Computer Networks 89 (2015) 107–122

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Survey Paper

The Capture-Recapture approach for population estimation in computer networks

Nicola Accettura [a,*], Giovanni Neglia [c], Luigi Alfredo Grieco [b]

[a] Berkeley Sensor & Actuator Center, University of California Berkeley, USA
[b] Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Italy
[c] Inria - EPI Maestro, Sophia-Antipolis Méditerranée, France

ARTICLE INFO

Article history:
Received 7 April 2014
Revised 20 July 2015
Accepted 22 July 2015
Available online 30 July 2015

Keywords:
Populations
Computer networks
Capture-Recapture
Maximum-likelihood

ABSTRACT

The estimation of a large population's size by means of sampling procedures is a key issue in many networking scenarios. Their application domains span from RFID systems to peer-to-peer networks; from traffic analysis to wireless sensor networks; from multicast networks to WLANs. The present contribution aims at illustrating and classifying in a coherent framework the main approaches proposed so far in the computer networks literature to deal with such a problem. In particular, starting from the methodologies proposed in ecological studies since the last century, this paper surveys their counterparts in the computer network domain, finding that many lessons can be gained from this insightful investigation. Capture-Recapture techniques are deeply analyzed to allow the reader to exactly understand their pros, cons, and applicability bounds. Finally, some open issues that deserve further investigations and could be relevant to afford estimation problems in next generation Internet are discussed for sake of completeness.

1. Introduction

In many networking problems, the optimization of a given service or system requires accurate estimates of the number of involved key items (whether they represent nodes, users, files, packets, flows, and so forth). For example, the estimation of the number of users sharing the same file in a peer-to-peer network [52] can be used for allowing a fine tuning of protocols' parameters and/or for monitoring purposes; the knowledge of the number of traffic flows handled by a router [74] is useful for enforcing Quality of Service (QoS) differentiation techniques or for improving network reliability; counting the number of RFID tags [68,72] is needed for inventory management.

For the sake of generality, the present contribution will refer to the term *population* to mean a large set of items (or *individuals*), whose cardinality cannot be easily inferred using plain counting procedures, due to wideness and/or variability of the set, thus requiring sophisticated estimation methods. Such terminology has been widely exploited in the computer networks literature by analogy with similar estimation problems in biological environments [65].

As a matter of fact, statistical approaches dealing with the estimation of the population size are based on sampling, i.e., on analyzing a subset of the entire population. The biometric community developed many approaches [63,65], some as old as the nineteenth century [44,51], to face the trade-off between the effectiveness of estimation methods and their computational complexity. *Most of the statistical approaches developed by the biometric community are framed into the Capture-Recapture (CR) methodology* [17,54], *which refers to the recognition of individuals recaptured in more than one sample and to the exploitation of such additional information for*

* Corresponding author. Tel.: +390805963301.
E-mail addresses: nicola.accettura@eecs.berkeley.edu, nicola.accettura@poliba.it (N. Accettura), alfredo.grieco@poliba.it (G. Neglia), giovanni.neglia@inria.fr (L.A. Grieco).

http://dx.doi.org/10.1016/j.comnet.2015.07.013
1389-1286/

deriving estimators more reliable than those based on the only knowledge of the sample size. Indeed, CR estimators could be Maximum likelihood, Bayesian, derived through hypothesis testing, etc. In details, the CR methodology assumes that the individuals caught in a sample can be captured again in following samples. In zoological contexts, this means that all animals captured in the first sample are marked and released in order to recognize them in subsequent catches. The employment of marking operations is the main reason why the *Capture-Recapture* approach is also referred to as *Mark-Recapture* in the literature.

It is worth observing that the *Capture-Recapture* methodology deals with centralized non-anonymous estimation strategies. In fact, after sampling the population, a central controller performs an estimation based on the knowledge of the gathered individuals. Centralized anonymous methods to estimate the population size in computer networks have been employed in [4,9,40]. As a counterpart, the strategy based on Bernoulli trials presented in [69] shows a distributed approach to the population size estimation in anonymous networks. It has to be noted that computability in anonymous networks is a very big issue [11,35] since with anonymous computations it is only possible to count probabilistically, even if the amount of randomness required is very little [20].

In the works introduced above, the information obtained by collected samples pertains only to their size: indeed, the sample size depends on the probability for each item in the population to be captured. Although these techniques are usually not very expensive in term of computation time and memory requirements, the convergence speed and the precision of the estimation process can be very low, and unsuitable for cases when system dynamics are fast. In fact, we point out that the information conveyed in a sample is more than just its size: indeed, knowing the identity of the individuals in each sample would help in improving both the accuracy and the speed of the estimation processes. In fact, tracking the capture history of each caught individual is useful for guessing insightful properties of the population evolution, in terms of arrivals and departures. In order to justify such evidence we show a numerical example. Let us assume that a given population is sampled repeatedly for performing size estimation and that the catching probability for each individual is $p = 1\%$. Considering the "lucky" case of all samples having size equal to 10, one expects that a fair estimate of the population size would be 1000. Actually, when inspecting the identity of the caught individuals in each sample, it is possible to find that either the same 10 individuals are caught in each sample or each individual is caught in a single sample. In both cases, a correct size estimation would be different from that provided by accounting only for the sample size. This simple example intuitively shows how *the information related to the identity of the elements caught in a sample can be exploited to provide a more reliable estimate and a faster convergence to the actual value of the population size*. The price to pay for this performance improvement is an increase of computational and storage requirements for handling the sampling history, which, in any case, remains often affordable by modern computing platforms.

As matter of fact, most of the identity-based estimators used in computer networks contexts are exactly framed in the *Capture-Recapture* approach, which is also the main focus of the present survey. Although Jesus et al. [37] collected and described some works both related to estimation problems in computer networks and dealing with CR sampling techniques, the aim of their survey was to analyze a wider spectrum of data aggregation techniques, without a specific focus on the *Capture-Recapture* theory as a whole. In this sense, the present contribution aims to shed some light on the *Capture-Recapture* methodology, while surveying its application in estimation problems related to computer networks and introducing those CR solutions easily deployable in more complex scenarios.

To help in understanding at a glance the statistical properties of the estimator surveyed in the following sections, Table 1 lists the network quantities evaluated in such works together with the underneath statistical approach exploited for the related estimation.

In order to gently introduce the *Capture-Recapture* methodology, firstly we note that a given population is modeled as *closed* [17], if its size does not change during the whole sampling process, or as *open* [53] (in the opposite case). The key assumption for a *closed* population is that no element is entering or leaving the population during sampling operations. Of course, it is easier to derive an estimator for a closed population, even though, in many circumstances, the assumption that the population is not varying during the sampling process is unrealistic. Contrariwise, the estimation of the size of an *open* population must take into account also the dynamics of the population, i.e., the arrival/departure rate during sampling stages. At the same time, it is worth to remark that this distinction is not always sharp, because, in some cases, estimators conceived for *closed* populations can be also adapted to dynamic contexts. Following this premise, Section 2 surveys *Capture-Recapture* estimators for *closed* populations, highlighting their properties and applicability in computer networks environments; afterwards, Section 3 introduces the most relevant *Capture-Recapture* methods for estimating the parameters related to *open* populations, focusing especially on the Jolly-Seber model [38,64] that we strongly believe will be the basis of population models for many future computer networks related estimation problems.

Then, Section 4 mentions some relevant related works, dealing with non-CR estimators. Finally, Section 5 draws conclusions, describing lessons learned, and explaining what in our humble opinion should still be done in the context of this research topic.

2. Capture-Recapture estimators for closed populations

All in all, the strategies framed into the CR methodology for the estimation of closed populations are mainly grouped in two categories: those dealing with only two samples, and those dealing with more than two samples. The first category includes the very basic CR strategies mainly used for a fast estimation based on a limited sampling capability. The second category includes a wide gamut of estimation techniques exploitable when sampling is not an issue, thus providing more accurate estimations. A first glance perspective on this categorization is sketched in the tree diagram

Table 1
Population estimates in computer networks and related statistical approach.

Quantities to be estimated	Section	Statistical approach		
		MLE	*Bayesian*	*Other*
Source data-rate of a flow handled by a router [14]	2.4.1	•		
Peer-to-peer network size [45]	2.1	•		
WSN size and scale of event [50]	2.1	•		
Number of peers in overlay networks [31]	2.2.3	•		
File's replicas in peer-to-peer networks [12,52]	2.3		•	
Detection of missing RFID tags [55]	2.3	•		
Traffic flows handled by a router [14]	2.4.1			•
RFID tags over unreliable radio channels [68]	2.4.2			•
Number of RFID tags [72]	3.1	•		
Membership size in multicast networks [4]	4		•	
Competing terminals in IEEE 802.11 networks [9]	4		•	
Network size estimation in anonymous networks [69]	4	•		
Tag population size in FSA protocols [40]	4	•		
Concurrent active flows in high-speed networks [74]	4			•

of Fig. 1. A classification of CR strategies according to the capture probability model is also represented.[1]

In details, each CR strategy was conceived assuming an underneath capture probability model [17]:

- The M_0 model assumes all elements having the same constant probability of being caught during the sampling process; the assumptions made for this model are too restrictive and unrealistic in most cases [48,53].
- The M_t model, instead, allows a time-varying catching probability but imposes that all elements have the same probability to be caught during each single sampling step.
- The model M_b is intended to describe the behavioral responses to capture, i.e., marked individuals have a different probability to be captured with respect to unmarked ones.
- The M_h model, finally, allows heterogeneity, i.e., elements have not the same probability of being caught.

Models based on the combinations of the above variations are also available, i.e., M_{tb}, M_{th}, M_{bh}, and M_{tbh}. Their comprehensive description and application to biological environments is given by Chao [17]. However, the behavioral responses to capture is in general out of the scope of estimation problems in computer networks, hence the estimators based on M_b, M_{tb}, M_{bh} and M_{tbh} models are not considered in the present contribution.

2.1. The Lincoln–Petersen Index

The corner stone of the CR methods was put by the *Lincoln–Petersen Index* [44], which is suitable for estimating the size of closed populations (the size is not changing between the two sampling phases) by means of two sampling rounds. All elements are captured with the same probability within each single round, although the catching probability can vary from the first to the second sample. In this sense, the underlying mathematical model is a M_t *Capture-Recapture* one.

[1] Each block in the diagram includes a reference to the subsection where the related estimator is described and discussed in the present contribution.

The formulation for the *Lincoln–Petersen Index* is:

$$\hat{N} = \frac{Mn}{m}, \quad m \neq 0 \tag{1}$$

where $\hat{N}$ is the estimator of the population size, M the number of elements belonging to the first sample, n the amount of elements in the second sample, and m the amount of elements found in both samples. The *Lincoln–Petersen Index* is used when the catching probabilities are not known a priori and it has an intuitive interpretation: the proportion between the size of the first sample and the whole population should be reflected by the proportion between the marked individuals found in the second sample and the size of the second sample itself. At the same time, this index is the MLE estimator based on the hypergeometric distribution.

Actually, it is possible to fix a priori the sample sizes M and n, while assuming that the catching probability remains the same for all individuals during each single sampling round. The accuracy of the *Lincoln–Petersen Index* based estimation depends on the size of the two samples M and n [58]. Let $1-\alpha$ be the desired level of confidence that the absolute difference between the true population size N and its estimate $\hat{N}$ will be smaller than a chosen level of accuracy A:

$$1-\alpha \leq \Pr\left[-A < \frac{\hat{N}-N}{N} < A\right]. \tag{2}$$

The hypergeometric distribution can be approximated by a normal one for a population size $N > 100$. With this outcome, an equivalent form for Eq. (2) is given [58] by:

$$1-\alpha = \phi\left(\frac{A}{1-A}\sqrt{\frac{nM(N-1)}{(N-n)(N-M)}}\right) - \phi\left(\frac{-A}{1+A}\sqrt{\frac{nM(N-1)}{(N-n)(N-M)}}\right) \tag{3}$$

where $\phi(z)$ is the cumulative unit normal distribution. If a rough estimate of the population size is available, it can be substituted to N in Eq. (3) to determine the samples' sizes M and n with a degree of freedom.

Although the *Lincoln–Petersen Index* is a MLE estimator, it is undefined when $m = 0$, i.e., when there is no element caught in both the samples. This problem was addressed

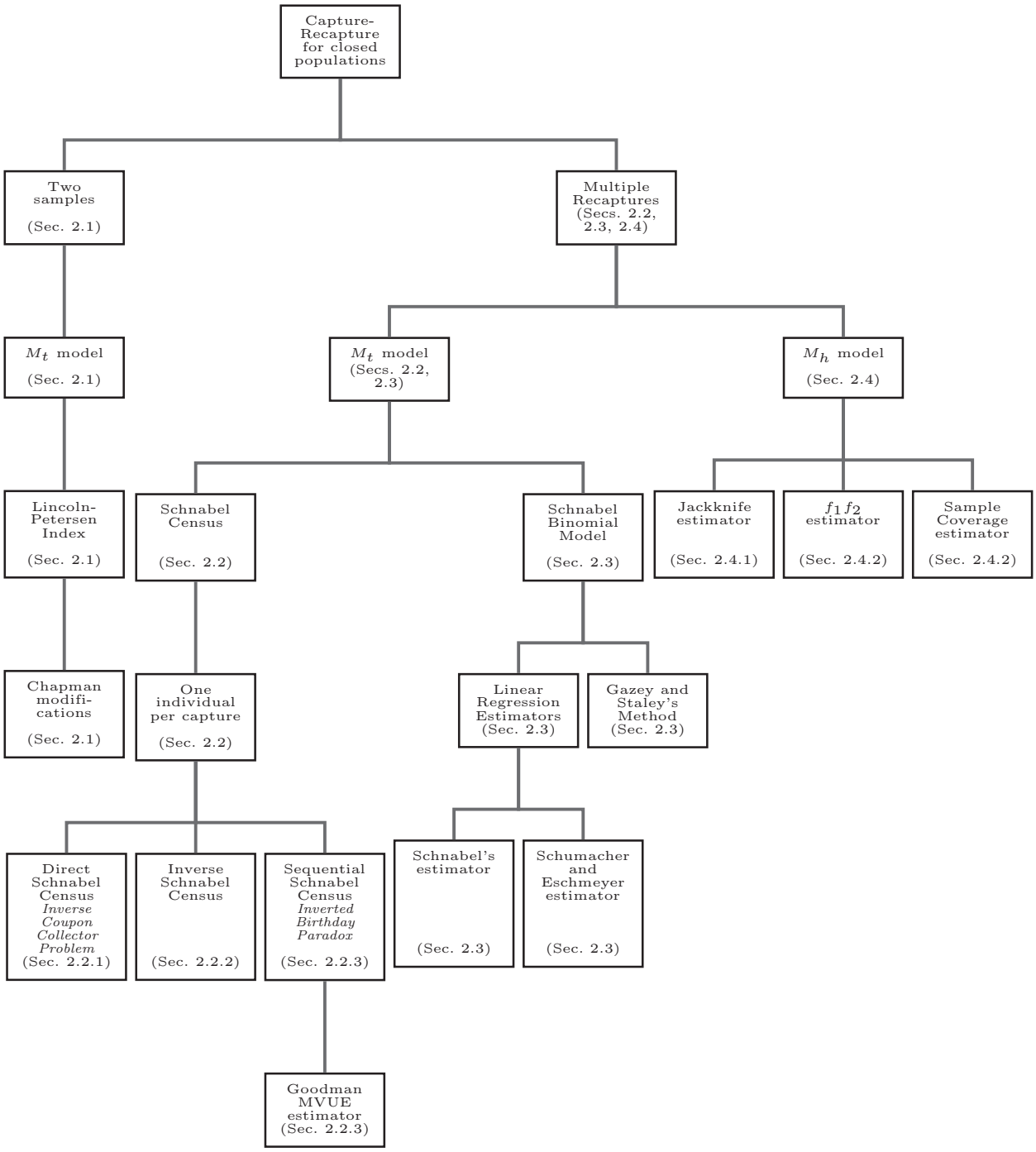

Fig. 1. Capture-Recapture estimators for closed populations.

by Chapman [18] by finding a modification of the *Lincoln–Petersen Index* able to estimate the population size also when $m = 0$:

$$\hat{N}_m = \frac{(M+1)(n+1)}{m+1}. \tag{4}$$

Under proper conditions, i.e., $nM/N > \log N$ and $n + M + 1 < N/2$, such estimator has a positive bias lower than 1, hence being essentially unbiased.

Example 1 (Estimation of the size of a peer-to-peer network.)**.** Eq. (4) has been employed for estimating the size of a peer-to-peer network [45]. In detail, a given source node starts two consecutive random walks by exchanging messages with other nodes on the underlying connection graph of a peer-to-peer network. The network topology is also supposed to remain unchanged between the two random walk explorations. Each random walk is supposed to collect a random subset of the nodes in the network, being therefore a

sample of the set of nodes in the network. For each random walk, the source node sends a gossip message to a subset of its neighbors imposing a value *TTL* for the time-to-live of the message itself. Other nodes, when receiving a gossip message, decrease the time-to-live, add their ID to the message, and decide to return a gossip-reply message (containing the list of polled nodes) to the source node if one of the following condition is fulfilled: (i) the number of neighbors is 1; (ii) the time-to-live is 0 and the gossip-message was not previously received. Otherwise, they forward the gossip message to a subset of their neighbors, except the sender one. As a matter of fact, the TTL value plays a key role for assuring equal sampling chance to all nodes. For this motivation, a simulation campaign has been performed on network topologies constructed according to both Waxman [71] and Barabasi and Albert [6] models, showing that if $TTL \geq 7$ in networks with 5000 nodes, the CR estimation converges to the actual network size. Finally, the authors show that the CR approach performs better with respect to other estimation techniques (e.g., the *Inverted Birthday Paradox* [8] surveyed later in Section 2.2.3) in terms of accuracy.

Example 2 (Estimation of WSN size and scale of an event.)**.** A modified version of the *Lincoln–Petersen Index* has also been employed in [50]. The authors consider a Wireless Sensor Network (WSN) and estimate the number of working sensors and the scale of spatial events, e.g., the diffusion of a fire or of a polluting agent (estimated as the number of nodes sensing the event).

To this end, first an algorithm for disseminating M seed messages almost uniformly into the network is used, according to the scheme presented in [7]. As a result, the proposed network protocol performs a *capture*, and the M nodes receiving a seed message will be considered as marked (they are called "tagged" in [50]).

Then, to estimate the size of the WSN network, an Inverse Sampling method is used: the network is queried until a predetermined amount m of marked nodes have been encountered during the process (i.e., a *recapture* is performed); the amount n of nodes queried during this process is also taken into account for executing the estimation. In details: the base station sends a message containing two fields, the first one initialized to m and the second to 0, to a randomly chosen target; when receiving the message, the target node increases the second field by 1 and, if it was marked during the *capture* phase (i.e., it holds a seed), decreases the first one by 1; the target node chooses another target node and forwards the message; this process continues until the first field in the message equals 0; finally, the last target node queried sends the value n contained in the second field, which was incremented hop-by-hop, to the base station. Assuming to know the number M of marked nodes, the authors estimate the proportion $P = M/N$ with respect to the unknown number N of nodes in the network by means of the unbiased estimator $\hat{P} = (m-1)/(n-1)$, whose coefficient of variation has an upper bound $(m-2)^{-1/2}$. Finally, an estimate of the population size N is found:

$$\hat{N} = \frac{M}{\hat{P}} = \frac{M(n-1)}{(m-1)}. \tag{5}$$

It has to be noted that the previous formulation is similar to that of the *Lincoln–Petersen Index*, but not the same, since the resulting estimator is based on the Inverse Sampling technique adopted.

To infer the scale of physically-connected events, the estimate $\hat{N}$ in (5) is used. Then, through the same dissemination algorithm, n nodes are uniformly queried in the network. Among them, only those that sensed the event provide a feedback to the base station. As a result, the base station will be aware of m nodes sensing the event out of n ones queried. The estimate of the number of nodes sensing the event is given by the following formulation:

$$\hat{M} = \frac{\hat{N}m}{n}, \tag{6}$$

which is still a MLE estimator based on the hypergeometric distribution.

2.2. The Multiple-Recapture Schnabel Census

The availability of multiple samples can be exploited for performing a more accurate size estimation of closed populations. To this aim, the *Schnabel Census* likelihood function can be employed for describing Multiple-Recapture experiments [24].[2] Indeed, a population with size N is assumed to be randomly sampled k times. Furthermore, the probability p_i that any individual is captured in the ith sampling period, with $i = 1, \ldots, k$, is not known a priori (the set of k catching probabilities is synthetically referred with $\{p_i\}$). In this way, the *Multiple-Recapture Schnabel Census* is an M_t *Capture-Recapture* model.

We denote by n_i the size of the ith sample. If each individual can be caught at time i with probability p_i independently from all the others, then n_i is a binomial random variable with parameters N and p_i. Another possibility is that sample sizes are fixed a priori and each capture phase ends when the target size is reached. We refer to the two situations respectively as the *random sample sizes case* and the *fixed sample sizes case*. In [24] they are simply called models A and B.

In general, each individual is featured by a specific capture history, which details in which sample the specific individual appears. Given the set of sample indices $S_k = \{1, 2, \ldots, k\}$, a capture history is then an element in the power set $\mathcal{P}(S_k)$. In this sense, the set of possible histories related to captured individuals is equal to $\mathcal{P}_c(S_k) = \mathcal{P}(S_k) \setminus \{\varnothing\}$, whose cardinality is equal to $2^k - 1$. Let us define u_ω, with $\omega \in \mathcal{P}_c(S_k)$, the number of individuals sharing the same capture history ω. Hence, the total number, r, of captured individuals is:

$$r = \sum_{\omega \in \mathcal{P}_c(S_k)} u_\omega. \tag{7}$$

Given the population size N, the sample sizes $\{n_i\}$, with $i \in S_k$, and the elements belonging to the set $\{u_\omega\}$, with $\omega \in \mathcal{P}_c(S_k)$, it is easy to understand that the following identity always holds:

$$n_i = \sum_{\substack{\omega \in \mathcal{P}_c(S_k) \\ i \in \omega}} u_\omega \quad for \quad 1, \ldots, k. \tag{8}$$

In the *random sample sizes case* introduced above, the probability to jointly obtain the sample sizes $\{n_i\}$ and the

[2] Unless otherwise specified, the formulas in this section are derived in [24].

arrangement $\{u_\omega\}$ for the capture histories is expressed by the following formula:

$$\Pr(\{u_\omega\},\{n_i\}) = \frac{N!}{(N-r)!\prod\limits_{\omega\in\mathcal{P}_c(S_k)} u_\omega!}\prod_{i=1}^{k}\left[p_i^{n_i}(1-p_i)^{N-n_i}\right], \tag{9}$$

that can be used for a joint MLE estimate of N and of the capture probabilities $\{p_i\}$. Usually the capture probabilities are considered as nuisance parameters. Interestingly, the MLE estimator of N is the same inferred in the *fixed sample sizes case* and is asymptotically featured by the same variance.[3]

The probability to observe the capture histories $\{u_\omega\}$ in the *fixed sample sizes case* is equal to the probability to observe the same histories in the *random sample sizes case* conditioned on the event that the sample sizes are equal to the fixed sizes chosen. By using Eq. (9) and $\Pr(\{n_i\}) = \prod_{i=1}^{k}\left[\binom{N}{n_i}p_i^{n_i}(1-p_i)^{N-n_i}\right]$, we obtain:

$$\begin{aligned}\Pr(\{u_\omega\}|\{n_i\}) &= \frac{\Pr(\{u_\omega\},\{n_i\})}{\Pr(\{n_i\})}\\ &= \frac{N!}{(N-r)!\prod\limits_{\omega\in\mathcal{P}_c(S_k)} u_\omega!}\frac{1}{\prod\limits_{i=1}^{k}\binom{N}{n_i}}.\end{aligned} \tag{10}$$

The MLE estimator for the population size N, inferred from Eq. (10) by means of the *ratio* method [23], is the value $\hat{N}$ solving the following equation:

$$h(N,r) = \prod_{i=1}^{k}(N-n_i) - N^{k-1}(N-r) = 0. \tag{11}$$

It is evident then the sample sizes $\{n_i\}$ and the total number of captured individuals r are sufficient statistics to estimate N. Distinguished marks are then not needed for each sample. At each stage, it is needed to mark only the unmarked individuals in the sample. Eq. (11) has only one solution greater than r maximizing the likelihood, except when r takes one of its extreme values, i.e., $r = \sum_{i=1}^{k} n_i$ or $r = \max_{i=1,\dots,k}(n_i)$, corresponding respectively to the two cases when each individual has been captured only once and when all the captured individuals appear in one sample. Let ρ denote the expected value of r:

$$\rho = \mathrm{E}[r] = N - \frac{1}{N^{k-1}}\prod_{i=1}^{k}(N-n_i), \tag{12}$$

the asymptotic behavior of the estimator $\hat{N}$ was described by Darroch [24] in terms of bias and variance in the limit process $N\to\infty$ and each $n_i\to\infty$, in such a way that the n_i/N ratios remain constant:

$$\begin{aligned}\mathrm{b}(\hat{N}) &\sim \frac{\left[\frac{k-1}{N} - \sum\limits_{i=1}^{k}\left(\frac{1}{N-n_i}\right)\right]^2 + \left[\frac{k-1}{N^2} - \sum\limits_{i=1}^{k}\left(\frac{1}{N-n_i}\right)^2\right]}{2\left[\frac{1}{N-\rho} + \frac{k-1}{N} - \sum\limits_{i=1}^{k}\left(\frac{1}{N-n_i}\right)\right]^2}\\ &= O(1)\end{aligned} \tag{13}$$

$$\mathrm{Var}[\hat{N}] \sim \frac{1}{\left[\frac{1}{N-\rho} + \frac{k-1}{N} - \sum_{i=1}^{k}\left(\frac{1}{N-n_i}\right)\right]} = O(N) \tag{14}$$

It is also worth to consider the confidence interval for the estimate of N. Noting that $\hat{N}$ is an increasing function of r, and assuming that r is approximately normally distributed around its expected value ρ, the confidence interval for the estimate $\hat{N}$ may be derived from the confidence interval of ρ:

$$1-\alpha \le \Pr\left[-A \le \frac{\rho - r}{\sqrt{\mathrm{Var}[r]}} \le A\right], \tag{15}$$

according to the procedure detailed in [24] ($\mathrm{Var}[r]$ is the variance of r and its expression is omitted for the sake of simplicity).

A case particularly important for the applications is when a single individual is captured at each stage, corresponding to a continuous sampling case [21,24]. This model is referred to as the *Direct Schnabel Census*. Until now we have assumed that the number of capture phases k is decided a priori. With sample sizes fixed to 1, it is also possible to stop sampling when a given number r of unique individuals is caught or when a given number $l = k - r$ of recaptures is reached. The corresponding models are respectively called the *Inverse Schnabel Census* and the *Sequential Schnabel Census*. Because of their practical importance, we devote the following sections to these variants of the Schnabel Census, highlighting their relation to other mathematical problems (i.e., the *Inverse Coupon Collector* and the *Inverted Birthday Paradox*) and describing some applications in computer networks.

2.2.1. Direct Schnabel Census

For the *Direct Schnabel Census*, all the formulas derived above hold with $n_i = 1$ for $i = 1, 2, \dots k$. Then $\prod_{i=1}^{k}\binom{N}{n_i} = \prod_{i=1}^{k}\binom{N}{1} = \prod_{i=1}^{k} N = N^k$. At the same time, it is easy to conclude that $\prod_{\omega\in\mathcal{P}_c(S_k)} u_\omega! = 1$, since there may not be two individuals with the same history and then $u_\omega \in \{0, 1\}$. Then, the likelihood function in Eq. (10) simplifies as follows:

$$\Pr(\{u_\omega\}) = \frac{N!}{(N-r)!N^k}. \tag{16}$$

The MLE estimate $\hat{N}$ of the population size can be obtained from Eq. (11), which becomes:

$$h(N,r) = (N-1)^k - N^{k-1}(N-r) = 0, \tag{17}$$

and it can be approximated as follows by considering $(1-\frac{1}{N})^N \approx e^{-1}$:

$$e^{-\frac{k}{N}} = 1 - \frac{r}{N}. \tag{18}$$

The following formulas for bias and variance can be obtained under the same approximation in the limit process $N\to\infty$ and $k\to\infty$, so that k/N is constant:

$$\mathrm{b}(\hat{N}) \sim \frac{k^2}{2N^2\left(e^{\frac{k}{N}} - 1 - \frac{k}{N}\right)^2} = O(1) \tag{19}$$

$$\mathrm{Var}[\hat{N}] \sim \frac{N}{\left(e^{\frac{k}{N}} - 1 - \frac{k}{N}\right)} = O(N). \tag{20}$$

[3] The mathematical properties of Eq. (9) are used for estimation problems dealing also with the dynamics of the population, i.e., elements leaving or entering in the population in-between the sampling steps [25].

The confidence intervals can be derived according to the same formulation of Eq. (15).

Eq. (16) , deriving from Eq. (10), expresses the probability of a specific capture history for each individual, even if, when each sample has unitary size, the capture histories only play a role through the total number of individuals caught r. It is useful to derive the probability of the set of all the capture histories where r individuals are caught. To this purpose, we can first calculate the probability of the set of all the capture histories that show the same number of individuals caught once, twice,..., k times. Let f_x denote the number of individuals caught x times, i.e., $\sum_i u_{\{i\}} = f_1$, $\sum_{i<j} u_{\{i,j\}} = f_2$, and so forth. The elements in the set $\{f_x\}$, with $x = 1, \ldots, k$, must satisfy the following relationships: $r = \sum_{x=1}^{k} f_x$ and $k = \sum_{x=1}^{k} x f_x$. The probability of not catching $n - r$ individuals and of obtaining the set $\{f_x\}$ can be calculated summing the probability of Eq. (16) over all values of $\{u_\omega\}$:

$$\frac{N!}{(N-r)!N^k} \frac{k!}{(1!)^{f_1}(2!)^{f_2} \ldots f_1! f_2! \ldots} \tag{21}$$

Summing (21) over all values of $\{f_x\}$, the probability of catching r individuals with k samples is given by the following expression:

$$\Pr(r \text{ individuals in } k \text{ samples}) = \frac{N!}{(N-r)!N^k} \{k\}r \tag{22}$$

where $\{k\}r$ is a Stirling number of the second kind.[4] For the *Direct Schnabel Census* Eq. (22) leads (obviously) to the same MLE estimator as Eq. (16), but the equation is useful to study the *Inverse* and *Sequential Schnabel Census*. While the use of Eq. (22) to estimate population size had already been proposed by Craig as early as 1953 [21], it was someway rediscovered in 1991 by Dawkins [26] who defined this as the *Inverse Coupon Collector Problem*, where the *Coupon Collector's Problem* is to find the average number of 1-sized samples to be collected in order to see at least one time all the individuals in a population.

2.2.2. Inverse Schnabel Census

The *Inverse Schnabel Census* corresponds to sampling continuously until a fixed amount, r, of individuals is caught. Therefore, the total number of samples needed k is a random variable. The rth individual will be caught in the kth capture if and only if $r-1$ individuals have been caught in the first $k-1$ captures and a new individual is caught in the kth one. Then by independence and using Eq. (22), it follows:

$$\Pr(k \text{ samples to find } r \text{ individuals}) = \frac{N!}{(N-r)!N^k} \{k-1\}r - 1. \tag{23}$$

The MLE estimator is the value of the population size N solving Eq. (17) also in this case, even if there is no solution $\hat{N} \geq r$, when the following condition is true:

$$k > r \sum_{i=1}^{r} \frac{1}{i} \tag{24}$$

[4] A Stirling number of the second kind is defined as: $\{a\}b = \frac{1}{b!} \sum_{i=0}^{b} (-1)^i \binom{b}{i} (b-i)^a$

Bias, and variance, in the limit process $N \to \infty$ and $r \to \infty$, so that r/N is constant, are:

$$\mathrm{b}(\hat{N}) = \frac{N \sum_{k=1}^{r-1} \frac{k}{(N-k)^3}}{\left[\sum_{k=1}^{r-1} \frac{k}{(N-k)^2}\right]^2} = O(1) \tag{25}$$

$$\mathrm{Var}[\hat{N}] = \frac{N}{\sum_{k=1}^{r-1} \frac{k}{(N-k)^2}} = O(N) \tag{26}$$

while confidence intervals are derived according to the same procedure of Eq. (15)), except that now $\hat{N}$ is monotone decreasing function of k.

2.2.3. Sequential Schnabel Census

In *Sequential Schnabel Census* the capture process is carried on until a given number of recaptures $l(=k-r)$ is achieved. Both the number of captures k and the number of individuals r are random variables. If the process has discovered r individuals by the time it stops, the total number of captures is $k = r + l$ and the last one has to be a recapture. Then, r individuals have been caught during the first $r + l - 1$ captures and one of the r individuals is recaptured in the kth sample. Using independence and Eq. (22), it holds

$$\Pr(r \text{ individuals caught to observe } l \text{ recaptures}) = \frac{N!r}{(N-r)!N^{r+l}} \left\{ {r+l-1 \atop r} \right\}. \tag{27}$$

The MLE estimation remains the same as in the previous cases, with $k = r + l$ and r being sufficient statistics [24]. However, a minimum variance unbiased estimator was derived by Goodman [33]:

$$\hat{N}_{MVUE} = \frac{k^2}{2(k-r)} \tag{28}$$

and expressed by Darroch [24] as follows:

$$\hat{N}_{MVUE} = \frac{\{k\}r}{\{k-1\}r}. \tag{29}$$

A similar reasoning is recognizable in the work of Bawa et al. [8], who inverted the *Birthday Paradox* for obtaining a population estimator. In details, according to the *Birthday Paradox*, the probability of having two persons born in the same day of the year in a group of 23 is $\sim 50\%$; an alternative formulation of the problem [46] states that if an N-sized population is repeatedly and randomly sampled with replacement, the number of trials k required for the first repetition of a sampled value has expectation equal to $\sqrt{2N}$. In such framework, the *Inverted Birthday Paradox* [8] exploits the number k of samples required to find the first repetition for estimating the population size as $\hat{N}_{IBP} = k^2/2$.

Example 3 (Estimation of the number of peers in overlay networks.)**.** Remarkably, Ganesh et al. [31] rediscovered the MVUE estimator of Eq. (28), by extending the theoretical arguments related to the *Inverted Birthday Paradox* (e.g., $\hat{N}_{IBP}$ corresponds to Eq. (28) when $k - r = 1$), and implemented it into the novel *Sample and Collide* technique for assessing the number of peers in an overlay network. In more details, they

show how to sample almost uniformly a network by means of a *Continuous Time Random Walk* (CTRW): an initiator node delivers on the network a sampling packet, which contains a timer value initialized to T; the sampling packet is forwarded until the timer value becomes less than 0, given that it is decremented at each hop i by a quantity equal to $\log(1/U)/d_i$, where U is picked uniformly in [0, 1] and d_i is the degree of the traversed node; the last node that receives the sampling packet sends its ID to the initiator node. Note that the sampling packet encounters with greater probability nodes with a higher degree: dividing the value to be subtracted from the timer by the degree guarantees that all nodes have the same chance to be caught when the timer expires. It is worth remarking that this trick was not used in the random walk technique introduced by the aforementioned work of Mane et al. [45], thus incurring the risk of non-uniform network sampling. Then, the *Sample and Collide* technique collects IDs by means of repeated CTRWs, stopping when the initiator has received IDs already caught earlier for exactly l times. Finally, the number k of performed CTRWs and l (i.e., $l = k - r$) are used for estimating the population size by means of Eq. (28), with l chosen as index for the estimation accuracy.

2.3. The Multiple-Recapture Schnabel's Binomial model

The *Schnabel Census* can be approximated by the *Schnabel's Binomial Model* [61], when the population size N is very large compared to the samples. Indeed, the population is sampled k times and the quantities inspected by the *Schnabel's Binomial Model* for each ith sample, with $i = 1, \ldots, k$, are: its size n_i; the number m_i of recaptures among the n_i individuals; the total number of individuals captured in the previous samples $M_i = \sum_{j=1}^{i-1}(n_j - m_j)$. Focusing on the generic ith sample, with $i = 2, \ldots, k$, the probability of recapturing m_i individuals, given n_i and M_i as fixed parameters, is modeled with the hypergeometric distribution and can be approximated with the binomial one, when N is very large compared to the sample size [5]:

$$\Pr(m_i|n_i, M_i) = \frac{\binom{M_i}{m_i}\binom{N-M_i}{n_i-m_i}}{\binom{N}{n_i}} \approx \approx \binom{n_i}{m_i}\left(\frac{M_i}{N}\right)^{m_i}\left(1-\frac{M_i}{N}\right)^{n_i-m_i}. \quad (30)$$

The *Schnabel's Binomial Model* is then the probability to collect a series of recaptures $m_2, m_3, \ldots, m_k$, given the knowledge of the sets $\{n_i\}$ and $\{M_i\}$, with $i = 1, \ldots, k$. In other words, it is the product of the probabilities of recaptures in each sample (see Eq. (30)) and it is expressed as follows:

$$\Pr(m_2, \ldots, m_k|\{n_i\}, \{M_i\}) = \prod_{i=2}^{k}\Pr(m_i|n_i, M_i) \approx \prod_{i=2}^{k}\binom{n_i}{m_i}\left(\frac{M_i}{N}\right)^{m_i}\left(1-\frac{M_i}{N}\right)^{n_i-m_i}. \quad (31)$$

The MLE estimator related to this model is the positive real root of the following equation [61]:

$$\sum_{i=2}^{k}\frac{(n_i - m_i)M_i}{N - M_i} = \sum_{i=2}^{k} m_i \quad (32)$$

although, if the values $\{M_i\}$ are very small compared to N, a first approximation to the solution of Eq. (32) is given by the so called Schnabel's estimate [61]:

$$\bar{N} = \frac{\sum_{i=2}^{k} n_i M_i}{\sum_{i=2}^{k} m_i} \quad (33)$$

Actually, the Schnabel's estimate can be obtained by means of linear regression methods as follows. According to the *Lincoln–Petersen Index*, the ratio of marked individuals m_i in a sample n_i should reflect the ratio of marked individuals M_i in the population N, and the error between these ratios is defined as $e_i = m_i/n_i - M_i/N$. A linear regression approach would employ a least-squares minimization of the errors, i.e., it would minimize $\sum_{i=2}^{k} w_i e_i^2$ with respect to N. The weights w_i are introduced since the variances of the errors e_i are not equal. The Schnabel's estimate of Eq. (33) is obtained by setting the weights proportional to the inverse of the variances, i.e., $w_i = n_i/M_i$ [65]. When departures from uniform random sampling are likely to occur, i.e., when the individuals are grouped or clustered, it is more efficient to set the weights proportional to the sample sizes [65]. Setting the weights $w_i = n_i$, the Schumacher and Eschmeyer's [62] estimator results as follows:

$$\bar{N}_{SE} = \frac{\sum_{i=2}^{k} n_i M_i^2}{\sum_{i=2}^{k} m_i M_i} \quad (34)$$

The *Schnabel's Binomial Model* has also been exploited for deducing the *Gazey and Staley Bayesian method* [32]. According to the Bayesian approach, the likelihood probability given by Eq. (31) is associated to a given prior distribution for the population size N in order to compute the posterior probability distribution $\Pr(N|m_1, \ldots, m_k)$. Since the distribution of the population size is unknown before the experiment is conducted, Gazey and Staley assume that such probability is represented by a "noninformative" discrete uniform distribution, i.e., the values for the population size in the set $\{N_j\}$, with $j = 1, \ldots, J$, are considered possible a priori with the same probability $\Pr(N_j) = 1/J$. The only condition to be imposed is that $\min(N_j) \geq M_k + n_k - m_k$. In this sense, the posterior probability for each population level N_j can be written as follows:

$$\Pr(N_j|m_1, \ldots, m_k) = \frac{\Pr(m_1, \ldots, m_k|N_j)\Pr(N_j)}{\sum_{l=1}^{J}\Pr(m_1, \ldots, m_k|N_l)\Pr(N_l)} = \frac{\prod_{i=1}^{k}\Pr(m_i|N_j)}{\sum_{l=1}^{J}\prod_{i=1}^{k}\Pr(m_i|N_l)}, \quad (35)$$

where the last equality follows from the probabilities $\Pr(N_j)$ being equal. Furthermore, Gazey and Staley show how such formulation can be put in a recursive form made by k successive steps, where, using the posterior distribution calculated at step $i - 1$, the posterior distribution in step i is estimated

based on information from the ith step. Such result is shown in the following expression:

$$\Pr(N_j|m_1,\ldots,m_i) = \frac{\Pr(m_i|N_j)\Pr(N_j|m_1,\ldots,m_{i-1})}{\sum_{l=1}^{J}\Pr(m_i|N_l)\Pr(N_l|m_1,\ldots,m_{i-1})}, \tag{36}$$

for $i = 1,\ldots,k$ and $j = 1,\ldots,J$, and $\Pr(N_j|m_1,\ldots,m_{i-1}) = \Pr(N_j) = 1/J$ for $i = 1$. The advantage of the *Gazey and Staley Bayesian method* lies in the explicit derivation of the probability distribution of the population size, which can be further manipulated for obtaining the population size estimator as $\hat{N}_{GS} = \mathrm{E}[N]$ and the related confidence intervals as well.

Example 4 (Estimation of the number of file replicas in P2P networks.)**.** A network application of the *Capture-Recapture* methods described above is presented in [12,52] to estimate the number N of replicas of the same file in an eDonkey peer-to-peer network. By analyzing the message exchanged among users in Nice (France) and other ones all around the world, it is assumed that: (i) the number of users does not significantly change in the time interval during which the estimation takes place; (ii) there is an equal probability to observe any user outside Nice. Then, the k *local* users in Nice interested to the same file f are identified and arbitrarily ordered. With reference to f, for each ith *local* user, with $i = 0,\ldots,k$, a sample is defined as the set of *non-local* users outside Nice that communicate with it. In details, for each ith sample the authors define: the sample size n_i; the number M_i of non-local users already seen in the previous samples $0,\ldots,i-1$; the number m_i of *non-local* users present in both M_i and n_i. A Bayesian approach derived from the *Gazey and Staley Bayesian method* is then used for estimating the population size N. More specifically, as in the *Gazey and Staley method*, the prior probability is assumed to be uniformly distributed over an interval of discrete values for the population size, although no assumptions are formulated on the width of such interval. In this sense, Eq. (35) is modified as follows:

$$\Pr(N|m_1,\ldots,m_k) = \frac{\Pr(m_1,\ldots,m_k|N)}{\sum_{n=N_{min}}^{N_{max}}\Pr(m_1,\ldots,m_k|n)} \tag{37}$$

where the minimum value for the population size is $N_{min} = M_k + n_k - m_k$ (similarly to the *Gazey and Staley method*), while the maximum population size N_{max} is computed so that $\Pr(N_{max}|m_1,\ldots,m_k) < \epsilon$ for a precision ϵ. In addition, it has been shown [12] that the likelihood probability $\Pr(m_1,\ldots,m_k|N)$ can be computed in a recursive fashion:

$$\frac{\Pr(m_1,\ldots,m_k|N)}{\Pr(m_1,\ldots,m_k|N-1)} = \frac{N}{N-M_k}\frac{\prod_{i=1}^{k}(N-n_i)}{N^k}. \tag{38}$$

Note that equating the previous relation to 1 would produce the same MLE population size estimator as Eq. (11) (i.e., it has been used the *ratio* method [23]). Hence, M_k is a sufficient statistics for performing the size estimation. Besides, noting that the probabilities $\Pr(m_1,\ldots,m_k|N)$ can be scaled arbitrarily in Eq. (37), the authors set $\Pr(m_1,\ldots,m_k|\hat{N}_{SE}) = 1$ for avoiding overflows, where $\hat{N}_{SE}$ is the estimation obtained from the *Schumacher and Eschmeyer's method*; the other probabilities $\Pr(m_1,\ldots,m_k|N)$, with $N \in [N_{min}, N_{max}]$, are computed recursively by means of Eq. (38).

Example 5 (Detection of missing RFID tags.)**.** The Schnabel's estimate is also employed in [55] for the detection of missing RFID tags. Following the investigation made in a previous work [36], some statistical methods are introduced to deal with the problem of missing RFID tags, while performing k reader sessions. Each reader session draws an n_i-sized sample of the population: the number m_i of recaptured tags in the sample and the total number M_i of read individual tags are then registered. The presence of a reliability layer is also assumed to obtain a running estimate of the probability p_M of having at least one tag missing. If the estimate $\hat{p}_M$ is higher than an acceptable threshold, then an additional reader session has to be initiated. In details, defining N as the total number of tags and q as the probability that a tag cannot be read during a session (e.g., because there is an obstacle between the tag and the reader), the estimate $\hat{p}_M$ is obtained according to the following relation:

$$\hat{p}_M = 1 - (1-\hat{q}^k)^{\hat{N}}. \tag{39}$$

where $\hat{N}$ and $\hat{q}$ are respectively the estimates of N and q. Such estimates are obtained by means of an iterative procedure which stops at the second iteration. Indeed, a first estimate $\bar{N}$ of the number of tags is performed according to Eq. (33). Instead, noting that the probability of error for each individual reader session can be estimated as $1 - n_i/\bar{N}$, a first estimate $\bar{q}$ of q is expressed as:

$$\bar{q} = \frac{1}{k}\sum_{i=1}^{k}\left(1 - \frac{n_i}{\bar{N}}\right). \tag{40}$$

Defining r as the number of distinct tags's found in k reader sessions, note that the expected value of r is $N(1-q^k)$. Hence, the ultimate estimate $\hat{N}$ of N is based on the estimate $\bar{q}$ of q as follows:

$$\hat{N} = \frac{r}{1-\bar{q}^k} \tag{41}$$

Instead, the last estimate $\hat{q}$ of q depends on $\hat{N}$ and it is derived by means of the same reasoning of Eq. (40). Finally, $\hat{N}$ and $\hat{q}$ are used in Eq. (39) for obtaining the searched estimate $\hat{p}_M$ of having at least one tag missing.

2.4. Heterogeneous capture probabilities

If the capture probabilities are heterogeneous, i.e. elements have not the same probability of being caught, the underlying model is referred to as a M_h *Capture-Recapture* one. Several estimators have been derived in this context. The main ones we are going to discuss have also been employed for estimation problems in computer networks.

2.4.1. The jackknife estimator

Estimating the population size by means of a M_h *Capture-Recapture* model would require a joint estimation of the capture probabilities, which are parameters not known a priori. However, it is possible to avoid the estimation of the capture probabilities, by using some sufficient statistics in a non-parametric approach. In this sense, the most known estimator employed for a M_h *Capture-Recapture* model is the *jackknife* one [13]: after collecting k samples of the population, the individuals found are grouped in k different sets,

with the ith set containing the elements seen i times; let f_i be the cardinality of the ith set, hereafter referred to as "capture frequency" [13], and representing a sufficient statistic aiming to ease the estimation of the population size without estimating the capture probabilities. Then, the *jackknife* estimator is computed as a linear combination of the capture frequencies f_i, with $i = 1 \dots k$, and its general expression is given in the following formulation:

$$\hat{N}_{JO} = a(k, O)_1 f_1 + a(k, O)_2 f_2 + \cdots + a(k, O)_k f_k \tag{42}$$

where $a(k, O)_i$ are the coefficients in the linear combination, each one depending on the number of samples k and on the parameter O, which represents the order of the estimation. In general, increasing O, the bias of $\hat{N}_{JO}$ decreases, but its variance is increased [13]. Therefore, there exist several mathematical expressions related to a given experiment, each one derived for a given order O of the estimation. As an example, we present the formulas corresponding to the first four orders of estimation:

$$\hat{N}_{J1} = \sum_{i=1}^{k} f_i + \frac{k-1}{k} f_1 \tag{43}$$

$$\hat{N}_{J2} = \sum_{i=1}^{k} f_i + \frac{2k-3}{k} f_1 - \frac{(k-2)^2}{k(k-1)} f_2 \tag{44}$$

$$\hat{N}_{J3} = \sum_{i=1}^{k} f_i + \frac{3k-6}{k} f_1 - \frac{3k^2-15k+19}{k(k-1)} f_2 + \frac{(k-3)^3}{k(k-1)(k-2)} f_3 \tag{45}$$

$$\hat{N}_{J4} = \sum_{i=1}^{k} f_i + \frac{4k-10}{k} f_1 - \frac{6k^2-36k+55}{k(k-1)} f_2 + \frac{4k^3-42k^2+148k-175}{k(k-1)(k-2)} f_3 - \frac{(k-4)^4}{k(k-1)(k-2)(k-3)} f_4 \tag{46}$$

The derivation of such expressions is quite complicated and it can be found in [13]. The right jackknife estimator is chosen by means of a hypothesis testing and according to a required significance level, e.g., 0.05. In details, one should identify the lowest order of estimation m, so that $\hat{N}_{Jm}$ has an associated significance level $P_m > 0.05$. However, as also suggested in [13], an interpolated estimate should be computed as linear combination between $\hat{N}_{Jm}$ and $\hat{N}_{Jm-1}$, in order to smooth the otherwise discrete nature of choosing exactly one among all the mathematical expressions for the jackknife estimator.

Example 6 (Estimation of the number of flows handled by a router.)**.** The CApture REcapture (CARE) queue management technique [14] is a viable fair bandwidth sharing scheme, which exploits the potentialities of the jackknife estimator among other features. Referring to the number of flows handled by a network router with N_{flows}, for each flow $i \in \{1, \dots, N_{flows}\}$, the major function of CARE is to adjust the flow rate to a fair bandwidth share, by dropping incoming packets with a probability d_i. In detail, CARE defines: the "packet count" M_i (proportional to the source rate of the flow i), as the number of packets related to the flow i and stored in the router buffer, having size N packets, and the "fair share size" N/N_{flows} (proportional to the bandwidth fair share), as the quota of the buffer size N that should be fairly occupied by each flow. When the rate of the flow i exceeds the bandwidth fair share, the ideal behavior of the adjustment mechanism performed by CARE is to drop packets with a probability given by the following formulation:

$$d_i = 1 - N/(M_i \cdot N_{flows}), \tag{47}$$

where the packet count M_i and the number of flows N_{flows} are respectively estimated with $\hat{M}_i$ and $\hat{N}_{flows}$ according to *Capture-Recapture* approaches.

A circular linked list of size n is considered for capturing the incoming packets with probability $p_{cap} = n/N$. Focusing on the flow i associated with the last captured packet, the m_i packets in the circular linked list pertaining to the flow i are counted. The probability distribution associated with the event "the packet count is equal to a given value M_i" is hypergeometric, and the MLE estimator $\hat{M}_i$ of the packet count is given by the following formula:

$$\hat{M}_i = \frac{N m_i}{n}, \tag{48}$$

whose expression is similar to that of the *Lincoln–Petersen Index* in Eq. (1). This is due to the fact that both formulas can be derived from the same assumption on the probability relationship, i.e., the probability to capture an individual in the population is equal to the probability to capture a marked individual among all marked individuals.

Instead, the number of flows N_{flows} is estimated by means of the jackknife estimator described above. This estimation is performed by identifying the underlying model as a M_h *Capture-Recapture* one. Indeed, packets pertaining to different flows have a different probability to be caught in the buffer. Specifically, the n packets stored in the circular linked list are considered as n samples with size equal to 1. The *jackknife* estimator is applied on those n samples, after evaluating the capture frequencies f_i, with $i = 1, \dots, n$. In this sense, the capture frequency f_i is the number of flows, whose representative packets in the circular linked list are exactly i.

The most important feature in this paper is the use of a fixed memory size for storing samples of the population, while applying the CR approach for solving two estimation problems: one with the jackknife estimator, the other one with an estimator similar to the *Lincoln–Petersen Index* as expressed by Eq. (48).

2.4.2. The $f_1 f_2$ estimator and the Sample Coverage one

Other two estimators for M_h *Capture-Recapture* models are those derived by Chao and Lee [15,16,42].

The jackknife estimator usually underestimates the population size if many individuals have very small capture probabilities so that they are caught only once or twice in the Multiple-Recapture experiments. For this reason, using the same variable environment of the jackknife estimator, the $f_1 f_2$ estimator [15,16] is expressed in the following formulation:

$$\hat{N}_{12} = r + \frac{f_1^2}{2 f_2} \tag{49}$$

where r is the total number of individuals caught in the population, as in Eq. (7). In [15,16] the $f_1 f_2$ estimator is shown to be a good approximation to the true value of the population size if the individuals caught in k samples are captured at most 2 times. The asymptotic variance of such estimator has the following expression:

$$\mathrm{Var}[\hat{N}_{12}] = f_2 \left[\frac{1}{4}\left(\frac{f_1}{f_2}\right)^4 + \left(\frac{f_1}{f_2}\right)^3 + \frac{1}{2}\left(\frac{f_1}{f_2}\right)^2 \right]. \quad (50)$$

The Sample Coverage estimator for a M_h *Capture-Recapture* model is given instead in the following formula [42]:

$$\hat{N}_{SC} = \frac{r}{C} + \frac{f_1}{C}\gamma^2 \quad (51)$$

where

$$C = 1 - \frac{f_1}{\sum_{i=1}^{k} i f_i}, \quad (52)$$

$$\gamma^2 = \max\left\{ \frac{\frac{r}{C} k \sum_{i=1}^{k} i(i-1) f_i}{(k-1)\left(\sum_{i=1}^{k} i f_i\right)^2} - 1, 0 \right\}. \quad (53)$$

The variance of such estimator can be computed approximately via the delta method [47] as:

$$\mathrm{Var}[\hat{N}_{SC}] = \sum_{i=1}^{k}\sum_{j=1}^{k} H_i H_j \,\mathrm{Cov}[f_i, f_j] \quad (54)$$

where $H_i = \frac{\partial \hat{N}_{SC}}{\partial f_i}$ and $\mathrm{Cov}[f_i, f_j]$ is given by:

$$\mathrm{Cov}[f_i, f_j] = \begin{cases} f_i\left(1 - \frac{f_i}{\hat{N}_{SC}}\right) & \text{if } i = j \\ -\frac{f_i f_j}{\hat{N}_{SC}} & \text{if } i \neq j \end{cases} \quad (55)$$

Example 7 (Estimated number of RFID tags over unreliable radio channels.)**.** Both of these techniques have been employed [68] in the estimation process of the number of RFID tags when communication occurs over unreliable radio channels. The frame-slotted ALOHA model of [41] is used by assumption. A tag receiving a query from the reader decides to advertise its presence with a given probability q. In that case, it chooses its transmission slot among the F ones in the frame, leveraging an hashing scheme that involves its own local identifier and a seed received in the query. The probability that the response is received by the reader depends also on the behavior of the radio channel. At the same time, the reader only needs to track whether each slot in the responding frame is idle or not, without accounting for collisions. From these premises, the algorithm proposed in [41] uses the estimate of the expected number of busy frame slots to estimate the cardinality of the tags' set.

In order to reduce the variance of the number of idle slots, the tag set is polled k repeated times with the same seed and the positions of occupied slots for each probe are recorded. Indeed, although the tag-to-slot mapping is fixed, the actual occupancy of the responding frames for the k probes may be different. In details, the probability that a slot supposed to be occupied is idle depends both on the actual number of tags selecting the slot and on the responding probabilities of those tags, which in turn depends also on the radio channel behavior. Such probability being different for all non-idle slots, a M_h *Capture-Recapture* model can be adopted to estimate the expected number of frame slots occupied. Indeed, the occupancy history of each slot in a frame across these multiple repeated probes corresponds to the statistics of capture history of marked individuals.

Finally, in order to reduce the variance of the estimator of the tag-set cardinality, which is due to the pseudo-random tag-to-slot mappings, m different seeds are used. Defining $\hat{N}_j$, with $j = 1, \ldots, m$, the estimate of the number of occupied slots under a perfect channel for the jth seed and performed with one between the CR estimator presented above, the authors employ the estimator shown in the following formula (and already presented in [41]) for evaluating the number of RFID tags:

$$\hat{N} = -\frac{F}{q}\log\frac{F - \frac{1}{m}\sum_{j=1}^{m}\hat{N}_j}{F}. \quad (56)$$

3. Capture-Recapture estimators for open populations

A population is considered *open* when the employed estimation approach assumes that individuals could join or leave the population itself during the sampling process. In this case, the mean service rate (i.e., the departure rate of elements from the population) and the mean input rate (i.e., the joining rate of elements in the population) must also be evaluated, since they directly affect the estimate of the population size.

Without loss of generality, if the population size does not change very quickly, the models for *closed* population presented so far can be employed also in contexts related to time-varying populations: the population size is assumed to stay constant during the sampling procedure, thus CR estimators for closed populations can be fairly applied. However, this assumption can be not satisfactory when the variation is significant during the sampling time interval. To face this issue, there are two main solutions: (i) employing CR estimators of *closed* populations on sliding time-windows; (ii) adopting CR estimators of *open* populations provided by the biometric literature, e.g., the *Jolly-Seber* model.

In the remaining part of this section we explain these two methods. For the first we describe some applications in the field of communication networks. To the best of our knowledge, the second has not been considered until now by the networking community, so we describe our personal view on future research developments.

3.1. Closed population estimators for open populations

The time interval during which the sampling process takes place can be considered as a sliding window over the time. Although a given population can be time-varying, i.e., there are individuals joining or leaving the population itself, it is possible to set the sliding window length small enough for considering the population size almost constant in that observation time interval. In this case closed population *Capture-Recapture* techniques can be employed for estimating the size of open populations. As a matter of fact, this

underlying assumption has been kept in some of the previously described works [31,52,74].

Example 8 (Estimation of the number of RFID tags.)**.** The aim of the work [72] is to estimate the number N of items still in stock for retailer application scenarios. In details, according to the *Lincoln–Petersen Index* two samples of the population are needed. In the application scenario described in [72], given a time window w, the samples are: (i) the number M of RFID tags read in w, properly accounted in a *readHistory* list; (ii) the number n of items sold in w, accounted in a *salesHistory* list. The size of the window is sufficiently large in order to have the two sample sizes meeting the requirements of Eq. (3). The simplest estimation could be done by counting the number m of items present in both samples and using the *Lincoln–Petersen Index* for evaluating the population size. However, this is not realistic, as shelves are constantly replenished in real scenarios: some bias is introduced in the estimation, since the replenished items will be detected instantaneously from a RFID tag reader, but some time will be needed for those items appearing also in the *salesHistory* list. To accomplish a non-biased estimation, the population size is properly corrected according to the replenishment history. The unbiasedness of the derived estimator is also proofed with theoretical arguments. The resulting key outcome of this work is the extension of the Lincoln–Petersen Index to the estimation of an open-population, since item replenishment corresponds to the assumption that individuals can enter the population.

3.2. The Jolly–Seber model

Since the last century, the estimation of *open* populations has been the key focus in the work of some statisticians. Specifically, a great effort has been spent for modelling time-varying populations in a MLE-based approach. A meaningful first step towards this ambitious aim was done by Darroch [25]: he exploited the Multiple-Recapture Census and its inherent properties that had been already described in a previous work [24]. In detail, Darroch extended the probability model of Eq. (9), surveyed before in Section 2.2, and derived the MLE estimators related to *open* populations with only departures or only joining elements. However, a likelihood function accounting for both deaths and immigration was unwieldy to manage.

In 1965, Jolly and Seber [38,64] faced this problem by introducing the model which takes their names, i.e., the *Jolly–Seber model*. Their approach is able to estimate only some model parameters in a MLE fashion, while the other parameters are reasonable combinations of the first ones. We recall that MLE estimators hold some desirable features, like asymptotically consistency, efficiency and normality. It is worth noting that the *Jolly–Seber* (JS) model does not make any assumption on the underlying evolutionary model for the population. In this sense, the estimation procedure provides several estimates for any population parameter (e.g., the population size or the number of individuals joining the population), one for each sampling time in a given observation window.

In detail, the JS model is able to estimate some population parameters by exploiting the information inherent to k samples randomly drawn over time from the population itself. Specifically, for a given ith sample, the following quantities are measured:

- n_i number of individuals caught in the ith sample;
- m_i number of marked individuals caught in the ith sample;
- $u_i = n_i - m_i$ number of individuals caught for the first time in the ith sample;
- R_i number of marked individuals released after the ith sample;
- r_i number of individuals from the release of R_i individuals which are subsequently recaptured;
- z_i number of different individuals caught before the ith sample which are not caught in the ith sample but are caught subsequently.

The values R_i take into account the possibility for individuals to die after being trapped. Therefore, those values are relevant in a biological setting, but they are probably less important for communication network estimation problems.

The population quantities estimated by the JS model are:

- N_i total number of individuals in the population until the ith sample;
- M_i total number of marked individuals in the population just before the ith sample;
- $U_i = N_i - M_i$ total number of individuals never seen before the ith sample;
- B_i number of new individuals joining the population in the interval between the ith and $(i+1)$th samples;
- p_i probability for each individual of being caught in the ith sample;
- ϕ_i probability of every marked individual of surviving from the ith to the $(i+1)$th sample.

In the remaining part of this paragraph, we will denote the related estimates with a hat.

The JS model is defined by a likelihood probability, which has to be maximized jointly with respect to the unknown parameters ϕ_i, p_i, and U_i, with $i = 1, \ldots, k$. At the same time, the quantities B_i, N_i, and M_i are treated as random variables and they are estimated by means of moment matching estimation [65]. However, the same estimators for the considered set of parameters and random variables can be obtained by an intuitive argument provided by Jolly [38] and herein reproduced.

As a first consideration, the proportion of the estimated marked individuals M_i w.r.t the estimated total number of individuals N_i is equal to the found proportion of the marked individuals n_i w.r.t the sample size n_i at the ith sampling time instant, leading to:

$$\hat{N}_i = \frac{\hat{M}_i n_i}{m_i} \tag{57}$$

which is consistent with the *Lincoln–Petersen Index*, although the number of marked individuals in the population is now an estimate, due to departures of individuals from the population (marked individuals can die or leave the population after being caught).

Secondly, the capture probability is the proportion of marked or total individuals alive at i that are captured in i:

$$\hat{p}_i = \frac{n_i}{\hat{N}_i} = \frac{m_i}{\hat{M}_i}. \tag{58}$$

An estimate of the actual number M_i of marked individuals at the ith sample (i.e., the individuals already captured and still surviving in the population) can be obtained by an intuitive argument too. Indeed, the ratio of marked individuals not seen at i and captured again at least once, i.e., $z_i/(\hat{M}_i - m_i)$, should be reflected by the ratio of individuals released after the ith sample and subsequently recaptured, i.e., r_i/R_i. The resulting proportion leads to:

$$\hat{M}_i = \frac{R_i z_i}{r_i} + m_i. \tag{59}$$

The survival rate estimator is obtained instead by considering the number of marked individuals in the population immediately after the ith sample as $M_i - m_i + R_i$. This expression accounts for the marked individuals not caught in the ith sample $(M_i - m_i)$ and for the number of individuals caught in the ith sample and released (R_i). A natural survival rate estimator is then computed as the proportion of the estimated number of marked individuals at the $(i+1)$th sample and the estimated number of marked individuals immediately after the ith sample:

$$\hat{\phi}_i = \frac{\hat{M}_{i+1}}{\hat{M}_i - m_i + R_i} \tag{60}$$

Further, the number of new individuals joining the population between the ith and the $(i+1)$th samples is only the difference between the estimated population size at sample $(i+1)$ and the expected number of survivors from i to $(i+1)$:

$$\hat{B}_i = \hat{N}_{i+1} - \hat{\phi}_i(\hat{N}_i - n_i + R_i). \tag{61}$$

Finally, the estimated total number of individuals never seen before the ith sample is given by:

$$\hat{U}_i = \hat{N}_i - \hat{M}_i. \tag{62}$$

It has to be noted that Eqs. (57)–(62) refer to each step in the sampling procedure, hence they neither express any prediction of the future states of the population, nor provide an evolutionary model of the population itself. Furthermore, as pointed out in [65], the estimates $\hat{M}_i$ and $\hat{N}_i$ are not *maximum-likelihood*, but they are simply used as intermediate steps in the calculation of the MLE ones, i.e., $\hat{\phi}_i$, $\hat{p}_i$ and $\hat{U}_i$. Also the estimate $\hat{B}_i$ is only valid if the survival probability ϕ_i is the same for all individuals and not just for the marked ones. These are the main reasons for considering the JS model as MLE-based more than a pure maximum-likelihood estimation technique.

No work has dealt with the *Jolly–Seber* model since now in the literature related to computer networks. As a matter of fact, a networking estimation problem would be more keen to accept some approximations on the population evolution description, which in turn would be quite unrealistic if adopted in studies of animal populations. In this sense, we believe that framing an underlying evolutionary model for a population in the *Jolly–Seber* approach should be investigated, thus opening an interesting research direction addressing population estimation problems in computer networks.

4. Related works

The present section mentions some relevant estimation techniques applied in computer networks contexts, although they do not deal with the *Capture-Recapture* approach.

Firstly, Kalman filtering [39] has been widely exploited for estimation problems in several engineering contexts, since it is the best linear filter having the smallest unconditioned error covariance [19]. Given an actual system, the Kalman filter reproduces it through a state model intended to make a prediction of the evolution of the system itself. Then, the Kalman filter works by comparing the output of the actual system with that of its state model in order to correct the state itself and achieve the best fitting state estimation. This approach can be fruitfully exploited for estimating a population. Indeed, if the state of the actual system under estimation is the population size and the measured output of that system is some quantity related the population itself, e.g., the size of a randomly picked sample, then the population size prediction can be corrected according to the difference between the measured sample size and the output of the reproduced system. It has to be noted that the Kalman filter is able to perform an estimation also in dynamical contexts, addressing the issues posed by the estimation of time-varying populations.

In this sense, the Kalman filter has been also employed for estimating the membership size in multicast networks [4]. The population of multicast terminals is assumed to evolve as an $M/M/\infty$ queuing system, hence terminals join with a rate λT (T is a "speeding up" factor for having a heavy traffic condition) and leave with a rate μ. A central controller polls all terminals at times iS $(i = 0, 1, \ldots)$, with $S > 0$, and terminals are assumed to reply in a synchronized fashion sending an acknowledgment with a small probability p. The found estimator for the multicast membership size has been demonstrated to be asymptotically unbiased.

Instead, the Extended version of the Kalman filter [19] has been employed for estimating the number of competing terminals in IEEE 802.11 networks [9]. Although a general assumption for the Extended Kalman filter applicability is that the state and the measurement noise have to be uncorrelated, the model formulation does not allow to fit this requirement. In this sense, the authors firstly propose the measurement noise to be non-linearly related to the previous state value. In addition, the authors point out that modeling the state noise as a stationary process would be not efficient in the given application scenario: low values for the state noise variance give accurate estimates in stationary conditions, although with a very long transient phase; contrariwise, a high value for the state noise variance allows a quick reaction to state changes, while implying a reduced accuracy in the estimation. The strategy proposed and employed by the authors to find a right trade-off in the choice of the state noise variance is to let it be time-varying: the noise variance is set to be 0 until the mean value of the innovation process, i.e., the difference between the actual system's and the observer's outputs, gets far from 0; in this last case, a change in the population size is detected, therefore the value of the state noise variance is set to a sufficiently large value, allowing the Kalman filter to move away from the former estimate and to quickly converge to a new one. The efficiency

of the "change detector" has been supported by simulation results.

From a statistical point of view [19], Kalman filtering is equivalent to adopt a Bayesian approach to estimation, thus providing asymptotically efficient and unbiased estimates [10,43]. Although this scheme efficiently performs on dynamic systems, it has strong guarantees only in some specific cases (e.g. linear systems with gaussian noise). In most estimation problems related to computer network, the observations are not in general affected by gaussian noise, so that the *Maximum-Likelihood Estimation* (MLE) [3,22,27–30,56] approach has also been investigated in the research literature. In fact, MLE techniques provide reliable parameter estimates, regardless of the prior probability distribution inherent to the observed phenomenon.

Other population estimation problems in computer networks have dealt with *Maximum-Likelihood Estimation* [27]. An example of MLE population estimators in computer networks is that presented in [69], where the authors have proposed a general distributed network size estimation strategy based on Bernoulli trials. This work is framed into the theory of consensus-algorithms for anonymous networks.[5] The authors have described how the MLE estimator for such networks can be found by exploiting the implications of the *Newton–Pepys problem* [66] (the probability of having at least one six when throwing six dice is greater than the probability of having at least two six when throwing twelve dice).

Another work dealing with an MLE approach is that presented in [40], where a MLE estimator of the tag population size in Frame Slotted ALOHA (FSA) protocols is derived. In particular, the authors have determined the exact probability distribution of the observable event space in FSA systems, thus enabling the MLE formulation of a slot-by-slot tag population estimator. However, the likelihood probability found has a complex expression. By means of graphical inspection, the authors have also shown that this likelihood probability has only one maximum, which can be, as a consequence, found using a gradient search.

For the sake of completeness, it has to be also mentioned the population size estimation technique proposed in [74] to enhance the performance of the Stabilized Random Early Detection (SRED) [49] mechanism for queue management. Although the estimation technique presented is classifiable as identity-based (i.e., individuals are recognized while sampling repeatedly the population, and this information is used for a better estimation), it is not framed in the *Capture-Recapture* theory. In particular, the authors have proposed a Hash-based Two-Level caching scheme (HaTCh) for estimating the number of concurrent active flows in high-speed networks. The authors have shown that SRED becomes inaccurate either when increasing the number of flows to estimate, or when TCP flows are mixed with non-responsive UDP flows; hence, to face this problem, they have proposed a two-level caching scheme, which uses hashing and a two-level caching mechanism to accurately estimate the number of active flows under various workloads. More specifically, an additional intermediate cache has been inserted in the caching scheme: it is conceived for isolating non-responsive UDP flows, which are characterized by a bigger workload with respect to the TCP ones.

[5] Consensus algorithms aim to obtain agreement amongst a number of agents for a single data value [57]. Besides, a network is anonymous if the agents' IDs are either not unique or not exploitable.

5. Concluding remarks

5.1. Lessons learned

The main objective of the present contribution is to shed some light on the effective spectrum of available techniques able to perform an efficient estimation of populations in computer networks. As shown throughout the paper, sampling a given population is more viable than counting individuals, and the statistical properties of the sampling procedure can be employed for inferring the size and the dynamics of that population. More interestingly, the identity of caught individuals can be registered, so that the capture history of each individual is an additional information that can be fruitfully exploited for obtaining more effective estimates.

Noticeably, we learned that the *Capture-Recapture* approach has been firstly conceived (since the last century) to afford biometric problems, such as the estimation of animal abundance in ecological environments. The computer networks scientific community has employed some *Capture-Recapture* estimators for evaluating the size of large populations.

We have also learned that the *Capture-Recapture* framework can face very different settings: a population can be closed or open; the probability for an individual to be captured can vary by time or not; individuals can have the same chance to be captured in each sampling step or not (allowing thus heterogeneity in the latter case). The biometric literature is so wide and diversified that a big research effort was required to select the works fitting also communication network estimation problems. As attested from the applications surveyed in this paper, the *Capture Recapture* approach is very desirable in the computer networks scientific community, although, to the best of our knowledge, no effort has been spent for collecting, selecting, and organizing the whole theory until now. The last insight is witnessed by considering that some works in computer networks rediscovered the same CR estimators already featured by the biometric community.

The key lesson learned comes from the investigation that we have done on the Jolly–Seber model for open populations. Albeit not yet used in computer networks at the present stage, such model could be very useful for better describing the estimation problems in computer networks surveyed in this paper.

In general, we hope the present contribution could be appreciated as a reference milestone for future investigations by readers interested to applying *Capture-Recapture* techniques to solve emerging issues in networking literature.

5.2. Research directions and future works

The investigations presented in this manuscript clearly shows that, in the frame of CR techniques, ML approaches can provide viable, effective, and lightweight estimators of large populations of individuals. Among the surveyed solutions, the Jolly–Seber model is the only one able to account

for open populations and multiple recaptures. As future research, this model could be enhanced by making it able to predict the future evolution of estimated populations. This research task could be tackled, similarly to what is done in [4] with the Kalman filter, by identifying the underlying model as an $M/M/\infty$ queueing system, and estimating jointly the size and dynamics of the population.

Overall, the CR estimators surveyed in this paper could be fruitfully used in future research activities related to several cutting edge research issues in networking literature, such as: (i) distributed caching mechanisms for content distribution networks and information centric networking protocols [73]; (ii) autonomic management systems in pervasive machine to machine and cyber physical systems made of billions devices [2,34]; (iii) nano-networks communications [1]; (iv) software defined networking approaches to network control plane [60,67]; (v) economic models in next generation mobile systems (i.e., 5G and beyond) [70]; (vi) cloud computing systems and applications [59].

Acknowledgments

This work was partially supported by the PON project RES NOVAE funded by the Italian MIUR and by the European Union (European Social Fund grant no. PON04a2-E).

References

[1] I.F. Akyildiz, J.M. Jornet, The internet of nano-things, Wireless Commun. IEEE 17 (6) (2010) 58–63.
[2] M.B. Alaya, S. Matoussi, T. Monteil, K. Drira, Autonomic computing system for self-management of machine-to-machine networks, in: Proceedings of the 2012 International Workshop on Self-Aware Internet of Things, ACM, 2012, pp. 25–30.
[3] J. Aldrich, R. A. Fisher and the making of maximum likelihood 1912–1922, Stat. Sci. 12 (3) (1997) 162–176.
[4] S. Alouf, E. Altman, C. Barakat, P. Nain, Optimal estimation of multicast membership, IEEE Trans. Signal Process. 51 (8) (2003) 2165–2176.
[5] N.T.J. Bailey, Improvements in the interpretation of recapture data, J. Anim. Ecol. 21 (1) (1952) 120–127.
[6] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[7] B.A. Bash, J.W. Byers, J. Considine, Approximately uniform random sampling in sensor networks, in: Proceeedings of the 1st International Workshop on Data Management for Sensor Networks: In Conjunction with VLDB 2004, ACM, 2004, pp. 32–39.
[8] M. Bawa, H. Garcia-Molina, A. Gionis, R. Motwani, Estimating Aggregates on a Peer-to-Peer Network, Technical report 2003-24, 2003.
[9] G. Bianchi, I. Tinnirello, Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network, in: INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies, vol. 2, IEEE, 2003, pp. 844–852.
[10] J. Bibby, H. Toutenburg, Prediction and Improved Estimation in Linear Models, Wiley, 1977.
[11] P. Boldi, S. Vigna, An effective characterization of computability in anonymous networks, in: Distributed Computing, Springer, Berlin, Heidelberg, 2001, pp. 33–47.
[12] P. Brown, S. Petrovic, A new statistical approach to estimate global file populations from local observations in the eDonkey P2P file sharing system, Ann. Telecommun. 66 (1) (2011) 5–16.
[13] K.P. Burnham, W.S. Overton, Estimation of the size of a closed population when capture probabilities vary among animals, Biometrika 65 (3) (1978) 625–633.
[14] M.K. Chan, M. Hamdi, An active queue management scheme based on a capture-recapture model, IEEE J. Select. Areas Commun. 21 (4) (2003) 572–583.
[15] A. Chao, Estimating the population size for capture-recapture data with unequal catchability, Biometrics 43 (4) (1987) 783–791.
[16] A. Chao, Estimating population size for sparse data in capture-recapture experiments, Biometrics 45 (2) (1989) 427–438.
[17] A. Chao, An overview of closed capture-recapture models, J. Agric. Biol. Environ. Stat. 6 (2) (2001) 158–175.
[18] D.G. Chapman, Some properties of the hypergeometric distribution with applications to zoological sample censuses, Univ. Calif. Publ. Stat. 1 (7) (1951) 131–159.
[19] C.K. Chui, G. Chen, Kalman Filtering: With Real-Time Applications, Springer, Verlag, 2009.
[20] B. Codenotti, P. Gemmell, P. Pudlak, J. Simon, On the amount of randomness needed in distributed computations, in: Proceedings of the 1997 International Conference on Principles of Distributed Systems, 1997, pp. 237–248.
[21] C.C. Craig, On the utilization of marked specimens in estimating populations of flying insects, Biometrika 40 (1–2) (1953) 170–176.
[22] H. Cramér, Mathematical Methods of Statistics, vol. 9, Princeton University Press, 1999.
[23] R.C. Dahiya, An improved method of estimating an integer-parameter by maximum likelihood, Am. Stat. 35 (1) (1981) 34–37.
[24] J.N. Darroch, The multiple-recapture census. I. Estimation of a closed population, Biometrika 45 (3–4) (1958) 343–359.
[25] J.N. Darroch, The multiple-recapture census. II. Estimation when there is immigration or death, Biometrika 46 (3–4) (1959) 336–351.
[26] B. Dawkins, Siobhan's problem: the coupon collector revisited, Am. Stat. 45 (1) (1991) 76–82.
[27] S.R. Eliason, Maximum Likelihood Estimation: Logic and Practice, vol. 96, Sage Publications, 1993.
[28] R.A. Fisher, On the "probable error" of a coefficient of correlation deduced from a small sample, Metron 1 (1921) 3–32.
[29] R.A. Fisher, On the mathematical foundations of theoretical statistics, Phil. Trans. Roy. Soc. Lond. Ser. A 222 (1922) 309–368.
[30] R.A. Fisher, Theory of statistical estimation, Math. Proc. Camb. Phil. Soc. 22 (05) (1925) 700–725.
[31] A.J. Ganesh, A.M. Kermarrec, E. Le Merrer, L. Massoulié, Peer counting and sampling in overlay networks based on random walks, Distrib. Comput. 20 (4) (2007) 267–278.
[32] W.J. Gazey, M.J. Staley, Population estimation from mark-recapture experiments using a sequential Bayes algorithm, Ecology 67 (4) (1986) 941–951.
[33] L.A. Goodman, Sequential sampling tagging for population size problems, Ann. Math. Stat. 24 (1) (1953) 56–69.
[34] L.A. Grieco, M.B. Alaya, T. Monteil, K. Drira, Architecting information centric ETSI-M2M systems, in: 2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), IEEE, 2014, pp. 211–214.
[35] J.M. Hendrickx, A. Olshevsky, J.N. Tsitsiklis, Distributed anonymous discrete function computation, IEEE Trans. Autom. Control 56 (10) (2011) 2276–2289.
[36] R. Jacobsen, K.F. Nielsen, P. Popovski, T. Larsen, Reliable identification of RFID tags using multiple independent reader sessions, in: 2009 IEEE International Conference on RFID, IEEE, 2009, pp. 64–71.
[37] P. Jesus, C. Baquero, P.S. Almeida, A survey of distributed data aggregation algorithms, IEEE Commun. Surv. Tutor. 17 (1) (2015) 381–404.
[38] G.M. Jolly, Explicit estimates from capture-recapture data with both death and immigration-stochastic model, Biometrika 52 (1–2) (1965) 225–247.
[39] R.E. Kalman, A new approach to linear filtering and prediction problems, J. Basic Eng. 82 (1) (1960) 35–45.
[40] B. Knerr, M. Holzer, C. Angerer, M. Rupp, Slot-wise maximum likelihood estimation of the tag population size in FSA protocols, IEEE Trans. Commun. 58 (2) (2010) 578–585.
[41] M. Kodialam, T. Nandagopal, W.C. Lau, Anonymous tracking using RFID tags, in: INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE, IEEE, 2007, pp. 1217–1225.
[42] S.M. Lee, A. Chao, Estimating population size via sample coverage for closed capture-recapture models, Biometrics 50 (1) (1994) 88–97.
[43] E.L. Lehmann, G. Casella, Theory of point estimation, vol. 31, Springer Science & Business Media, 1998.
[44] F.C. Lincoln, Calculating Waterfowl Abundance on the Basis of Banding Returns, U.S. Department of Agriculture, 1930.
[45] S. Mane, S. Mopuru, K. Mehra, J. Srivastava, Network Size Estimation in a Peer-to-peer Network, Technical report 05-030, 2005.
[46] R. Motwani, P. Raghavan, Randomized Algorithms, Cambridge University Press, 1995.
[47] G.W. Oehlert, A note on the delta method, Am. Stat. 46 (1) (1992) 27–29.
[48] D.L. Otis, K.P. Burnham, G.C. White, D.R. Anderson, Statistical inference from capture data on closed animal populations, Wildlife Monogr. 62 (1978) 1–135.
[49] T.J. Ott, T.V. Lakshman, L.H. Wong, Sred: stabilized red, in: INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 3, IEEE, 1999, pp. 1346–1355.

[50] S.L. Peng, S.S. Li, X.K. Liao, Y.X. Peng, N. Xiao, Estimation of a population size in large-scale wireless sensor networks, J. Comput. Sci. Technol. 24 (5) (2009) 987–997.
[51] C.G.J. Petersen, The yearly immigration of young plaice into the Limfjord from the German Sea, Rep. Danish Biol. Station 6 (1896) 1–48.
[52] S. Petrovic, P. Brown, A new statistical approach to estimate global file populations in the eDonkey P2P file sharing system, in: Teletraffic Congress, 2009. ITC 21 2009. 21st International, IEEE, 2009, pp. 1–8.
[53] K.H. Pollock, J.D. Nichols, C. Brownie, J.E. Hines, Statistical inference for capture-recapture experiments, Wildlife Monogr. 107 (1990) 3–97.
[54] K.H. Pollock, Capture-recapture models, J. Am. Stat. Assoc. 95 (449) (2000) 293–296.
[55] P. Popovski, K. Fyhn, R.M. Jacobsen, T. Larsen, Robust statistical methods for detection of missing RFID tags, IEEE Wireless Commun. 18 (4) (2011) 74–80.
[56] C. Radhakrishna Rao, Information and accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Math. Soc. 37 (3) (1945) 81–91.
[57] W. Ren, R.W. Beard, E.M. Atkins, A survey of consensus problems in multi-agent coordination, in: Proceedings of the 2005 American Control Conference, 2005, IEEE, 2005, pp. 1859–1864.
[58] D.S. Robson, H.A. Regier, Sample size in Petersen mark–recapture experiments, Trans. Am. Fisher. Soc. 93 (3) (1964) 215–226.
[59] M.N.O. Sadiku, S.M. Musa, O.D. Momoh, Cloud computing: opportunities and challenges, IEEE Potentials 33 (1) (2014) 34–36.
[60] S. Salsano, N. Blefari-Melazzi, A. Detti, G. Morabito, L. Veltri, Information centric networking over SDN and OpenFlow: architectural aspects and experiments on the OFELIA testbed, Comput. Netw. 57 (16) (2013) 3207–3221.
[61] Z.E. Schnabel, The estimation of total fish population of a lake, Am. Math. Monthly 45 (6) (1938) 348–352.
[62] F.X. Schumacher, R.W. Eschmeyer, The estimate of fish population in lakes or ponds, J. Tennessee Acad. Sci. 18 (1943) 228–249.
[63] C.J. Schwarz, G.A.F. Seber, Estimating animal abundance: review III, Stat. Sci. 14 (4) (1999) 427–456.
[64] G.A.F. Seber, A note on the multiple-recapture census, Biometrika 52 (1/2) (1965) 249–259.
[65] G.A.F. Seber, The Estimation of Animal Abundance and Related Parameters, Blackburn Press, Caldwell, New Jersey, 2002.
[66] S.M. Stigler, Isaac Newton as a probabilist, Stat. Sci. 21 (3) (2006) 400–403.
[67] R. Strijkers, M.X. Makkes, C. de Laat, R. Meijer, Internet factories: creating application-specific networks on-demand, Comp. Netw. 68 (2014) 187–198.
[68] W.K. Sze, W.C. Lau, RFID counting over unreliable radio channels—the capture-recapture approach, in: 2010 IEEE International Conference on Communications (ICC), IEEE, 2010, pp. 1–6.
[69] D. Varagnolo, G. Pillonetto, L. Schenato, Distributed cardinality estimation in anonymous networks, IEEE Trans. Autom. Control 59 (3) (2014) 645–659.
[70] X. Wang, M. Chen, T. Taleb, A. Ksentini, V. Leung, Cache in the air: exploiting content caching and delivery techniques for 5G systems, IEEE Commun. Mag. 52 (2) (2014) 131–139.
[71] B.M. Waxman, Routing of multipoint connections, IEEE J. Select. Areas Commun. 6 (9) (1988) 1617–1622.
[72] L. Weiss Ferreira Chaves, E. Buchmann, K. Böhm, Tagmark: Reliable estimations of RFID tags for business processes, in: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, pp. 999–1007.
[73] G. Xylomenos, C. Ververidis, V. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. Katsaros, G. Polyzos, A survey of information-centric networking research, IEEE Commun. Surv. Tutor. 16 (2) (2014) 1024–1049.
[74] S. Yi, X. Deng, G. Kesidis, C.R. Das, Technique for estimating the number of active flows in high-speed networks, ETRI J. 30 (2) (2008) 194–204.

Nicola Accettura is currently Post-doc at "University of California Berkeley," Berkeley, CA, USA. His main research interests include Internet of Things (IoT) and statistical modelling for communication networks. He received his Laurea (Bachelor's degree) in Computer Systems Engineering in 2004 and his Laurea Specialistica (Master's degree) in Telecommunications Engineering in 2007, both with honors, from "Politecnico di Bari," Italy. He obtained his Dottorato di Ricerca (PhD) in Electronics Engineering from "Scuola Interpolitecnica di Dottorato" (SIPD) and "Politecnico di Bari," Italy, in February 2013. From February 2012 to August 2012, he has been visiting student at the MAESTRO team of INRIA Sophia-Antipolis (France).

Giovanni Neglia received the Master's degree in electronic engineering and Ph.D. degree in telecommunications from the University of Palermo, Palermo, Italy, in 2001 and 2005, respectively. He has been a Researcher with the Maestro team, INRIA, Sophia Antipolis, France, since September 2008. In 2005, he was a Research Scholar with the University of Massachusetts, Amherst, visiting the Computer Networks Research Group. Before joining INRIA, he was a Postdoctorate with the University of Palermo and an External Scientific Advisor with the Maestro team, INRIA. His research focuses on modeling and performance evaluation of computer networks and proposals of new mechanisms to improve their performance.

Luigi Alfredo Grieco is an assistant professor (tenured in 2008) at "Politecnico di Bari". He has been (March-June 2009) a visiting researcher at INRIA - Sophia Antipolis, France, working on Internet measurements. He has been (Oct.-Nov. 2013) a visiting researcher at LAAS-CNRS, Toulouse, France, working on Machine-to-Machine communications. His main research interests include also: resource allocation in packet switching networks and Future Internet architectures. He authored more than 100 papers, published on international venues of great renown. Currently, he is an Editor for the IEEE TVT journal (Top Associate Editor 2012) and an Executive Editor for the ETT journal, Wiley.