Computer Networks xxx (2015) xxx–xxx

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

Q1 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 xxx

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

Please cite this article as: N. Accettura et al., The Capture-Recapture approach for population estimation in computer networks, Computer Networks (2015), http://dx.doi.org/10.1016/j.comnet.2015.07.013