Network cardinality estimation using max consensus: the case of Bernoulli trials

Riccardo Lucchese   Damiano Varagnolo   Jean-Charles Delvenne   Julien Hendrickx

***Abstract*— Interested in scalable topology reconstruction strategies with fast convergence times, we consider network cardinality estimation schemes that use, as their fundamental aggregation mechanism, the computation of bit-wise maxima over strings. We thus discuss how to choose optimally the parameters of the information generation process under frequentist assumptions on the estimand, derive the resulting Maximum Likelihood (ML) estimator, and characterize its statistical performance as a function of the communications and memory requirements. We then numerically compare the bitwise-max based estimator against lexicographic-max based estimators, and derive insights on their relative performances in function of the true cardinality.**

***Index Terms*— distributed estimation, size estimation, bitwise max consensus, quantization effects, peer-to-peer networks.**

I. Introduction

Information on the topology of a communication network may be instrumental in distributed applications like optimization and estimation tasks. For example, in distributed regression frameworks, knowing the number of active sensors allows to correctly weight prior information against evidence of the data [1]. Moreover, continuously estimating the number of active nodes or communication links corresponds to monitoring the network connectivity, and thus to being able to trigger network reconfiguration strategies [2].

The focus is then to understand how to distributedly perform topology reconstruction given devices with bounded resources (e.g., battery / energy constraints, communication costs, etc.). Of course, considering different trade-offs leads to different optimal strategies. Here we are motivated by real-world applications such as vehicular traffic estimation and specifically consider the case of peer-to-peer networks where all the participants are required to: *i)* share the same final result (and thus the same view of the network); *ii)* keep the communication and computational complexity at each node uniformly bounded in time; *iii)* reach consensus on the estimates using the smallest number of communications possible.

This work is supported by the Framework Programme for Research and Innovation Horizon 2020 under grant agreement n. 636834 (DISIRE), and the Swedish research council Norrbottens Forskningsråd.

Riccardo Lucchese and Damiano Varagnolo are with the Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden. Emails: { riccardo.lucchese | damiano.varagnolo }@ltu.se. Jean-Charles Delvenne and Julien Hendrickx are with the Institute of Information and Communication Technologies, Electronics and Applied Mathematics, Université catholique de Louvain, Louvain-la-Neuve, Belgium. Emails: { jean-charles.delvenne | julien.hendrickx }@uclouvain.be

Since aggregation mechanisms scale better than flooding or epidemic protocols (at the cost of some loss of information) [3], [4], the aforementioned objectives are usually addressed using order statistics consensus aggregation mechanisms (like max, min, and combinations of them). Natural questions are then: which one is the scheme that leads to topology estimators that are optimal in Mean Squared Error (MSE) terms? And what are the fundamental limitations of information aggregation for topology estimation purposes, i.e., what can be estimated and what not?

Towards answering what is the maximum achievable accuracy of aggregation-based estimators, here we focus on max-consensus strategies and pursue to characterize the fundamental properties of aggregating maxima for cardinality estimation purposes.

Literature review: if agents of a network are not constrained to keep their communication and memory requirements fixed at every iteration, then it is known that one can reconstruct the whole topology of a network by both exchanging tables of the agents IDs, if these IDs are unique, or using simple randomized techniques to generate these IDs [5]. If instead communications and memory requirements have to stay constant in time, and IDs are not guaranteed to be unique, there exists no algorithm that always computes correctly with probability one, in finite time and with a bounded average bit complexity even just the size of the network [6], [7].

These results motivate the existence of probabilistic counting algorithms, where agents estimate the size of their network by either performing different actions based on the perceived events (as in interval sampling, capture-recapture or random walks [8], [9], [10], [11], [12], [13]) or performing the same actions in parallel (as in the case where all the agents are required to share the same knowledge) [14], [15] [16].

The particular scenario considered in this manuscript is usually approached endowing each agent with (possibly non-unique) IDs and letting then the network compute opportune statistics of these IDs. Estimators of this kind have three building blocks: 1) an initialization phase, where the local memory $\boldsymbol{y}_i$ of each agent i is initialized locally using some probabilistic mechanism; 2) an aggregation phase, where the network distributedly computes an opportune function of the initial $\boldsymbol{y}_i$'s and eventually reaches consensus on a value $\boldsymbol{y}$; 3) an estimation phase, where each agent infers the size of the network from $\boldsymbol{y}$.

Aggregating the $\boldsymbol{y}_i$'s using average consensus is then known to lead to estimators whose statistical performance improve either linearly [17], [18], [19], [20] or exponentially [21] with the size of $\boldsymbol{y}_i$'s (depending on how the $\boldsymbol{i}_i$'s are initialized). Averaging nonetheless has the big drawback of slow convergence dynamics (a property that is inherited from the underlying averaging process).

Aggregating the $\boldsymbol{y}_i$'s using order statistics consensus (e.g., max-consensus) has the advantage of converging in a smaller number of communication steps then is required by an averaging process. Specifically, the computation of maxima over the $\boldsymbol{y}_i$'s can be performed in two different ways: 1) using a lexicographic order, if the $\boldsymbol{y}_i$ represent a real number (or a vector thereof); 2) bitwise, when the $\boldsymbol{y}_i$ are viewed as a string of bits.

The properties of estimation strategies using the lexicographic order have been analyzed in the literature and variants of these schemes have been proposed to address specific tasks. Statistical characterizations can be found in [22], [23], [20], [24], and have been improved in [25] by exploiting the aggregation of order statistics (i.e., computing the k-th biggest maximum of the various $\boldsymbol{y}_i$ instead of just the maximum value. This leads to an estimator that is a perfect counter for small networks and with the same estimation performance of the aforementioned methods for big networks). [26], [27] instead exploit temporal repetitions of the max-consensus strategy to build estimators that are tailored for dynamic networks with size changing in time.

In contrast, the literature on bitwise strategies is not so abundant: at the best of our knowledge the unique manuscript is [28] where the authors generate the $\boldsymbol{y}_i$'s with Bernoulli trials similarly to what we propose here, but both derive a different estimator (cf. the following statement of contributions) and do not consider the optimal design of the Bernoulli parameters.

Statement of contributions: we consider network size estimation based on bitwise max-consensus strategies. This focus is motivated by the fact that the literature dealing with lexicographic max consensus is at the best of our knowledge neglecting the discrete nature of the $\boldsymbol{y}_i$'s and obtains approximate results that are based on the assumptions that the $\boldsymbol{y}_i$'s are absolutely continuous r.v.s; in other words the literature ignores quantization effects. With analyzing bitwise max-consensus schemes we thus both begin accounting for the discrete nature of the $\boldsymbol{y}_i$'s and work towards understanding the performance limitations of computing maxima bitwise or lexicographically. Our contributions are thus:

- extending [28] by considering potentially non-identically distributed bits, and determining the optimal Bernoulli rates using frequentist assumptions in (13);
- obtaining the novel ML estimator (18), different from the one in [28], characterizing its statistical properties in Propositions 2 and 3, and verifying that it practically reaches its Cramér-Rao (C-R) bound;
- comparing bitwise and lexicographic estimators and collecting numerical evidence on which strategy is optimal in Sec. VII.

Organization of the manuscript: Sec. II introduces our assumptions, while Sec. III formally casts the cardinality estimation problem. Sections IV, V and VI address different aspects of the estimation problem, by respectively designing the structure of parameters dictating the information generation scheme, determining the functional structure of the estimator, and characterizing its statistical performances. Sec. VII then compares the performance of our bitwise-max estimator with that of lexicographic-max strategies. Finally, Sec. VIII collects a few concluding remarks and discusses future directions.

II. Background and Assumptions

We model a distributed network as a connected undirected graph $G = (V, E)$ comprising $N = |V|$ collaborating agents. We assume that the network operates within the following shared framework:

Memory model: the generic agent $i \in V$ avails locally of a memory storage of M-bits that is represented by the vector

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i,1} & y_{i,2} & \dots & y_{i,M} \end{bmatrix}^T \in \{0,1\}^M . \tag{1}$$

Communication model: time is partitioned into an ordered set of equally lasting intervals indexed by $t = 0, 1, 2, \ldots$, each referred to as an "epoch". During each epoch, randomly, uniformly and i.i.d. during the epoch, each agent $i \in V$ broadcasts its whole $\boldsymbol{y}_i$ to all its neighbors through a perfect channel (i.e., without collisions, delays, or communication errors).

Aim of the agents: to estimate the cardinality of the network N while being subject to the following constraints:

C1) obtain the same estimate when the algorithm terminates (i.e., letting $\widehat{N}_i$ denote the final estimate for the generic agent i, it is required that $\widehat{N}_i = \widehat{N}$, $\forall i \in V$);

C2) obtain this estimate in d epochs, where d is the network's diameter (notice that in our synchronous protocol d is the minimum number of epochs such that information generated at any node is propagated to the remaining nodes in the network).

We moreover assume that N is unknown but deterministic. Agents have no a-priori knowledge on the network topology and thus on its cardinality except for an upper bound on the network size, i.e., there exists a number $N_{\max}$ such that $N \leq N_{\max}$ and $N_{\max}$ is available to the network.

III. Problem Formulation: Size estimation with Bernoulli trials

Statistical size estimation schemes that are based on aggregation strategies share the following common structure:

1) during initialization, each agent independently initialize its memory $\boldsymbol{y}_i$ extracting a value from a probability distribution $\mathbb{P}$ that is independent of N;
2) then agents aggregate the various $\boldsymbol{y}_i$ (i.e., distributedly compute a function of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$) and reach consensus on a final $\boldsymbol{y}$;
3) since N parameterizes the previous aggregation process, N becomes statistically identifiable through $\boldsymbol{y}$.

Thus, even if the $\boldsymbol{y}_i$'s do not depend statistically on N, $\boldsymbol{y}$ does, so that $\boldsymbol{y}$ conveys statistical information on N.

The design of size estimators is then possible on 3 levels: 1) which $\mathbb{P}$ to use to initialize the $\boldsymbol{y}_i$'s; 2) which aggregation scheme to use; 3) how to map the final aggregate $\boldsymbol{y}$ into a point estimate $\widehat{N}$ of N.

As for the first design level, we consider the specific $\mathbb{P}$ for which the $\boldsymbol{y}_i$'s are initialized bit-wise, i.e., for which each smallest atom of available information is initialized independently. More specifically, we assume that each M-dimensional memory $\boldsymbol{y}_i = [y_{i,1}, \ldots, y_{i,M}]$ is initialized with M i.i.d. Bernoulli samples, i.e., with

$$y_{i,m} = \begin{cases} 1 & \text{with probability } 1-\theta_m \\ 0 & \text{with probability } \theta_m \end{cases} \qquad m = 1, \ldots, M\,. \tag{2}$$

As for the second design level, we consider the bit-wise max consensus of the $\boldsymbol{y}_i$'s, an aggregation operation that eventually yields (in finite time and at each agent) the vector

$$\boldsymbol{y} = [y_1, \ldots, y_M]^T, \quad y_m := \max_{i \in V} \{y_{i,m}\}, \quad m = 1, \ldots, M \tag{3}$$

with probability

$$\mathbb{P}[\boldsymbol{y}\,;\, N, \boldsymbol{\theta}] = \prod_{\{m \,:\, y_m = 1\}} (1 - \theta_m^N) \prod_{\{m \,:\, y_m = 0\}} \theta_m^N \tag{4}$$

with $\boldsymbol{\theta} := [\theta_1, \ldots, \theta_M]$. The generated information $\boldsymbol{y}$ is thus statistically dependent on the unknown network cardinality N, so that N is statistically identifiable through $\boldsymbol{y}$.

As for the third design level, given our lack on a-priori knowledge on N, we make the classical choice of letting $\widehat{N}$ be the ML estimator of N given $\boldsymbol{y}$.

From these considerations arise the following three questions:

Q1) what is the functional structure of $\widehat{N}$?
Q2) What is the $\boldsymbol{\theta}$ that minimizes the MSE of $\widehat{N}$?
Q3) Does $\widehat{N}$ have some optimality property?

IV. Designing $\boldsymbol{\theta}$

Before answering Q1 we proceed to answer Q2. Our approach to the design of $\boldsymbol{\theta}$ in (4) is then to consider the so-called C-R inequality [29, Eq. 4.1.61], i.e., the fact that the smallest variance that can be achieved by any estimator $\widehat{N}(\boldsymbol{y})$ of N given $\boldsymbol{y}$ is bounded below. Specifically, under mild assumptions holding in our framework, it holds that

$$\operatorname{var}\left(\widehat{N}(\boldsymbol{y})\right) \geq \frac{\left(1 + \dfrac{\partial \mathbb{E}\left[\widehat{N}(\boldsymbol{y}) - N\right]}{\partial N}\right)^2}{\mathcal{I}(N; \boldsymbol{\theta})} \tag{5}$$

where $\mathcal{I}(N; \boldsymbol{\theta})$ is the Fisher Information (FI) [29, Def. 4.1.4] about N given $\boldsymbol{y}$, i.e.,

$$\mathcal{I}(N; \boldsymbol{\theta}) \quad := \mathbb{E}\left[\left(\frac{\partial \ln \mathbb{P}[\boldsymbol{y}\,;\, N, \boldsymbol{\theta}]}{\partial N}\right)^2\right]. \tag{6}$$

Neglecting the bias term, (5) implies immediately that a small FI $\mathcal{I}$ induces estimators with high variance.

Our choice is then to consider the bias term negligible, select that $\boldsymbol{\theta}$ that minimizes the worst C-R bound over all the possible N's, and thus to solve

$$\boldsymbol{\theta}^* := \arg \max_{\boldsymbol{\theta} \in (0,1)^M} \min_{N \in \{1, \ldots, N_{\max}\}} \mathcal{I}(N; \boldsymbol{\theta})\,. \tag{7}$$

Instrumental to (7), we notice the following lemma:

Lemma 1 For any fixed N,

$$\boldsymbol{\theta}^*(N) := \arg \max_{\boldsymbol{\theta} \in (0,1)^M} \mathcal{I}(N; \boldsymbol{\theta}) = \left[\alpha^{1/N}, \ldots, \alpha^{1/N}\right] \tag{8}$$

where

$$2 - \frac{2 - \ln \alpha}{\alpha} = 0 \quad \Rightarrow \quad \alpha \approx 0.2031878699 \ldots \tag{9}$$

Proof (of Lemma 1) Since the y_m's in (3) are independent,

$$\begin{aligned} \mathcal{I}(N; \boldsymbol{\theta}) &= \sum_{m=1}^{M} \mathbb{E}\left[\left(\frac{\partial \ln \mathbb{P}[y_m\,;\, N, \theta_m]}{\partial N}\right)^2\right], \\ &= \sum_{m=1}^{M} \frac{\theta_m^N (\ln \theta_m)^2}{1 - \theta_m^N}. \end{aligned} \tag{10}$$

Thus to maximize (10) it is sufficient to maximize the single term

$$\frac{\theta^N (\ln \theta)^2}{1 - \theta^N} =: i(\theta, N),$$

and this already implies that the vector $\boldsymbol{\theta}^*(N)$ must have all entries identical. Defining

$$\omega(\theta, N) := 2 - \frac{2 + \ln \theta^N}{\theta^N} \tag{11}$$

it follows that

$$\frac{\partial i(\theta, N)}{\partial \theta} = \frac{\omega(\theta, N) \ln \theta}{\theta(1 - \theta^{-N})^2}, \tag{12}$$

i.e., $i(\theta, N)$ is maximized for $\theta^N = \alpha$ with α satisfying condition (9). $\diamondsuit$

Given that $\alpha^{1/N}$ is decreasing with N, it then follows immediately from Lemma 1 that (7) is attained by

$$\boldsymbol{\theta}^* = \left[\alpha^{1/N_{\max}}, \ldots, \alpha^{1/N_{\max}}\right]. \tag{13}$$

V. The ML estimator

Given (13), in what follows we assume $\theta_m = \theta$, for $m = 1, \ldots, M$, and analyze the estimation strategy for a generic $\theta \in (0, 1)$. Thus (2) specializes to

$$y_{i,m} = \begin{cases} 1 & \text{with probability } 1-\theta \\ 0 & \text{with probability } \theta \end{cases} \qquad m = 1, \ldots, M, \tag{14}$$

while the joint distribution of $\boldsymbol{y}$ in (4) simplifies to

$$\mathbb{P}[\boldsymbol{y}\,;\, N] = \prod_{\{m \,:\, y_m = 1\}} (1 - \theta^N) \prod_{\{m \,:\, y_m = 0\}} \theta^N\,. \tag{15}$$

It is a classic result showing that the sample average

$$\overline{y} = \overline{y}(\boldsymbol{y}) := \frac{\sum_{m=1}^{M} y_m}{M}, \tag{16}$$

is a minimal complete sufficient statistic for N. We may in fact write (4) in terms of $\overline{y}$ as $\mathbb{P}[\boldsymbol{y}\,;\,N] = (1-\theta^N)^{M\overline{y}} \cdot \theta^{NM(1-\overline{y})}$, so that, conditionally on the sample average, the probability of observing a given $\boldsymbol{y}$ is independent of θ^N (indeed one can regard $\overline{y}$ as the main output of the bitwise aggregation scheme (3)).

Starting then from the score of N

$$\ell(\overline{y}; N) := \frac{\partial \ln \mathbb{P}[\overline{y}\,;\,N]}{\partial N} = \left(1 - \frac{\overline{y}}{1-\theta^N}\right) M \ln\theta\,, \quad (17)$$

the ML estimator follows as

$$\begin{aligned} \widehat{N}(\overline{y}) &:= \arg\max_{\overline{N}\in[1,N_{\max}]} \mathbb{P}\left[\overline{y}\,;\,\overline{N}\right] \\ &= \begin{cases} 1 & \text{if } \overline{y} \leq 1-\theta \\ \log_\theta(1-\overline{y}) & \text{if } 1-\theta < \overline{y} < 1-\theta^{N_{\max}} \\ N_{\max} & \text{otherwise.} \end{cases} \end{aligned} \quad (18)$$

We notice that in the derivation of the ML estimator we relaxed the integer constraint $N \in \{1,\dots,N_{\max}\}$ by extending the search interval to the real segment $[1, N_{\max}]$. Indeed, while a real size parameter does not match perfectly our information generation scheme, considering the unconstrained estimator (18) allows us to devise closed-form performance characterizations.

We also notice that by extending $\widehat{N}(\cdot)$ to be defined over $[0,1]$ instead of over $\{0, 1/M, 2/M, \dots, 1\}$, and letting

$$\vartheta := 1 - \theta^N, \qquad 1 \leq N \leq N_{\max} \quad (19)$$

be the success rate of each of the generic experiment y_m, it holds $\widehat{N}(\vartheta) = \log_\theta(1-\vartheta) = N$ (indeed the empirical success rate $\overline{y}$ is a consistent estimator of the success rate ϑ). This motivates (18) also as an intuitive estimator of the network size.

VI. Characterization of $\widehat{N}(\overline{y})$

On one hand, the distribution of the ML estimator (18) can be numerically computed for every θ given the fact that $M\overline{y} \sim \mathrm{Bin}\left(M, 1-\theta^N\right)$. On the other hand, there is no dedicated literature reporting closed form characterizations of logarithms of binomial random variables. Since a comprehensive analysis of those variables is beyond the scope of this paper, we resort to a simplified statistical characterization of the ML estimator $\widehat{N}(\overline{y})$ w.r.t. the classical performance indexes

$$\mathbb{E}\left[\frac{\widehat{N}-N}{N}\right], \qquad \mathrm{var}\left(\frac{\widehat{N}-N}{N}\right). \quad (20)$$

Proposition 2 For all $1 \leq N \leq N_{\max}$,

$$\left|\mathbb{E}\left[\widehat{N}\right] - N\right| \leq O\left(\frac{1}{M}\right). \quad (21)$$

Proof (of Prop. 2) Recall that in our assumptions N is an unknown but fixed parameter. Let then $\widetilde{N}(\cdot)$ be a smooth approximation of $\widehat{N}(\cdot)$, i.e., a function $\widetilde{N} : \mathbb{R} \mapsto \mathbb{R}$ satisfying

$$\widetilde{N}(\vartheta) = \widehat{N}(\vartheta)\,, \quad \widetilde{N}(\overline{y}) = \widehat{N}(\overline{y})\,, \quad \overline{y} = 0, \frac{1}{M}, \frac{2}{M}, \dots, 1 \quad (22)$$

for all the potential outcomes $\overline{y}$, and that is endowed for every $\mathcal{Y} \in [0,1]$ with k-th order derivatives

$$\widetilde{N}^{(k)}(\mathcal{Y}) := \frac{\partial \widetilde{N}(\mathcal{Y})}{\partial \mathcal{Y}}, \qquad \forall\, k \geq 1. \quad (23)$$

Notice that such $\widetilde{N}(\cdot)$ can be chosen within the ring of polynomials with degree at most $M+1$.

Consider now the Taylor expansion of $\widetilde{N}(\cdot)$ around the success rate ϑ in (19), valid in the whole unitary segment $[0,1]$ by construction since $\widetilde{N}$ is smooth [30, p. 286]. This means that at the points where (22) holds we may rewrite $\widehat{N}(\cdot)$ in terms of the Taylor expansion of $\widetilde{N}(\cdot)$, i.e.,

$$\widehat{N}(\overline{y}) = \widetilde{N}(\vartheta) - (\overline{y}-\vartheta)\widetilde{N}^{(1)}(\vartheta) + \frac{(\overline{y}-\vartheta)^2}{2}\widetilde{N}^{(2)}(\zeta) \quad (24)$$

where $\zeta = \zeta(\overline{y})$ in the remainder is a real number between ϑ and $\overline{y}$.

Noticing that, by construction, $\widetilde{N}(\vartheta) = N$, and taking the expectation on both sides of (24) w.r.t. $\overline{y}$ yields then

$$\mathbb{E}\left[\widehat{N}\right] = N + \frac{c_1}{M}\widetilde{N}^{(1)}(\vartheta) + \frac{c_2}{2M^2}\widetilde{N}^{(2)}(\zeta) \qquad (25)$$

with

$$c_k := M^k \mathbb{E}\left[(\overline{y}-\vartheta)^k\right] \quad (26)$$

and

$$c_1 = 0, \quad c_2 = M\vartheta(1-\vartheta). \quad (27)$$

We thus recover the assertion by considering that the derivatives $\widetilde{N}^{(k)}(\vartheta)$ are continuous in the compact $[0,1]$, and that the coefficients c_k are finite. $\diamondsuit$

To assess the role of the term $O(1/M)$ in (21) and of the derivative of the bias appearing in the C-R bound (5) we plot in Figures 1 and 2 numerical evaluations of the interested quantities computed through an opportune Monte Carlo (MC) scheme.

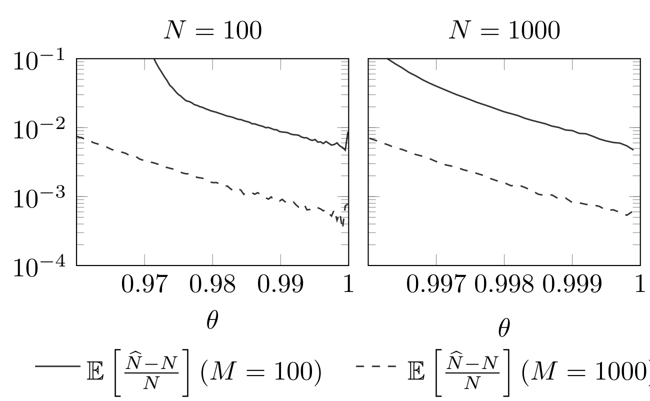

Fig. 1: MC evaluation (10^6 runs for each θ) of the relative error mean of $\widehat{N}$ for $N_{\max} = 2000$ and different values of N, M and θ.

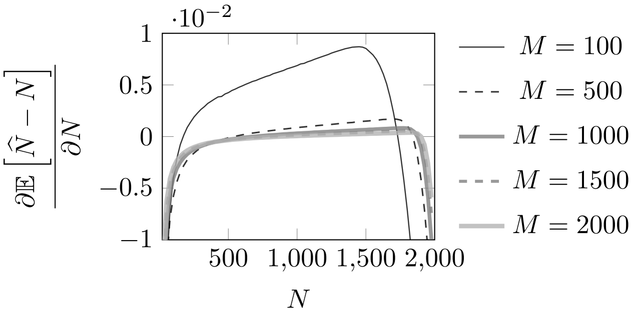

Fig. 2: MC evaluation (10^8 runs for each θ) of the derivative of the bias appearing in the C-R bound (5).

Proposition 3 For all $1 \leq N \leq N_{\max}$,

$$\begin{aligned} \operatorname{var}\left(\widehat{N}\right) & \leq \frac{1-\theta^N}{M\theta^N(\ln\theta)^2} + O\left(\frac{1}{M^2}\right) \\ & = \left(\mathcal{I}\left(N;\theta\right)\right)^{-1} + O\left(\frac{1}{M^2}\right) \end{aligned} \tag{28}$$

Proof (of Prop. 3) Reasonings similar to the proof of Prop. 2 provide a lower bound on $\mathbb{E}\left[\widehat{N}\right]$, an upper bound on $\mathbb{E}\left[\widehat{N}^2\right]$, and thus inequality (28) through the equivalence

$$\operatorname{var}\left(\widehat{N}\right) = \mathbb{E}\left[\widehat{N}^2\right] - \mathbb{E}\left[\widehat{N}\right]^2. \tag{29}$$

$\diamondsuit$

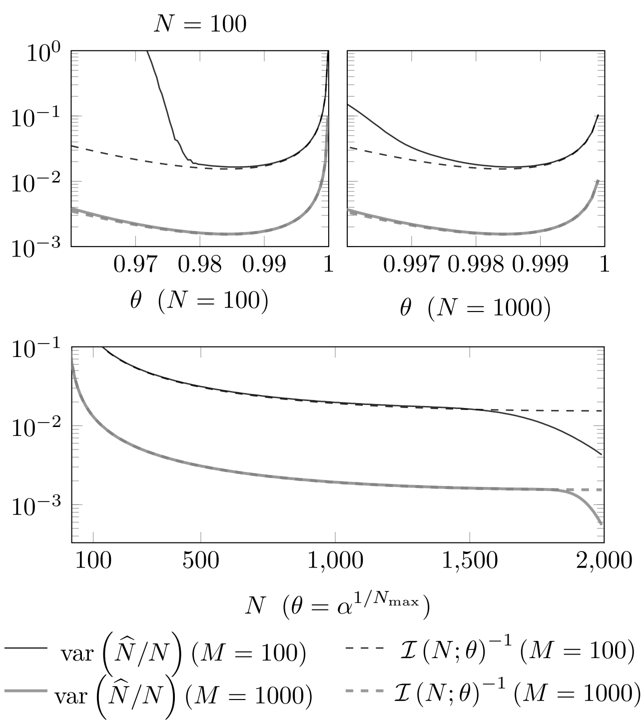

Fig. 3: In solid lines, the MC evaluation (10^6 runs for each θ) of the relative error variance of $\widehat{N}$ for $N_{\max} = 2000$ and different values of N, M and θ. The dashed lines correspond to first term in the right-hand side of (28).

To assess the role of the term $O(1/M^2)$ in (28) we plot in Fig. 3 both numerical evaluations of the performance index $\operatorname{var}\left(\widehat{N}/N\right)$ (computed through an opportune MC scheme) and the inverse of the FI, i.e., $\mathcal{I}\left(N;\theta\right)^{-1}$, in (10). Together, the Figures 2 and 3 show that the actual variance of the novel estimator $\widehat{N}$ practically reaches the C-R bound (5).

Remark 4 We stress that the statistical performance of $\widehat{N}\left(\overline{y}\right)$ reported in Propositions 2 and 3 do not depend on the precise communication topology. Indeed, different topologies may just lead to different convergence times, and not different statistics on the estimate.

VII. Simulations

Here we corroborate the characterization reported in Propositions 2 and 3 through a numerical analysis. Specifically, we compare the estimator $\widehat{N}\left(\overline{y}\right)$ in (18) against other estimation strategies with equivalent convergence times and bounded memory requirements. To this aim we consider synthetic networks with variable sizes and study how the performance of $\widehat{N}$ compares against the max-consensus based estimator considered in [20], [22], [27], [4]. This estimator, here called N_{uni}, can be implemented on top of the synchronous framework of Sec. III through the following specifics (c.f. also the general discussion of Sec. III):

i) every i-th agent initializes a local vector $\boldsymbol{w}_i = \begin{bmatrix} w_{i,1} & \dots & w_{i,K} \end{bmatrix} \in \mathbb{R}^K$ by extracting a K-sample from the uniform distribution $\mathcal{U}\left[0,1\right]$;
ii) agents distributedly aggregate K maxima entry-wise (rather then bit-wise). The consensus vector resulting from this process is denoted by
$$\boldsymbol{w} = \left[w_1, \dots, w_K\right], \quad w_k = \max_{1 \leq i \leq N} \left\{w_{i,k}\right\}. \tag{30}$$
iii) agents locally compute the ML estimator of N given $\boldsymbol{w}$ through
$$N_{\text{uni}} = N_{\text{uni}}(\boldsymbol{w}) := \begin{cases} 1 & \text{if } \chi(\boldsymbol{w}) \leq 1 \\ \chi(\boldsymbol{w}) & \text{if } 1 < \chi(\boldsymbol{w}) < N_{\max} \\ N_{\max} & \text{otherwise} \end{cases} \tag{31}$$
where
$$\chi(\boldsymbol{w}) := \frac{K-1}{-\sum_k \ln w_k}. \tag{32}$$

It is known that if the above estimator relies on r.v.s $w_{i,k}$ with absolutely continuous distributions then it is irrelevant from which exact absolutely continuous distributions one extracts [20, Prop. 7]. E.g., sampling a Gaussian distribution would lead to an alternative estimator with the same statistical performance of N_{uni}. Moreover, assuming $N_{\max} = +\infty$ leads to [20, Eq. (9)]

$$\operatorname{var}\left(N_{\text{uni}}\right) = \frac{N^2}{K-2}. \tag{33}$$

We then notice that the literature dedicated to (31) usually neglects addressing the problem of how to optimally encode each $w_{i,k}$ with a finite number of bits. Nonetheless, to compare $\widehat{N}$ and N_{uni} in terms of estimation performance vs. memory usage we should address this issue. Since at the best of our knowledge there is currently no dedicated literature on this problem, we consider the most simple (and most unfair to $\widehat{N}$) comparison approach, namely we evaluate the performance of N_{uni} without considering any quantization effects.

Specifically, to compare the statistical performance of $\widehat{N}$ against N_{uni} we:

1) assume that $\widehat{N}$ uses M bits;
2) consider several versions of N_{uni}, denoted with $N_{\text{uni}}^{(b)}$ for $b = 2, 3, \dots$ and with b denoting how many bits one would use to encode a single w_k in (30). This means that the generic $N_{\text{uni}}^{(b)}$ uses $K = \text{ceil}\,(M/b)$ different w_k's – but at the same time we consider these w_k's as non-quantized. In other words, we let $N_{\text{uni}}^{(b)}$ operate on more scalars as b decreases but then we completely discard the negative effects of quantization and let $N_{\text{uni}}^{(b)}$ exploit absolutely continuous r.v.s..

We thus computed numerically scheme the performance of $\widehat{N}$ and of the various $N_{\text{uni}}^{(b)}$, and then compared them graphically in Fig. 4. By construction, both estimators converge at the same time and ideally require the same communication resources; the metric used to compare the two strategies is the variance of the relative estimation error.

The figure highlights an interesting numerical result: for any b, $\widehat{N}$ has smaller error variance than $N_{\text{uni}}^{(b)}$ when N is large, while it performs worse when N is small. This suggests that there may be a size $\overline{N}$, possibly function of $N_{\max}$, M and b, for which if $N > \overline{N}$ then using $\widehat{N}$ leads to smaller error variances, while if $N < \overline{N}$ then it is better to use $N_{\text{uni}}^{(b)}$.

This intuition is motivated by the following argument: selecting $\theta = \theta^*$ as in (13), neglecting the term $O(1/M^2)$ in (28) (cf. Fig. 3), and equating the approximated $\text{var}\left(\widehat{N}\right)$ to $\text{var}\left(N_{uni}^{(b)}\right)$ in (33) (with $K = \text{ceil}\,(M/b) \approx M/b$) leads to an identity of the form

$$\frac{\alpha^{-(N/N_{\max})} - 1}{(N/N_{\max})^2} = (\ln \alpha)^2 \, \frac{Mb}{M - 2b} \qquad (34)$$

where the left-hand side of the equation is strictly decreasing in N, while the right-hand side is constant. The rule-of-thumb (34) would then confirm that for each $N_{\max}$, M and b there exists a value $\overline{N}$ for which if $N < \overline{N}$ then $N_{\text{uni}}^{(b)}$ performs better, while if $N > \overline{N}$ then $\widehat{N}$ does.

Nonetheless, we stress that both in our simulations in Fig. 4 and in the reasoning that led to (34), only $\widehat{N}$ considers the quantized nature of $\boldsymbol{y}$, while the various $N_{\text{uni}}^{(b)}$ do not. We thus expect that actual implementations of $N_{\text{uni}}^{(b)}$ will perform worse then what is shown, i.e., that the variance $\text{var}\left(N_{\text{uni}}^{(b)}\right)$ in (33) represents a lower bound on the attainable performance of actual implementations of $N_{\text{uni}}^{(b)}$.

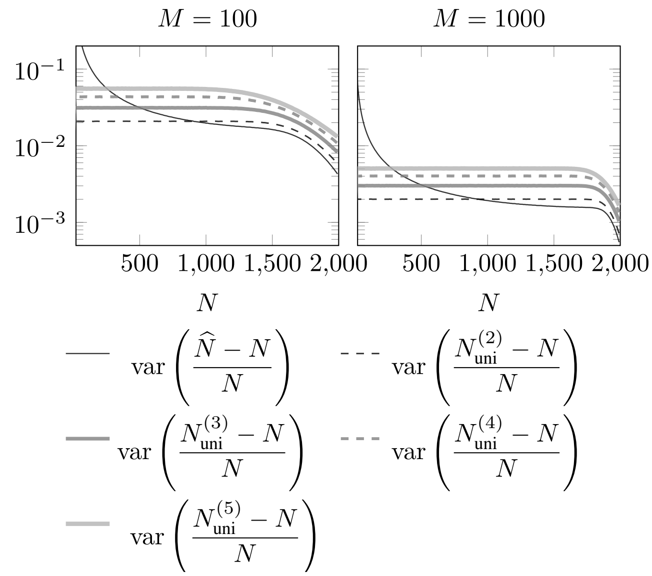

Fig. 4: MC evaluation (10^6 runs for point) of the statistical performance of $\widehat{N}$ and $N_{\text{uni}}^{(b)}$ for $N_{\max} = 2000$ and different values of M, N. $N_{\text{uni}}^{(b)}$ denotes the estimator N_{uni} when the number of real scalars $w_{i,k}$ stored at the i-th node is $K = \text{ceil}\,(M/b)$.

VIII. Conclusions

We aimed at improving the effectiveness of topology inference techniques that aggregate information using max-consensus schemes, starting from the consideration that agents exchange information that is intrinsically quantized. We thus departed from the literature, that usually analyzes schemes based on lexicographic max-consensus operations, and considered strategies that are based on bitwise max-operations.

In particular, we considered frequentist assumptions on the estimand (i.e., we considered the estimand network size N to be a deterministic, unknown but fixed quantity) and then characterized that particular estimation scheme where each bit of the information generated during the initialization of the algorithm is generated independently. We notice that the frequentist assumption is fundamental for our discoveries, since it leads to design the information generation scheme so that the final a-consensus quantity has maximal Fisher information content – a property that we found to hold when each bit is generated as an i.i.d. Bernoulli trial.

Characterizing the resulting estimation scheme in terms of its statistical performance shows then what we consider being the major contribution of this manuscript: bitwise max-operations are meaningful to build practical estimators, since their MSE is *often* favorable against the MSEs of estimators based on lexicographic computations of maxima (given the same number of bits exchanged during the consensus protocol). Nonetheless the bitwise scheme seems to be not *always* favorable, since lexicographic strategies potentially perform better for small network sizes N.

Our major result thus opens more questions than how many it closes: first of all, it calls for a precise analytical characterization of when bitwise-max strategies are better than lexicographic ones. Moreover it calls for exploring also Bayesian approaches, where the estimand N is assumed to be a r.v. with its own prior distribution. Indeed we noticed that having a good initial guess of the estimand N can be exploited to direct the generation of the initial information, and leads to final estimates with better statistical indexes. Bayesian scenarios are also intrinsically connected to practical situations, e.g., when estimation rounds are continuously repeated for network monitoring purposes so that information on the estimand is accumulated from one step to the next one.

REFERENCES

[1] 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.

[2] R. Lucchese, D. Varagnolo, and K. H. Johansson, "Distributed detection of topological changes in communication networks," in *IFAC World Congress*, 2014.

[3] R. Van Renesse, "The importance of aggregation," in *Future Directions in Distributed Computing*, 2003, pp. 87–92.

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

[5] B. Codenotti, P. Gemmell, P. Pudlák, and J. Simon, "On the Amount of Randomness Needed in Distributed Computations," CNR Italy, Tech. Rep., 1997.

[6] 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.

[7] 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.

[8] J. Lemiesz, M. Kardas, and M. Kuty\lowski, "On Distributed Cardinality Estimation: Random Arcs Recycled," in *Proceedings of the Twelfth Workshop on Aanlytic Alogrithmics and Combinatorics*. SIAM, 2015, pp. 129–137.

[9] E. Mane, E. Mopuru, K. Mehra, and J. Srivistava, "Network Size Estimation In A Peer-to-Peer Network," University of Minnesota, Department of Computer Science, Tech. Rep., 2005.

[10] D. Kostoulas, D. Psaltoulis, I. Gupta, K. Birman, and A. Demers, "Decentralized schemes for size estimation in large and dynamic groups," in *Proceedings - Fourth IEEE International Symposium on Network Computing and Applications, NCA 2005*, vol. 2005, 2005, pp. 41–48.

[11] L. Massouliè, 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.

[12] A. J. Ganesh, A.-M. Kermarrec, E. L. Merrer, and L. Massoulié, "Peer counting and sampling in overlay networks based on random walks," *Distrib. Comput.*, vol. 20, pp. —-, 2007.

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

[14] F. Garin and Y. Yuan, "Distributed privacy-preserving network size computation: A system-identification based method," in *52nd IEEE Conference on Decision and Control*. IEEE, Dec. 2013, pp. 5438–5443.

[15] I. Shames, T. Charalambous, C. N. Hadjicostis, and M. Johansson, "Distributed Network Size Estimation and Average Degree Estimation and Control in Networks Isomorphic to Directed Graphs," in *Allerton Conference on Communication Control and Computing*, 2012.

[16] F. Morbidi and A. Y. Kibangou, "A Distributed Solution to the Network Reconstruction Problem," *Systems and Control Letters*, vol. 70, pp. 85–91, 2014.

[17] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, 2003.

[18] M. Jelasity and A. Montresor, "Epidemic-style proactive aggregation in large overlay networks," *24th International Conference on Distributed Computing Systems, 2004. Proceedings.*, 2004.

[19] P. Jesus, C. Baquero, and P. S. Almeida, "Flow updating: Fault-tolerant aggregation for dynamic networks," *Journal of Parallel and Distributed Computing*, 2015.

[20] D. Varagnolo, G. Pillonetto, and L. Schenato, "Distributed cardinality estimation in anonymous networks," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 645–659, 2014.

[21] D. Varagnolo, L. Schenato, and G. Pillonetto, "A variation of the Newton-Pepys problem and its connections to size-estimation problems," *Statistics & Probability Letters*, vol. 83, no. 5, pp. 1472–1478, 2013.

[22] C. Baquero, P. S. 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, Apr. 2012.

[23] J. Lumbroso, "An optimal cardinality estimation algorithm based on order statistics and its full analysis," in *International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms*, 2010.

[24] 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.

[25] R. Lucchese and D. Varagnolo, "Networks cardinality estimation using order statistics," in *American Control Conference*, 2015.

[26] H. Terelius, D. Varagnolo, and K. H. Johansson, "Distributed size estimation of dynamic anonymous networks." in *IEEE Conference on Decision and Control*, 2012.

[27] M. Albano, N. Pereira, and E. Tovar, "How many are you (an approach for the smart dust world)?" in *2013 IEEE 1st International Conference on Cyber-Physical Systems, Networks, and Applications (CPSNA)*. IEEE, Aug. 2013, pp. 101–105.

[28] 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.

[29] M. Basseville and I. V. I. V. Nikiforov, *Detection of Abrupt Changes: theory and application*. Prentice-Hall, Apr. 1993.

[30] G. H. Hardy, *A Course of Pure Mathematics*, 10th ed. Cambridge: Cambridge University Press, 1967.