# The estimation of the number of lost multi-copy documents: a new type of informetrics theory



EGGHE, Leo & Proot, G. (2007) The estimation of the number of lost multi-copy documents: a new type of informetrics theory. In: JOURNAL OF INFORMETRICS, 1(4). p. 257-268.

DOI: 10.1016/j.joi.2007.02.003
Handle: http://hdl.handle.net/1942/1784

# The estimation of the number of lost multi-copy documents: a new type of informetrics theory

by

L. Egghe. Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium
and
G. Proot. Universiteit Antwerpen (UA), Stadscampus, Prinsstraat 13, B-2000 Antwerpen, Belgium

## ABSTRACT

A probabilistic model is presented to estimate the number of lost multi-copy documents, based on retrieved ones. For this we only need the number of retrieved documents of which we have one copy and the number of retrieved documents of which we have two copies. If we also have the number of retrieved documents of which we have three copies then we are also able to estimate the number of copies of the documents that ever existed (assumed that this number is fixed over all documents). Simulations prove the stability of the model.

The model is applied to the estimation of the number of lost printed programmes of Jesuit theatre plays in the Provincia Flandro-Belgica before 1773. This Jesuit province was an

[1] Permanent address
E-mail addresses: leo.egghe@uhasselt.be, goran.proot@ua.ac.be
Key words and phrases: multi-copy document, book historical law.

administrative entity of the order, which was territorially slightly larger in extent than present day Flanders, the northern, Dutch-speaking part of Belgium.

It is noted that the functional model $P_j$ for the fraction of retrieved documents with j copies is a size-frequency function satisfying $(P_{j+1}/P_j)/(P_j/P_{j-1})<1$ for all j. It is further noted that the "classical" size-frequency functions are different: Lotka's function satisfies the opposite inequality and the decreasing exponential one gives always 1 for the above ratio, hence showing that we are in a new type of informetrics theory.

We also provide a mathematical rationale for the "book historical law" stating that the probability to lose a copy of a multi-copy document (i.e. an edition) is an increasing function of the size of the edition.

The paper closes with some open problems and a description of other potential applications of this probabilistic model.

# I. Introduction

Most printed documents appear in several copies. The number of these copies is usually high. Indeed: books usually appear in hundreds or thousands of copies and the same is true for printed journals. They are spread out over the world (for international literature) or over one country or region (for literature with local interest). Other examples of multi-copy documents are (non-exhaustively!): newspapers, music scores, "In Memoriam"-cards, theatre plays, … . Even non-documentary objects fall in the category of having multiple copies: engravings, etches, pieces of art produced by a factory, furniture, tools, cars, stamps and many other collector's items. To fix the idea we will, however, continue to use the terminology "multi-copy documents".

Typical for these multi-copy documents is that, at the time of production, we have a "complete" set of copies (whatever their amount is) and that from that time on copies can be lost: the further we are away from the production time (i.e. $t = 0$ ) the higher the probability that a copy of a document is lost. Here we look at a cumulative time-period $[0,t]$ and we do

not claim anything about the probability, in a time-period $[t, t + \Delta t]$, that a copy will be lost: this can increase in t, say in the case of old material which has not much value (e.g. newspapers, commercial printings, …) but this probability can also decrease in t at a certain moment as is e.g. the case for precious materials as old printings kept in a museum or a library.

Lost copies do not always imply that the document as such is lost: the latter means that all copies of the document are lost. Dependent on the application we can have that the majority of documents are lost or that lost documents are very rare. Of very precious old printings (e.g. of books) it will occur only very rarely that all copies are lost (or destroyed, e.g. by fire). The topic that lead us to this article is an example where the majority of multi-copy documents are lost, namely in the case of printed programmes of Jesuit theatre plays edited before 1773 in the Provincia Flandro-Belgica. This Jesuit province encompassed secondary schools in 18 cities, which nowadays belong to the Nord de France (Dunkerque, Cassel, Bergues and Bailleul), the southern part of the Netherlands (Breda, Maastricht, 's-Hertogenbosch and Roermond) and present-day Flanders (Aalst, Antwerpen, Bruges, Brussels, Gent, Halle, Ieper, Kortrijk, Mechelen and Oudenaarde).

These printed programmes of theatre plays were not considered as precious (certainly at the time of their performance) and many copies are lost. For a certain number of theatre programmes, all copies are lost or destroyed, which in most cases means that all information about the play itself is lost too. One reason for this can be that, in those times, paper was very scarce and one re-used the paper of many of these printed copies of theatre plays. For more on this historical problem we refer the reader to Proot and Egghe (2007).

This intriguing historical case was the origin for this paper which will treat this problem in a general way: based on "what we have", i.e. some found copies of multi-copy documents, is it possible to predict the number of lost documents, i.e. of which we do not have a single printed copy anymore ? It will turn out that only the knowledge of the number of documents of which we have one copy and the knowledge of the number of documents of which we have two copies, is already enough to estimate the number of lost documents. This will be done in the next section where we will also show that the method is very stable: this will be done by performing simulations of lost copies on a corpus of which we know the size. In the same

section, the model will be applied to the data that we have on found Jesuit theatre plays. Since these plays originally were printed in at least 150 copies (going up to 850 copies – see further) this application shows that in this case the results are almost independent of this (unknown) number a of copies.

The third section is then devoted to establishing a model to estimate this unknown number a of copies (especially for low values of a this is needed in order to apply the model in the second section). It turns out that we now also need the number of documents of which we found three copies: this number is of course known but needs to be large enough in order to yield reliable estimates for the number a.

The fourth section considers the in the second section proved formulae for the fraction $P_j$ of documents of which we have j copies $(j = 1,2,3,...)$. Here we show that this size-frequency function satisfies

$$Q_j = \frac{\frac{P_j + 1}{P_j}}{\frac{P_j}{P_j - 1}} < 1$$

for all $j \geq 2$. This proves that the informetrics theory based on the function $j \to P_j$ cannot be Lotkaian since for these functions we always have $Q_j > 1$. The decreasing exponential function is between these two types of informetrics theories since here we have $Q_j = 1$ for all j. This section concludes that we encountered a new type of informetrics theory.

The fifth section gives a (partial) explanation of the so-called book historical law (see Proot and Egghe (2007) for some historical references) stating that the probability to lose a copy of a document is an increasing function of the size of the edition.

The sixth section formulates some open problems concerning this model and discusses some possible applications to examples of multi-copy documents (or even not-printed objects), which were briefly mentioned in the beginning of this introductory section.

# II. The model

The model is not time-dependent. We suppose that we have a situation of documents of which a copies were produced (printed) some time ago (the precise timing of this is not needed in the model). We do not need to know the exact value of the variable a: we will treat a as a parameter and we will evaluate the results (and the possible need to know the value of a) later. Now we look at the present time and count the number of found documents of which we have i copies $(i = 1,2,3,...)$. Can (some of) these numbers predict the number of lost documents, i.e. documents of which all a copies are lost ?

The used probabilistic methods are elementary and can e.g. be found in Canavos (1984) or Grimmett and Stirzaker (1985).

The basic (unknown) number is p which we define as the probability for a copy to be lost $(0 < p < 1)$. It is the unknown number being the division of the number of lost copies (in total) by the total number of copies that ever existed.

Since p is the probability for a copy to be lost we can, using this unknown number, determine the fraction of the documents of which we still have $j \in \{0,1,2,...,a\}$ copies left. This is denoted by $P_j$ and equals

$$P_j = \binom{a}{j} p^{a-j} (1-p)^j \qquad (1)$$

Note that this formula also comprises the cases where no copies are left, i.e. the fraction $P_0$ of lost documents:

$$P_0 = p^a \qquad (2)$$

and also comprises the case where no copies are lost, i.e. the fraction $P_a$ of documents of which we have all copies:

$$P_a = (1-p)^a \qquad (3)$$

Note that (1) treats all document probabilities since

$$\sum_{j=0}^{a} P_j = 1 \tag{4}$$

as is readily seen.

Note that $P_0$ is the fraction of lost documents (unknown but this is the fraction we are looking for) and that $P_1$ and $P_2$ are also unknown but that $\frac{P_2}{P_1}$ is known: indeed, denote by N the unknown total number of documents that ever existed, then

$$\frac{P_2}{P_1} = \frac{NP_2}{NP_1} \tag{5}$$

which is the division of two known numbers: the number of documents of which we have two copies found and the number of documents of which we have found one copy. Note that the unknown N cancels in (5). But, using (1) for $j = 1$ and $j = 2$ we find

$$P_1 = ap^{a-1}(1-p) \tag{6}$$

$$P_2 = \frac{a(a-1)}{2}p^{a-2}(1-p)^2 \tag{7}$$

, hence (5) reduces to

$$\frac{P_2}{P_1} = \frac{a-1}{2}\frac{1-p}{p} \tag{8}$$

Solving equation (8) for p gives

$$p = \frac{1}{1+\frac{2P_2}{(a-1)P_1}} \tag{9}$$

Formula (9) in the formula (2) for $P_0$ yields

$$P_0 = \left[1 + \frac{2P_2}{(a-1)P_1}\right]^{-1} \tag{10}$$

In this formula, as said above, $\frac{P_2}{P_1}$ is known but the parameter a is unknown. How to determine the value of a will be the topic of Section III. Now we will apply this model to the case of Jesuit theatre programmes and it will turn out (lucky as we are!) that, with these practical data (and probably in much more occasions), $P_0$ is almost constant in a. The data are as follows:

- We have 714 documents (editons of theatre programmes) with 1 copy,
- We have 82 documents with 2 copies,
- We have 4 documents with 3 copies,
- We have 3 documents with 4 copies and
- We have 1 document with 5 copies
- We have no documents with 6 or more copies,

totalling to 804 found documents (theatre plays). Hence, based on (5) we have

$$\frac{P_2}{P_1} = \frac{82}{714} \tag{11}$$

It is historically known that small Jesuit colleges printed between 150 and 200 copies of each theatre programme while large Jesuit colleges printed between 680 and 850 copies of the programmes for their theatrical performances. Though these are large differences in the value of a it will turn out that, due to the fact that a, in any case, is large, it has almost no influence on the value of $P_0$. Indeed, using (11) we find for $P_0$ (formula (10)): for $a = 150$:

$$P_0 = \left( \frac{1}{1 + \frac{2}{149} \frac{82}{714}} \right)^{150} = 0.7936955$$

hence 79.4% of all plays is lost. For $a = 200$ this gives

$$P_0 = \left( \frac{1}{1 + \frac{2}{199} \frac{82}{714}} \right)^{200} = 0.7939673$$

hence still 79.4%. Even for $a = 750$ we have

$$P_0 = \left( \frac{1}{1 + \frac{2}{749} \frac{82}{714}} \right)^{750} = 0.7945627$$

being 79.5% and the same for $a = 850$.

So $P_0$ is very stable in a and we can conclude that we lost about 79.4% (or 79.5%) of all editions of theatre programmes. Even for $a \to \infty$ we can calculate $P_0$ based on the general formula (10).

**Proposition**:

$$\lim_{a \to \infty} P_0 = e^{-\frac{2P_2}{P_1}} \tag{12}$$

**Proof**: By (10):

$$\lim_{a \to \infty} P_0 = \lim_{a \to \infty} \left( \frac{1}{1 + \frac{2P_2}{P_1} \frac{1}{a - 1}} \right)^a$$

$$= \lim_{b \to \infty} \left( \frac{1}{1 + \frac{2P_2}{P_1} \frac{1}{b}} \right)^b \tag{13}$$

, denoting $b = a - 1$ and remarking that

$$\lim_{a \to \infty} \frac{1}{1 + \frac{2P_2}{P_1} \frac{1}{a - 1}} = 1$$

Denoting $B = \frac{2P_2}{P_1}$ it follows that

$$\lim_{b \to \infty} \ln \left( \frac{1}{1 + \frac{2P_2}{P_1} \frac{1}{b}} \right)^b = \lim_{b \to \infty} \frac{\ln \left( \frac{1}{1 + \frac{B}{b}} \right)}{\frac{1}{b}} = - B \tag{14}$$

by the l'Hôspital's rule. By (13) and (14) we now have

$$\lim_{a \to \infty} P_0 = e^{- B} = e^{- \frac{2P_2}{P_1}} . \qquad \square$$

For the value (11) of $\frac{P_2}{P_1}$ this gives

$$\lim_{a \to \infty} P_0 = e^{- 2\frac{82}{714}} = 0.7947785$$

hence still under 79.5%, so very stable!

These calculations show that only low values of a have an influence on $P_0$, e.g. for $a = 5$ we have, with (11): $P_0 = 0.7564083$, hence about 75.6%. In this case it might be necessary to calculate a from the data. This will be executed in Section III.

**Note**:

The here established method works for all $a \in \mathbb{N}$ from $a = 2$ onwards, i.e. for real multi-copy documents. Indeed, if $a = 1$ we only have $P_0 = p$ and $P_1 = 1 - p$ which is not enough to determine $P_0$. Already from $a = 2$ onwards is the method working: for $a = 2$ we have $P_0 = p^2$, $P_1 = 2p(1 - p)$, $P_2 = (1 - p)^2$, hence the model, using $\frac{P_2}{P_1}$, can be executed (deriving p from $\frac{P_2}{P_1}$ and then putting p in $P_0 = p^2$).

Finally, we have to estimate the number of lost documents. Let us denote by $N_l$ this number and by $N_f$ the number of found documents. Since we denoted by N the total number of documents that ever existed, we clearly have

$$N = N_f + N_l \tag{15}$$

We have

$$NP_0 = N_l \tag{16}$$

$$N(1 - P_0) = N_f \tag{17}$$

by definition, hence

$$N_l = N_f \frac{P_0}{1 - P_0} \tag{18}$$

This general formula can be applied to our case of Jesuit theatre plays where we found $P_0 \approx 0.794$. Recall that $N_f = 804$. Hence

$$N_1 \approx 804\frac{0.794}{0.206} \text{ pieces (i.e. editions of theatre programmes)}$$

or

$$N_1 \approx 3{,}099 \text{ pieces} \quad (19)$$

The total number of pieces that ever existed is, hence, estimated by (15):

$$N \approx 804 + 3{,}099 \text{ pieces}$$

$$N \approx 3{,}903 \text{ pieces} \quad (20)$$

We leave open a mathematical theory to calculate the confidence intervals for $P_0$ (or $N_1$ or N) but we have executed several simulations of random samples in abstracts "copies" of "pieces" (i.e. editions of theatre programmes). The experiments show that we can have a high confidence in the numbers above. The results are as follows. We explored the reliability of the mathematical model as follows. In a database, we created 10 different fictitious corpora, each corpus containing a different number of editions (from 1,000 until 10,000). Every edition is present in 150 copies ($a = 150$). Every copy of each edition is represented by one record in the database. The largest corpus of 10,000 edition counts therefore 1,500,000 unique records, the smallest one 150,000 records or "copies of editions". Firstly, we had the computer pick out 1,000 records (or copies) at random from every corpus. Then the sample was analysed: how many editions were presented by one copy, how many by two copies and so on (see Table 1). This exercise provided us with the values for $P_1$ and $P_2$, needed in formula (12).

Table 1. Distribution of editions and copies (corpus = 10,000 editions, a = 150, n = 1,000)

| | | *Number of editions* | *Number of copies* |
|---|---|---|---|
| $P_1$ | 906 | pieces are found with | 1 copy |
| $P_2$ | 44 | pieces are found with | 2 copies |
| $P_3$ | 2 | pieces are found with | 3 copies |
| $P_4 \ldots P_a$ | no | pieces are found with | 4 or more copies |

The total number of found editions is 952. These numbers result for $P_0$ in

$$P_0 = \left( \frac{1}{1 + \frac{2}{149} \cdot \frac{44}{906}} \right)^{150} = 0.90867535$$

Using this result in formula (15) and in formula (18) results in an estimated total number (N) of 10,223 editions. As the corpus from which the sample is taken counts 10,000 editions, this sample gives a very precise idea of its size.

For each of these 10 corpora, we repeated this exercise 30 times and took 30 samples of 1,000 records (copies) in order to get an estimation of the precision of the estimations in practise. The results of the tests are presented in Table 2.

Table 2. Estimation of N based on 30 random samples with n = 1,000 (a = 150)

| *Size corpus* | *Average result for N* | *Min. value for N* | *Max. value for N* | *Standard deviation* |
|---|---|---|---|---|
| 10,000 | 10,073 | 8,074 | 12,178 | 1,278 |
| 9,000 | 9,007 (8,801) | 6,437 (6,437) | 14,978 (11,096) | 1,563 (1,101)* |
| 8,000 | 7,890 | 6,520 | 11,278 | 971 |
| 7,000 | 6,947 | 5,070 | 9,099 | 753 |
| 6,000 | 6,126 | 5,238 | 8,117 | 644 |
| 5,000 | 5,102 | 4,165 | 6,485 | 530 |
| 4,000 | 4,059 | 3,183 | 5,456 | 469 |
| 3,000 | 3,010 | 2,449 | 3,389 | 215 |
| 2,000 | 2,011 | 1,796 | 2,425 | 143 |
| 1,000 | 998 | 873 | 1,106 | 56 |

* These values are due to one sample. The numbers between brackets give the results when this sample is omitted.

The size of the samples being constant ( n = 1,000 ), it is obvious that they are more exact, the more the size of the corpus decreases. For corpora between 1,000 and 3,000 editions, the standard deviation is very low and even the found minimum and maximum values for N give a reasonable indication of the real size of the corpus.[2] Of course, these results are average results of each time 30 samples. Figs. 1 and 2 show the individual results of each sample for a corpus of 3,000 editions (left) and 10,000 editions.

Fig. 1. Corpus of 3,000 editions

Fig. 2. Corpus of 10,000 editions

Overall, the results of the individual estimations are very satisfactory.

[2] Only one sample out of 300 gave a totally wrong image of the estimated size, see Table 2. When we leave that sample out, we get a normal series for the corpus containing 9,000 editions.

In practice, it is most of the time impossible to control the size of the sample. In general, scholars try to take the sample as large as possible. It is useful to know when the size of a sample is sufficient. We tested the variation of the size of the sample in relation to a constant corpus of 4,000 editions each consisting of 150 copies.

Table 3. Estimation of N based on samples with a decreasing size (corpus of 4,000 editions, a = 150)

| *Size sample* | *Average result for N* | *Min. value for N* | *Max. value for N* | *Standard deviation* |
|---|---|---|---|---|
| 1,000 | 4,059 | 3,183 | 5,456 | 469 |
| 900 | 4,022 | 3,286 | 5,034 | 392 |
| 800 | 3,915 | 3,176 | 6,000 | 582 |
| 700 | 3,902 | 3,186 | 4,833 | 428 |
| 600 | 4,059 | 2,895 | 5,943 | 652 |
| 500 | 4,254 | 2,856 | 6,284 | 848 |
| 400 | 3,829 | 2,933 | 5,727 | 770 |
| 300 | 4,470 | 2,345 | 8,647 | 1,342 |
| 200 | 4,842 | 1,616 | 19,670 | 3,417 |
| 100* | - | - | - | - |

*Samples with n = 100 result sometimes in $P_2 = 0$, where the model requires a value $P_2 > 0$.

The greater the sample sizes, the more exact the estimations. Table 3 shows that the exactness of random samples between 700 and 1,000 is quite high: the standard deviation is between 9.7 and 14.9% of the found average result for *N*. Decreasing the sample to 600, 500 or 400 has already tangible consequences: the estimated size of the corpus shows an average mistake of ca. 20% of the found average result for *N*, which is, - under circumstances – acceptable. Smaller samples result in unreliable $(200 < n < 300)$ or even unworkable $(n = 100)$ estimations.

Table 4 and 5 can serve as a guideline for scholars to evaluate the precision of a concrete sample.

**Table 4**. Average result for N of each time 30 samples taken at random with a = 150

| | *Size sample* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Size corpus* | 1,000 | 900 | 800 | 700 | 600 | 500 | 400 | 300 | 200 |
| 1,000 | 998 | 990 | 1,001 | 1,030 | 1,013 | 1,006 | 1,024 | 1,026 | 1,060 |
| 2,000 | 2,011 | 2,023 | 2,016 | 2,028 | 2,048 | 1,962 | 2,031 | 2,095 | 2,308 |
| 3,000 | 3,010 | 3,017 | 3,045 | 3,187 | 2,969 | 3,087 | 3,251 | 3,421 | 4,021 |
| 4,000 | 4,059 | 4,022 | 3,915 | 3,902 | 4,059 | 4,254 | 3,829 | 4,470 | 4,842 |
| 5,000 | 5,102 | 5,013 | 5,253 | 5,151 | 4,977 | 5,106 | 5,517 | 5,093 | * |
| 6,000 | 6,126 | 6,171 | 5,983 | 6,304 | 6,221 | 6,476 | 6,218 | 7,023 | 7,261 |
| 7,000 | 6,947 | 7,190 | 7,661 | 6,962 | 7,208 | 7,754 | 7,745 | 8,978 | * |
| 8,000 | 7,890 | 8,226 | 8,716 | 8,792 | 8,099 | 8,036 | 9,165 | 9,571 | * |
| 9,000 | 9,007 | 9,379 | 9,230 | 9,503 | 9,426 | 10,357 | 9,681 | 12,058 | * |
| 10,000 | 10,073 | 10,242 | 10,162 | 10,516 | 11,749 | 10,937 | 12,136 | 12,705 | * |

*Some samples have P2 = 0

**Table 5**. Standard deviation (in %) of N mentioned in **table 4**.

| | Size sample | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Size corpus* | 1,000 | 900 | 800 | 700 | 600 | 500 | 400 | 300 | 200 |
| 1,000 | 56 | 65 | 76 | 72 | 102 | 97 | 134 | 179 | 245 |
| 2,000 | 143 | 187 | 213 | 245 | 200 | 265 | 417 | 432 | 835 |
| 3,000 | 215 | 234 | 384 | 440 | 292 | 509 | 648 | 1,022 | 2,308 |
| 4,000 | 469 | 392 | 582 | 428 | 652 | 848 | 770 | 1,342 | 3,417 |
| 5,000 | 530 | 645 | 887 | 939 | 967 | 1,399 | 1,434 | 1,763 | * |
| 6,000 | 644 | 830 | 913 | 1,318 | 1,029 | 1,763 | 1,246 | 2,946 | 4,103 |
| 7,000 | 753 | 775 | 1,684 | 1,197 | 1,331 | 1,636 | 3,478 | 7,703 | * |
| 8,000 | 971 | 1,229 | 1,659 | 1,447 | 1,720 | 2,127 | 3,823 | 4,933 | * |
| 9,000 | 1,563 | 1,468 | 1,777 | 2,845 | 2,538 | 3,022 | 4,381 | 4,223 | * |
| 10,000 | 1,278 | 1,704 | 1,723 | 2,137 | 3,421 | 3,750 | 4,743 | 9,763 | * |

*Some samples result sometimes in $P_2 = 0$, where the model requires a value $P_2 > 0$.

Each number in these tables is the result of 30 samples. In the upper left edge of table 4, the average estimation of N is presented for the smallest corpus (i.e. 1,000 editions of each 150 copies) and the largest sample (i.e. 1,000 copies taken at random): 998. The corresponding number in table 5 indicates the standard deviation of that estimation: 56, or 5.6% of the found average result for *N*. That means that a sample of 1,000 copies of a corpus of 1,000 editions gives a very good indication of N.
On the other hand, the tables indicate that a sample of 300 copies for an estimated corpus of 10,000 editions is very unreliable. The average result of 30 samples for N still gives 12,705 editions, but the standard deviation of these estimations is 9,763!

These simulations provide us with a guideline for the interpretation of the result based on our sample of Jesuit theatre programmes. The estimation of the total production of theatre programmes bearing the same characteristics as the programmes in the sample, led to the number of $3{,}903$ editions (cf. (20)). Given the fact that the sample consisted of 907 single copies ($714 + 2*82 + 3*4 + 4*3 + 5$ copies, see above), this estimation is probably quite exact (cf. table 5). From a methodological point of view, this information is of high importance. It means that, if we analyse the theatrical production of the Jesuits in Flanders before 1773 on the basis of retrieved programmes, we may extrapolate the results to a group of pieces five times larger than the retrieved number.

The next section is devoted to a mathematical – probabilistic model to determine a, the number of copies per document.

# III. Determination of the number of copies per document

In this section we assume that a, the number of copies per document, is small enough to have an influence on $P_0$ (see formula (10)) and hence should be calculated. Clearly, besides formula (8) (to determine p) we need another equation. This is best done by calculating the value of $\frac{P_3}{P_2}$, from (1). We have (7) for $P_2$ and (1) yields

$$P_3 = \frac{a(a-1)(a-2)}{6} p^{a-3}(1-p)^3 \tag{21}$$

So

$$\frac{P_3}{P_2} = \frac{a-2}{3}\frac{1-p}{p} \tag{22}$$

Results (8) and (22) yield:

$$\frac{\frac{P_3}{P_2}}{\frac{P_2}{P_1}} = \frac{2}{3}\frac{a-2}{a-1}$$

from which a easily follows:

$$a = \frac{\frac{3P_1P_3}{P_2^2} - 4}{\frac{3P_1P_3}{P_2^2} - 2} \tag{23}$$

Note that $P_1$, $P_2$, $P_3$ are not known but that $\frac{P_2}{P_1}$ (see formula (5)) and similarly, $\frac{P_3}{P_2}$ are known from the retrieved data:

$$\frac{P_3}{P_2} = \frac{NP_3}{NP_2} \qquad (24)$$

is the division of the known number of documents of which we have three copies found by the known number of documents of which we have two copies found. The determination of p is as in the previous section (formula (9)) and hence $P_0$ (formula (10)) is completely determined.

Unfortunately, in our example of Jesuit theatre programmes, the number $NP_3 = 4$ is too low to be a stable value in these formulae. We have

$$\frac{P_3}{P_2} = \frac{4}{82} \qquad (25)$$

and this yields, together with (11):

$$a = 3.7557376$$

an unrealistic number. Still, formula (10) gives $P_0 = 0.7403156$, about 5% lower than in the previous section (based on higher values of a) which still can be considered as rather stable.

We want to underline that the above model certainly is useable in case one has not lost many documents in which case the number $NP_3$ will be high and trustable. Note that in the case of Jesuit theatre programmes the number p is extremally high, being around $p = 0.999$ for whatever value of $a \geq 200$ and around $p = 0.998$ for $a = 150$ (use formula (9) and (11) to establish this). This is the reason why, although at least 150 copies of Jesuit theatre programmes existed (and in many cases even up to 850), we hardly found any plays with 3 or more copies: in short: “almost all copies have been destroyed”.

# IV. Informetric properties of the function $j6P_j$

The function $j \to P_j$ as given by (1) is what we call in informetrics a size-frequency function (cf. Egghe (2005)): it expresses the number (or rather the fraction, the difference is only a factor N as expressed in (5)) of documents of which we still have (or found) $j \in \{1,2,3,...\}$ copies (here we do not consider the case $j = 0$ anymore). In general informetrics terminology, we could say that $P_j$ expresses the fraction of sources with j items (cf. again Egghe (2005)).

In classical informetrics one is then automatically thinking of a "classical" size-frequency model, e.g. the law of Lotka

$$P_j = \frac{C}{j^{\alpha}} \tag{26}$$

$(C > 0, \ \alpha > 1)$ or of a decreasing power law

$$P_j = b_0 b^j \tag{27}$$

$(b_0 > 0, \ 0 < b < 1)$. It is clear that relation (1) is not of these types (it is the binomial distribution)! But here, a more fundamental result can be derived.

Let us look at the indicators $\frac{P_{j+1}}{P_j}$ $(j = 1,2,3,...)$. For the model developed in this article we find readily

$$\frac{P_{j+1}}{P_j} = \frac{a - j}{j+1}\frac{1-p}{p} \tag{28}$$

Let us define the relative indicators

$$Q_j = \frac{\frac{P_{j+1}}{P_j}}{\frac{P_j}{P_{j-1}}} = \frac{P_{j+1}P_{j-1}}{P_j^2} \tag{29}$$

for $j = 2,3,\ldots$ . Hence we have, for our model here, by (28):

$$Q_j = \frac{j}{j+1}\frac{a-j}{a-j+1} \tag{30}$$

Since here the requirement $j+1 \leq a < \infty$ is clear (since a is the maximum number of copies, available at the start), we see, by (30) that

$$\frac{j}{2(j+1)} \leq Q_j < \frac{j}{j+1} \tag{31}$$

for all $j = 2,3,\ldots$ . We find, e.g. $Q_2 \in \left[\frac{1}{3}, \frac{2}{3}\right)$ and so on, but for all j we have that

$$\frac{1}{3} \leq Q_j < 1 \tag{32}$$

If we calculate $Q_j$ for the law of Lotka (26) we have, as is readily seen

$$Q_j = \frac{\frac{C}{(j+1)^{\alpha}}\frac{C}{(j-1)^{\alpha}}}{\left(\frac{C}{j^{\alpha}}\right)^2}$$

$$Q_j = \left(\frac{j^2}{j^2-1}\right)^{\alpha} > 1 \tag{33}$$

for all j. Hence $Q_j$ occupies a disjoint range when compared with (32)!

For the decreasing exponential function (27), we readily see that

$$Q_j = 1 \tag{34}$$

for all j, hence being (strictly) between the cases (32) and (33).

The above is a proof that the informetric data $P_j$ on the number of documents with j copies is non-Lotkaian and not exponential. The inequality (32) expresses that the decline in $\frac{P_{j+1}}{P_j}$ as a function of j is much faster for our present model than for the Lotkaian or exponential model. This is a remarkable conclusion: finding missing copies of documents, no matter what p (the probability for a copy to be lost) is, is an activity which leads to very fast declining values of $P_j$, the fraction of documents with j copies recovered. This means that it is, relatively, very hard to find multiple copies of a single document.

The above shows that our present model belongs to a new type of two-dimensional informetrics theory. While Lotkaian informetrics describes a two-dimensional informetrics theory of growth of sources and items (cf. Egghe (2005)), the present model describes a two-dimensional informetrics theory of "what is left", hence a two-dimensional informetrics theory of aging (or obsolescence) – here in the sense of recovering copies of documents, hence also describing the loss of copies and, consequently, when all copies of a document are lost, the loss of documents. Whether or not this two-dimensional model of aging can also be applied to the more "classical" topic of aging in terms of citation analysis, is left as an open problem.

# V. A rationale for the book historical law

The book historical law says (Willard (1943) but see also other references in Proot and Egghe (2007)): The probability to save a copy of an edition is reversely proportional to the size of the edition. We carefully checked the literature on this subject and noticed that this law has not been formulated in a more accurate way, let alone that it has been proved. We therefore formulate the above "law" as follows: The probability to save a copy of an edition is a decreasing function of the size of the edition. Equivalently, and using the parameters p and a in this article, we can state the book historical law as:

Book historical law: The probability p to lose a copy is an increasing function of the size a of the edition.

We will now give a partial explanation of this expected regularity, not taking into account other variables such as temporary interest of documents or, simple, the money value of documents. In the sequel we will show that p is an increasing relation (to be explained further) of a.

Denoting $\frac{P_2}{P_1} = x$, we have, by (9):

$$p = \frac{1}{1 + \frac{2}{a-1}x} \tag{35}$$

In practise we can assume that $x < 1$; in fact (11) shows that $x = 1$, a logical fact. So, if we let x vary in $]0,1[$, we have that

$$\frac{1}{1 + \frac{2}{a-1}} \leq p \leq 1 \tag{36}$$

showing that p has an (evident) upper bound in 1 and a lower bound

$$f(a) = \frac{1}{1 + \frac{2}{a-1}} \tag{37}$$

which is a concavely increasing function since $f'(a) > 0$, $f''(a) < 0$ for all $a \geq 2$, the absolute lower bound of a (since we deal with $P_2$). Formulae (36) and (37) imply that the relation between p and a (describing the book historical law) is as in Fig. 3.

Fig. 3. The relation between p (the probability to lose a copy) and a (the size of the edition) is given by the shaded area.

Fig. 3 shows that, when a is low, we can have values of p in the range $\left[\frac{1}{3},1\right]$ (maximally) but for larger values of a we see that p can only be large (close to 1). So, the higher a, the more limited is the range in which p can vary and the higher this range is situated, giving a partial explanation of the book historical law. Note that this explanation could only be given based on the boundedness of $\frac{P_2}{P_1}$ (here by 1, but higher bounds could serve as well). Hence a high value of a (the size of the collection) forces p (the probability to lose a copy) to be high. The intuition for this is clear: a high value of a implies that it will be difficult to have pieces with a low number of copies, unless p is very high (close to 1). The fact that $\frac{P_2}{P_1} = 1$ expresses that we have relatively more pieces with 1 copy than with 2 copies which can only be understood when p is large.

# VI. Conclusions and open problems: suggestions for further research

By means of found copies of multi-copy documents we were able to estimate the number of lost documents and hence also to estimate the total number of multi-copy documents that ever existed. This probabilistic theory shows that the numbers are relatively independent of the number a of copies per document as long as a is not very small: in the other case the theory is complemented with a formula to estimate the value of a.

Simulations show that the estimated number of lost documents is very stable. These simulations are executed by random sampling in the copies, where we know in advance the total number of documents.

We applied the model to the case of Jesuit theatre programmes in which case $a \geq 150$ (and where a can even go up to 850). As mentioned above, these large values of a (number of printed copies of theatre plays) guarantee a stable percentage of lost plays, estimated in this case around 80%.

It is clear that this theory could be applied to other cases of multi-copy documents. One could study the problem of estimating the number of lost documents in case the documents are precious. Here a will be smaller in which case Section III can be used to estimate a (needed since, for smaller values of a, $P_0$, being the fraction of lost documents, is more dependent of a). But in this case, p will also be smaller (being the probability to lose a copy) implying that, in this case, one has more documents (than in the case of Jesuit theatre programmes) of which more (i.e. 3, 4,…) copies are found (i.e. not lost), making the estimate of a more reliable (see Section III).

We remarked that the size-frequency function $j \to P_j$ (fraction of documents for which we found j copies) that we encountered in this theory satisfies the inequality (for all $j \geq 2$)

$$Q_j = \frac{\frac{P_{j+1}}{P_j}}{\frac{P_j}{P_{j-1}}} < 1$$

while we have the opposite inequality for Lotkaian size-frequency functions and while we always have $Q_j = 1$ for decreasing exponential size-frequency functions, hence noting that we are in a new type of informetrics theory, describing loss (or rather recovery) of items of sources. We leave it for further study whether this model can also be used for the description of two-dimensional aging in the "classical" sense: the decline of citations in time.

We also gave a partial rationale for the book historical law: The probability to lose a copy of a document is an increasing function of the size of the edition.

It is our hope that this model will be applied in many other (varying) examples of multi-copy documents (and even multi-copy objects as described in the introductory section), hereby further testing the stability of the probabilistic model. The further development of this non-Lotkaian informetrics theory is also a challenge.

# References

G.C. Canavos (1984). Applied probability and statistical Methods. Little, Brown and Company, Boston, USA.

L. Egghe (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK.

G.R. Grimmett and D.R. Stirzaker (1985). Probability and random Processes. Clarendon Press, Oxford, UK.

G. Proot and L. Egghe (2007). The estimation of editions on the basis of survivals: printed programmes of Jesuit Theatre Plays in the Provincia Flandro-Belgica (before 1773). With a note on the "bookhistorical law", forthcoming.

O.M. Willard (1943). The survival of English books printed before 1640: a theory and some illustrations. The Library, Fourth Series, 23, 171-190.