A model for measuring the congestion in library shelves.

Peer-reviewed author version

A MODEL FOR MEASURING THE CONGESTION IN LIBRARY SHELVES

by

L. Egghe

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

ABSTRACT

A model for measuring the congestion in library shelves after j years ($j \in \mathbb{N}$) is obtained by taking j-fold convolutions of the distributions that describe the yearly growth of literature (e.g. periodicals, books on a certain topic, ...) From this one can estimate the expected number of critical points in the shelf, after j years. One can also calculate the probability that there will be m ($m \in \mathbb{N}$) critical points after j years.
The paper closes by examining two concrete cases.

I.   Introduction.

This paper studies the problem of estimating (in function of time) the number of critical points in shelves that occur because of the growth of literature. More concretely we will focus on periodical shelves and investigate how long it takes before the available space for a periodical will disappear. But we can mention here that the same models can be used in the study of book shelves where, e.g., the books are ordered per subject and a certain amount of space is available, per subject (e.g. per UDC, Dewey class or whatever). One can even think of applying these congestion models to the case of a service desk where a staff member gives services, divided by "empty" time periods. Where one has consecutive services, congestion has occurred. This problem is postponed for later study. For the rest of this paper we will focus on the congestion problem in periodical shelves.

Suppose we start at a certain time t=0. This is the starting point in which the shelves have been organized (by human intervention) in such a way that, according to the average growth per year, each periodical has an available shelf space of a years ($a \in \mathbb{N}$). It is indeed clear that each periodical must be given the same space, relative to its average growth per year, otherwise we create non-optimal situations. The value of a is dependent on the number of shelves available at time t and is fixed but arbitrary in this paper.

We note that it is not important to know for a periodical how many back-sets are available : back-sets occupy a fixed length on the shelves (and might influence the initial choice of the value of the constant a) but do not play any role in the growth model that is under study here.

Suppose alltogether we have N periodicals. We will make two simplifying assumptions :
1. This set of N periodicals remains the same in the period under study. So all these periodicals are kept current and no new periodicals are added.
2. The average growth of each periodical (as measured in shelf space occupied) remains constant in the period under study.

It is certainly so that these assumptions are not always true in practise. In a forthcoming paper we intend to add the complexity, caused by not assuming 1 or 2, to the model as developed here. This model hence serves as a basic machine on which more concrete situations can be grafted. Of course assumptions 1 and 2 are not completely unrealistic. One might decide to keep new periodicals on a separate shelf until a new time period of reshelving arrives. In this case the present model can be applied twice (with different parameters). In case of the cancellation of subscriptions one can argue that our models here are upper bounds for the congestion problems that can occur. As to assumption 2 we can note that most journals satisfy this. In our test case of library and information science periodicals we only found a few counterexamples to assumption 2. In fact in some cases we found the opposite effect, hence where the average growth of a periodical diminishes. JASIS is an example of a steady growing annual growth (in terms of occupied shelf length), while e.g. Journal of Documentation has the opposite effect. But in most cases, we have that assumption 2 is correct.

Suppose that, for each periodical (represented by the numbers $i\in\{1,...,N\}$), the number μ_i denotes its average growth, e.g. the number of mm added each year, on the average. The real growth of each periodical i, however, is a random variable which we assume to be the normal distribution around μ_i. Let σ_i^2 denote the variance of the annual growth of periodical i. Hence we assume that the annual growth of periodical i is ruled by the normal distribution $N(\mu_i,\sigma_i^2)$, which density we denote by f_i :

$$ f_i(x) = \frac{1}{\sqrt{2\pi}\ \sigma_i}\ e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2} \tag{1} $$

for $x\in\mathbb{R}$ although, since $\mu_i>0$, most of the density is obtained for $x>0$.

We supposed that at our starting point $t=0$, each periodical i has an available shelf space equal to $a\mu_i$, the average growth of this periodical in a years. The numbers μ_i and σ_i are easy to estimate from the shelf (see appendix B).

First we will develop a probabilistic model that will answer the following questions
1. For each $j\in\mathbb{N}$ (hence j can be $<a$, $=a$ or $>a$), what is the expected number of periodicals that have used all the available space that was reserved for them (i.e. that are congested), after j years ?
2. What is the probability to have, after $j\in\mathbb{N}$ years, $m\in\{1,...,N\}$ periodicals for which congestion has occurred ?

This will be studied in the next two sections. The last section is devoted to two practical test cases.

II. The expected number of congested periodicals after j years (j∈N).

So the annual growth of each periodical $i \in \{1,...,N\}$ is ruled by the normal distribution $N(\mu_i, \sigma_i^2)$ with density (1). Let $j \in N$ and denote by $z_{i,j}$ the growth (in shelf length) of periodical i during the first j years. Now

$$z_{i,j} = \sum_{k=1}^{j} x_{i,k} , \qquad (2)$$

where $x_{i,k}$ denotes the growth of periodical i in year $k \in \{1,...,j\}$.

We can assume that the growth in year k is independent of that in year k' ($k \neq k'$). It is then well-known (see e.g. Feldman and Fox (1991), p. 347) that the distribution of $z_{i,j}$ is the j-fold convolution of the distributions of $x_{i,1},...,x_{i,j}$. Hence

P(periodical i grows $z_{i,j}$ in the first j years)

$$= P\left(\sum_{k=1}^{j} x_{i,k} = z_{i,j} \right)$$

$$= \underbrace{(f_i \circledast ... \circledast f_i)(z_{i,j})}_{j \text{ times}} , \qquad (3)$$

where f_i is given by (1) and where $\circledast$ denotes convolution. Convolutions have been increasingly used in information science. We refer to the following papers that use convolution in the building of mathematical models : Egghe (1993, 1994), Egghe, Rao and Rousseau (1995), Egghe and Rousseau (1998, 1999). Note also that Rousseau (1998) is a review article on this topic.

Formula (3) is the j-fold convolution of normal distributions. Hence we have - see e.g. Feldman and Fox (1991), p. 348, that (3) equals

$$\frac{1}{\sqrt{2\pi} \sqrt{j} \, \sigma_i} e^{-\frac{1}{2} \frac{(z_{i,j} - j\mu_i)^2}{j\sigma_i^2}} \qquad (4)$$

, again a normal distribution with average $j\mu_i$ and variance $j\sigma_i^2$. Periodical i will be congested after j years with a probability

$$P \text{ (i has a growth } z_{i,j} > a\mu_i \text{ in j years).} \qquad (5)$$

But $z_{i,j} > a\mu_i$ if and only if

$$\frac{z_{i,j} - j\mu_i}{\sqrt{j} \, \sigma_i} > \frac{(a-j)\mu_i}{\sqrt{j} \, \sigma_i} \qquad (6)$$

and this is distributed in a standard normal way. Denoting by Φ its well-known cumulative distribution function we obtain that (5) equals

$$1 - \Phi\left(\frac{(a-j)\mu_i}{\sqrt{j} \, \sigma_i} \right) \qquad (7)$$

From this it also follows that, denoting by E(j) the expected number of congested periodicals after j years, we have

$$E(j) = \sum_{i=1}^{N} \left(1 - \Phi \left(\frac{(a-j)\mu_i}{\sqrt{j}\ \sigma_i} \right) \right).$$ (8)

Properties of E(j)

1. For $j=a$ we find $E(a)=\frac{N}{2}$ (since $\Phi(0)=\frac{1}{2}$). Hence, after $j=a$ years, 50% of the periodicals is congested (if one does not interfere).

2. The behaviour of E(j) in function of j is as in Fig. 1. One can easily show that E(j) increases and that $\exists j_0 < a$ such that this increase is convex for $j < j_0$ and concave for $j > j_0$.
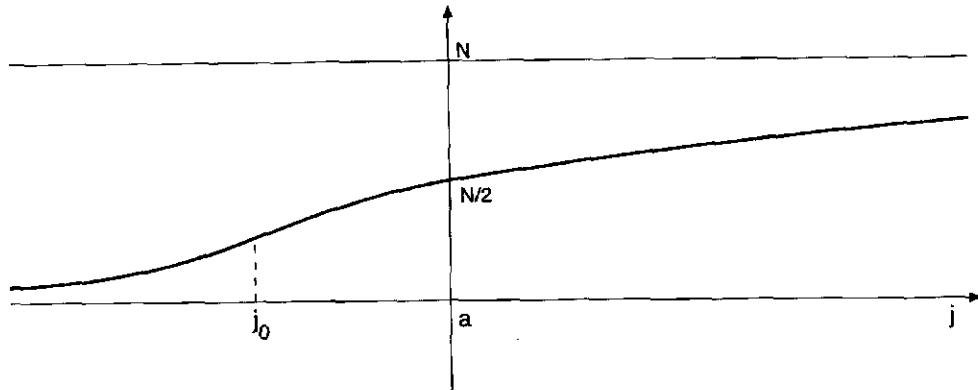

Fig. 1. The graph of E(j)

This graph describes the evolution of the average congestion of periodical shelves in function of time. For its derivation, see appendix A where one presents an equation of degree 3 for the calculation of the osculation point j_0. For this, the assumption that

$$V = \frac{\sigma_i}{\mu_i}$$ (9)

is a constant in i is adopted. Appendix B investigates this and concludes from experiments that this is acceptable, certainly as a first approximation. Note that V is the so-called variation coefficient, i.e. a measure of relative dispersion. As proved in Egghe and Rousseau (1990, 1991), V is the best concentration measure and is used in econometrics, biometrics and informetrics.

3. If V above is constant, we find that

$$E(j) = N(1 - \Phi_j),$$ (10)

where we denote

$$\Phi_j = \Phi \left(\frac{a-j}{\sqrt{j}\ V} \right)$$ (11)

(cf. (8) above).

III.   The probability to have m congested periodicals (m∈{1,...,N}) after j years (j∈ℕ).

For i∈{1,...,N}, it follows from (7) that the probability that _only_ periodical i will be congested after j years is

$$\left(1 - \Phi\left(\frac{(a-j)\mu_i}{\sqrt{j}\ \sigma_i} \right) \right) \prod_{\substack{k=1 \\ k \ne i}}^{N} \Phi\left(\frac{(a-j)\mu_k}{\sqrt{j}\ \sigma_k} \right). \tag{12}$$

More generally, the probability that, after j years, m∈{1,...,N} periodicals are congested, is given by

$$\varphi_m(j) = \sum_{\text{all } A_m} \prod_{\ell \in A_m} \left(1 - \Phi\left(\frac{(a-j)\mu_\ell}{\sqrt{j}\ \sigma_\ell} \right) \right) \prod_{\substack{k=1 \\ k \notin A_m}}^{N} \Phi\left(\frac{(a-j)\mu_k}{\sqrt{j}\ \sigma_k} \right) \tag{13}$$

where $\sum_{\text{all } A_m}$ is the summation over all possible m-subsets from {1,...,N}, i.e. sets containing m elements from {1,...,N}.

From now on we again adopt the assumption (9). This yields for (13), using (11) that

$$\varphi_m(j) = \binom{N}{m} (1-\Phi_j)^m \ \Phi_j^{N-m} \tag{14}$$

Properties of $\varphi_m(j)$

1. $\varphi_m(j)$, as function of j, has a maximum for $j=j_m$ such that $\Phi_{j_m} = 1 - \dfrac{m}{N}$ (for the proof we refer the reader to appendix C) in which case we find

$$\varphi_m(j_m) = \binom{N}{m} \left(\frac{m}{N} \right)^m \left(1 - \frac{m}{N} \right)^{N-m} \tag{15}$$

$$= \frac{N!}{N^N} \cdot \frac{m^m}{m!} \cdot \frac{(N-m)^{N-m}}{(N-m)!} \tag{16}$$

We now use Stirling's formula for n high :

$$n! \approx \sqrt{2\pi n}\ n^n\ e^{-n} \tag{17}$$

yielding (since N and N-m are high)

$$\varphi_m(j_m) \approx \sqrt{\frac{N}{N-m}}\ \frac{m^m}{m!e^m} \tag{18}$$

If also m is high, we have

$$\varphi_m(j_m) \approx \sqrt{\frac{N}{2\pi m(N-m)}}. \tag{19}$$

$\varphi_m(j_m)$ represents the highest possible probability that we find m congested periodicals in the shelf (measured over time j∈ℕ).

2. $\varphi_m(j)$ as a (discrete) function of m (formula (14)) is a binomial distribution. It is well-known that its mode is given as follows :

    –   if $N(1-\Phi_j)+1-\Phi_j\in N$, then this as well as $N(1-\Phi_j)-\Phi_j$ are the two modes (i.e. the two entire consecutive numbers for which $N(1-\Phi_j)$ is in between),

    –   if $N(1-\Phi_j)+1-\Phi_j\notin N$ (most likely situation) then the (unique) mode is given by the entire number m_0 that is closest to $N(1-\Phi_j)$.

In any case, since $N>>\Phi_j$ and $N>>1-\Phi_j$ we have that the mode of $\varphi_m(j)$ (for m variable) is approximated by

$$N(1-\Phi_j). \tag{20}$$

So, not only is $N(1-\Phi_j)$ the expected number of congested periodicals (previous section) but it is also the most likely number of congested periodicals, at any time j. These are in fact properties of the binomial distribution.

Notes

1. According to our notations one should find

$$E(j) = \sum_{m=0}^{N} m\varphi_m(j) = N(1-\Phi_j) \tag{21}$$

which can readily be checked in the case of formula (14) (also well-known for the binomial distribution). So this argument gives a second proof for formula (10).

2. The average number of years that it takes to have $m\in\{1,...,N\}$ congested journals is given by

$$\sum_{j=0}^{\infty} j\varphi_m(j). \tag{22}$$

However, we were unable to find an analytic, manageable, form for formula (22).

3. It is easy to see that, if X is the r.v., denoting the time j in which the first journal starts being congested, we have

$$P(X=j) = \Phi_{j-1}^{N}(1-\Phi_j^{N}). \tag{23}$$

For j high we have (since then Φ_{j-1}, Φ_j are very small)

$$P(X=j) \approx \Phi_{j-1}^{N}. \tag{24}$$

Problem : calculate E(X).

IV. An example

In our example of N=13 periodicals in library and information science (appendix B) we found that all values of V were between 0.1 and 0.2 with average $\overline{V}$ =0.14. We will use this value for V as well as a=5 years as an example, corresponding to the real case in my

library. We obtain the following table, for $j=1,\ldots,8$ for the values of $\Phi_j = \Phi\left(\dfrac{5-j}{\sqrt{j}\,0.14}\right)$ (cf.(11))

j	1	2	3	4	5	6	7	8
Φ_j	1	1	1	0.9997	0.5	0.0018	0	0

As we can see here, the evolution of the congestion is almost deterministic, due to the fact that $V=0.14$ is very low (very little fluctuations from year to year). This is a good situation : one only has to intervene just before $j=5$.

Let us now see what happens if $V=0.14$ is changed into $V=0.5$ and $N=100$. We keep $a=5$ years. Now we obtain the following table for $\Phi_j = \Phi\left(\dfrac{5-j}{\sqrt{j}\,0.5}\right)$, $j=1,\ldots,10$.

j	1	2	3	4	5	6	7	8	9	10
Φ_j	1	0.99998	0.9896	0.8413	0.5	0.1835	0.0655	0.017	0.0038	0.0008

This yields the following values for E(j), based on equation (10).

j	1	2	3	4	5	6	7	8	9	10
E(j)	0	0.002	1.04	15.87	50	81.65	93.45	98.3	99.62	99.92

Note that these values also represent the mode of $\varphi_m(j)$, by (20).

The decision is clear : one must intervene soon after $j=3$ years. The reader can verify the shape of E(j) as determined in section II. Here we find, using equation (A7) in x and $j_0 = \dfrac{1}{x}$, that the osculation point $j_0 \approx 4.5 < a = 5$ as predicted by the model. The above Φ_j-values can also be used to determine $\varphi_m(j)$. We calculate this for the important years $j=3$ and $j=4$.

$$\varphi_m(3) = \binom{100}{m} (0.0104)^m (0.9896)^{100-m}$$

$$\varphi_m(4) = \binom{100}{m} (0.1587)^m (0.8413)^{100-m}.$$

This yields some concrete probabilities to have m congested periodicals after j (=3 or 4) years : $\varphi_1(3)=0.369$, $\varphi_2(3)=0.192$, $\varphi_3(3)=0.066$ for $j=3$ (note that the mode is at $m=1 \approx 1.04 = E(3)$, as found in the model) and $\varphi_{13}(4)=0.085$, $\varphi_{14}(4)=0.100$, $\varphi_{15}(4)=0.108$, $\varphi_{16}(4)=0.108$, $\varphi_{17}(4)=0.100$ (note that the modes are at $m=15$ and 16 and that $E(4)=15.87$ is between them, as found in the model).

This shows that the found formulae are calculable and that the model works.

References

Egghe, L. (1993)
Consequences of Lotka's law in the case of fractional counting of authorship and of first author counts. *Mathematical and Computer Modelling* 18(9): 63-77.

Egghe, L. (1994)
Special features of the author-publication relationship and a new explanation of Lotka's law based on convolution theory. *Journal of the American Society for Information Science* 45(6): 422-427.

Egghe, L., Rao, I.K. Ravichandra and Rousseau, R. (1995)
On the influence of production on utilization functions : obsolescence or increased use ? *Scientometrics* 34(2): 285-315.

Egghe, L. and Rousseau, R. (1990)
Elements of concentration theory. In : L. Egghe and R. Rousseau eds., *Informetrics 89/90*. Proceedings of the Second International Conference on Bibliometrics, Scientometrics and Informetrics (London, Canada, 1989), 97-137. Amsterdam : Elsevier.

Egghe, L. and Rousseau, R. (1991)
Transfer principles and a classification of concentration measures. *Journal of the American Society for Information Science* 42(7): 479-489.

Egghe, L. and Rousseau, R. (1998)
A theoretical study of recall and precision using a topological approach to information retrieval. *Information Processing and Management* 34(2/3): 191-218.

Egghe, L. and Rousseau, R. (1999)
The influence of publication delays on the observed aging distribution of scientific literature. Preprint.

Feldman, D. and Fox, M. (1991)
Probability. The Mathematics of Uncertainty, New York: Marcel Dekker.

Rousseau, R. (1998)
Convolutions and their applications in information science. *Canadian Journal of Library and Information Science*, to appear.

Appendix A

Denote

$$g_i(j) = \frac{(a-j)\mu_i}{\sqrt{j}\ \sigma_i} \tag{A1}$$

then $g_i'(j)<0$ and $g_i''(j)>0$ as is readily seen. Furthermore

$$E'(j) = -\sum_{i=1}^{N} \Phi'\left(\frac{(a-j)\mu_i}{\sqrt{j}\ \sigma_i}\right) g_i'(j) > 0 \ . \tag{A2}$$

Also

$$E''(j) = -\sum_{i=1}^{N} \left[\Phi''\left(\frac{(a-j)\mu_i}{\sqrt{j}\ \sigma_i}\right) g_i'^2(j) + \Phi'\left(\frac{(a-j)\mu_i}{\sqrt{j}\ \sigma_i}\right) g_i''(j)\right] . \tag{A3}$$

From the fact that

$$\Phi'(x) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}x^2} \tag{A4}$$

$$\Phi''(x) = \frac{-x}{\sqrt{2\pi}}\ e^{-\frac{1}{2}x^2} \tag{A5}$$

and the above it follows that $E'(j) > 0$ always. Since

$$\Phi''\left(\frac{(a-j)\mu_i}{\sqrt{j}\ \sigma_i}\right) \quad \begin{cases} < 0 & (j<a) \\ > 0 & (j>a) \end{cases} \tag{A6}$$

and since $\Phi'\left(\frac{(a-j)\mu_i}{\sqrt{j}\ \sigma_i}\right) > 0$ always, it follows that $E''(j) < 0$ for $j \geq a$ as well as for $j_0 \leq j \leq a$, $\exists j_0 < a$. Since obviously $\lim_{j\to\infty} E(j) = 0$ we have that j_0 can be choosen so that $E''(j) > 0$ for $j < j_0$.

This proves the shape of Fig. 1. The value of j_0 follows from the equation $E''(j_0) = 0$. Let us suppose that (9) is valid. Then it is a simple calculation to see that $\frac{1}{j_0}$ is the solution of the equation

$$1 + (a+V)x + a(3V-a)x^2 - a^3x^3 = 0 \tag{A7}$$

Appendix B

Investigation of the values $\frac{\sigma_i}{\mu_i}$.

We checked the periodicals in the LUC-library on library and information science. We kept only these periodicals for which at least 10 volumes where available and corrected for varying (increasing or decreasing) growth, where it was apparent (note that varying growth was not included in our model). This is done in diverse ways. If a periodical changed the number of volumes per year, only the last stable period (with a constant number of volumes) was kept. If a periodical experienced a continuously varying growth (as e.g. is the case with JASIS) we corrected for this using the following technique.

Suppose we want to correct the cloud of points $(x_1,y_1),...,(x_M,y_M)$ for varying growth. We first determine the best fitting straight line $y = \alpha + \beta x$, i.e. the line for which

$$\sum_{j=1}^{M} (y_j - (\alpha + \beta x_j))^2 \tag{B1}$$

is minimal (as e.g. for regression lines) but now limited to the constraint

$$y_1 = \alpha + \beta x_1. \tag{B2}$$

Indeed, since we want to measure the change in growth, we assume that the first value is correct (fixed growth at the starting time $t_0 = x_1$). The above problem is a constraint extremum problem and is solved using the Lagrangian function ($\lambda = $ Lagrange multiplicator)

$$g(\alpha,\beta,\lambda) = \sum_{j=2}^{M} (y_j - (\alpha + \beta x_j))^2 + \lambda(y_1 - (\alpha + \beta x_1)) \tag{B3}$$

(note that $\sum_{j=1}^{M}$ in (B1) reduces to $\sum_{j=2}^{M}$ because of (B2)).

The solution of this problem is easy (system of linear equations in α and β) :

$$\alpha = y_1 - \beta x_1, \tag{B4a}$$

where

$$\beta = \frac{\left(\dfrac{1}{M-1}\displaystyle\sum_{j=2}^{M} x_j y_j\right) - x_1\left(\dfrac{1}{M-1}\displaystyle\sum_{j=2}^{M} y_j\right) - y_1\left(\dfrac{1}{M-1}\displaystyle\sum_{j=2}^{M} x_j\right) + x_1 y_1}{\left(\dfrac{1}{M-1}\displaystyle\sum_{j=2}^{M} x_j^2\right) - 2x_1\left(\dfrac{1}{M-1}\displaystyle\sum_{j=2}^{M} x_j\right) + x_1^2} \qquad (B4b)$$

The calculation of this is rather easy : one makes tables of the values of x_j, y_j, x_j^2 and $x_j y_j$ ($j=2,\ldots,M$) and calculates the averages. These are the values needed between brackets in (B4b).

Once this is done, we correct the values (x_j, y_j). For x_j taking consecutive entire number values : replace y_j by ($j=1,\ldots,M$)

$$y_j' = y_j - \beta(j\text{-}1). \qquad (B5)$$

The points (x_j, y_j'), $j=1,\ldots,M$ represent the yearly fluctuations of the periodical growth under the assumption that the average growth is constant (on the level of the first year).

Example : the journal JASIS has the following sequence of (bound) volume thickness (as available in LUC) in mm :

x_j	1	2	3	4	5	6	7	8	9	10	11	12	13
y_j	32	33	32	32	33	43	54	58	44	54	53	64	70

Applying the above method this yields $\beta=2.594$ and $\alpha=29.406$. Note that only β is needed in formula (B5). This yields the corrected data (correction for increasing growth) :

x_j	1	2	3	4	5	6	7	8	9	10	11	12	13
y_j'	32	30.41	26.81	24.22	22.62	30.03	38.44	39.84	23.25	30.65	27.06	35.47	38.87

This set of data gives $\mu=30.74$ and $\sigma=5.96$ (use the unbiased estimator for the dispersion) yielding $V=0.19$. In total we checked 13 journals and in all cases we found that $V\in[0.1;0.2]$, where as μ could vary between 24.6 mm and 110.0 mm. We may, therefore, conclude that the assumption that $v=\frac{\sigma}{\mu}$ is approximately a constant is acceptable within our fixed group of periodicals, certainly as a first approximation. We would wellcome a theoretical rationale for it and also further experimentation in other sets of periodicals.

Appendix C

We use formula (14) to study the function $\varphi_m(j)$. It is easy to show that

$$\varphi_m'(j) = \binom{N}{m}\Phi_j^{N-m-1}(1-\Phi_j)^{m-1}g'(j)\Phi_j'(N-m-N\Phi_j) \qquad (C1)$$

(g as in (A1), dropping the index i, by assumption (9)).
Hence $\varphi_m'(j)=0$ if and only if

$$\Phi_j = \Phi\left(\frac{a-j}{\sqrt{j}\ V}\right) = 1 - \frac{m}{N}. \tag{C2}$$

In this point (call it j_m) we have that

$$\varphi_m''(j_m) = -\binom{N}{m}\Phi_j^{N-m-1}(1-\Phi_j)^{m-1}g'^2(j)N\Phi_j'^2 \tag{C3}$$

$$< 0,$$

always.

In conclusion : the probability for having m congested periodicals (m fixed) after j years is maximal if $j=j_m$, where $\Phi\left(\frac{a-j_m}{\sqrt{j_m}\ V}\right) = 1 - \frac{m}{N}$. This maximal value is, according to (14) equal to

$$\varphi_m(j_m) = \binom{N}{m}\left(\frac{m}{N}\right)^m\left(1-\frac{m}{N}\right)^{N-m} \tag{C4}$$