that the error probability $\alpha(p, S)$ is not known a priori since S is not known. To optimize p we can thus exploit the following classical frequentists approach: the optimal p is the one that minimizes the largest error over all possible $S \leq S_{\max}$. The strategy is then to compute p^* as the minimizer of

$$\begin{aligned} \alpha^*(p, S_{\max}) &:= \max_{S \in \{1, \ldots, S_{\max}\}} \alpha(p, S) \\ &= \max_{S \in \{1, \ldots, S_{\max}\}} \mathbb{P}\left[\widehat{S} \neq S \ ; \ S, p\right] . \end{aligned}$$

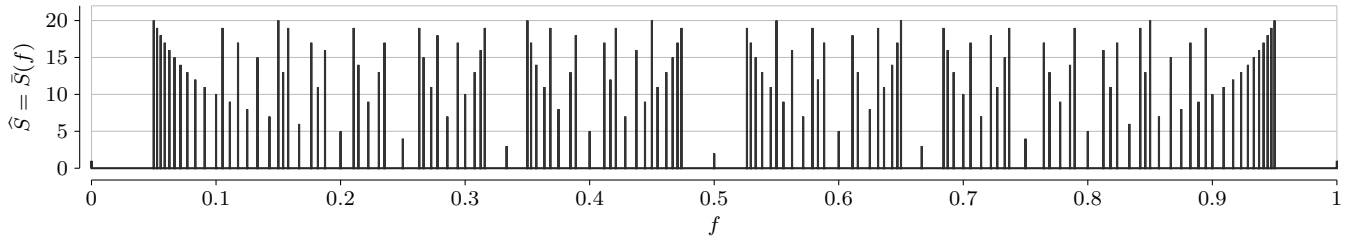

Fig. 2. ML estimator $\widehat{S}$ as a function of f for $S_{\max} = 20$.

The solution is thus

$$\begin{cases} p^*(S_{\max}) := \operatorname{argmin}_{p\in[0,1/2]} \alpha^*(p, S_{\max}) \\ \bar{\alpha}(S_{\max}) := \alpha^*\left(p^*(S_{\max}), S_{\max}\right) \end{cases} \tag{12}$$

where we restricted to the interval $[0, 1/2]$ since $\alpha(p, S)$ is symmetric with respect to $p = 1/2$.

Analytical expressions for the previous quantities are not available, but some considerations can be extrapolated from classical results on the distribution of what are called the *totatives* of the actual S. More precisely, the totatives of S are the positive integers smaller than S and relatively prime to S. The totient function, usually denoted with $\phi(S)$ and also called the *Euler phi-function*, indicates the number of totatives of S. Noticeably, in our case $\phi(S) = |\mathcal{F}_S|$, i.e. $\phi(S)$ indicates also the cardinality of the set $\mathcal{F}_S$. The totient function has been well studied in the context of Number Theory [33, p. 15] [7], [34, Chap. 8].

A result that is extremely useful for our purposes is given in [7], where the author showed that for high S,the distribution of the totatives of S in $\{1, \dots, S\}$ is approximatively uniform. An other important result is the bound offered in [34, Thm. 8.7]:

$$\phi(S) > \frac{S}{e^{\gamma} \log\log S + \frac{3}{\log\log S}} \tag{13}$$

where $\gamma \simeq 0.577$ is the so-called Euler-Mascheroni constant. By numerical inspection, from (13) it follows $\phi(S)/S > 0.15$ for $S \leq 10^{10}$, i.e. for all the networks that are meaningful for our approach. With simple numerical computations it is easy to check that if $p \in (0.25, 0.75)$ then $\mathbb{P}\left[f \in \mathcal{F}_S \,;\, S, p\right] > 0.15$ for all S, i.e., for the same p's, $\alpha(p, S) < 0.85$. An example of this fact can be seen in Fig. 3.

As it can be seen in Fig. 3, the choice for $p \in (0.25, 0.75)$ not very critical. In fact, for p in this interval, the gap between the maxima and minima of $\max_S \alpha(p, S)$ is small. The important point is that the worst probability of error, which is a non-decreasing function of $S_{\max}$, is bounded away from one for reasonably large $S_{\max}$. An important consideration is that if S is prime then $\phi(S) = S - 2$. Moreover the unique cases implying $\widehat{S} \neq S$ happen when the y_i's are all zeros or all ones. As a consequence, for prime S's $\alpha(p, S)$ can be extremely close to zero.

In Fig. 4 we plot how $\alpha^*\left(\frac{1}{2}, S_{\max}\right)$ and $\bar{\alpha}(S_{\max})$ depend on $S_{\max}$. As noticed before, both the quantities always stay

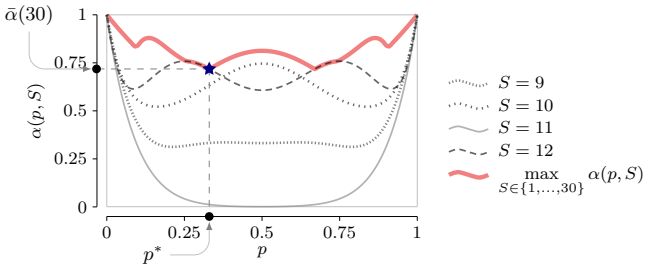

Fig. 3. $\alpha(p, S) = \mathbb{P}\left[\widehat{S} \neq S \,;\, p\right]$ as a function of p for various values of S, and $\alpha^*(p, 30) = \max_{S\in\{1,\dots,30\}} \alpha(p, S)$. The optimal point $(p^*, \bar{\alpha}(30)) \approx (0.33, 0.72)$ is indicated by the blue star and the black dots on the axes. The properties of the totient function assure that if p is sufficiently far from 0 and 1 then $\alpha(p, S) < 0.85$ uniformly in S.

below 0.85, with $\bar{\alpha}(S_{\max})$ being just a little better than $\alpha^*\left(\frac{1}{2}, S_{\max}\right)$, specially for large S. In this paper we do not further analyze the connections between $\phi(S_{\max})$ and the changes of p^* and α^*, and consider them as future works.

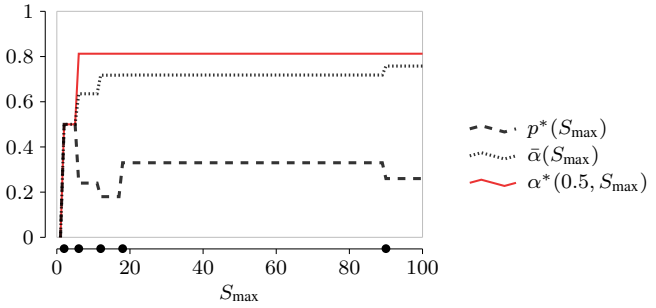

Fig. 4. Dependency of $p^*(S_{\max})$, $\bar{\alpha}(S_{\max})$ and $\alpha^*(0.5, S_{\max})$ on $S_{\max}$. Circles on the abscissas axis indicate for which $S_{\max}$ the former quantities change. This happens when the increase of $S_{\max}$ implies to consider an S having a particularly low $\phi(S)$.

C. The multiple generations case: $M > 1$

If agents generate a single sample $y_i \sim \mathcal{B}(p)$, the error probability $\mathbb{P}\left[\widehat{S} \neq S \,;\, p\right]$ might be equal to $\bar{\alpha}(S_{\max})$. Since this probability of error might be fairly high, the estimation performance seems to be extremely poor. The question is thus: is it possible to ameliorate the performance letting the agents generate M i.i.d. values $y_{i,1}, \dots, y_{i,M} \sim \mathcal{B}(p)$ and compute, by means of average-consensus strategies, $f_m = \frac{1}{S}\sum_i y_{i,m}$, for $m = 1, \dots, M$?

Let $\mathbf{f} := (f_1, f_2, \dots, f_M)$. The ML estimator $\widehat{S}$ can be computed as follows: since the various $y_{i,m}$ are independent, the likelihood can be factorized, i.e.

$$\mathbb{P}[\mathbf{f}\ ;\ S,p] = \prod_{m=1}^{M} \mathbb{P}[f_m\ ;\ S,p]\ . \tag{14}$$

This factorization implies that the novel hypothesis space is the intersection of the hypotheses spaces $\mathcal{I}_{f_m}$. In other words, S has non-zero likelihood only if $S \in \bigcap_{m=1}^{M} \mathcal{I}_{f_m}$. With the same Occam's razor interpretation of Sec. II-A, the ML estimator $\widehat{S}$ is still the simplest plausible hypothesis:

Proposition 6. Given the likelihood in (14), then the ML estimator is given by

$$\begin{aligned} \widehat{S}(\mathbf{f}) &:= \arg\max_{S \in \bigcap_{m=1}^{M} \mathcal{I}_{f_m}} \mathbb{P}[\mathbf{f}\ ;\ S,p] \\ &= \min\left(\bigcap_{m=1}^{M} \mathcal{I}_{f_m}\right) \\ &= \mathrm{LCM}\left(\bar{S}(f_1), \dots, \bar{S}(f_M)\right) \end{aligned} \tag{15}$$

for every $p \in [0,1]$, where $\mathrm{LCM}(\cdot)$ is the least common multiple operator. This strategy cannot overestimate S and is also biased, i.e.

$$\widehat{S}(\mathbf{f}) \le S, \qquad \mathbb{E}_f\left[\widehat{S}(\mathbf{f})\right] < S \text{ for } S \ge 2\ .$$

The computation of the parameter p that minimizes the worst error probability of the novel ML estimator can be made mimicking the derivations made for the scenario $M = 1$. Alternatively, it is possible to exploit the analysis of the previous section to obtain an interesting result: start defining $\bar{S}(f_m)$ conformably to (7), then define

$$\widetilde{S}(\mathbf{f}) := \max\left(\min_m \mathcal{I}_{f_m}\right) = \max\left\{\bar{S}(f_1), \dots, \bar{S}(f_M)\right\}\ . \tag{16}$$

$\widetilde{S}(\mathbf{f})$ is s.t. $\widetilde{S}(\mathbf{f}) \le \widehat{S}(\mathbf{f}) \le S$, and this implies that the probabilities of error for the two estimators $\widetilde{S}(\mathbf{f})$ and $\widehat{S}(\mathbf{f})$ satisfy

$$\mathbb{P}\left[\widehat{S}(\mathbf{f}) \ne S\ ;\ S,p\right] \le \mathbb{P}\left[\widetilde{S}(\mathbf{f}) \ne S\ ;\ S,p\right]\ .$$

E.g., if $S = 6$ and $M = 2$ then the event $f_1 = \frac{1}{2}$, $f_2 = \frac{1}{3}$ leads to $\widehat{S}(2,3) = \mathrm{LCM}(2,3) = S$, while $\widetilde{S}(2,3) = \max\{2,3\} = 3 \ne S$.

Consider now that

$$\begin{aligned} \mathbb{P}\left[\widetilde{S}(\mathbf{f}) \ne S\ ;\ S,p\right] &= \mathbb{P}\left[\bar{S}(f_m) \ne S\ \forall m\ ;\ S,p\right] \\ &= \left(\mathbb{P}\left[\bar{S}(f_m) \ne S\ ;\ S,p\right]\right)^M \\ &= \left(\alpha(p,S)\right)^M\ . \end{aligned}$$

This indicates that the probability that $\widehat{S}$ is wrong exponentially vanishes with the amount of information that agents exchange. More formally:

Proposition 7. If $y_{i,1}, \dots, y_{i,M} \sim \mathcal{B}(p^*)$, with $p^* = p^*(S_{\max})$ and $\bar{\alpha} = \bar{\alpha}(S_{\max})$ defined in (12) then

$$(1-p^*)^{S_{\max} M} \le \mathbb{P}\left[\widehat{S}(\mathbf{f}) \ne S\ ;\ p^*, M\right] \le (\bar{\alpha})^M \tag{17}$$

Speculations on the nature of $\bar{\alpha}$ and p^* lead to the conclusion that the upper bound in (17) is pessimistic. Consider in fact that $(\alpha^*)^M$ represents the worst-case probability of the event $\widehat{S}_m \ne S$ for all m. As shown in a previous example, this is a necessary but not sufficient condition for the event $\widehat{S} \ne S$. Now increasing M implies that the number of the cases where $\widehat{S} = S$ even if $\widehat{S}_m \ne S$ also increases. This eventually leads to discrepancies between $(\alpha^*)^M$ and the actual $\mathbb{P}\left[\widehat{S} \ne S\ ;\ p^*, M\right]$. Fig. 5 supports this intuition, and shows that $\mathbb{P}\left[\widehat{S} \ne S\ ;\ p^*, M\right]$ appears to decay to zero faster than what indicated in Prop. 7.

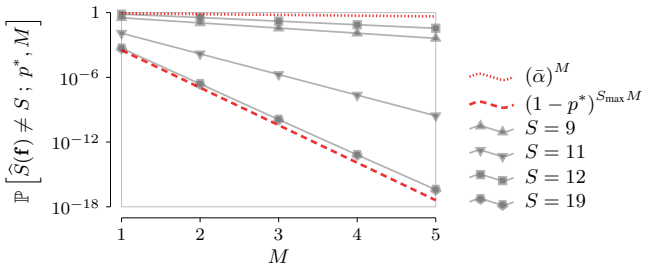

Fig. 5. $\mathbb{P}\left[\widehat{S} \ne S\ ;\ p^*, M\right]$ as a function of M and for various values of S, for the case $S_{\max} = 20$, and its lower and upper bounds described in (17).

III. Numerical considerations

Up to now we exploited assumptions 3, that in practical implementations will surely be violated. The natural question is then what are the effects of these violations.

Let $f_{i,m}$, $m = 1, \dots, M$, $i = 1, \dots, S$ be the quantities actually computed by the various agents under realistic assumptions such as packet losses, quantization effects, and finite numbers of communications. (Notice that we still assume the existence of the bound $S_{\max}$.) Modeling the effects of a finite number of consensus iterations and the effects of packet losses with multiplicative errors $\delta_{i,m}$, and modeling quantization errors as additive noises $\Delta_{i,m}$, we obtain the model

$$f_{i,m} = (1+\delta_{i,m})\, f_m + \Delta_{i,m} \quad m = 1, \dots, M,\ i = 1, \dots, S\ . \tag{18}$$

Assume then the knowledge on upper bounds $\delta_{\max}$ and $\Delta_{\max}$ s.t.

$$|\delta_{i,m}| < \delta_{\max}, \quad |\Delta_{i,m}| < \Delta_{\max}\ . \tag{19}$$

Let the agents map the actually computed averages $f_{i,m}$ into an element $f^q_{i,m}$ of the alphabet $\mathbb{F}_{S_{\max}}$ via the natural relation

$$f^q_{i,m} := \arg\min_{f \in \mathbb{F}_{S_{\max}}} |f - f_{i,m}|\ . \tag{20}$$

It is then possible to prove the following:

Proposition 8. If $\delta_{\max} + \Delta_{\max} < \dfrac{1}{2S_{\max}(S_{\max}-1)}$ then $f^q_{i,m} = f_m$.

Prop. 8 assures that if errors $\delta_{\max}$ and $\Delta_{\max}$ are sufficiently small then the measured averages $f_{i,m}$ are mapped into

the elements $f_{i,m}^q$ that would have been computed under idealized conditions. Then $\widehat{S}$ will give exactly the same results that would be obtained in absence of errors.

Vice versa, if the errors are not sufficiently small then the various $f_{i,m}$ *can* be mapped incorrectly. Due the strong nonlinearities of the map 2 and of the operator $\mathrm{LCM}(\cdot)$ in (15), the differences between the computed $\widehat{S}$ and the results that would be obtained in absence of errors literally explode. The result is that if the conditions expressed in Prop. 8 cannot be assured, then the estimation strategy here proposed is *useless*.

IV. Conclusions

In this paper we shown how it is possible to estimate the size of an anonymous network computing the averages of a set of independent Bernoulli trials. Despite its simplicity, the derived estimator has several very appealing qualities: first of all, its implementation on actual devices is naïve. Moreover, thanks to the fact that the probability of error vanishes exponentially with the number of generated Bernoulli trials, its estimation performances can be extremely good even exchanging a very small amount of information.

All these nice properties come among with a bittersweet fact that reduces the field of applicability of the strategy to small networks. In fact, due to the discontinuous and strongly nonlinear nature of estimation map, the estimator is either insensible to numerical errors or completely useless.

This naturally leads to consider the following important future works: first of all it is necessary to understand how it is possible to push the sensitivity phase transition point as far as possible. Moreover a requisite for actual implementations is the development of strategies detecting (and possibly correcting) these sensitivity problems.

References

[1] M. Yamashita and T. Kameda, "Computing on an anonymous network," in *Proceedings of the seventh annual ACM Symposium on Principles of distributed computing*, 1988, pp. 117 – 130.

[2] P. Boldi and S. Vigna, "Computing vector functions on anonymous networks," in *Proceedings of the 16th annual ACM symposium on Principles of distributed computing*, 1997.

[3] ——, "An effective characterization of computability in anonymous networks," *Lecture Notes in Computer Science*, vol. 2180 / 2001, pp. 33 – 47lo, 2001.

[4] E. Kranakis, "Symmetry and computability in anonymous networks: A brief survey," in *3rd International Colloquium on Structural Information & Communication Complexity*, N. Santoro and P. G. Spirakis, Eds. Siena, Italy: Carleton Scientific, June 1996.

[5] D. Varagnolo, G. Pillonetto, and L. Schenato, "Distributed statistical estimation of the number of nodes in sensor networks," in *IEEE Conference on Decision and Control*, December 2010.

[6] S. M. Stigler, "Isaac Newton as a probabilist," *Statistical Science*, vol. 21, no. 3, pp. 400 – 403, August 2006.

[7] D. H. Lehmer, "The distribution of totatives," *Canadian Journal of Mathematics*, vol. 7, pp. 347 – 357, 1955.

[8] R. Ruggles and H. Brodie, "An empirical approach to economic intelligence in World War II," *Journal of the American Statistical Association*, vol. 42, no. 237, pp. 72 – 91, March 1947.

[9] C. Estan and G. Varghese, "New directions in traffic measurement and accounting," in *ACM SIGCOMM'02 Conference*, 2002.

[10] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proceedings of the 10th annual conference on Internet measurement*, 2010.

[11] C. Gkantsidis, M. Mihail, and A. Saberi, "Random walks in peer-to-peer networks: algorithms and evaluation," *Performance Evaluation*, vol. 63, no. 3, pp. 241 – 263, March 2006.

[12] B. F. Ribeiro, "On the design of methods to estimate network characteristics," Ph.D. dissertation, University of Massachusetts - Amherst, May 2010.

[13] L. Massoulié, E. L. Merrer, A.-M. Kermarrec, and A. Ganesh, "Peer counting and sampling in overlay networks: random walk methods," in *25th annual ACM symposium on Principles of distributed computing*, 2006.

[14] D. Kostoulas, D. Psaltoulis, I. Gupta, K. P. Birman, and A. J. Demers, "Active and passive techniques for group size estimation in large-scale and dynamic distributed systems," *The Journal of Systems and Software*, vol. 80, no. 10, pp. 1639 – 1658, October 2007.

[15] K. Horowitz and D. Malkhi, "Estimating network size from local information," *Information Processing Letters*, vol. 88, no. 5, pp. 237 – 243, December 2003.

[16] G. S. Manku, "Routing networks for distributed hash tables," in *Proceedings of the twenty-second annual symposium on Principles of distributed computing*, 2003.

[17] G. Seber, *The estimation of animal abundance and related parameters*. London: Charles Griffin & Co., 1982.

[18] S.-L. Peng, S.-S. Li, X.-K. Liao, Y.-X. Peng, and N. Xiao, "Estimation of a population size in large-scale wireless sensor networks," *Journal of Computer Science and Technology*, vol. 24, no. 5, pp. 987 – 997, September 2009.

[19] M. H. Hansen and W. N. Hurwitz, "On the theory of sampling from finite populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 333 – 362, 1943.

[20] W. W. Piegorsch and A. J. Bailer, *Analyzing environmental data*. Wiley, January 2005.

[21] M. Sirken and I. Shimizu, "Population based establishment sample surveys: The Horvitz-Thompson estimator," *Survey Methodology*, vol. 25, no. 2, pp. 187 – 191, December 1999, statistics Canada, Catalogue No. 12-001.

[22] M. M. Salehi, "Comparison between Hansen-Hurwitz and Horvitz-Thompson estimators for adaptive cluster sampling," *Environmental and ecological statistics*, vol. 10, no. 1, pp. 115 – 127, 2003.

[23] S. Petrovic and P. Brown, "A new statistical approach to estimate global file populations in the eDonkey P2P file sharing system," in *21st International Teletraffic Congress*, September 2009.

[24] M. Bawa, H. Garcia-molina, A. Gionis, and R. Motwani, "Estimating aggregates on a peer-to-peer network," Stanford InfoLab, Tech. Rep., April 2003.

[25] E. Cohen, "Size-estimation framework with applications to transitive closure and reachability," *Journal of Computer and System Sciences*, vol. 53, no. 3, pp. 441 – 453, December 1997.

[26] D. Mosk-Aoyama and D. Shah, "Fast distributed algorithms for computing separable functions," *IEEE Transactions on Information Theory*, vol. 7, no. 7, pp. 2997 – 3007, July 2008.

[27] J. Cichoń, J. Lemiesz, W. Szpankowski, and M. Zawada, "Two-phase cardinality estimation protocols for sensor networks with provable precision," in *IEEE Wireless Communications and Networking Conference*, Paris, France, April 2012.

[28] C. Baquero, P. S. Almeida, R. Menezes, and P. Jesus, "Extrema propagation: Fast distributed estimation of sums and network sizes," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 4, pp. 668 – 675, April 2012.

[29] I. Cidon and Y. Shavitt, "Message terminating algorithms for anonymous rings of unknown size," *Information Processing Letters*, vol. 54, no. 2, pp. 111 – 119, April 1995.

[30] J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, "Distributed anonymous discrete function computation and averaging," arXiv:1004.2102v2 (math.OC), Tech. Rep., April 2010.

[31] F. Fagnani and S. Zampieri, "Randomized consensus algorithms over large scale networks," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 4, pp. 634 – 649, May 2008.

[32] ——, "Asymmetric randomized gossip algorithms for consensus," in *Proceedings of the 17th IFAC World Congress*, July 2008.

[33] N. Koblitz, *A Course in Number Theory and Cryptography*, 2nd ed. Springer-Verlag, 1994.

[34] E. Bach and J. O. Shallit, *Algorithmic Number Theory: Efficient algorithms*. The MIT Press, 1996.