# Distributed cardinality estimation in anonymous networks

Damiano Varagnolo *Member, IEEE*, Gianluigi Pillonetto *Member, IEEE*, Luca Schenato *Member, IEEE*

***Abstract*—We consider estimation of network cardinality by distributed anonymous strategies relying on statistical inference methods. In particular, we focus on the relative Mean Square Error (MSE) of Maximum Likelihood (ML) estimators based on either the maximum or the average of $M$-dimensional vectors randomly generated at each node. In the case of continuous probability distributions, we show that the relative MSE achieved by the max-based strategy decreases as $1/M$, independently of the used distribution, while that of the average-based estimator scales approximately as $2/M$. We then introduce a novel strategy based on the average of $M$-dimensional vectors locally generated from Bernoulli random variables. In this case, the ML estimator, which is the Least Common Multiple (LCM) of the denominators of the irreducible fractions corresponding to the $M$ elements of the average vector, leads to an MSE which goes exponentially to zero as $M$ increases. We then discuss the effects of finite precision arithmetics in practical dynamic implementations. Numerical experiments reveal that the MSE of the strategy based on Bernoulli trials is two order of magnitude smaller than that based on continuous random variables, at a price of one order of magnitude slower convergence time.**

***Index Terms*—Size estimation, sensor networks, distributed estimation, privacy-preservation, number of nodes, number of agents, anonymous networks, consensus.**

## I. Introduction

In the last decades we have been witnessing the success of networked systems composed from hundreds to millions of electronic devices or intelligent nodes, also called agents, that are capable of interaction and cooperation. Examples are the mobile telephony, the Internet, and more recently the wireless sensor networks and social networks. As a consequence, there has been a shift from the design of centralized architectures to decentralized and distributed ones in order to improve scalability, robustness to failure and structural flexibility. But beside performance and scalability, many distributed architectures also need to ensure substantial security and privacy. These peculiarities may be in conflict, since the preservation of the anonymity of the nodes contrasts with the necessity of cooperation to achieve a desired objective. Eventually, this kind of dichotomies pose challenging questions in terms of architecture and algorithmic design even for simple operations.

In this work we focus on the problem of computing the size of a network, i.e., the number of nodes composing it, under privacy constraints. The estimation of this quantity can be very important for disparate examples: a first one is the detection of topological changes, such as the disconnection of a part of the network into non-connected subcomponents. An other is the estimation of how many people in a group likes a specific item without explicitly disclosing the opinion of each person. As a consequence, this problem has attracted considerable research interests. The difficulty of this problem strongly depends on the assumptions and features of the network. In fact, if nodes are provided with unique IDs then the task can be accomplished with centralized or hierarchical algorithms in finite time [1]. Differently, if IDs are not present or are not unique to each node then these ID-based strategies cannot be used.

The framework analyzed in this work is exactly the one of the so-called *anonymous networks* [2], where nodes are usually assumed not to have unique IDs, but to be identical and to possess little information on the topology, e.g., often just a bound on the network size. This framework is often used to obtain computability proofs for distributed algorithms, see, e.g., [3]. Concerning our estimation problem, Cidon et al. [4], see also Hendrickx et al. [5], proved that the size of an anonymous network cannot be estimated correctly with probability one using algorithms that terminate in finite time and with bounded computational complexity. It is instead possible to estimate this quantity admitting the possibility of errors. The focus is then on finding estimators having low errors probabilities and suitable computational schemes.

In this work, we are interested in addressing the problem of size estimation where nodes have bounded computational, memory and bandwidth resources. Also we are interested in purely distributed strategies where each node executes the same operations, i.e., where there is no leader or overlay structures. Moreover, we assume that each node has a very limited knowledge of the network topology as in ad-hoc and mobile networks.

We notice that this size-estimation framework is important in several distributed applications. For example in ad-hoc Wireless Sensor Networks (WSNs) [6] pairing the initial number of nodes (that might be known at the time of deployment) and consequent estimates can be used to detect potential disconnections. Applications arise also in distributed estimation contexts, e.g., [7], where the knowledge of the number of measurements (or, equivalently, agents) helps to obtain better estimation performance. Finally, indications on the size of the network can be instrumental for coordination of robotic agents, that may take different actions based on how many they are [8].

### A. Literature review

The estimation of the size of a network, and more generally the size of a group, using statistical inference is an old problem that can be traced back at least to the *German tank problem* [9], i.e., the problem of estimating the production of German tanks from the serial numbers of the captured ones. Most of the strategies are based on sampling a subset of the whole available information, an approach that is motivated by the possible costs associated to querying the whole network. A typical example is the estimation of network traffic flow, which is usually computed from data packets that are randomly sampled from the main stream [10].

An important family of such sampling methods is based on *random walk strategies* [11], [12] which rely on passing a token through the network to collect information each time it visits an agent. In particular, in [13], two different approaches are proposed to estimate the number of peers in network. The first is based on the return-time, i.e., the number of steps made by the random walk of the token to return to the sender. The second is based on the time-to-vanish, i.e., the number of steps required for a counter present in the token to become zero: to connect the vanishing time with the network size, the counter is decreased every time it is passed by a quantity that

D. Varagnolo is with the School of Electrical Engineering, KTH Royal Institute of Technology, Osquldas Väg 10, SE-100 44 Stockholm, Sweden. Email: damiano@kth.se. G. Pillonetto and L. Schenato are with the Department of Information Engineering, University of Padova, via Gradenigo 6/b, 35131 Padova, Italy. Emails: { 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, by the Swedish Research Council, and by the Knut and Alice Wallenberg Foundation.

is (stochastically) proportional to the number of neighbors of the receiving node. Statistical properties of the return-time and time-to-vanish are then used to infer the network size.

Another approach within the class of sampling methods has been proposed for networks of nodes provided with *random IDs* [14], [15]. For example, it is possible to map the random ID of each agent into the real segment $[0, 1]$. Then, an agent interested in estimating the network size asks who have its ID belonging to a certain interval $I \subset [0, 1]$. Finally, the network size is inferred from the answer and the size of $I$. We notice that the size of $I$ dominates the (stochastic) amount of information to be exchanged and the performance of the algorithm.

A third strategy is based on the so-called *capture-recapture strategies*, sometimes referred also as *Lincoln-Petersen methods* [16]. The approach is to let a master disseminate a certain number of messages called "seeds", that are propagated through the network. The master then queries a certain number of nodes asking whether they hold a seed or not. From the number of seeds in the set of queried nodes it is possible to estimate the size of the network [17]. These capture-recapture methods are strongly connected to the estimation of population totals through sampling of finite populations [18] where $S$ individuals are characterized by some weights $y_i$ from which it is possible to define the population total $\tau := \sum_{i=1}^{S} y_i$. A classical estimation approach within this framework is then to select $k$ individuals from the whole set under the assumption that the selection probabilities are $p_i$, and then use the so-called *Hansen-Hurwitz estimator* [19] $\widehat{\tau} := \frac{1}{k} \sum_{i=1}^{k} \frac{y_i}{p_i}$. In the network size estimation problem $y_i = 1$ and $p_i$ is the inverse of the quantity to be estimated. A similar population total estimator is the so-called Horvitz-Thompson estimator [20], defined as $\widehat{\tau} := \sum_{i=1}^{k} \frac{y_i}{\pi_i}$ where $\pi_i$ is an suitable modification of the previous $p_i$'s, see, e.g., [21] for detailed insights. We notice that capture-recapture methods have also strong links with the so-called *inverted birthday paradox*. The direct birthday paradox (or *birthday problem*) stems from the fact that in a group of 23 people the probability that two persons have the same birthday date is approximatively $\frac{1}{2}$. The inverse problem is that, knowing how many people have the same birthday, it is possible to estimate the size of the group. Papers exploiting these ideas are, e.g., [22].

There are also specific approaches which leverage the peculiarities of the environment where the network operates. For example, in [23] underwater communications networks are studied: in this framework it is necessary to avoid nodes interferences, thus the estimation strategies let only a fraction of nodes to respond to queries. The estimation of the network size is then performed based on the number of received answers. Another ad-hoc strategy example is given by [24] where it is shown that in Master/Slave ad-hoc networks random walks strategies may perform better if they consider if tokens are sent by masters or slaves. Finally, in practical applications sometimes the size-estimation task is demanded to suitable ad-hoc devices rather than to ad-hoc strategies. A typical approach in this context is to let a mobile agent move and interrogate the static nodes with suitable queries, as did, e.g., in [25], [26].

### B. Contribution

The contribution of this work is twofold. The first is to extend the results of [27], [28] which proposed a strategy, in the following called Exponential-Maximum Strategy (EMS), that estimates the sum of a set of numbers from the maximum of a set of exponentially distributed random numbers. In particular, (i) we show that the size estimator proposed in [28] is the Maximum Likelihood (ML) estimator, (ii) we provide the closed form expression of the distribution of this estimator, from which it is possible to compute exact confidence intervals and all the statistical moments, (iii) we prove that, when considering continuous random variables, the distribution of this estimator is independent of the distribution of the generated random numbers, (iv) we show that the relative Mean Square Error (MSE) scales as $1/M$ where $M$ is the number of trials locally generated by each node, (v) we analyze the sensitivity of the estimator to quantization errors. We notice that similar alternative strategies have been proposed recently, see, e.g., [29], [30], [31], which are shown to provide smaller estimation errors. Nonetheless all preserve the same $1/M$ scaling in terms of relative MSE. We also show that using continuous random variables as before but substituting the computation of maxima with the computation of averages leads to ML estimators with relative MSEs that scale approximately as $2/M$.

Our second contribution sprouts then from substituting continuous random variables with discrete ones. We propose in fact a novel strategy, based on the average of $M$-dimensional vectors whose components are Bernoulli trials independently generated by the various nodes. We refer to this strategy as the Bernoulli-Average Strategy (BAS). We show that in this case the ML estimate of the network size is the Least Common Multiple (LCM) of the denominators of the irreducible fractions corresponding to the $M$ elements of the average vector. Incidentally, the BAS is strongly connected with the Newton-Pepys problem [32], and its MSE depends on the distribution of the totatives of the network size, i.e., on how the integers that are smaller than the network size and co-prime to it are distributed in $\mathbb{N}_+$ [33]. Exploiting this key fact we show that the relative MSE of the BAS decreases exponentially in $M$, a feature that is, is to the best of our knowledge, unique in the available literature.

We also address the problem of finite precision arithmetics in practical implementations of both strategies. We thus show that BAS is insensitive to quantization errors as long as the quantization is sufficiently fine.

Finally, we provide realistic numerical simulations based on dynamic strategies for wireless sensor networks. In particular, we explore asynchronous broadcast communications protocols that are guaranteed to converge in finite time with probability one: for the EMS we consider the implementation proposed in [28], while for the BAS we consider the ratio average consensus proposed in [34]. The simulations then confirm that BAS outperforms EMS in terms of MSE of at least two order of magnitudes, at the price of a slower convergence rate.

The structure of the paper is the following. In Section II we formulate the problem in mathematical terms, then propose the EMS and BAS algorithms in Sections III and IV. Then we consider the effects of finite precision arithmetics in Section V and propose practical implementations (among with Monte Carlo (MC) characterizations) in Section VI. Finally we summarize in Section VII the main differences between the estimators and draw some concluding remarks.

**Remark 1.** The results obtained in the paper are coherent with the following impossibility result from [4], see also [5]:

**Theorem 2** (Thm. 9 [4])**.** 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).

The coherency is given by the fact that all the strategies that we will propose are estimators whose probabilities of error are not zero, even if they can be made arbitrarily close to zero.

## II. Problem formulation

We model a network with a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{1, \ldots, S\}$ is the set of nodes composing the network and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of the communication links between the nodes. We assume that the graph $\mathcal{G}$ is undirected, i.e., $(i, j) \in \mathcal{E} \Leftrightarrow (j, i) \in \mathcal{E}$, and not time-varying. Hereafter we moreover assume $S$ to be a deterministic but unknown parameter.

In our framework we impose the distributed estimation strategy to rely only on local communications with limited coordination among nodes. Through the paper we will use a strategy based on the following 3-steps paradigm: 1) nodes locally generate a set of random data, 2) then they distributedly compute a function that takes as inputs the locally generated data, 3) finally, from the output of this computation, nodes locally estimate $S = |\mathcal{V}|$. More formally, we will consider the following strategy:

1) each node $i = 1, \ldots, S$ locally generates $M \in \mathbb{N}_+$ i.i.d. random values $y_{i,m} \in \mathbb{R}$, $m = 1, \ldots, M$, using a probability density $p(\cdot)$ that is identical among all nodes and that does not depend neither on the actual number of nodes $S$ nor on the number of generated values $M$. Each node is thus endowed with the vector $\boldsymbol{y}_i = [y_{i,1},\ y_{i,2},\ \ldots,\ y_{i,M}]^T$;
2) nodes distributedly compute a $M$-dimensional vector $\boldsymbol{f} = [f_1,\ f_2,\ \ldots,\ f_M]^T \in \mathbb{R}^M$ starting from the various local $M$-dimensional vectors $\boldsymbol{y}_i = [y_{i,1},\ y_{i,2},\ \ldots,\ y_{i,M}]^T$. More precisely, each $f_m$ is computed from the set $\{y_{1,m}, \ldots, y_{S,m}\}$ using an suitable function $F : \mathbb{R}^S \to \mathbb{R}$, i.e.,
$$f_m = F(y_{1,m}, \ldots, y_{S,m}) \tag{1}$$
where $F$ can be computed through local communication and simple operations. Some examples for (1) are: $f_m$ is the arithmetic mean of $\{y_{1,m}, \ldots, y_{S,m}\}$, its maximum, its minimum or its variance (see, e.g., [35]). In the following we will use superscripts like “ave” and “max” to denote particular instances of $F$ and $f_m$. We notice that $F(\cdot)$ does not depend on the index $m$;
3) since the joint probability of the values $f_m$’s depends on $S$, each agent can locally compute an estimate $\widehat{S}$ of $S$ using $\boldsymbol{f} \in \mathbb{R}^M$ via statistical inference. This is done through a function $\widehat{S} : \mathbb{R}^M \to \mathbb{R}$ generally indicated with
$$\widehat{S} := \Psi(f_1, \ldots, f_M)\ . \tag{2}$$
Notice that while $S$ is deterministic, $\widehat{S}$ is a random variable.

This general strategy is illustrated in Figure 1, where local and distributed operations are highlighted with suitable gray rectangles.

To compare different choices of the parameters $p(\cdot)$, $F(\cdot)$ and $\Psi(\cdot)$, it is necessary to define appropriate measures of performance. A typical choice is given by the Mean Square Error (MSE), $\mathbb{E}\left[\left(S - \widehat{S}\right)^2\right]$, where $\widehat{S}$ is a generic estimator of $S$.

Consider then that the estimators that we consider depend on the generating p.d.f. $p(\cdot)$, the consensus function $F$ and the estimator $\Psi$. Ideally we thus would like to minimize the MSE over all the possible choices of the triple $(p, F, \Psi)$ over all the values that $S$ may assume, but this is a formidable infinite dimensional problem. In this work we focus on special classes of the triple $(p, F, \Psi)$ and study the behavior of the MSE in these classes, to get insights on the optimization problem for the general case. More specifically, we constrain the analysis to the case where $\Psi$ is a Maximum Likelihood (ML) estimator. In general, ML estimators are inadmissible in the MSE sense (see, e.g., the so-called James-Stein estimator [36]) and thus suboptimal. Nonetheless, they have two favorable properties: a) they are consistent and thus asymptotically optimal, and b) they lead to close form solutions on which it is possible to speculate and highlight tradeoffs between computational complexity and estimation performance.

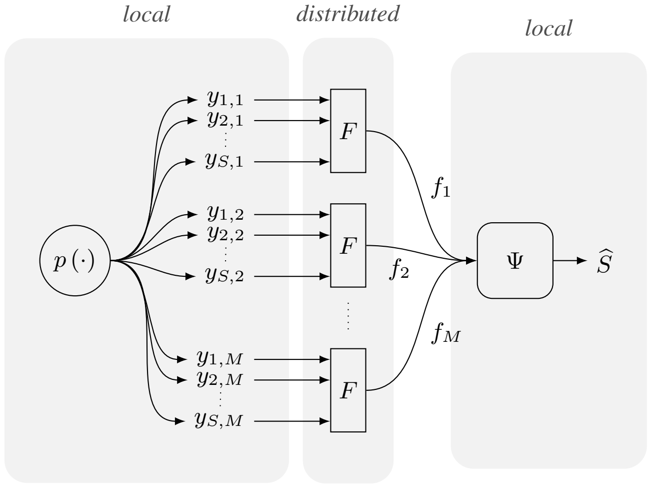


Figure 1: Graphical representation of the estimation strategy for the number of sensors $S$ when the various $y_{i,m}$ are generated using a probability distribution $p(\cdot)$.

## III. Estimation using the maximum function

Let $F$ be the maximum, i.e.,
$$f_m^{\max} = F_{\max}(y_{1,m}, \ldots, y_{S,m}) = \max_i \{y_{i,m}\}\ , \tag{3}$$
and consider for now the following assumptions, removed in the following:

**Assumption 3.** There are no quantization effects, i.e., numbers are represented by an unlimited number of bits.

**Assumption 4.** Consensus algorithms are performed using an infinite number of consensus steps.

**Assumption 5.** Communication among nodes is reliable, i.e., there is no packet loss.

For the purpose of this section, we recall two basic results on order statistics [37]. Let $S$ to be the number of elements of the sample $y_{1,m}, \ldots, y_{S,m}$, and $f_m^{(k)}$ to be its $k$-th order statistic. Let every $y_{i,m}$ be i.i.d. with $p(a)$ its probability density evaluated in $a$, and with $P(a)$ its probability distribution evaluated in $a$, i.e., $P(a) = \int_{-\infty}^{a} p(x)\, dx$. Then
$$p_{f_m^{(k)}}(a) = \frac{S!}{(k-1)!(S-k)!} P(a)^{k-1} (1 - P(a))^{S-k}\, p(a)\ . \tag{4}$$

Consider now data $y_{i,m}$ to be uniformly distributed, i.e., $y_{i,m} \sim \mathcal{U}[0, 1]$. The probability density of the (duplicate-insensitive) $S$-th order statistic $f_m^{\max}$ is given by (4) and is equal to $p(f_m^{\max}\ ;\ S) = S\big(f_m^{\max}\big)^{S-1}$ for all $m$. Since the various $f_m^{\max}$’s are independent
$$p(f_1^{\max}, \ldots f_M^{\max}\ ;\ S) = \prod_{m=1}^{M} p(f_m^{\max}\ ;\ S) = S^M \prod_{m=1}^{M} \left(f_m^{\max}\right)^{S-1}\ . \tag{5}$$

It follows that the ML estimator for $S$ is given by
$$\begin{aligned} \widehat{S} &= \Psi_{\text{ML}}(f_1^{\max}, \ldots, f_M^{\max}) \\ &:= \arg\max_S p(f_1^{\max}, \ldots, f_M^{\max}\ ;\ S) \\ &= \left( \frac{1}{M} \sum_{m=1}^{M} -\log(f_m^{\max}) \right)^{-1} \end{aligned}\ . \tag{6}$$

Defining $z := -\log\left(f_m^{\max}\right)$, it is easy to see that $z$ is an exponential random variable with rate $S$, i.e.,

$$p\left(z\ ;\ S\right) = \begin{cases} Se^{-Sz} & \text{if } z \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Recall also that the sum of $M$ i.i.d. exponential random variables with rate $S$ is a Gamma random variable with shape $M$ and scale $\frac{1}{S}$. $\left(\widehat{S}\right)^{-1}$ is thus a scaled version of this sum of exponentials, thus $\frac{MS}{\widehat{S}} \sim \operatorname{Gamma}\left(M, 1\right)$, that implies $\frac{\widehat{S}}{MS} \sim \operatorname{Inv-Gamma}\left(M, 1\right)$. This translates into

$$p\left(\widehat{S}\ ;\ S, M\right) = \Gamma\left(M\right)^{-1} \frac{1}{\widehat{S}} \left(\frac{MS}{\widehat{S}}\right)^{M} \exp\left(-\frac{MS}{\widehat{S}}\right) \tag{8}$$

from which it follows, for $M > 2$,

$$\mathbb{E}\left[\widehat{S}\ ;\ M\right] = \frac{SM}{M-1}, \quad \operatorname{var}\left(\widehat{S}\ ;\ M\right) = S^2 \frac{M^2}{(M-1)^2(M-2)}\ , \tag{9}$$

$$\mathbb{E}\left[\left(\frac{S-\widehat{S}}{S}\right)^2\ ;\ M\right] = \frac{M^2+M-2}{(M-1)^2(M-2)}\ . \tag{10}$$

These results can be summarized in the following proposition:

**Proposition 6.** Assume $y_{i,m} \sim \mathcal{U}\left[0,1\right]$ and consider the max-consensus scenario, i.e., $F = F_{\max}$. Then the ML estimator is given by

$$\widehat{S} = \Psi_{\mathrm{ML}}\left(f_1^{\max}, \ldots, f_M^{\max}\right) = \left(\frac{1}{M}\sum_{m=1}^{M} -\log\left(f_m^{\max}\right)\right)^{-1} \tag{11}$$

and, using Landau notation,

$$\lim_{M \to +\infty} \mathbb{E}\left[\widehat{S}\ ;\ S, M\right] = S$$

$$\mathbb{E}\left[\left(\frac{S-\widehat{S}}{S}\right)^2\ ;\ S, M\right] = \frac{1}{M} + o\left(\frac{1}{M}\right) \quad \text{for } M \gg 1\ .$$

The relative MSE in this case thus scales as $1/M$. This is similar to the MSE achievable by means of random walks strategies, see, e.g., [13]. However the strategy presented here may be preferred in some practical situations. In fact, in our framework estimates are full parallel, i.e., for each time each agent has a local estimate that becomes more and more accurate as the time passes. In random walk strategies, instead, one must first obtain an estimate with a random walk and then relay this information to all other nodes.

A reasonable question is then whether there are different random generation distributions $p(\cdot)$ other than the uniform one that can lead to better performance. Interestingly, the answer to this question is negative as shown in the following proposition, which shows an equivalence, in terms of MSE, of all the random variables whose cumulative distribution $P(\cdot)$ is absolutely continuous:

**Proposition 7.** Let $\mathcal{P}$ be the class of random variables whose cumulative distribution $P(\cdot)$ is absolutely continuous and let $p\left(\cdot\right)$ be the corresponding probability density. Let also $\widehat{S}(p, F_{\max}, \Psi)$ be a generic estimator as in (2).

Then

$$\begin{aligned} \min_{\Psi, p \in \mathcal{P}} \mathbb{E}\left[\left(S - \widehat{S}(p, F_{\max}, \Psi)\right)^2\right] = \\ \min_{\Psi} \mathbb{E}\left[\left(S - \widehat{S}(\mathcal{U}\left[0,1\right], F_{\max}, \Psi)\right)^2\right]\ . \end{aligned} \tag{12}$$

*Proof.* Let $x$ be a random variable with probability density $p_x\left(\cdot\right) \in \mathcal{P}$ and cumulative distribution $P_x(\cdot)$. Letting $y = P_x(x)$ it follows that $y \sim \mathcal{U}\left[0,1\right]$ since $\mathbb{P}\left[y \leq a\right] = \mathbb{P}\left[P_x(x) \leq a\right] = a$. Moreover there exists a map $P_x^{-1}(\cdot)$ such that $x = P_x^{-1}(y)$ almost surely. Now, let $\Psi_x$ be any generic function used to compute $\widehat{S}_x = \Psi_x(f_1^{\max}, \ldots, f_M^{\max})$, where $f_m^{\max} = \max_i\{x_{i,m}\}$ and $x_{i,m} \sim p_x(\cdot)$. Let also $\widetilde{f_m^{\max}} = \max_i\{\widetilde{x}_{i,m}\}$ with $\widetilde{x}_{i,m} \sim \mathcal{U}\left[0,1\right]$ and define $\widehat{S}_y = \Psi_y\left(\widetilde{f_1^{\max}}, \ldots, \widetilde{f_M^{\max}}\right) := \Psi_x\left(P_x^{-1}\left(\widetilde{f_1^{\max}}\right), \ldots, P_x^{-1}\left(\widetilde{f_M^{\max}}\right)\right)$. Then, using the monotonicity property of the max-function $F_{\max}$ and the fact that the probability densities of $f_m^{\max}$ and $P_x^{-1}(\widetilde{f_m^{\max}})$ differ at most on a set whose Lebesgue measure is zero, it is immediate to see that $\widehat{S}_x$ and $\widehat{S}_y$ have the same probability distribution over the space where $S$ may assume its values. Hence, $P\left(\widehat{S}(p_x, F_{\max}, \Psi_x)\right) = P\left(\widehat{S}(\mathcal{U}\left[0,1\right], F_{\max}, \Psi_y)\right)$. Since this holds for any $\Psi_x$, it must hold also for the minimizer of (12) and this concludes the proof. ∎

This proposition basically states that there is no advantage in using random number generators different from the uniform distribution in terms of achievable performance for a large class of distributions. For example, this class includes all the commonly used distributions such as the exponential distribution proposed in [28], the Gaussian distribution, and the beta distribution.

Another interesting property of the maximum function is that if the goal is to estimate the inverse of the number of nodes $S^{-1}$, then the ML estimator coincide with the Minimum Variance Unbiased Estimator (MVUE) estimator:

**Proposition 8.** Let $p = \mathcal{U}\left[0,1\right]$ and $\widehat{S^{-1}} = \Phi\left(f_1^{\max}, \ldots, f_m^{\max}\right)$, where the $f_m^{\max}$'s are obtained under a max-consensus scenario as in (3). Then the MVUE $\Phi_{\mathrm{MV}}$ for $S^{-1}$ is its ML estimator, i.e.,

$$\begin{aligned} \Phi_{\mathrm{MV}} \quad & := \arg\min\nolimits_{\Phi} \mathbb{E}\left[\left(S^{-1} - \widehat{S^{-1}}\right)^2\ ;\ M, S\right] \\ & = \arg\max\nolimits_{S^{-1}} p\left(f_1^{\max}, \ldots, f_M^{\max}\ ;\ S\right) =: \Phi_{\mathrm{ML}}\ . \\ & \text{s.t. } \mathbb{E}\left[\Phi\right] = S^{-1} \end{aligned}$$

Moreover

$$\widehat{S^{-1}} = \Phi_{\mathrm{MV}}\left(f_1^{\max}, \ldots, f_M^{\max}\right) = \frac{1}{M}\sum_{m=1}^{M} -\log\left(f_m^{\max}\right)$$

and

$$\mathbb{E}\left[\widehat{S^{-1}}\ ;\ M, S\right] = S^{-1}\ ,$$

$$\mathbb{E}\left[\left(\frac{S^{-1} - \widehat{S^{-1}}}{S^{-1}}\right)^2\ ;\ M, S\right] = \operatorname{var}\left(\frac{\widehat{S^{-1}}}{S^{-1}}\ ;\ M, S\right) = \frac{1}{M}\ .$$

*Proof.* (11) indicates that $\Phi_{\mathrm{ML}} = \frac{1}{\widehat{S}} = \frac{1}{M}\sum_{m=1}^{M} -\log\left(f_m^{\max}\right)$ and $\widehat{S^{-1}} \sim \operatorname{Gamma}\left(M, (MS)^{-1}\right)$. Thus $\mathbb{E}\left[\widehat{S^{-1}}\ ;\ M, S\right] = S^{-1}$, i.e., the ML estimator is unbiased, and $\operatorname{var}\left(\widehat{S^{-1}}\right) = \frac{S^{-2}}{M}$. We then show that the estimator is efficient, since it achieves the Cramér-Rao lower bound, and thus that it is Minimum Variance. From the likelihood (5) it follows that the Fisher Information for $S$ is

$$I(S) := \mathbb{E}\left[-\frac{\partial^2}{\partial S^2}\log p\left(f_1^{\mathrm{ave}}, \ldots, f_M^{\mathrm{ave}}\ ;\ S, M\right)\right] = \frac{M}{S^2}.$$

Considering then the Fisher Information reparametrization rule, i.e., if the transformation $S = \mu(S')$ is differentiable then the Fisher Information for $S'$ is $I\big(\mu(S')\big)\left(\frac{\partial\mu(S')}{\partial S'}\right)^2$, it follows that the Fisher Information for $S' = S^{-1}$ is $\frac{M}{S^{-2}}$. Thus $\operatorname{var}\left(\widehat{S^{-1}}\right) = \left(I(S^{-1})\right)^{-1}$, i.e., the ML estimator is efficient and thus MVUE.

∎

### A. Discussion

We now comment on the results presented above. The first observation is that the MV estimator $\widehat{S^{-1}} = \Phi\left(f_1, \ldots, f_M\right)$ can be decomposed into simpler blocks. In fact, all the quantities $f_m$ are passed through the same nonlinear function $\psi : \mathbb{R} \to \mathbb{R}$ transforming each $f_m$ into an unbiased estimate $\widehat{S_m^{-1}} := \psi\left(f_m\right),\ m = 1, \ldots, M$ of $S^{-1}$. More specifically, under the max-consensus we have $\psi\left(\cdot\right) = -\log\left(\cdot\right)$.

Now, since the $f_m$ are uncorrelated, also the $\widehat{S_m^{-1}}$ are uncorrelated. This implies that, to obtain the global estimate $\widehat{S^{-1}}$ using all the available information, the various $\widehat{S_m^{-1}}$ have simply to be combined through an arithmetic mean, i.e.,

$$\widehat{S^{-1}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{S_m^{-1}} \ .$$

Since the ML estimator for the number of nodes $\widehat{S}$ is simply $\widehat{S} = \frac{1}{\widehat{S^{-1}}}$, then the estimate for $S$ can be immediately obtained.

The second observation is that no prior information about $S$ has been exploited. Indeed, we have just assumed $S \in \mathbb{R}_+$ and not $S \in \mathbb{N}_+$ which is some sort of prior information. A possible generalization is to consider a Maximum A Posteriori (MAP) estimator if prior information about $S$ is available, which of course may provide better estimates. This generalization has been considered in [38] but it is not reported here in the interest of space.

The third observation is that the probability distributions not considered in Proposition 7 are relative to discrete or mixed random variables. However, estimators as in (2) based on discrete or mixed random variables and with $F = F_{\text{max}}$ are not going to provide better MSE scalings. E.g., consider for simplicity the case where the $y_{i,m}$'s are Bernoulli and i.i.d., i.e.,

$$y_{i,m} = \begin{cases} 1 & \text{with probability } 1-\theta \\ 0 & \text{with probability } \theta \end{cases} \qquad m = 1, \ldots, M \quad (13)$$

for an opportune $\theta$, so that, applying $F_{\text{max}}$,

$$f_m = \begin{cases} 1 & \text{with probability } 1-\theta^S \\ 0 & \text{with probability } \theta^S \end{cases} \qquad m = 1, \ldots, M. \quad (14)$$

Also this case, similar to [39], has an ML estimator of $S$

$$\widehat{S} = \log_\theta \left(1 - \frac{\sum_{m=1}^{M} f_m}{M}\right) \qquad (15)$$

that is characterized by a relative MSE scaling as $\beta / M$, where the constant $\beta$ depends also on the choice of the Bernoulli parameter $\theta$.

In the next section we will then show that using the average function it is possible to overcome this limit, and achieve estimators with relative MSEs scaling exponentially in $M$.

## IV. Estimation using the average function

Consider now the consensus function $F$ to be the average, i.e.,

$$f_m^{\text{ave}} = F_{\text{ave}}\left(y_{1,m}, \ldots, y_{S,m}\right) = \frac{1}{S} \sum_{i=1}^{S} y_{i,m} \qquad m = 1, \ldots, M \ . \quad (16)$$

Similarly to the previous section, we start with ML strategies based on continuous density distributions and show that, as in the $F_{\text{max}}$ case, the relative MSE scales proportionally with $1/M$. We then move to discrete distributions and show that, differently from the $F_{\text{max}}$ case, using Bernoulli distributions leads to drastic improvements in the MSE scalings.

### A. Continuous distributions

Consider again Assumptions 3, 4 and 5, and a zero-mean normal distribution for the generation of the data $y_{i,m}$, i.e., $y_{i,m} \sim \mathcal{N}\left(0, 1\right)$, implying thus $f_m^{\text{ave}} \sim \mathcal{N}\left(0, S^{-1}\right)$ for all $m$. With the same arguments used in Section III it is possible to obtain the following result:

**Proposition 9.** Assume $y_{i,m} \sim \mathcal{N}\left(0, 1\right)$ and consider the average-consensus scenario, i.e., $F = F_{\text{ave}}$. Then the ML estimator is given by

$$\widehat{S} = \Psi_{\text{ML}}\left(f_1^{\text{ave}}, \ldots, f_M^{\text{ave}}\right) = \left(\frac{1}{M} \sum_{m=1}^{M} \left(f_m^{\text{ave}}\right)^2\right)^{-1} \qquad (17)$$

and, using Landau notation,

$$\lim_{M \to +\infty} \mathbb{E}\left[\widehat{S}\ ;\ S, M\right] = S;$$

$$\mathbb{E}\left[\left(\frac{S - \widehat{S}}{S}\right)^2 \ ;\ S, M\right] = \frac{2}{M} + o\left(\frac{1}{M}\right) \quad \text{for } M \gg 1 \ .$$

This strategy thus provides an MSE worse than the one obtained using strategy (6), based on max consensus. Since the distributed computation of averages is also much slower than computing maxima, using $F_{\text{ave}}$ does not seem a sound choice. In fact, one could be tempted to state that also in this case it is not possible to do much better than the $1/M$ scaling, at least asymptotically in $S$. In fact, the probability density $p_{f_m}(\cdot)$ of the average $f_m$ is $p_{f_m}(\cdot) = p(\cdot) \ \star \ \cdots \ \star \ p(\cdot)$ where $\star$ indicates the convolution operator, applied thus $S$ times.

Assume then $p(\cdot)$ to be zero-mean and unit-variance. Since $p_{f_m}(\cdot)$ converges to $\mathcal{N}\left(0, S^{-1}\right)$ in distribution as $S$ goes to infinity, under reasonable regularity assumptions one would expect the performance to be equivalent of that obtained by starting with a Gaussian density. Actually this reasoning does not hold in general, since if $p(\cdot)$ is very different from a Gaussian distribution and if $S$ is small then the central limit approximation does not hold. Indeed, as shown below, using a discrete distribution with a weak prior information, namely the knowledge of a bound on the size $S \leq S_{\text{max}}$, it is possible to construct an estimator whose relative MSE decreases to zero exponentially fast with the number of experiments $M$.

### B. Bernoulli trials: $M = 1$

Consider then the case where the $y_{i,m}$ are Bernoulli random variables. In this section, in addition to Assumptions 3, 4, 5, we also include the following:

**Assumption 10.** There exists an upper bound on the number of nodes that actually constitutes the network, i.e., $S_{\text{max}} \in \mathbb{N}_+$ s.t. $S \leq S_{\text{max}}$ is known.

We start analyzing the case $M = 1$, i.e., when each agent generates only one scalar. Assume then that each agent $i$ locally generates $y_i \sim \mathcal{B}\left(p\right)$ i.i.d., where $\mathcal{B}\left(p\right)$ indicates the Bernoulli distribution with success probability $p$. It follows that

$$\sum_{i=1}^{S} y_i \sim \text{Bin}\left(S, p\right)$$

where $\text{Bin}\left(S, p\right)$ is the binomial distribution of $S$ experiments with success probability $p$. Let

$$f^{\text{ave}} := \frac{1}{S} \sum_{i=1}^{S} y_i \qquad (18)$$

be the result of an average-consensus process that nodes perform on the various $y_i$ under assumptions 3, 4 and 5. Being

$\left(\sum_{i=1}^{S} y_i\right) \in \{0,\dots,S\}$, it follows that it must be $f^{\text{ave}}S \in \{0,\dots,S\}$. In other words, with $S$ unknown, $f^{\text{ave}}$ must belong to the finite set

$$\mathbb{F}_{S_{\text{max}}} := \left\{ \frac{k}{S} \text{ s.t. } k=0,\dots,S, \text{ and } S=1,\dots,S_{\text{max}} \right\} . \tag{19}$$

Once $S$ is fixed, $f^{\text{ave}}$ corresponds to the exact fraction of nodes that generated ones. In mathematical terms, the probability mass function for $f^{\text{ave}}$ is

$$\mathbb{P}\left[f^{\text{ave}}\ ;\ S,p\right] = \begin{cases} \begin{pmatrix} S \\ f^{\text{ave}}S \end{pmatrix} p^{f^{\text{ave}}S}\,(1-p)^{S-f^{\text{ave}}S} \\ \quad \text{if } f^{\text{ave}}S \in \mathbb{N}_+, \quad f^{\text{ave}} \in [0,1] \\ 0 \quad \text{otherwise} \end{cases} \tag{20}$$

with the interpretation that $\mathbb{P}\left[f^{\text{ave}}\ ;\ S,p\right] \neq 0$ if and only if there exists a network of $S$ nodes compatible with the observed data, i.e., s.t. exactly $f^{\text{ave}}S$ of those generated $y_i = 1$ while the rest generated $y_i = 0$.

If $f^{\text{ave}}$ is observed and $p$ is known, (20) represents the likelihood as a function of $S$. Let us define the set $\mathcal{I}_{f^{\text{ave}}}$ of $S$ for which the likelihood is strictly positive, i.e.,

$$\begin{aligned} \mathcal{I}_{f^{\text{ave}}} \quad := & \quad \left\{S \,|\, \mathbb{P}\left[f^{\text{ave}}\ ;\ S,p\right] > 0\right\} = \left\{S \,|\, f^{\text{ave}}S \in \mathbb{N}_+\right\} = \\ = & \quad \left\{S = \ell\overline{S} \,|\, \ell \in \mathbb{N}_+, f^{\text{ave}}\overline{S} = \overline{k} \text{ and } (\overline{k},\overline{S}) \text{ are coprime}\right\} . \end{aligned} \tag{21}$$

Notice that for any average $f^{\text{ave}}$ generated according to the randomized strategy proposed in this section, the variable $\overline{S} = \overline{S}\left(f^{\text{ave}}\right)$ introduced by definition (21) is unique and does not depend on $p$. Moreover $S \in \mathcal{I}_f$, therefore the true number of nodes $S$ must be a multiple of $\overline{S}$. Figure 2 shows a graphical representation of an instance of the likelihood (20). In the figure the ML estimate is $\overline{S}$.

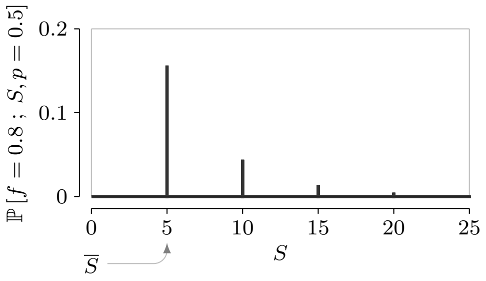


Figure 2: Graphical representation of the likelihood given in (20) for $f^{\text{ave}} = 0.8$, $p = 0.5$. In this case, $\overline{S} = 5$.

This is not a consequence of the particular realizations of the $y_{i,m}$. Indeed, it turns out that $\overline{S}$ always corresponds to the ML estimator for $S$, as formally stated in the following proposition:

**Proposition 11.** Given the likelihood in (20), the ML estimator is given by

$$\widehat{S}(f^{\text{ave}};p) := \arg\max_{S\in\mathcal{I}_{f^{\text{ave}}}} \mathbb{P}\left[f^{\text{ave}}\ ;\ S,p\right] = \min \mathcal{I}_{f^{\text{ave}}} = \overline{S}\left(f^{\text{ave}}\right) \tag{22}$$

for every $p \in [0,1]$. Moreover, it cannot overestimate the true number of nodes $S$ and is also biased, i.e.,

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

where the expectation is w.r.t. the r.v. $f^{\text{ave}}$.

*Proof.* In this proof we will indicate the coprime representation $\widehat{k} = f^{\text{ave}}\widehat{S}$ with $k = fS$ for ease of notation. We want to prove that $\mathbb{P}\left[f\ ;\ S,p\right] \geq \mathbb{P}\left[f\ ;\ \nu S,p\right]$, i.e., that

$$\begin{pmatrix} S \\ k \end{pmatrix} p^k (1-p)^{S-k} \geq \begin{pmatrix} \nu S \\ \nu k \end{pmatrix} p^{(\nu k)} (1-p)^{(\nu S - \nu k)} \tag{23}$$

for every $\nu \in \mathbb{N}$ and $p \in [0,1]$. Exploiting

$$\max_{p\in[0,1]} p^{(\nu k)}(1-p)^{(\nu S-\nu k)} = \frac{k^{(\nu k)}\,(S-k)^{(\nu S-\nu k)}}{S^{(\nu S)}}$$

(23) can be rewritten as

$$\frac{\begin{pmatrix} \nu S \\ \nu k \end{pmatrix}}{\begin{pmatrix} S \\ k \end{pmatrix}} \frac{k^{(\nu-1)k}\,(S-k)^{(\nu-1)(S-k)}}{S^{(\nu-1)S}} \leq 1 \ . \tag{24}$$

We show now that (24) holds true by induction on $S$.

- *Base case*: it is immediate to check that (24) holds true for every $k = S$.
- *Inductive step*: assume (24) holds true for $S, k$. Since

$$\frac{\begin{pmatrix} \nu(S+1) \\ \nu k \end{pmatrix}}{\begin{pmatrix} S+1 \\ k \end{pmatrix}} = \frac{\begin{pmatrix} \nu S \\ \nu k \end{pmatrix}}{\begin{pmatrix} S \\ k \end{pmatrix}} \cdot \frac{\prod_{j=1}^{\nu-1}(\nu S + j)}{\prod_{j=1}^{\nu-1}(\nu S - \nu k + j)} \ ,$$

$$\frac{k^{(\nu-1)k}\,(S+1-k)^{(\nu-1)(S+1-k)}}{(S+1)^{(\nu-1)(S+1)}} = \frac{k^{(\nu-1)k}\,(S-k)^{(\nu-1)(S-k)}}{S^{(\nu-1)S}} \cdot \triangle$$

with

$$\triangle := \left(\frac{S-k+1}{S+1}\right)^{(\nu-1)} \left(\frac{S-k+1}{S-k}\right)^{(\nu-1)(S-k)} \left(\frac{S}{S+1}\right)^{(\nu-1)S} \ ,$$

to prove the inductive step it is sufficient to prove that

$$\frac{\prod_{j=1}^{\nu-1}(\nu S + j)}{\prod_{j=1}^{\nu-1}(\nu S - \nu k + j)} \cdot \triangle \leq 1 \ .$$

Introduce the change of variable $x := S - k$, $x = 0 \Leftrightarrow k = S$, $x = S \Leftrightarrow k = 0$. Then (IV-B) can be rewritten as

$$g(x) := \frac{\prod_{j=1}^{\nu-1}(\nu S{+}j)}{\prod_{j=1}^{\nu-1}(\nu x{+}j)} \left(\frac{x{+}1}{S{+}1}\right)^{(\nu-1)} \left(1{+}\frac{1}{x}\right)^{(\nu-1)x} \left(\frac{S}{S{+}1}\right)^{(\nu-1)S} .$$

Since $g(S) = 1$, to prove the inductive step it is sufficient to show that $g(x)$ is non-decreasing in $x \in [0,S]$. To prove this, we consider $h(x) := \log\left(\overline{g}(x)\right)$ where $\overline{g}(x)$ is $g(x)$ deprived of the terms not depending on $x$, i.e.,

$$h(x) = -\sum_{j=1}^{\nu-1} \log\left(\nu x + j\right) + (\nu-1)(x+1) + (\nu-1)x\log\left(1+\frac{1}{x}\right) \ .$$

Considering its derivatives

$$h'(x) = -\sum_{j=1}^{\nu-1} \frac{1}{\nu x + j} + (\nu-1)\log\left(1+\frac{1}{x}\right)$$

$$h''(x) = -\sum_{j=1}^{\nu-1} \frac{\nu}{(\nu x + j)^2} - (\nu-1)\frac{1}{x(x+1)}$$

we observe that $h'(0) = +\infty$, $h'(+\infty) = 0$, $h''(x) \leq 0$ for all $x$. Being thus $h'(x) \geq 0$ for all $x$, $h(x)$ is monotonically increasing. Thus $g(x)$ is monotonically increasing, eventually implying the inductive step to hold true. ■

Proposition 11 assures that $\widehat{S}$ is the ML estimator for $S$ independently of the Bernoulli parameter $p$, i.e., $\widehat{S} = \overline{S}(f^{\text{ave}})$. This is coherent with Occam's razor interpretations of (22), i.e., with the fact that if $f$ is the measured fraction of nodes that generated ones, then $\widehat{S}$ is the smallest ("simplest") and hence the most probable network that could have generated that fraction $f^{\text{ave}}$. This is indeed consistent with the fact that the ML estimator sometimes underestimates the true number of nodes $S$. The following corollary can be derived using the same techniques of Proposition 11.

**Corollary 12.**

$$\mathbb{P}\left[f^{\text{ave}}\ ;\ \nu\widehat{S}, p\right] \geq \mathbb{P}\left[f^{\text{ave}}\ ;\ \kappa\nu\widehat{S}, p\right] \quad \forall \kappa, \nu \in \mathbb{N},\ p \in [0,1]\ .$$

Interestingly, corollary 12 is connected to 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 sixes when throwing twelve dice [32].

We notice that the map $\widehat{S} = \overline{S}(f^{\text{ave}})$ is extremely nonlinear, as it can be seen in Figure 3. Without assumption 10, this map is defined over the positive rational numbers in $[0, 1]$, i.e., $f^{\text{ave}} \in \mathbb{Q}_+ \cap [0, 1]$. Under assumption 10, $\widehat{S}$ is defined over the set $\mathbb{F}_{S_{\text{max}}}$ defined in (19).

The procedure just described provides an algorithm to compute the ML estimator under the average-consensus scenario when samples are generated from independent Bernoulli trials of success probability $p$. Although the ML estimator does not depend explicitly on $p$, its performance does. Since $p$ is a design variable we are interested in optimizing it. First, we need to define a performance index. In this context, a sensible choice is the estimator error probability

$$\alpha(p, S) := \mathbb{P}\left[\widehat{S} \neq S\ ;\ S, p\right] = \mathbb{P}\left[f \notin \mathcal{F}_S\ ;\ S, p\right] \tag{25}$$

where

$$\mathcal{F}_S := \{f^{\text{ave}} \mid f^{\text{ave}} S = k \text{ with } (k, S) \text{ coprime}\}\ .$$

The right hand side of (25) follows from the observation that $\widehat{S} = \overline{S}(f^{\text{ave}}) = S$ if and only if $f^{\text{ave}} S = k$ and the pair $(k, S)$ is coprime. This provides a numerical procedure for computing the estimator error probability $\alpha(p, S)$: first, compute the set $\mathcal{F}_S$ which does not depend on $p$, and then compute the error probability $\mathbb{P}\left[f^{\text{ave}} \notin \mathcal{F}_S\ ;\ S, p\right]$ exploiting (20).

Since $S$ is not known, also error probability $\alpha(p, S)$ is not known a priori, therefore a classical frequentists approach is to consider the worst-case scenario by computing the largest error over all possible $S \leq S_{\text{max}}$, i.e., to consider

$$\alpha^*(p, S_{\text{max}}) := \max_{S \in \{1, \dots, S_{\text{max}}\}} \alpha(p, S) = \max_{S \in \{1, \dots, S_{\text{max}}\}} \mathbb{P}\left[\widehat{S} \neq S\ ;\ S, p\right]$$

and then to compute one of the minimizers $p^*$ of this error probability and its corresponding error probability, i.e.,

$$p^*(S_{\text{max}}) := \arg\min_{p \in [0, 1/2]} \alpha^*(p, S_{\text{max}}) \tag{26}$$

$$\overline{\alpha}(S_{\text{max}}) := \alpha^*\left(p^*(S_{\text{max}}), S_{\text{max}}\right)\ . \tag{27}$$

In (27) we used the fact that $\alpha(p, S)$ is symmetric with respect to the point $p = 1/2$ and therefore the minimization can be restricted to the interval $[0, 1/2]$. Although analytical expressions for $\alpha(p, S)$, $\alpha^*(p, S_{\text{max}})$, $\overline{\alpha}(S_{\text{max}})$ and $p^*(S_{\text{max}})$ are not available, some considerations can be extrapolated from results on the distribution of the totatives numbers. More precisely, the totatives of a positive integer $S$ are the positive integers that are relatively prime to $S$ and not bigger than $S$. The totient function, usually denoted with $\phi(S)$, indicates the number of totatives of $S$. In our case $\phi(S) = |\mathcal{F}_S|$, i.e., $\phi(S)$ indicates also the cardinality of the set $\mathcal{F}_S$. The function $\phi(S)$ is usually called the *Euler phi-function* and it has been well studied in the context of Number Theory [40, p. 15] [33], [41, Chap. 8]. As shown in [33], for high $S$, the distribution of the totatives of $S$ in $\{1, \dots, S\}$ is approximatively uniform. Moreover the number of the totatives of $S$ can be bounded exploiting [41, Thm. 8.7]

$$\phi(S) > \frac{S}{e^{\gamma} \log\log S + \frac{3}{\log\log S}} \tag{28}$$

where $\gamma \simeq 0.577$ is the so-called Euler-Mascheroni constant. From (28) it is possible to obtain, by numerical inspection, $\phi(S)/S > 0.15$ for $S \leq 10^{10}$, i.e., for all the networks with meaningful size. With numerical computations it is also possible to show that, for the same range of $S$'s, if $p \in (0.25, 0.75)$ then $\mathbb{P}\left[f^{\text{ave}} \in \mathcal{F}_S\ ;\ S, p\right] > 0.15$. This implies that, for these $S$'s, $\alpha(p, S) < 0.85$ uniformly in $S$ and $p \in (0.25, 0.75)$. An example of this fact is shown in Figure 4.

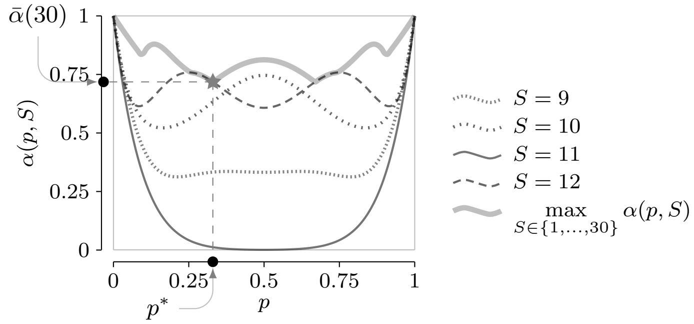


Figure 4: $\alpha(p, S) = \mathbb{P}\left[\widehat{S} \neq S\ ;\ p\right]$ as a function of $p$ for various values of $S$, and $\alpha^*(p, 30) = \max_{S \in \{1, \dots, 30\}} \alpha(p, S)$. The star indicates the optimal point $(p^*, \overline{\alpha}(30)) \approx (0.33, 0.72)$. We can notice that if $S$ is prime, e.g., $S = 11$, then $\alpha(p, S)$ can be close to zero. This is due to the fact that, for prime $S$'s, $\phi(S) = S - 2$. We notice that $\phi(S)$ is particularly low for $S$ whose prime factors are just 2 and 3.

As it can be seen in Figure 4, 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 $S_{\text{max}}$, is bounded away from one for reasonably large $S_{\text{max}}$.

In Figure 5 we plot how $\alpha^*\left(\frac{1}{2}, S_{\text{max}}\right)$ and $\overline{\alpha}(S_{\text{max}})$ depend on $S_{\text{max}}$. As noticed before, both the quantities stay always below 0.85, with $\overline{\alpha}(S_{\text{max}})$ being just a little better than $\alpha^*\left(\frac{1}{2}, S_{\text{max}}\right)$, specially for large $S$. The analytical connection between $\phi(S_{\text{max}})$ and changes of $p^*$ and $\alpha^*$ are beyond the scope of this paper and will be considered in future extensions.

### C. Bernoulli trials: $M > 1$

When each agent generates 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 $\overline{\alpha}(S_{\text{max}})$, which is fairly high, and thus estimation performance can be extremely poor. In this section we see how, acting on $M$, performance can achieve good rates. Assume then that nodes generate $M$ i.i.d. values $y_{i,1}, \dots, y_{i,M} \sim \mathcal{B}(p)$ from which they compute $f_m^{\text{ave}} := \dfrac{1}{S} \displaystyle\sum_i y_{i,m}$, $m = 1, \dots, M$ by means of average-consensus strategies. We define $\boldsymbol{f}^{\text{ave}} := [f_1^{\text{ave}}, \dots, f_M^{\text{ave}}]^T$. The ML estimator $\widehat{S}$ can

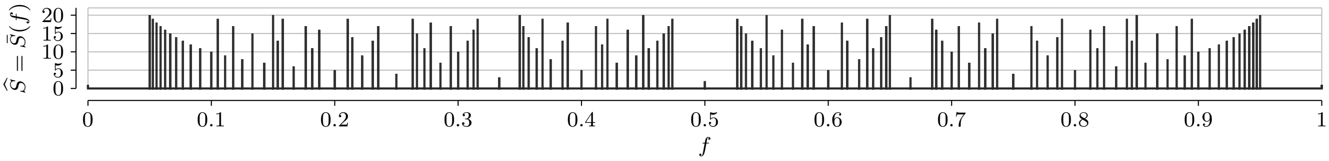


Figure 3: ML estimator $\widehat{S}$ as a function of $f$ for $S_{\text{max}} = 20$.

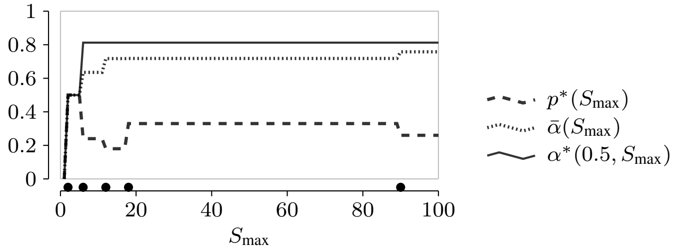


Figure 5: Dependency of $p^*(S_{\text{max}})$, $\overline{\alpha}(S_{\text{max}})$ and $\alpha^*(0.5, S_{\text{max}})$ on $S_{\text{max}}$. Circles on the abscissas axis indicate for which $S_{\text{max}}$ the former quantities change. We notice that the quantities vary only when the increase of $S_{\text{max}}$ implies to consider an $S$ having very few totatives, i.e., a particularly low $\phi(S)$.

be computed as follows: the independence between the various $y_{i,m}$ implies that the likelihood can be written as

$$\mathbb{P}\left[\boldsymbol{f}^{\text{ave}}\ ;\ S, p\right] = \prod_{m=1}^{M} \mathbb{P}\left[f_m^{\text{ave}}\ ;\ S, p\right] \tag{29}$$

which is non zero only if $S$ belongs to the intersection of the hypotheses spaces $\mathcal{I}_{f_m^{\text{ave}}}$, i.e., $S$ has non-zero likelihood only if $S \in \bigcap_{m=1}^{M} \mathcal{I}_{f_m^{\text{ave}}}$. This observation indirectly suggests how to compute the ML estimator $\widehat{S}$ in the more general scenario of multiple experiments $M$ following the same Occam's razor interpretation of Section IV-B:

**Proposition 13.** Given the likelihood in (29), then the ML estimator is given by

$$\begin{aligned}\widehat{S}(\boldsymbol{f}^{\text{ave}}) &:= \arg\max_{S \in \bigcap_{m=1}^{M} \mathcal{I}_{f_m^{\text{ave}}}} \mathbb{P}\left[\boldsymbol{f}^{\text{ave}}\ ;\ S, p\right] \\ &= \min\left(\bigcap_{m=1}^{M} \mathcal{I}_{f_m^{\text{ave}}}\right) = \text{LCM}\left(\overline{S}(f_1^{\text{ave}}), \ldots, \overline{S}(f_M^{\text{ave}})\right)\end{aligned} \tag{30}$$

for every $p \in [0, 1]$, where LCM $(\cdot)$ is the least common multiple operator. Moreover, it cannot overestimate the true number of nodes $S$ and is also biased, i.e.,

$$\widehat{S}(\boldsymbol{f}^{\text{ave}}) \leq S, \qquad \mathbb{E}_{\boldsymbol{f}^{\text{ave}}}\left[\widehat{S}(\boldsymbol{f}^{\text{ave}})\right] < S \quad \text{for} \quad S \geq 2\ .$$

*Proof.* To prove Proposition 13 it must be shown that

$$\mathbb{P}\left[f_1^{\text{ave}}, \ldots, f_M^{\text{ave}}\ ;\ \widehat{S}, p\right] \geq \mathbb{P}\left[f_1^{\text{ave}}, \ldots, f_M^{\text{ave}}\ ;\ \kappa\widehat{S}, p\right] \tag{31}$$

for all $\kappa \in \mathbb{N}$ and $p \in [0, 1]$. It follows immediately that $\widehat{S} = \nu_m \widehat{S}_m$ for an suitable $\nu_m \in \mathbb{N}$. Inequality (31) can then be proved considering that

$$\mathbb{P}\left[f_1^{\text{ave}}, \ldots, f_M^{\text{ave}}\ ;\ \widehat{S}, p\right] = \prod_{m=1}^{M} \mathbb{P}\left[f_m^{\text{ave}}\ ;\ \nu_m \widehat{S}_m, p\right]\ ,$$

$$\mathbb{P}\left[f_1^{\text{ave}}, \ldots, f_M^{\text{ave}}\ ;\ \kappa\widehat{S}, p\right] = \prod_{m=1}^{M} \mathbb{P}\left[f_m^{\text{ave}}\ ;\ \kappa\nu_m \widehat{S}_m, p\right]\ ,$$

and considering that corollary 12 can be used for element-by-element inequalities so that

$$\prod_{m=1}^{M} \mathbb{P}\left[f_m^{\text{ave}}\ ;\ \nu_m \widehat{S}_m, p\right] \geq \prod_{m=1}^{M} \mathbb{P}\left[f_m^{\text{ave}}\ ;\ \kappa\nu_m \widehat{S}_m, p\right]\ .$$

■

The computation of the optimal probability $p$ to minimize the error probability of the novel ML estimator is even more difficult than the one of the scenario $M = 1$, however some bounds can be obtained based on the analysis of the previous section. In fact notice that, if $\overline{S}(f_m^{\text{ave}})$ is defined in conformity to (22), then

$$\widetilde{S}(\boldsymbol{f}^{\text{ave}}) := \max\left(\min_{m} \mathcal{I}_{f_m^{\text{ave}}}\right) = \max\left\{\overline{S}(f_1^{\text{ave}}), \ldots, \overline{S}(f_M^{\text{ave}})\right\}$$

is a valid estimator that has the property $\widetilde{S}(\boldsymbol{f}^{\text{ave}}) \leq \widehat{S}(\boldsymbol{f}^{\text{ave}}) \leq S$. This implies that the error probability for the two estimators satisfy

$$\mathbb{P}\left[\widehat{S}(\boldsymbol{f}^{\text{ave}}) \neq S\ ;\ S, p\right] \leq \mathbb{P}\left[\widetilde{S}(\boldsymbol{f}^{\text{ave}}) \neq S\ ;\ S, p\right]\ .$$

For example, if $S = 6$ and $M = 2$, then the event $f_1^{\text{ave}} = \frac{1}{2}$, $f_2^{\text{ave}} = \frac{1}{3}$ leads to $\widehat{S}(2, 3) = \text{LCM}\,(2, 3) = S$ and $\widetilde{S}(2, 3) = \max\{2, 3\} = 3 \neq S$.

Since

$$\begin{aligned}\mathbb{P}\left[\widetilde{S}(\boldsymbol{f}^{\text{ave}}) \neq S\ ;\ S, p\right] &= \mathbb{P}\left[\overline{S}(f_m^{\text{ave}}) \neq S\ \forall m\ ;\ S, p\right] \\ &= \left(\mathbb{P}\left[\overline{S}(f_m^{\text{ave}}) \neq S\ ;\ S, p\right]\right)^M = \left(\alpha(p, S)\right)^M\end{aligned}$$

the error probability of the estimator $\widehat{S}$ exponentially decreases to zero with the number of experiments $M$. This observation is the basis to obtain the following bounds for the probability of error and the MSE of the ML estimator:

**Proposition 14.** Let $y_{i,1}, \ldots, y_{i,M} \sim \mathcal{B}(p^*)$, with $p^* = p^*(S_{\text{max}})$ and $\overline{\alpha} = \overline{\alpha}\,(S_{\text{max}})$ defined in (27). Then

$$(1 - p^*)^{S_{\text{max}} M} \leq \mathbb{P}\left[\widehat{S}(\boldsymbol{f}^{\text{ave}}) \neq S\ ;\ p^*, M\right] \leq (\overline{\alpha})^M \tag{32}$$

$$(1 - p^*)^{S_{\text{max}} M} \leq \mathbb{E}\left[\left(\widehat{S}(\boldsymbol{f}^{\text{ave}}) - S\right)^2 ; p^*, M\right] \leq (S_{\text{max}} - 1)^2\, (\overline{\alpha})^M. \tag{33}$$

*Proof.* • *case* $\mathbb{P}\left[\widehat{S} \neq S\ ;\ p^*, M\right] \leq (\alpha^*)^M$: a necessary condition for the event $\widehat{S} \neq S$ is $\widehat{S}_m \neq S$ for all $m$, thus

$$\mathbb{P}\left[\widehat{S} \neq S\ ;\ p^*, M\right] \leq \mathbb{P}\left[\widehat{S}_1 \neq S, \ldots, \widehat{S}_M \neq S\ ;\ p^*\right]\ .$$

Being moreover the various $\widehat{S}_m$'s independent, we obtain

$$\mathbb{P}\left[\widehat{S} \neq S\ ;\ p^*, M\right] \leq \prod_{m=1}^{M} \mathbb{P}\left[\widehat{S}_m \neq S\ ;\ p^*\right] \leq (\alpha^*)^M\ .$$

• *case* $(1 - p^*)^{S_{\text{max}} M} \leq \mathbb{P}\left[\widehat{S} \neq S\ ;\ p^*, M\right]$: In general, $f_m^{\text{ave}} = 0\ \forall m \Rightarrow \widehat{S} \neq S$. Being this a sufficient condition,

$$(1 - p^*)^{SM} = \mathbb{P}\left[f_1^{\text{ave}} = 0, \ldots, f_M^{\text{ave}} = 0\right] \leq \mathbb{P}\left[\widehat{S} \neq S\ ;\ p^*, M\right]\ .$$

The case is then proved considering that, $\forall S \in [1, S_{\max}]$, $(1-p^*)^{S_{\max}M} \leq (1-p^*)^{SM}$.

- *case* $\mathbb{E}\left[\left(\widehat{S}-S\right)^2 ; p^*, M\right] \leq (S_{\max}-1)^2 (\alpha^*)^M$: derives immediately from the inequality

$$\mathbb{E}\left[\left(\widehat{S}-S\right)^2 ; p^*, M\right] \leq (S_{\max}-1)^2 \, \mathbb{P}\left[\widehat{S} \neq S \; ; \; p^*, M\right] \; .$$

- *case* $(1-p^*)^{S_{\max}M} \leq \mathbb{E}\left[\left(\widehat{S}-S\right)^2 ; p^*, M\right]$: derives immediately from the inequality

$$\mathbb{P}\left[\widehat{S} \neq S \; ; \; p^*, M\right] \leq \mathbb{E}\left[\left(\widehat{S}-S\right)^2 ; p^*, M\right] \; .$$

■

Due to the nature of $\overline{\alpha}$ and $p^*$, the upper bounds in (32) and (33) are pessimistic. For example, as shown in Figure 6, $\mathbb{P}\left[\widehat{S} \neq S \; ; \; p^*, M\right]$ appears to decay to zero faster than what indicated in Proposition 14. The interpretation is the following: $(\alpha^*)^M$ represents the worst-case probability of the event $\widehat{S}_m \neq S$ for all $m$, that is a necessary but not sufficient condition for the event $\widehat{S} \neq S$. As soon as $M$ increases, the number of the cases where $\widehat{S} = S$ even if $\widehat{S}_m \neq S$ increases, and this leads to discrepancies between $(\alpha^*)^M$ and the actual $\mathbb{P}\left[\widehat{S} \neq S \; ; \; p^*, M\right]$. The same reasonings can be applied to $\mathbb{E}\left[\left(\widehat{S}-S\right)^2 ; p^*, M\right]$.

In any case (33) implies directly that the relative MSE for the current estimator scales at worst as $\overline{\alpha}^M$. Eventually this scaling is thus intrinsically different from the ones of the estimators based on max consensus described in Section III, proportional to $1/M$.

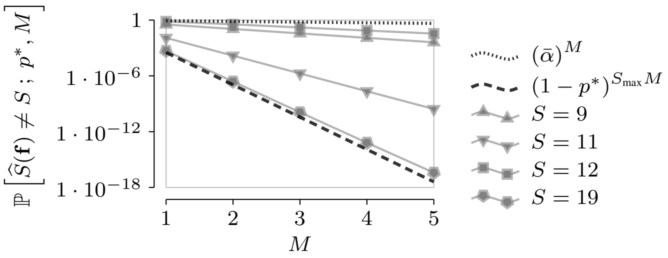


Figure 6: $\mathbb{P}\left[\widehat{S} \neq S \; ; \; p^*, M\right]$ as a function of $M$ and for various values of $S$, for the case $S_{\max} = 20$, and its lower and upper bounds described in (32).

## V. Effects of finite precision arithmetics

In practical implementations Assumptions 3, 4 and 5 will surely be violated, as mentioned at the end of Section II. Here we analyze the effects of additive errors that model the effects of quantization issues. We start by noticing that, based on the specific scenarios and adopted algorithms, such errors can be estimated a-priori (see, e.g., [42], [43], [44], [45]). We then assume that, if $f_{i,m}$, $m = 1, \ldots, M$, $i = 1, \ldots, S$ are the actual quantities computed by the various nodes, then

$$f_{i,m} = f_m + \Delta_{i,m} \qquad m = 1, \ldots, M, \; i = 1, \ldots, S \tag{34}$$

where the $\Delta_{i,m}$'s is the quantization error. Under uniform quantization, for example, these errors are bounded by an opportune $\Delta_{\max}$, i.e., $|\Delta_{i,m}| < \Delta_{\max}$.

### A. Max-consensus case with uniform distribution

Consider that the Jacobian of transformation (6) is

$$\nabla \widehat{S}(f_1, \ldots, f_M) = \left[ \frac{1}{f_1} \; \cdots \; \frac{1}{f_M} \right] \frac{1}{\sum_{m=1}^{M} -\log(f_m)} \widehat{S}(f_1, \ldots, f_M) \, .$$

Exploiting then model (34) and using a first-order Taylor expansion of $\widehat{S}$ around the error-free values $f_1, \ldots, f_M$ it follows that

$$\frac{\widehat{S}(f_{i,1}, \ldots, f_{i,M}) - \widehat{S}(f_1, \ldots, f_M)}{\widehat{S}(f_1, \ldots, f_M)} \approx \frac{\sum_{m=1}^{M} \Delta_{i,m} f_m^{-1}}{\sum_{m=1}^{M} -\log(f_m)} \, . \tag{35}$$

In the worst case expressed by bound $|\Delta_{i,m}| < \Delta_{\max}$ it follows that

$$\left| \frac{\Delta \widehat{S}}{\widehat{S}} \right| \lesssim \Delta_{\max} \frac{\sum_{m=1}^{M} f_m^{-1}}{\sum_{m=1}^{M} -\log(f_m)} \, .$$

This worst case scenario can thus be approximatively analyzed considering the behavior of the random variables $\widehat{S}(f_1, \ldots, f_M)$ and

$$\frac{\sum_{m=1}^{M} f_m^{-1}}{\sum_{m=1}^{M} -\log(f_m)} = \frac{\frac{1}{M}\sum_{m=1}^{M} f_m^{-1}}{\frac{1}{M}\sum_{m=1}^{M} -\log(f_m)} \, .$$

Recalling the results of Section III, $-\log(f_m)$ is an exponential random variable with rate $S$, while the density of $f_m$ is given by $p(f_m \; ; \; S) = S f_m^{S-1}$, implying that $\mathbb{E}\left[f_m^{-1} \; ; \; S\right] = \frac{S}{S-1}$. Exploiting (9), (10) and the i.i.d.-ness of the $-\log(f_m)$'s and of the $f_m^{-1}$'s, it holds that

$$\widehat{S}(f_1, \ldots, f_M) \xrightarrow[M \to +\infty]{\mathbb{P}} S$$

$$\frac{1}{M} \sum_{m=1}^{M} -\log(f_m) \xrightarrow[M \to +\infty]{\mathbb{P}} \mathbb{E}\left[-\log(f_m) \; ; \; S\right] = \frac{1}{S}$$

$$\frac{1}{M} \sum_{m=1}^{M} f_m^{-1} \xrightarrow[M \to +\infty]{\mathbb{P}} \mathbb{E}\left[f_m^{-1} \; ; \; S\right] = \frac{S}{S-1}$$

where the arrows indicate convergence in probability. Thus, in the max-case scenario and for $M$ sufficiently large, the (approximated) worst case bound reads as follows

$$\left| \frac{\Delta \widehat{S}}{\widehat{S}} \right| \lesssim \Delta_{\max} \frac{S^2}{S-1} \approx S \Delta_{\max} \, . \tag{36}$$

which describes again a smooth relation between the relative estimation error and the upper bounds on the computational errors.

### B. Average-consensus case with Bernoulli distribution

In this case it is possible to prove the following:

**Proposition 15.** Let the actually computed averages $f_{i,m}$ be mapped into an element $f_{i,m}^q$ of the alphabet $\mathbb{F}_{S_{\max}}$ via the relation $f_{i,m}^q := \arg\min_{f \in \mathbb{F}_{S_{\max}}} |f - f_{i,m}|$ . where $\mathbb{F}_{S_{\max}}$ was defined in (19). If $\Delta_{\max} < \dfrac{1}{2 S_{\max}(S_{\max}-1)}$ then $f_{i,m}^q = f_m$ .

*Proof.* Given $|\Delta_{i,m}| < \Delta_{\max}$, it holds that

$$|f_m - f_{i,m}| = |f_m - (1 + \delta_{i,m}) f_m - \Delta_{i,m}| \leq \Delta_{\max} \, .$$

The proposition then is proved as soon as $\Delta_{\max}$ is assured to be at most half of the minimal distance between the elements in $\mathbb{F}_{S_{\max}}$. Given definition (19), the latter quantity can be in general expressed as

$$\left| \frac{k_1}{S_1} - \frac{k_2}{S_2} \right| = \left| \frac{S_2 k_1 - S_1 k_2}{S_1 S_2} \right|$$

with $S_1, S_2 \in 1, \ldots, S_{\max}$, $k_1 \in 1, \ldots, S_1$, and $k_2 \in 1, \ldots, S_2$. The smallest distance corresponds thus to the choice $S_1 = S_{\max}$,

$S_2 = S_{\text{max}} - 1$, $k_1 = k_2 = 1$, thus to $\left(S_{\text{max}}\left(S_{\text{max}} - 1\right)\right)^{-1}$, and this eventually proves the proposition. ■

The intuition behind the previous proposition is that if $\Delta_{\text{max}}$ is sufficiently small then the measured average $f_{i,m}$ is mapped into the correct element $f^q_{i,m}$, while if not then $f_{i,m}$ *can* be mapped incorrectly. The consequences are the following: let the estimator $\widehat{S}$ to be defined on the various $f^q_{i,m}$ rather than $f^q_{i,m}$, i.e., let $\widehat{S} = \widehat{S}\left(f^q_{i,1}, \ldots, f^q_{i,M}\right)$ rather than $\widehat{S} = \widehat{S}\left(f_{i,1}, \ldots, f_{i,M}\right)$. If now $\Delta_{\text{max}}$ is sufficiently small then $\widehat{S}$ is exactly the same $\widehat{S}$ that would be obtained in absence of errors. Vice versa, if $\Delta_{\text{max}}$ is sufficiently big, then the difference between the computed $\widehat{S}$ and the $\widehat{S}$ that would be obtained in absence of errors may literally explode, due the strong discontinuous nature of the map of Figure 3 and of the operator $\text{LCM}\left(\cdot\right)$ in (30).

The strategy can nonetheless be implemented in real networks. As an illustrative example, assume for simplicity the quantization of the $y_{i,m}$ to be uniform in $[0, 1]$ and the convergence to the average values to be up to the size of the quantization bins (see, e.g., [45]). Given Proposition (15), for a network of 200 nodes the quantization error must be at most $1.25 \cdot 10^{-5}$, that can be achieved using just 17 bits per each $y_{i,m}$, while for a network of 1000 nodes then the quantization error must be at most $5 \cdot 10^{-7}$, that can be achieved using just 21 bits.

## VI. Numerical experiments

We now propose practical implementation procedures that do not rely on Assumption 4. We then evaluate their performance with opportune Monte Carlo (MC) analyses.

### A. Implementation

We focus on two specific asynchronous broadcast max and average consensus schemes suitable for Wireless Sensor Networks (WSNs), summarized in Algorithms 1 and 2 (the latter inspired by [34]). Notice that we do not extensively analyze the multitude of average consensus algorithms present in literature, being this beyond the purpose of the paper. We recall that we model the network with the graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{1, \ldots, S\}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of the communication links ($\mathcal{G}$ is assumed to be undirected, i.e., $(i, j) \in \mathcal{E} \Leftrightarrow (j, i) \in \mathcal{E}$). We also let $\mathcal{V}_i$ be the set of neighbors of node $i$ ($i \notin \mathcal{V}_i$), and $d_i := |\mathcal{V}_i|$ its cardinality.

**Algorithm 1** Maximum consensus

```
(initialization) for $i = 1, \ldots, S$, $m = 1, \ldots, M$ let $f^{\text{max}}_{i,m}(0) = y_{i,m} \sim \mathcal{U}[0, 1]$ i.i.d.
for t = 1, 2, ... do
    (node extraction) select $i \in \mathcal{V}$ (i.i.d. and with uniform extraction probability)
    for $j \in \mathcal{V}_i$ do
        (update of main variables: neighbors) for $m = 1, \ldots, M$ do
            $f^{\text{max}}_{j,m}(t+1) = \max\left\{f^{\text{max}}_{i,m}(t), f^{\text{max}}_{j,m}(t)\right\}$    (37)
    for $j \notin \mathcal{V}_i$ do
        (update of main variables: remaining nodes) for $m = 1, \ldots, M$ do
            $f^{\text{max}}_{j,m}(t+1) = f^{\text{max}}_{j,m}(t)$    (38)
```

**Algorithm 2** Average consensus

```
(initialization) for $i = 1, \ldots, S$, $m = 1, \ldots, M$ let $\widetilde{f}^{\text{ave}}_{i,m}(0) = y_{i,m} \sim \mathcal{B}(p)$ i.i.d., $z_{i,m}(0) = 1$, $f^{\text{ave}}_{i,m}(0) = \frac{\widetilde{f}^{\text{ave}}_{i,m}(0)}{z_{i,m}(0)}$
for t = 1, 2, ... do
    (node extraction) select $i \in \mathcal{V}$ (i.i.d. and with uniform extraction probability)
    (update of auxiliary variables: extracted node) for $m = 1, \ldots, M$ do
        $\widetilde{f}^{\text{ave}}_{i,m}(t+1) = \frac{1}{1+d_i}\widetilde{f}^{\text{ave}}_{i,m}(t) + \sum_{j \in \mathcal{V}_i} \frac{1}{1+d_j}\widetilde{f}^{\text{ave}}_{j,m}(t)$    (39)
        $z_{i,m}(t+1) = \frac{1}{1+d_i} z_{i,m}(t) + \sum_{j \in \mathcal{V}_i} \frac{1}{1+d_j} z_{j,m}(t)$    (40)
    for $j \in \mathcal{V}_i$ do
        (update of auxiliary variables: neighbors) for $m = 1, \ldots, M$ do
            $\widetilde{f}^{\text{ave}}_{j,m}(t+1) = \frac{1}{1+d_i}\widetilde{f}^{\text{ave}}_{i,m}(t) + \frac{d_j}{1+d_j}\widetilde{f}^{\text{ave}}_{j,m}(t)$    (41)
            $z_{i,m}(t+1) = \frac{1}{1+d_i} z_{i,m}(t) + \frac{d_j}{1+d_j} z_{j,m}(t)$    (42)
    for $j \notin \mathcal{V}_i$ do
        (update of auxiliary variables: remaining nodes) for $m = 1, \ldots, M$ do
            $\widetilde{f}^{\text{ave}}_{j,m}(t+1) = \widetilde{f}^{\text{ave}}_{j,m}(t) \qquad z_{i,m}(t+1) = z_{i,m}(t)$    (43)
    (update of main variables: all the nodes) for $j = 1, \ldots, S$, $m = 1, \ldots, M$ do
        $f^{\text{ave}}_{j,m}(t+1) = \frac{\widetilde{f}^{\text{ave}}_{j,m}(t+1)}{z_{j,m}(t+1)}$    (44)
```

Notice that, assuming strongly connected communication networks and finely quantized data, the uniform random node selection process ensures finite time convergence in probability of the local variable of both Algorithms (see [28] for Algorithm 1 and [46] for Algorithm 2), i.e.,

$$\mathbb{P}\left[\exists \tau \mid f^{\text{max}}_{i,m}(t) = f^{\text{max}}_{m}, \forall t \geq \tau\right] = 1,$$

$$\mathbb{P}\left[\exists \tau \mid f^{\text{ave}}_{i,m}(t) = f^{\text{ave}}_{m}, \forall t \geq \tau\right] = 1, \forall i, \forall m$$

where $f^{\text{max}}_m = \max_i \{y_{i,m}\}$ and $f^{\text{ave}}_m = \frac{1}{S}\sum_{i=1}^{S} y_{i,m}$.

It is then convenient to redefine following local dynamic estimators:

$$\widehat{S}_i^{UM}(t) := \left(\frac{1}{M}\sum_{m=1}^{M} -\log\left(f^{\text{max}}_{i,m}(t)\right)\right)^{-1}$$

$$\widehat{S}_i^{BA}(t) := \text{LCM}\left(\overline{S}\left(f^{\text{ave}}_{i,1}(t)\right), \ldots, \overline{S}\left(f^{\text{ave}}_{i,M}(t)\right)\right)$$

where $\overline{S}(\cdot)$ is defined in (21) and where the superscripts $UM$ and $BA$ indicate that the Uniform-Maximum strategy and the Bernoulli-Average strategy, respectively. Since the arguments of these estimators converge in finite time in probability, the also the estimators inherit finite time convergence in probability. More formally,

$$\mathbb{P}\left[\exists \tau \mid \widehat{S}_i^{UM}(t) = \Psi_{\text{ML}}\left(f^{\text{max}}_1, \ldots, f^{\text{max}}_M\right), \forall t \geq \tau\right] = 1$$
$$\mathbb{P}\left[\exists \tau \mid \widehat{S}_i^{BA}(t) = \widehat{S}\left(\boldsymbol{f}^{\text{ave}}\right), \forall t \geq \tau\right] = 1, \forall i$$

where the right hand side of the equalities correspond to the asymptotic estimators defined in (11) and (30), respectively.

### B. Stopping criteria

The dynamic implementation of the estimators requires the definition of opportune stopping criteria. We notice that the approaches that can be followed are mainly two: the first is to estimate an *a-priori* stopping time $\tau$ that guarantees reaching consensus with an arbitrarily large probability. For example, [28] and [42] provide bounds for such stopping times for the max-consensus and the average-consensus, respectively, that depend on the graph connectivity properties such as the conductance or the spectral gap. However, these values are either not known in advance or, even if known, they lead to very conservative bounds with little practical use.

The second approach is to define *a-posteriori* stopping times based on the observed evolution of the local estimates $\widehat{S}_i(t)$. Considering that the quantized consensus implementations described in this section eventually converge in a finite number of steps, we propose the following heuristic, which exhaustive analysis is beyond the scope of this paper: each node $i$ counts how many times it performed a communication step (thus either step 3 or 5 in Algorithm 1 or step 4 or 6 in Algorithm 2) and then stops (i.e., become *not selectable* in step 3) if the local estimate $\widehat{S}_i(t)$ has not changed in a certain user-defined number of steps $T$. We notice that it is meaningful to let this interval depend on the current estimate, e.g., be of $\left\lceil T_m \widehat{S}_i(t) + T_q \right\rceil$ communication steps, with $T_m$ and $T_q$ some user-defined quantities. Nonetheless in our experiments this dependency was in general not incrementing the convergence performance.

### C. Monte Carlo analysis

We consider random geometric graphs of $S = 40$ nodes obtained by uniform placement in the $[0,1] \times [0,1]$ square, with nodes with uniform communication radius $\rho = 0.3$ (a realization is shown in Figure 7a). All transmitted data and local variables are encoded into 16 bits arithmetics. We also assume the knowledge of $S_{\max} = 80$ as an upper bound on the size of the network, and we choose $T = 10$ as the stopping parameter for both Algorithm 2 and Algorithm 1.

Figures 7b and 7c show respectively the evolution of a typical realization of some of the local estimates $\widehat{S}_i^{UM}(t)$ and $\widehat{S}_i^{BA}(t)$ for the network depicted in Figure 7a, as a function of the number of effective communication steps of each node, as defined in Section VI-B. As expected, the estimates obtained with the max-consensus based strategy are monotonically increasing, and in general do not converge to the true size $S$. Differently, the ones obtained with the Bernoulli-trials based strategy show a peculiar behavior: before reaching consensus, the $f_{i,m}^{\text{ave}}(t)$'s might be associated to fractions whose denominators are not factors of $S$. Due to the multiplicative effects of the underlying Least Common Multiple (LCM) operation in (30), $\widehat{S}_i^{BA}(t)$ might thus *temporarily* be much bigger than both $S$ and $S_{\max}$. The situation $\widehat{S}_i^{BA}(t) > S_{\max}$, a clear sign that the consensus has not yet converged, is thus handled in our simulations by putting $\widehat{S}_i^{BA}(t) = 0$. The time evolution shows that initially $\widehat{S}_i^{BA}(t) = 0$, but as soon as the local estimators provide a feasible output, i.e., $0 < \widehat{S}_i^{BA}(t) \leq S_{max}$, then this is the exact network size $S$ in most of the realizations.

Figures 8a and 8c plot the empirical distributions of the convergence times of the estimators $\widehat{S}_i^{UM}(t)$ and $\widehat{S}_i^{BA}(t)$ obtained from 1000 MC experiments with $M = 1$ and $M = 5$. We notice that the convergence times do not strongly depend on $M$, and that the max-consensus based method requires much fewer communication steps than the Bernoulli-trials based estimator, typically an order of magnitude smaller. The latter algorithm compensates this slower convergence, inherited by the convergence properties of the average-consensus algorithm, with an extremely higher accuracy of the estimates. Consider in fact Figures (8b) and (8d), showing the empirical distributions of the estimates $\widehat{S}_i^{UM}$ and $\widehat{S}_i^{BA}$ at the stopping times obtained from 1000 MC experiments with $M = 1, 5$. Here it is immediate to notice how the estimates $\widehat{S}_i^{UM}$ have a large variance for both $M = 1, 5$, while $\widehat{S}_i^{BA}$ for $M = 5$ provided the exact network size $S = 40$ for 97.3% of the times.

To highlight this different behavior of the estimation error we plot in Figure 9 the relative MSEs of the two estimators for various $M$'s (1000 MC runs for each $M$). As expected, the MSE of the Bernoulli-Average Strategy (BAS) estimator decays exponentially with $M$, while the one of the max-consensus based estimator scales as $1/M$. We also notice that $\widehat{S}_i^{BA}$ always outperforms $\widehat{S}_i^{UM}$ of at least two orders of magnitude.

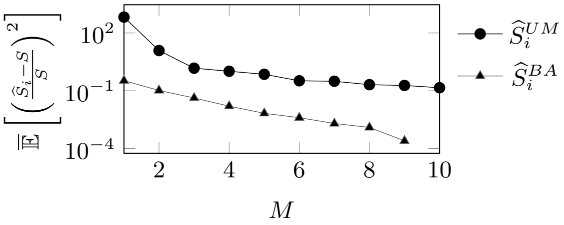


Figure 9: Comparison of the empirical MSEs of the estimators $\widehat{S}_i^{UM}$ and $\widehat{S}_i^{BA}$ for different values of $M$ and 1000 MC runs per $M$. The plot does not show the MSE of $\widehat{S}_i^{BA}$ for $M = 10$, since in our MC simulations this estimator always detected correctly $S$.

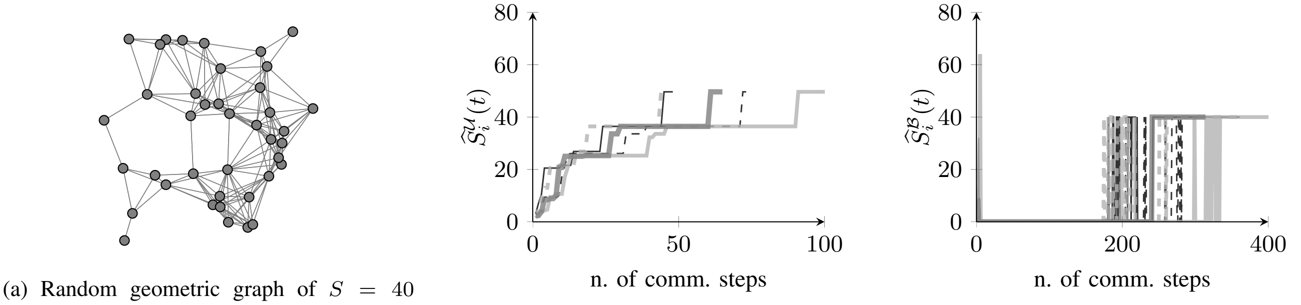
## VII. Conclusions and future works

In this work we characterized two strategies for estimating the size of anonymous networks, based on first generating a set of i.i.d. random numbers, then computing either their maximum or average, and then exploiting the statistical correlation between these quantities and the network size.

We characterized the intrinsic differences between the maximum and the average consensus strategies. More precisely, we showed that for the maximum consensus strategy the variance of the relative estimation error is intrinsically proportional to the inverse of the number of samples and is independent of the particular density chosen to generate the data. For the average consensus strategy we instead showed that when using discrete distributions, more specifically Bernoulli trials, the probability of returning a wrong value of the network size goes to zero exponentially with the number of samples. However, this desirable property comes at the price of a slower convergence, since distributed algorithms for computing averages are intrinsically slower than distributed algorithms for computing maxima.

It has been also shown how these distinct estimation strategies exhibit drastically different sensitivities to numerical errors. In fact, when the random numbers are realizations of continuous random variables the sensitivity is proportional to the amplitude of the numerical errors. Instead, when the random numbers follow Bernoulli trials, the estimation process is either insensible or completely unreliable, depending again on the amplitude of the numerical errors.

This work leads to numerous plausible future research directions. One is the implementation of fast average consensus algorithms such the diffusive methods to reduce the converge time of the Bernoulli-Average strategy. Another is the application of the algorithms for real time tracking, network size change detection, and more generally network topology discovery applications. For instance, the algorithms here proposed could be used to check if a certain network is more likely to be circulant than a star by looking at *how* and *how fast* the estimates have been obtained.

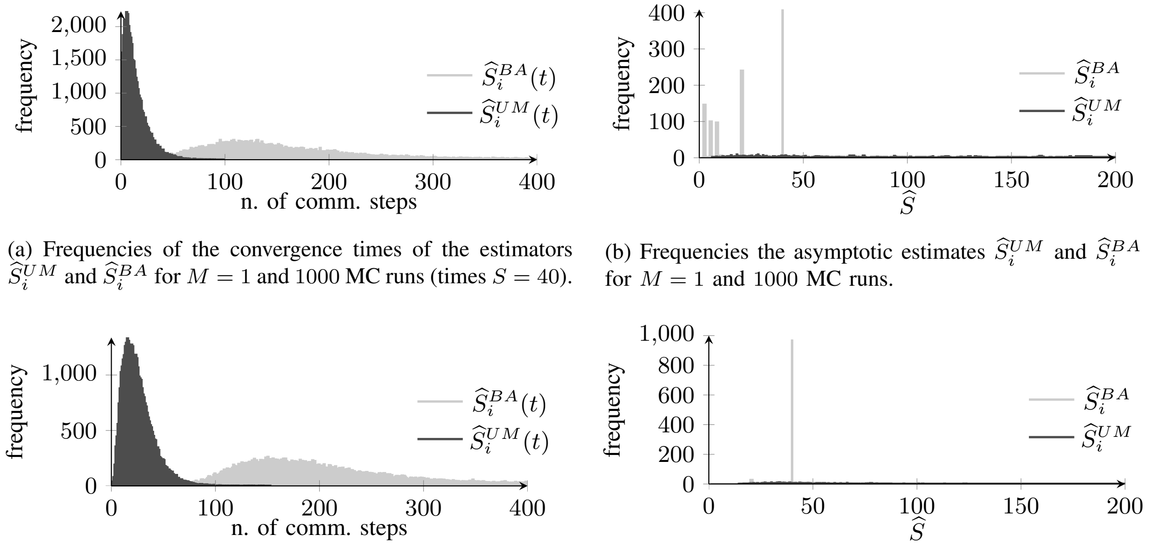


(a) Random geometric graph of $S = 40$ nodes in $[0,1] \times [0,1]$ and with communication radius $\rho = 0.3$.

(b) Evolution of the local estimates $\widehat{S}_i^{UM}(t)$ $(i = 1, \ldots, 5)$ for $M = 5$.

(c) Evolution of the local estimates $\widehat{S}_i^{BA}(t)$ $(i = 1, \ldots, 5)$ for $M = 5$.

Figure 7: Example of a single Monte-Carlo run with $M = 5$.



(a) Frequencies of the convergence times of the estimators $\widehat{S}_i^{UM}$ and $\widehat{S}_i^{BA}$ for $M = 1$ and 1000 MC runs (times $S = 40$).

(b) Frequencies the asymptotic estimates $\widehat{S}_i^{UM}$ and $\widehat{S}_i^{BA}$ for $M = 1$ and 1000 MC runs.

(c) Frequencies of the convergence times of the estimators $\widehat{S}_i^{UM}$ and $\widehat{S}_i^{BA}$ for $M = 5$ and 1000 MC runs (times $S = 40$).

(d) Frequencies the asymptotic estimates $\widehat{S}_i^{UM}$ and $\widehat{S}_i^{BA}$ for $M = 5$ and 1000 MC runs.

Figure 8: Comparison of the asymptotic properties of the estimators $\widehat{S}_i^{UM}(t)$ and $\widehat{S}_i^{BA}(t)$ for different values of $M$.

## References

[1] T. M. Shafaat, A. Ghodsi, and S. Haridi, "A Practical Approach to Network Size Estimation for Structured Overlays," *Self-organizing Systems*, vol. 5343, pp. 71–83, 2008.

[2] 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.

[3] P. Boldi and S. Vigna, "An Effective Characterization of Computability in Anonymous Networks," *Lecture Notes in Computer Science*, vol. 2180 / 200, pp. 33 – 47, 2001.

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

[5] J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, "Distributed anonymous discrete function computation," *IEEE Transactions on Automatic Control*, vol. 56, no. 10, pp. 2276–2289, Oct. 2011.

[6] C. Raghavendra, K. Sivalingam, and T. Znati, *Wireless Sensor Networks*. Springer, 2006.

[7] D. Varagnolo, G. Pillonetto, and L. Schenato, "Distributed parametric and nonparametric regression with on-line performance bounds computation," *Automatica*, vol. 48, no. 10, pp. 2468–2481, 2012.

[8] F. Bullo, J. Cortés, and S. Martínez, *Distributed control of robotic networks*. Princeton University Press, 2009.

[9] 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, Mar. 1947.

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

[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, Mar. 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 *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, 2006, pp. 123–132.

[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, Oct. 2007.

[15] K. Horowitz and D. Malkhi, "Estimating Network Size from Local Information," *Information Processing Letters*, vol. 88, no. 5, pp. 237–243, Dec. 2003.

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

[17] 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, Sept. 2009.

[18] 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.

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

[20] M. Sirken and I. Shimizu, "Population based establishment sample surveys: The Horvitz-Thompson estimator," *Survey Methodology*, vol. 25, no. 2, pp. 187–191, Dec. 1999.

[21] 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.

[22] 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*, Sept. 2009.

[23] M. Howlader, M. R. Frater, and M. J. Ryan, "Estimating the Number and Distribution of the Neighbors in an Underwater Communication Network," in *SENSORCOMM*, Aug. 2008.

[24] R. Ali, S. S. Lor, and M. Rio, "Two algorithms for network size estimation for master/slave ad hoc networks," in *IEEE 3rd International Symposium on Advanced Networks and Telecommunication Systems*, Dec. 2009.

[25] A. Leshem and L. Tong, "Estimating sensor population via probabilistic sequential polling," *IEEE Signal Processing Letters*, vol. 12, no. 5, pp. 395–398, May 2005.

[26] L. Ching, "Distributed Algorithms for Dynamic Topology Construction and Their Applications," Ph.D. dissertation, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 2004.

[27] 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, Dec. 1997.

[28] 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.

[29] P. Jesus, C. Baquero, and P. S. Almeida, "A Survey of Distributed Data Aggregation Algorithms," University of Minho, Tech. Rep., 2011.

[30] P. S. S. Almeida, C. Baquero, and A. Cunha, "Fast Distributed Computation of Distances in Networks," in *IEEE Conference on Decision and Control*, Nov. 2012.

[31] J. Cichon, 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, Apr. 2012.

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

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

[34] C. N. Hadjicostis and T. Charalambous, "Average Consensus in the Presence of Delays and Dynamically Changing Directed Graph Topologies," *IEEE Transactions on Automatic Control (under review)*, Oct. 2012.

[35] J. Cortés, "Distributed algorithms for reaching consensus on general functions," *Automatica*, vol. 44, no. 3, pp. 726–737, Mar. 2008.

[36] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proceedings of the Third Berkeley Symposyum on Mathematical Statistics and Probability*, vol. 1, no. 4, pp. 197–206, 1956.

[37] H. A. A. David and H. N. N. Nagaraja, *Order Statistics*. Wiley series in Probability and Statistics, 2003.

[38] 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*, Atlanta, USA, Dec. 2010, pp. 1498–1503.

[39] J. Cichon, J. Lemiesz, and M. Zawada, "On Cardinality Estimation Protocols for Wireless Sensor Networks," *Ad-hoc, mobile, and wireless networks*, vol. 6811, pp. 322–331, 2011.

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

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

[42] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized Gossip Algorithms," *IEEE Transactions on Information Theory / ACM Transactions on Networking*, vol. 52, no. 6, pp. 2508–2530, June 2006.

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

[44] F. Fagnani and S. Zampieri, "Average Consensus with Packet Drop Communication," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 102–133, Jan. 2009.

[45] T. Aysal, M. Coates, and M. Rabbat, "Distributed Average Consensus With Dithered Quantization," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4905–4918, Oct. 2008.

[46] R. Carli, F. Fagnani, P. Frasca, and S. Zampieri, "Gossip consensus algorithms via quantized communication," *Automatica*, vol. 46, no. September, pp. 70–80, 2010.

**Damiano Varagnolo** (M'08) received the Dr. Eng. degree in automation engineering and the Ph.D. degree in information engineering from the University of Padova respectively in 2005 and 2011. He was research engineer at Tecnogamma S.p.A., Treviso, Italy during 2006-2007 and visiting scolar researcher at UC Berkeley in 2010. Currently he is a post-doctoral scholar at KTH, Royal Institute of Technology, Stockholm. His interests include statistical learning, distributed optimization, distributed non-parametric estimation techniques, multi-agent systems, and control of HVAC systems.

**Gianluigi Pillonetto** (M'03) was born on January 21, 1975 in Montebelluna (TV), Italy. He received the Doctoral degree in Computer Science Engineering cum laude from the University of Padova in 1998 and the PhD degree in Bioengineering from the Polytechnic of Milan in 2002. In 2000 and 2002 he was visiting scholar and visiting scientist, respectively, at the Applied Physics Laboratory, University of Washington, Seattle. From 2002 to 2005 he was Research Associate at the Department of Information Engineering, University of Padova. Since 2005 he has been Assistant Professor of Control and Dynamic Systems at the Department of Information Engineering, University of Padova. His research interests are in the field of system identification, stochastic systems, deconvolution problems, nonparametric regularization techniques, learning theory and randomized algorithms.

**Luca Schenato** (M'04) received the Dr. Eng. degree in electrical engineering from the University of Padova in 1999 and the Ph.D. degree in Electrical Engineering and Computer Sciences from the UC Berkeley, in 2003. From January 2004 till August 2004 he was a post-doctoral scholar at U.C. Berkeley. Currently he is Associate Professor at the Information Engineering Department at the University of Padova. His interests include networked control systems, multi-agent systems, wireless sensor networks, swarm robotics and biomimetic locomotion. Dr. Schenato has been awarded the 2004 Researchers Mobility Fellowship by the Italian Ministry of Education, University and Research (MIUR), and the 2006 Eli Jury Award in U.C. Berkeley. He currently serves as Associate Editor for IEEE Transactions on Automatic Control.