A New Statistical Approach to Estimate Global File Populations in the eDonkey P2P File Sharing System

Sanja Petrovic
Orange Labs,
Sophia-Antipolis, France
Email: petrovic.sanja@gmail.com

Patrick Brown
Orange Labs,
Sophia-Antipolis, France
Email: patrick.brown@orange-ftgroup.com

***Abstract*—In this paper, we propose a new statistical approach, also known in biology under the name *capture-recapture methods* in order to estimate global population statistics from local observations. Evaluating population sizes in P2P systems has received much attention lately as these may be useful to set system parameters, to derive other system statistics or to predict system performance. As these systems are very large, encompassing several millions of users and since they are highly distributed estimating population sizes is a challenging task. More precisely we are interested in estimating the number of file replicas in the system, i.e., the size of the population of users possessing given files. To this end we propose a capture-recapture method which is both computationally efficient and accurate. The method proposed allows deriving global population statistics from local and time limited observations. We apply the method on a measurement data set of several days on a residential network. We compare the results obtained from direct counting procedures with those derived with the proposed methodology.**

I. Introduction

The estimation of size of populations is an important problem studied in P2P literature. The estimation of these sizes in distributed applications can be very useful for setting parameters or for monitoring purposes. But, the problem turns out to be very challenging since these systems can be very big with several millions of users, and highly distributed without any global information.

A means of estimating the size of the population possessing a given file is to use search engines or crawlers (see e.g. [1], [2], [3]). These systems are usually based on inspecting the clients' content or inspecting the servers' databases. However, considering the actual size of the eDonkey network, which implies several millions of clients (4-5 millions) with several hundreds of servers, such methods are very intensive in resource utilization. Also servers try to protect themselves from these systematical queries and will stop answering to them.

Other methods observe traffic at one point in the network and base their evaluation on the *direct counting* of the observed users having the considered file in order to estimate relative file populations (see e.g. [4]). However the resulting evaluations may be biased. Direct counting will give close to complete population levels for less popular files while the evaluation will be strongly underestimated for popular files. Extending the observation period may allow observing larger populations but this will be at the cost of detecting population size changes during the observation period.

As an alternative to these methods, we propose in this paper a statistical approach for estimating the size of the population from a limited number of local observations. Our approach is based on *capture-recapture methods*, which were originally developed in the context of wildlife biological studies [5], [6]. Instead of basing the estimation solely on the union of all peers observed possessing a given characteristic, the idea is to derive supplementary information from set intersections. The sets are interpreted as random samples. The general idea of capture-recapture methods is that, by taking several population samples, and based on information of how many individuals were re-sampled and how many of them were not, it is possible to obtain an estimate of the size of the global population.

We applied the proposed approach on our measurement data acquired in the residential network in the Nice area (France).

Since our measurement data were limited to a subset of eDonkey users only, and since users communicate with a limited number of other peers, we could observe only a limited part of the complete eDonkey population. The proposed approach helped us to overcome this limitation, with limited time and resources, and to derive global information concerning file populations and their dynamics. Note that we focus on the eDonkey P2P file sharing system but the method can be easily generalized to all other distributed P2P file sharing systems.

The paper is organized as follows. In section II we first introduce the capture-recapture principle followed by a brief review of its early applications in computer science. In section III we describe a set of capture-recapture methods which allow deriving increasing estimation accuracy with the number of peers downloading or uploading a given file. These methods are used in section IV on the traffic observations performed in Nice and the obtained results are discussed.

II. The capture-recapture methodology

A. Foundation

Although the first recorded use of this method has been attributed to Laplace, the more systematic use of capture-recapture methods has been introduced by ecologists in the end of the 19th century as a means of estimating the size of wildlife populations. More recently, it became popular in other areas such as epidemiology and social sciences. In epidemiology,

typical applications include estimating illegal drug addicts, people infected with HIV, and so on. While in demography, it has been used to estimate the size of hard-to-count human populations.

The foundation of capture-recapture methods is in a simple probability problem called the *birthday paradox* [7] which evaluates the probability for two samples taken from a common set to share a common element. This probability is non negligible even with not so moderate size samples. As an example, in a group of 23 people the probability that two persons have the same birthday date is at least $\frac{1}{2}$.

Contrarily, the inverse problem, called the *inverted birthday paradox* [7], estimates the size of the population based on the number of identical samples observed between two different sample sets, where samples are taken uniformly and at random. The inverse problem forms the basis for more complex capture-recapture methods which estimate the size of the population based on more than two sample sets. The efficiency of the method relies on the non negligible probability of finding repeated samples.

There exist many different capture-recapture models that can be classified into two groups: open population models and closed population models. The population will be open if its composition may significantly change during the observation period due to new arrivals or definitive departures. If not the population will be closed.

The simplest capture-recapture method, the *Petersen method* [5], applies to two capture sets performed on the same global population. In the first one, M_1 individuals are observed and marked so as to be recognized later on, then they are returned to the global population. Later, a second capture is performed with C_2 individuals of which R_2 are recaptures, i.e., they were marked in the the first capture. Assuming that the population is closed, that the both observations (series) are random samples, and that all individuals have the same probability to be observed, we should have:

$$\frac{M_1}{\widehat{N}} = \frac{R_2}{C_2}, \tag{1}$$

where $\widehat{N}$ is the population estimate. Equation (1) simply states that the proportion of marked individuals in the total population equals the proportion of marked individuals in the C_2 population. So, we have:

$$\widehat{N} = \frac{C_2 M_1}{R_2}. \tag{2}$$

We will see in section III that this estimation corresponds to the maximum likelihood estimate. We will also see how one can obtain more accurate estimates by using more than two sets of samples.

B. Methods related to the capture-recapture methods in computer science

Capture-recapture methods have not been very much exploited in computer science although brute counting is often resource intensive. We give two recent examples where such methods have been proposed. The first example consists in using the inverted birthday paradox to estimate the total number of peers in a distributed P2P network. In [8] and [9] the authors use the number of peers encountered in a random walk before a peer is seen for the second time to estimate the global population size. In [9] the accuracy of the estimator is improved by continuing the random walk until a predefined number of peers are reencountered. The second example [10] consists in proposing to use capture recapture models for open populations to estimate appearance and lifetime of web pages. Both examples were tested on simulations but not run on real systems to estimate populations sizes.

III. Deriving file population sizes

The eDonkey network consists of users who gather to exchange different files. Each user joins the network with requests to download one or several files and, in the same time, the user can share several other files. With respect to a given file we may define two populations of users: the population of users that have an incomplete copy (i.e., *population of downloaders*), and the population of users that have the complete file (i.e., *population of seeds*). Note that due to some practical constraints, a user wishing to download a file f can only contact a limited number of users sharing that file even if there exists a large population. The subset of contacted users is random, while its size is influenced by limitations of the user's operating system and by the number of other files that he is downloading. The randomness of the subset of contacted peers is a necessary condition for the system to efficiently share the load on the different uploaders.

We are interested in estimating the correct order of magnitude of the peer population having a given file f in the eDonkey network, for f covering the largest possible proportion of shared files. Due to our measurement setting, we are able to observe only users in a specific geographical area and their relevant eDonkey messages (both sent and received) with all other users in the eDonkey network. These messages allow us to determine subsets of distant users with respect to the file their message concerns and with respect to the local peer with which they are communicating.

In this section we show how to apply the capture-recapture methodology to estimate global files population from these local observations.

A. Assumptions, notations and principle

In this paper, we assume that:

($\mathcal{H}1$) The population is closed.

($\mathcal{H}2$) The observations are independent, i.e., there is an equal probability for each user outside Nice to be observed.

Although eDonkey peers arrive and depart from the system, their presence in the system lasts typically several hours or days. Since we estimate populations over short time periods, typically hours, we will consider that assumption ($\mathcal{H}1$) is justified. Assumption ($\mathcal{H}2$) is difficult to verify. We have not noticed evident correlations between samples such as alternations of non overlapping and highly overlapping samples.

An a posteriori confirmation of $(\mathcal{H}2)$ is that long term brute counting of populations converges to our estimations. Note also that choosing random peers for download is a guarantee of efficient resource usage in the network and a highly plausible strategy of eDonkey servers.

Let us now consider a given file f and the users in Nice interested in f (whether it be for download or upload). With respect to our observations these are the users in Nice which we have seen exchanging information concerning f with other users during the period of observation. For each such user *u* in Nice we can define a sample as the set of users outside of Nice (i.e., *non-local* users) that communicate with *u* concerning file f. So, we introduce the following notations:

- N = The size of the file population that we wish to estimate;
- $\widehat{N}$ = The estimated file population size;
- $S+1$ = The number of samples (i.e. the numbers of users in Nice having exchanged pertinent information concerning file f).

We may arbitrarily order the samples and index them starting from 0. Then for each sample i in $1, 2, .., S$, it is possible to define the following notations:

- M_i (as ***M**arked*) = The total number of non-local users observed in the previous samples $1, 2, .., i-1$;
- C_i (as ***C**aptured*) = The number of non-local users observed in the i-th sample;
- R_i (as ***R**ecaptured*) = The number of non-local users in the i-th sample that have already been observed in one of the previous samples $1, 2, .., i-1$.

Under the assumptions $(\mathcal{H}1)$ and $(\mathcal{H}2)$, the probability that R_i users were observed previously among C_i, conditioned on population size (N), is given by the hypergeometric distribution [5]:

$$P(R_i|N) = \left(\begin{array}{c} M_i \\ R_i \end{array} \right) \left(\begin{array}{c} N-M_i \\ C_i-R_i \end{array} \right) / \left(\begin{array}{c} N \\ C_i \end{array} \right). \quad (3)$$

Under the assumption $(\mathcal{H}2)$ the probability of observing a sequence R_i $i = 1, 2, .., S$ of *recaptures* in a sequence C_i $i = 1, 2, .., S$ of *captures* conditioned on N is then:

$$P(R_1, R_2, ..., R_S|N) = \prod_{i=1}^{S} P(R_i|N). \quad (4)$$

Equations (3) and (4) are the likelihood of observing R_i, respectively $R_1, R_2, ..., R_S$ recaptures given N.

A classical approach to estimate population sizes from recapture statistics is to rely on the maximum likelihood estimator, i.e. the unknown parameter N is estimated as the one maximizing the likelihood. The Petersen estimate of Equation (2), can be shown to maximize the likelihood (3) in the case $S = 1$.

Before going on let us note that the hypergeometric distribution (3) may be approximated by the binomial distribution when the true size of the population is large compared to the observed population:

$$P(R_i|N) \approx \left(\begin{array}{c} C_i \\ R_i \end{array} \right) \left(\begin{array}{c} \frac{M_i}{N} \end{array} \right)^{R_i} \left(\begin{array}{c} 1-\frac{M_i}{N} \end{array} \right)^{C_i-R_i}. \quad (5)$$

This consists in assuming that the recapture probability stays equal to $\frac{M_i}{N}$ all through sample i, i.e. sampling is performed with replacement.

The ratio $\frac{R_i}{C_i}$ gives an estimate of the recapture probability $\frac{M_i}{N}$. Using the binomial distribution (5) we obtain the average and the variance of this estimate:

$$E\left[\frac{R_i}{C_i}\right] = \frac{M_i}{N},\ var\left[\frac{R_i}{C_i}\right] = \frac{M_i}{NC_i}\left(1-\frac{M_i}{N}\right). \quad (6)$$

We present three methods to estimate population sizes from more than two sample sets: a Bayesian method by *Gazey and Staley* and two linear regression methods: the *Schnabel's binomial model* and the *Schumacher and Eschmeyer's regression model* (see [11], [12]). The method we propose will consist in obtaining a preliminary estimation by a linear regression method, then using this result to efficiently calculate a more accurate estimation by a Bayesian method.

B. Bayesian model

The idea of the *Gazey and Staley Bayesian method* [13] is to assume an a priori probability distribution for N. The posterior distribution is then evaluated using the conditional probabilities (4) and Bayes' theorem. In the absence of additional information a discrete uniform distribution may be used as a priori distribution. In [13] the possible population size values are restricted to K regularly distributed discrete values $(N_k,\ k = 1, 2, .., K)$, thus:

$$P(N_k) = K^{-1}, \quad (7)$$

with the condition that the smallest population size must be greater than or equal to the total number of observed individuals, i.e., $min(N_i) \geq M_S + C_S - R_S$.

Applying Bayes' theorem with equations (4) and (7), we obtain:

$$\begin{aligned} P(N_k|R_1, R_2, .., R_S) &= \frac{P(N_k)P(R_1, R_2, .., R_S|N_k)}{\sum_{l=1}^{K} P(N_l)P(R_1, R_2, .., R_S|N_l)} \\ &= \frac{\prod_{i=1}^{S} P(R_i|N_k)}{\sum_{l=1}^{K} \prod_{i=1}^{S} P(R_i|N_l)}. \end{aligned} \quad (8)$$

In [13], the authors apply Bayes' theorem in S successive steps, using posterior distribution calculated in step $i-1$ to estimate posterior distribution in step i based on information from the ith sample. At step i:

$$P_i(N_k) = \frac{P(R_i|N_k)P_{i-1}(N_k)}{\sum_{l=1}^{K} P(R_i|N_l)P_{i-1}(N_l)}, \quad (9)$$

for i=1,..,S and k=1,2,..,K, with $P_0(N_k) = \frac{1}{K}$ and $P_i(N_k) = P(N_k|R_1, R_2, .., R_i)$. The advantage of this method lies in the explicit derivation of the distribution of N from which $\widehat{N} = E(N)$ and confidence intervals may be derived. Also, the plots of the successive posterior distributions can be used as a visual diagnostic to verify the assumption $(\mathcal{H}1)$ when samples are taken in successive periods of time. However, this method requires an intensive computation and one must choose the K discrete possible values for N. We will present an efficient implementation of a Bayesian method. We next present two methods based on linear regressions.

C. Linear regression models

The two models presented therein are based on a linear regression of the samples.

The Petersen model expresses that the ratio $\frac{R_i}{C_i}$ gives an estimate of the recapture probability $\frac{M_i}{N}$. Both methods presented consist in estimating $\frac{1}{N}$ with a regression analysis. Plotting $\frac{R_i}{C_i}$ with respect to M_i we should obtain points scattered around a line (ideally passing through the origin) with slope $\frac{1}{N}$:

$$\frac{R_i}{C_i} = \frac{1}{N} M_i + error_i.$$

According to the least square approximation, we estimate the slope ($\frac{1}{N}$) as the value minimizing the square of errors $\sum \left(\frac{R_i}{C_i} - \frac{1}{N} M_i \right)^2 = \sum error_i^2$:

$$\hat{N}^{-1} = \sum_{i=1}^{S} \frac{R_i}{C_i} M_i / \sum_{i=1}^{S} M_i^2.$$

This estimator assumes that the errors have the same variance. If it is not the case one may choose to perform the least-square fitting of the straight line with weights w_i. In which case the estimator of $1/N$ and its variance are given by:

$$\hat{N}^{-1} = \sum w_i \frac{R_i}{C_i} M_i / \sum w_i M_i^2, \tag{10}$$

$$var\left(\frac{1}{\hat{N}}\right) = \frac{\sum w_i \left(\frac{R_i}{C_i}\right)^2 \sum w_i M_i^2 - \left(\sum w_i \frac{R_i}{C_i} M_i\right)^2}{\left(\sum_{i=1}^{S} w_i M_i^2\right)^2 (S-1)}. \tag{11}$$

The methods considered differ in the chosen weights.

Schnabel's method: : Schnabel's evaluation can be obtained by setting the weights to the inverse of the variance (6). Thus the variances of all errors $w_i \times error_i$ are equal. In practice the weights are set to $w_i = C_i/M_i$ assuming $M_i \ll N$. Note that it is the relative values of the weights which are important so that the unknown factor N in Equation (6) is not needed to set the weight values. The Schnabel estimate for $1/N$ is:

$$\hat{N}^{-1} = \sum R_i / \sum C_i M_i.$$

Schumacher and Eschmeyer's regression method: In the Schumacher and Eschmeyer's method [14] the weights are set to C_i. There may be a tendency for tagged groups of peers to be grouped or clustered. This method is recommended for use when such departures from randomness are probable (i.e., when ($\mathcal{H}2$) is not completely satisfied). The estimate is then:

$$\hat{N}^{-1} = \sum R_i M_i / \sum C_i M_i^2.$$

The variance estimator of the reciprocal of the population density ($\frac{1}{N}$) is given by Equation (11) with $w_i = C_i$.

The linear regression methods rely on general statistical properties and the validity of the confidence limits which may be obtained from the expressions of variance (11) require a large number of samples. We choose to use these methods to obtain orders of magnitude for the following Bayesian method which we propose in the context of eDonkey population estimations.

D. Proposed method

In this section we propose to use a Bayesian method to evaluate population sizes and to obtain reliable confidence intervals. We derive some properties of the the estimation which justify its usage and show it may be efficiently implemented. In contrary to the method proposed in [13] we consider all possible values of population sizes for the a priori probability. The expression used to calculate the probability of the population being size n is then:

$$P(n|\overline{R}) = \frac{r(n)}{\sum_{k=N_{min}}^{\infty} r(k)}. \tag{12}$$

where we set $N_{min} = M_S + C_S - R_S$, $P(n|\overline{R}) = P(n|R_1, R_2, .., R_S)$ and $r(n) = \prod_{i=1}^{S} P(R_i|n)$. Note that the probabilities (12) but also (8) and (9) are completely defined by the fact that for any two $n, k \geq N_{min}$, $\frac{P(n|\overline{R})}{P(k|\overline{R})} = \frac{r(n)}{r(k)}$ and the probabilities sum to one. This property allows to prove the following:

Proposition 1. *Both Bayesian estimations obtained by direct calculation with Equation (8) and with an iterative calculation with Equations (9) produce the same probabilities.*

In the context of the eDonkey samples which are collected over identical time periods there is no advantage to calculate the probability distributions iteratively. To derive our results we will use a direct calculation (using Equation (12)) starting from uniform a priori probabilities over all integers. This is justified by the following property.

Proposition 2. *The distribution of population sizes defined by Equation (12) may be calculated if and only if at least two recaptures have been observed, i.e.* $\sum_{i=1}^{S} R_i \geq 2$.

A necessary and sufficient condition for the existence of $P(n|\overline{R})$ is the convergence of the denominator of (12). From the definition of $r(n)$ and the asymptotic expression (5) for $P(R_i|n)$ we obtain $r(n) = O(n^{-\sum_{i=1}^{S} R_i})$. This proves the proposition. This is a weak condition as we hardly expect to be able to estimate correctly the total population size if less than two recaptures are detected.

We only need to calculate the $r(n)$ by a constant factor. Note that $C_i = M_{i+1} - M_i + R_i$. Expressions $r(n)$ may then be efficiently calculated iteratively by noting that for each i:

$$\frac{P(R_i|n)}{P(R_i|n-1)} = \frac{(n-M_i)(n-C_i)}{n(n-M_i-C_i+R_i)} = \frac{(n-M_i)(n-C_i)}{n(n-M_{i+1})}$$

Finally we obtain:

$$\frac{r(n)}{r(n-1)} = \prod_{i=1}^{S} \frac{P(R_i|n)}{P(R_i|n-1)} = \frac{n}{n-M_{S+1}} \frac{\prod_{i=0}^{S}(n-C_i)}{n^{S+1}}$$

This proves the following proposition:

Proposition 3. *The order in which the samples are numbered does not influence the posterior probabilities* $P(n|\overline{R})$.

From the last equation we may show (see the Appendix):

Proposition 4. *The probabilities $P(n|\overline{R})$ grow to a maximum value then decline to zero as n tends to infinity.*

For numerical evaluations, to avoid overflows we set $r(\widehat{N}_{SE}) = 1$ where $\widehat{N}_{SE}$ is the estimation obtained from the *Schumacher and Eschmeyer's regression method*, assuming $\widehat{N}_{SE}$ is close to the index n_0 maximizing $r(.)$.

The expressions $r(n)$ are calculated for all n between N_{min} and N_{max}, where N_{max} is obtained by the approximate convergence test $r(n)/\sum_{N_{min}}^{n} r(i) < \varepsilon$ for a precision ε. Recall that $r(n)$ decreases faster than n^{-2}. The resulting computing time for estimating the population sizes on a one hour observation period (i.e. 2000 files) on a 1.60GHz computer is less than one minute.

IV. Experiments

A. Measurement methodology

Our measurements were performed on a residential ADSL network with 20000 users in a subset of the Nice area (France). We defined these users as *local* users, and *non-local* or *distant* users the remaining hosts. We performed passive measurements. The analysis was applied on the saved data off-line. Inspecting only specific eDonkey signaling messages, we may observe information concerning which files users possess, if they possess a complete copy (in which case they are seeds) or a partial copy (in which case they are downloaders) also which files they were requesting. We may have incomplete information as we rely on the messages peers decide to exchange. Note that data is made anonymous and there is no stable storage of connections after the analysis.

We conducted several measurements and chose the representative data sets done during four days (from Oct 25, 2007 to Oct 29, 2007). For this data set, we identified 28793 different files (i.e., files requested by users). In order to apply capture-recapture methods, we need to have at least two peers in Nice requesting the same file. Thus, the methods we propose can be applied to 10174 files considering the given data set.

B. Estimating the number of downloaders and seeds with capture-recapture methods

We used the proposed Bayesian capture-recapture method presented in Section III-D to estimate the number of downloaders and seeds for files in the system. For each file a sample is the set of distant peers in contact with a peer from Nice to upload or download the file. We further distinguish seeds from downloaders.

The estimated numbers of downloads and seeds per file in the eDonkey network are shown on Figure 1, where x and y denote respectively the number of downloaders and the number of seeds of the same file. We note a correlation between the number of seeds and the number of downloaders: file storage is related to file requests. However we note that the dispersion may be important.

In order to analyze the obtained result, we represent these estimates differently. In Figure 2 we represent the distribution of the values of the ratio $\frac{y}{x}$ for each file, i.e. the seeders

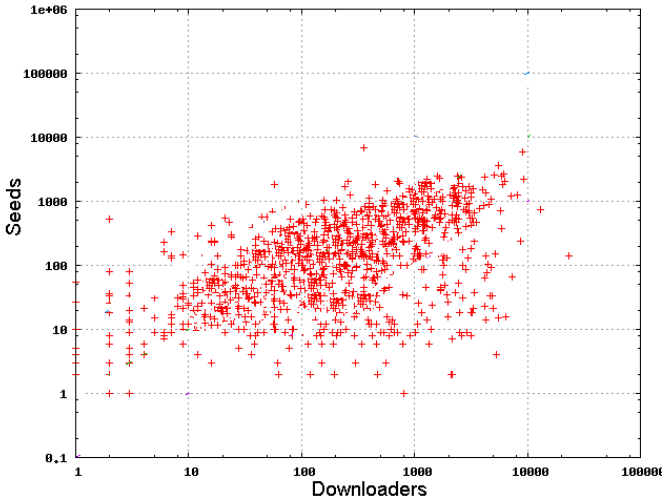

Fig. 1. Number of seeds (y) with respect to number of downloaders (x). This plot correspond to the data set captured Oct 28, 2007 between 7PM and 8PM.

to downloaders ratio. The typical life cycle of a file usually comprises three phases which may have various lengths: A starting phase for newly introduced files with less seeds than downloaders (let's say $\frac{y}{x} < 0.1$), followed with a phase when these numbers are of the same order ($0.1 < \frac{y}{x} < 10$), and eventually it finishes with a phase with more seeds than downloaders ($\frac{y}{x} > 10$) corresponding to *old* files sufficiently appreciated to be kept on users' disks.

Let us compare the distributions obtained during different lengths of time, i.e. one hour and one day. On longer times of observation, we notice a decrease of the percentage of the first group with more downloaders, x, than seeds, y. This can be explained by the fact we detect files with high download rates on short time intervals because they produce frequent signalling messages. On the contrary we need longer observations times to detect signalling messages concerning files rarely downloaded. This seems to indicate we correctly estimate a file population when we can observe exchanges concerning the file. Short time periods however do not give us a clear indication on the number of non popular files.

On Figure 3 we represent the distribution of the total file population including both downloaders and seeds ordered in decreasing file rank. Here again we see that the estimations after four hours of observation do not give larger population estimations than those derived from the 8 pm peak hour. The estimations seem reliable concerning the files observed on the period. We may expect that the differences of estimation from one hour to the other give a reliable estimation of the populations changes over the period. This shows the advantage of the proposed method which allows deriving global file populations sizes over short time periods.

Note also that similarly to previous studies [15], [4], we obtain that the distribution of file population in the eDonkey system is not Pareto distributed. A starting flat region is followed by a fast decrease, i.e., the popularity of the most popular files is of the same order of magnitude whereas the popularity of less popular files declines very fast.

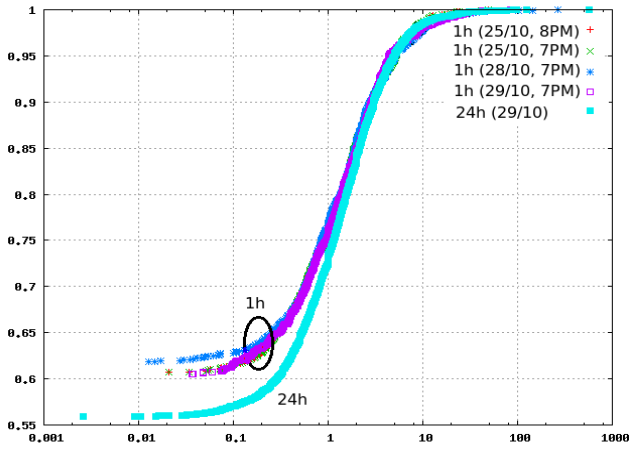

Fig. 2. Distribution of the ratios of seeds and downloaders per file ($\frac{y}{x}$) as a function of a file rank.

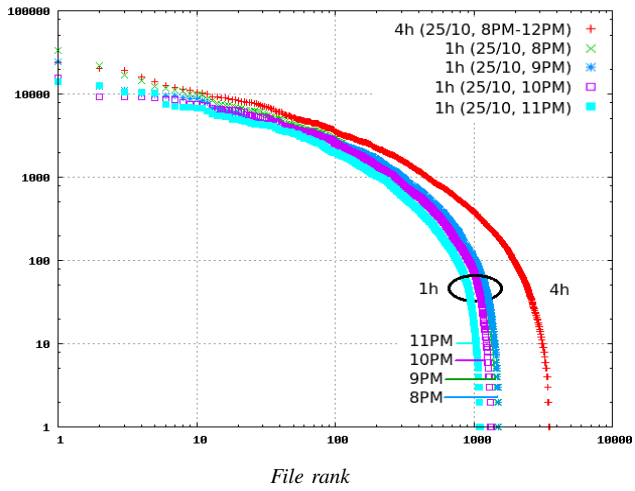

Fig. 3. Distribution of the sum of downloaders and seeds per file ($x + y$) as a function of a file rank.

C. Confidence intervals

In Figures 4 and 5 we plot the population size estimations and the 95% confidence intervals obtained using the Bayesian method and compare them to population sizes obtained by counting non local users. The confidence intervals depend on the number of peers in Nice uploading or downloading the file and on the number of recaptures. In Figure 4 we plotted results for files which interested at least three peers in Nice. In Figure 5 we plotted results for files which interested only two peers in Nice and for which more than three recaptures were observed. We observe that confidence intervals are usually relatively small: orders of magnitude are correctly evaluated. However isolated large confidence intervals appear at all popularity levels. The method requires a sufficient number of recaptures to constrain confidence intervals. We conclude that as soon as more than three users share the same file in Nice we can estimate the population size quite precisely without the need for extensive observations, while we only obtain an order of magnitude if the file is shared by two people in Nice during the observation period.

We observe that for the most popular files simple counting underestimates the true size of the population by a factor of 100 or more. Also simple counting results do not fall inside confidence intervals. Even for less popular files the accuracy of counting is very variable. Thus observing a small number of (non local) peers does not guarantee a precise evaluation of the peer population by simple counting.

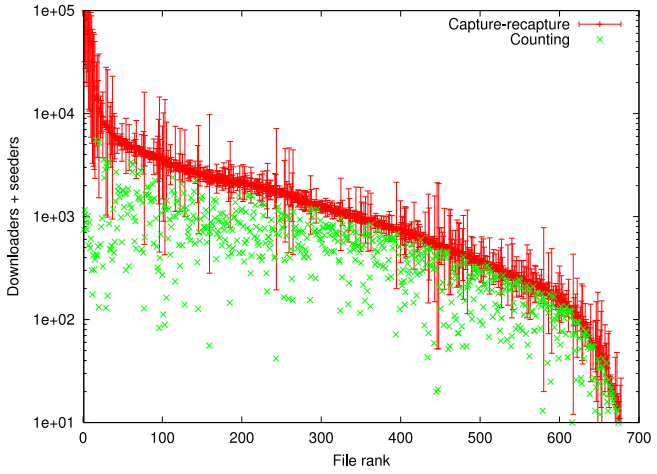

Fig. 4. Estimation of downloaders (x) plus seeds (y) with 95 % confidence intervals for files shared by more than two users in Nice, and comparison with direct counting. One-hour data at $7PM$, Oct 28, 2007.

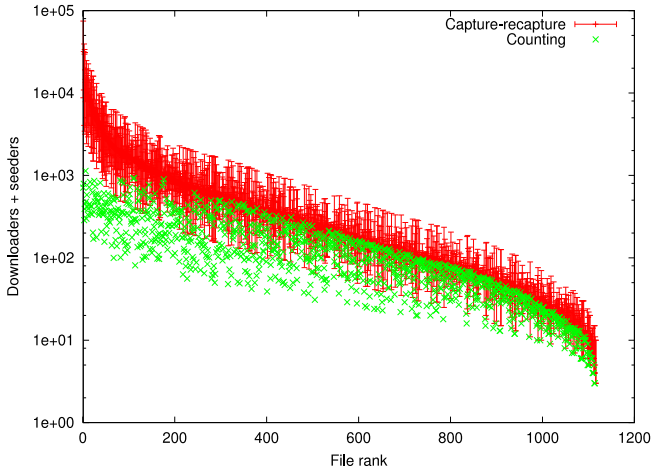

Fig. 5. Estimation of downloaders (x) plus seeds (y) with 95 % confidence intervals for files shared by two users in Nice, and comparison with direct counting. One-hour data at $7PM$, Oct 28, 2007.

D. Number of downloaders and seeds for different files as a function of time

By calculating the number of downloaders and seeds for each file and for each hour, and by using the proposed approach, one can obtain the evolution of the number of downloaders and seeds of a particular file during the observation time interval. An insight on the dynamics of file populations can be obtained.

In Figure 6 and Figure 7 show the population evolution for two popular files at the moment of the capture which lasts from Friday 26th of October 2007 evening to Monday 29th.

Both files represent films with approximate size 700 MBytes. In Figure 6 we show the current number of users in Nice (A) and the estimated number of users in the network either downloading or possessing the complete file. Note that we use the left axis labels for (B) which reaches a peak of 10000 users on Sunday evening. On the same period 300 users have the file in the subset of Nice users.

In Figure 7 we show for the second file the total number of users (seeds and downloaders) observed in Nice, (A), as well as the number of estimated non local downloaders (B) and seeds (C). Here again we observe a peak on Sunday evening. In addition we observe that the number of seeds in the network is clearly increasing which is a possible indication that the file has been recently introduced and is in its expansion phase.

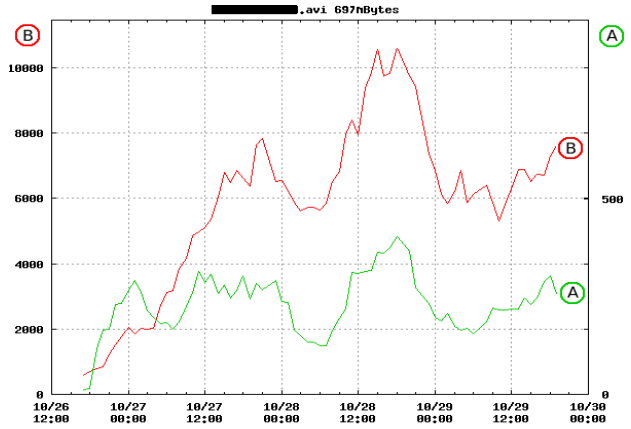

Fig. 6. Example of a film with a size of 700 MBytes. (A) The file population in Nice. (B) The estimated file population in the eDonkey network.

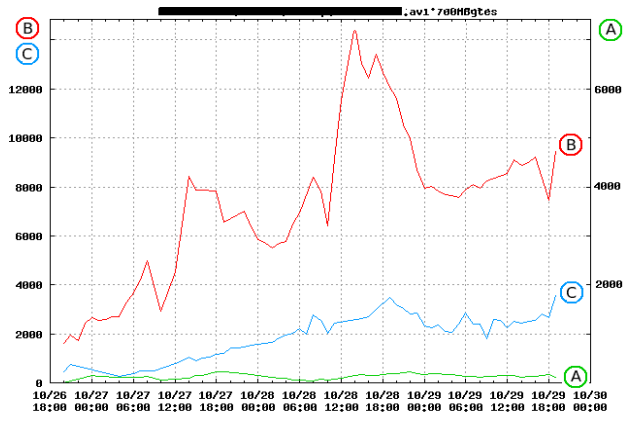

Fig. 7. Example of a film with a size of 700 MBytes. (A) The file population in Nice. (B) The estimated population of downloaders of the file in the eDonkey network. (C) The estimated population of seeds of the file in the eDonkey network.

E. Comparing the Capture-recapture method with Counting

In Figure 8 we compare the capture-recapture estimation with simple counting results. The figure shows the distribution of population sizes over two one-hour periods (i.e. $12AM$ and $7PM$) and over a 24 hour period encompassing the two previous periods. The results concern downloaders. They reveal that on short observation periods, it is not possible to obtain global file population sizes by counting. Direct counting results seem to suggest that population sizes were of the same order at $12AM$ and $7PM$. Capture-recapture methods show that the population size is in fact twice larger at $7PM$ at least for popular files. We see by counting peers on a 24 hour period we obtain a slight underestimation of population sizes. As mentioned previously, for capture-recapture methods, long observation periods reveal a larger number of less popular files while popular file populations are correctly estimated over short and partial observations. Thus a single daily (peak) hour observation over a limited number of peers is sufficient to trace population size trends over long periods at least for popular files.

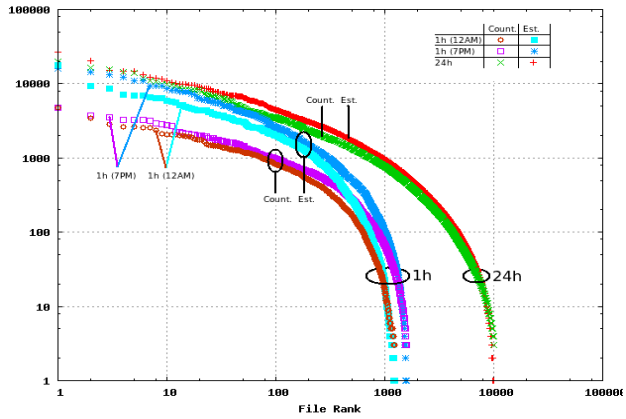

Fig. 8. File request popularity distribution: results comparison of the estimation method and the direct counting method

On Figure 9, we compare population sizes obtained by counting and by our estimation on an hourly basis over four days for a given file. We see that a simple scaling will not reproduce the global population fluctuation. In particular the counting method misses the evening peak traffic on the first day. In the same way the counting method does not detect the daily variation in the last 24 hours while it appears clearly with the capture-recapture estimate.

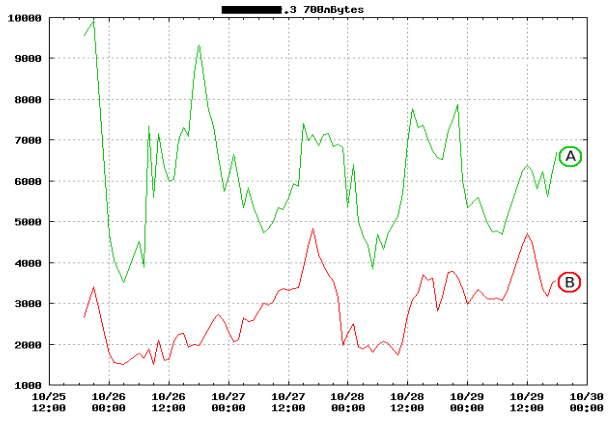

Fig. 9. Difference between number of file replicas for a given film in the network obtained with capture-recapture method (A), and with the direct counting method (B).

These comparisons show the advantage of the proposed

capture-recapture methodology especially for short observation periods on partial observations. This approach is less demanding in the terms of memory requirements and volume of traffic analyzed. Also, population fluctuations may be studied on short time scales.

V. Conclusion

In this paper, we propose a new statistical approach, also known in biology under the name *capture-recapture methods* in order to estimate global population statistics from local observations. The proposed approach consists in a capture-recapture method based on Bayes' theorem which offers a good compromise between cost and accuracy. We apply the proposed approach on our measurement data set obtained on a residential network.

The comparison with existing solutions in practice shows the advantage of the proposed capture-recapture approach especially for short observation periods. The analysis on our traffic observations show that simple counting procedures may produce underestimated populations sizes by more than two orders of magnitude for most popular files. For less popular files the difference seems unpredictable. Thus the proposed method seems useful even for estimating relative population sizes. Population statistics may thus be obtained with less memory requirements and by analyzing smaller volumes of traffic. Also, the possibility of obtaining global population sizes in short observation periods makes it a promising approach to study the dynamics of population sizes.

Estimating the global population statistics can be an end in itself but it may also be useful for setting parameters, controlling the system behavior or as input for performance models. In [16] population statistics were used both to estimate downloading times observed on real systems and as input to a model.

We feel these methods which have been used extensively to estimate living populations are a useful tool to estimate *populations* in computer science where the appearance of very large distributed systems raises obstacles against brute force counting procedures.

Appendix

Proof of proposition 4. Denote $M = M_{S+1}$. We prove that the ratio $\frac{r(n)}{r(n-1)}$ is larger than one below a threshold n_0 and smaller beyond. It is sufficient to prove that for $x \in [0, \frac{1}{M}[$

$$f(x) = \ln\left(\frac{1}{1-xM}\prod_{i=0}^{S}(1-xC_i)\right)$$

is negative in a neighborhood of 0 and changes sign only once. Note that $\lim_{x\to 0^+} f(x) = 0$. Also:

$$f'(x) = \frac{M}{1-xM} - \sum_{i=0}^{S}\frac{C_i}{1-xC_i}.$$

We have $f'(0) = M - \sum_{i=0}^{S} C_i < 0$ as the number of different peers obtained in the end, M, must be smaller than the sum of peers obtained in each sample set, $\sum_{i=0}^{S} C_i$. This proves that $f(x) < 0$ on some interval $]0, \varepsilon[$. As $\lim_{x\to\frac{1}{M}} f(x) = \infty$, $f(x)$ changes sign at least once. We now prove that $f(x)$ is convex at any point where its derivative is positive or null. The derivative must be positive or null when $f(x)$ becomes positive for the first time. Thus its derivative will stay positive from then on which will finish the proof. To prove convexity at points of positive derivative we write:

$$f"(x) = \left(\frac{M}{1-xM}\right)^2 - \sum_{i=0}^{S}\left(\frac{C_i}{1-xC_i}\right)^2.$$

Let $a = \frac{M}{1-xM} > 0$ and $b_i = \frac{C_i}{1-xC_i} > 0$. Then:

$$f'(x) > 0 \Rightarrow a > \sum_{i=0}^{S} b_i \Rightarrow a^2 > (\sum_{i=0}^{S} b_i)^2 > \sum_{i=0}^{S} b_i^2 \Rightarrow f"(x) > 0.$$

References

[1] S. Handurukande, A. Kermarrec, F. L. Fessant, L. Massoulié, and S. Patarin, "Peer sharing behaviour in the edonkey network, and implications for the design of server-less file sharing systems," in *EuroSys'06*, Leuven, Belgium, April 2006.

[2] D. Stutzbach and R. Rejaie, "Understanding churn in peer-to-peer networks," in *Internet Measurement Conference*, October 2006.

[3] M. Steiner, E. W. Biersack, and T. En Najjary, "Actively monitoring peers in KAD," in *IPTPS'07, 6th International Workshop on Peer-to-Peer Systems*, Bellevue, USA, February 2007.

[4] L. Plissonneau, J.-L. Costeux, and P. Brown, "Detailed analysis of edonkey transfers on adsl," in *2nd EuroNGI Conference on Next Generation Internet Design and Engineering*, Valencia, Spain, April 2006.

[5] G. Seber, *The estimation of animal abundance and related parameters*. 2nd ed.,London: Charles Griffin & Co., 1982.

[6] C. Schwarz and G. Seber, "Estimating animal abundance: review III," *Statistical Science*, vol. 14, pp. 427–56, 1999.

[7] W. Feller, *An Introduction to Probability Theory and Its Applications*. 3rd ed.,New York: Wiley, 1968, vol. 1.

[8] M. Bawa, H. Garcia-Molina, A. Gionis, and R. Motwani, "Estimating aggregates on a peer-to-peer network," Dept. of computer science, Stanford University, Technical report, 2003.

[9] L. Massoulié, E. L. Merrer, A.-M. Kermarrec, and A. Ganesh, "Peer counting and sampling in overlay networks: random walk methods," in *PODC '06: Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*, New York, NY, USA, 2006, pp. 123–132.

[10] I. Anagnostopoulos, P. Stavropoulos, G. Kouzas, C. Anagnostopoulos, and D. D. Vergados, "Estimating the evolution of categorized web page populations," in *ICWE '06: Workshop proceedings of the sixth international conference on Web engineering*. New York, NY, USA: ACM, 2006, p. 13.

[11] W. E. Ricker, *Computation and interpretation of biological statistics of fish populations*. Fisheries Research Board of Canada, 1975, vol. Bulletin 191.

[12] C. J. Krebs, *Ecological Methodology*. New York: Harper and Row, 1989.

[13] W. Gazey and M. Staley, "Population estimation from mark-recapture experiments using a sequential bayes algorithm," *Ecology*, vol. 67, pp. 941–951, 1986.

[14] F. X. Schumacher and R. W. Eschmeyer, "The estimate of fish population in lakes or ponds," *Journal of the Tennessee Academy of Science*, no. 18, pp. 228–249, January 1943.

[15] F. L. Fessant, S. B. Handurukande, A.-M. Kermarrec, and L. Massoulié, "Clustering in peer-to-peer file sharing workloads," in *IPTPS*, ser. Lecture Notes in Computer Science, vol. 3279. Springer, 2004, pp. 217–226.

[16] S. Petrovic, "Towards a better understanding of emule," Ph.D. dissertation, University of Nice–Sophia Antipolis, 2008.