Percolation as a model for informetric distributions: fragment size distribution characterised by Bradford curves

Non Peer-reviewed author version

Bogaert, Jan; ROUSSEAU, Ronald & Van Hecke, Piet (2000) Percolation as a model for informetric distributions: fragment size distribution characterised by Bradford curves. In: Scientometrics, 47(2). p. 195-206.

DOI: 10.1023/A:1005678707987
Handle: http://hdl.handle.net/1942/821

PERCOLATION AS A MODEL FOR INFORMETRIC DISTRIBUTIONS: FRAGMENT SIZE DISTRIBUTION CHARACTERISED BY BRADFORD CURVES

JAN BOGAERT*, RONALD ROUSSEAU †, PIET VAN HECKE***

** University of Antwerp (UIA), Department of Biology,*

Universiteitsplein 1, 2610 Wilrijk, Belgium

*** KHBO, Zeedijk 101, B-8400 Oostende, Belgium*

and UIA, Department of Library and Information Science,

Universiteitsplein 1, 2610 Wilrijk, Belgium

Abstract

It is shown how Bradford curves, i.e. cumulative rank-frequency functions, as used in informetrics, can describe the fragment size distribution of percolation models. This interesting fact is explained by arguing that some aspects of percolation can be interpreted as a model for the success-breeds-success or cumulative advantage phenomenon. We claim, moreover, that the percolation model can be used as a model to study (generalised) bibliographies. This article shows how ideas and techniques studied and developed in informetrics and scientometrics can successfully be applied in other fields of science, and vice versa.

† Corresponding author.

1. Introduction

The notions 'informetric' or 'bibliometric' distributions refer to a set of mathematical representations and formulations of regularities observed in bibliographies, lists of authors, citation lists, and similar data [1]. R. and S. Rousseau [1] have shown that these regularities occur not only in library and information, or scientometric settings, but almost everywhere in the sciences, social sciences and humanities, including the word usage of popular song texts.

In this article, we focus on the so-called Leimkuhler representation, using Bradford curves [2]. Bradford curves represent cumulative rank-frequency distributions, usually (following Bradford [3]) represented on a semi-log scale. We have shown that the equation

$$R(r) = \frac{K}{2-\alpha}\left(M^{2-\alpha} - \left(M^{1-\alpha} + \frac{r(\alpha-1)}{K} \right)^{\frac{\alpha-2}{\alpha-1}} \right) \qquad (1)$$

where K, M and α are parameters, describes these Bradford curves quite well [2,4,5]. The symbol R(r) denotes the cumulative number of items produced by the r most productive sources. The parameter α is the most important one, being the same as the exponent in the Lotka distribution:

$$f(y) = \frac{C}{y^{\alpha}}, \quad y = 1, \ldots, y_{\max} \qquad (2)$$

where f(y) denotes the number of sources with production y. In the case that $\alpha = 2$, the denominator in (1) becomes zero and, hence this expression is not valid

anymore. Then a Bradford curve is described through Leimkuhler's function, i.e. a function of the form:

$$R(r) = a \ln(1 + b\, r) \qquad (3)$$

with parameters a and b.

2. Percolation maps in physics and in ecological studies

The percolation map is a two-dimensional square lattice with m x m cells. For every lattice cell, the probability to be occupied is denoted as p ($0 < p < 1$). Pixels that are vertical or horizontal neighbours form clusters of connected cells. It has been shown that, for an infinite lattice, cells will form a continuous infinite cluster spanning two sides of the map from $p > 0.5928$ on [6]. This value $p = p_c = 0.5928$ is known as the critical probability. In physical applications it corresponds to a phase transition [6,7,8]. For $p < p_c$, a lot of small clusters exist, spatially spread over the square lattice. Examples of percolation maps are presented in Fig.1. This figure clearly illustrates the effect of an increase in p. Higher p-values result in larger clusters with a higher degree of connectivity and in a smaller number of clusters. If $p > p_c$ a cluster connecting opposite sites of the map is observed.

Insert Fig.1 about here

Percolation theory finds its origins in physics. One of its aims is to study the flow of particles or energy through a porous lattice of grid cells. In general one may say that percolation is one of the best (and simplest) models for disordered media. To support this claim Grimmett [8] begins his book with four types of disordered physical

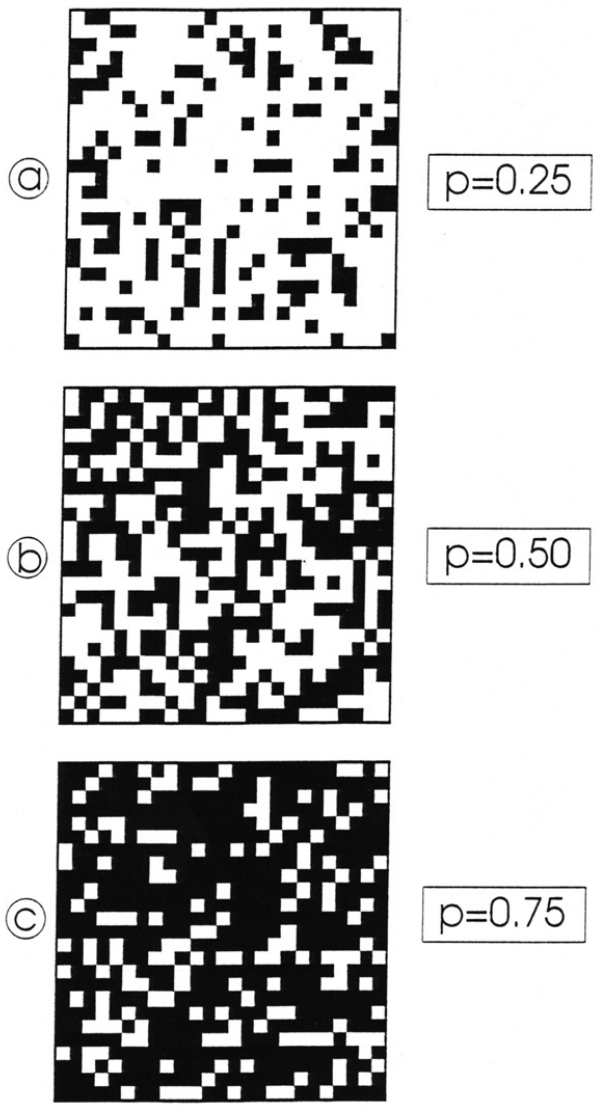

Fig.1 The effect of the percolation probability p on cluster formation in percolation maps (25 x 25). (a) p = 0.25; (b) p = 0.50; (c) p = 0.75.

systems, emphasising in each the role of the percolation model. These examples are: disordered electrical networks, ferromagnetism, epidemics and the manufacture of silicon wafers for microchips. The occurrence of critical phenomena such as phase transitions are central to the appeal of percolation.

Percolation maps have found their way in the ecological literature too. There they are often used to simulate habitat fragmentation [9], where they are considered as neutral models in landscape ecological research [10-13]. Indeed, in little more than a decade fragmentation, the breaking up of large habitat or land areas into smaller parcels, has become an environmental issue of world-wide proportion [14]. It can be considered as one of the most severe processes to depress biodiversity. The number and size of the remaining fragments will determine the number of species that can support sufficiently large populations to persist within each fragment and the number of subpopulations among fragments [15-18]. The smaller the fragmented blocks, the more the density of populations decreases and the risk of extinction grows. It, moreover, leads to geographical isolation, and therefore diminishes the recolonisation probability [19]. In these ecological applications, a percolation map with probability p is considered as a landscape with a particular degree of fragmentation. Hence, highly fragmented landscapes are characterised by the smallest p-values.

In this article we will show how informetric models can help understanding this kind of ecological and physical modelling. In particular we will relate our function R(r) to the cluster size distribution. It is yet another attempt to bridge the gap between the information sciences, in particular informetrics, and other fields of science [20].

3. Simulation procedure, results and curve fitting

Ninety nine percolation maps (m = 500, probability increment of 0.01) were generated using a FORTRAN-77 program. For random numbers, the Manly algorithm was used [21]. One seed number was replaced by the last four numbers of the clock time (system time in seconds). Patch recognition and area calculation (expressed as number of pixels) were executed using the geographical information system GRASS 4.1 (Geographical Resource Analysis Support System)[22]. Patch area data were converted to frequency and cumulative rank-frequency distributions.

A non-linear least squares algorithm was used to find the parameters of best fitting Bradford curves. These parameter values are shown in the appendix. Results of this fitting exercise were excellent until p = 0.57. For p = 0.58 and p = 0.59 visual inspection showed that the results were poor. The whole procedure broke down from p = 0.60 on. This shows that for these p-values the resulting patch distribution can not be described by a Bradford curve. This is also clear from the following example (p = 0.75). For p = 0.75 there is one enormous patch of 187883 pixels, while the second largest one contains only 12 pixels. It is clear that such inequalities in production do not fit into the classical informetric framework. Figs 2, 3 and 4 illustrate three interesting cases: for p = 0.10 we have a Bradford curve with a rising tail (α = 2.93), for p = 0.40 we have a classical Leimkuhler curve (α is almost 2; hence we have fitted the following Leimkuhler function: $R(r) = 24907.46 \ln(1 + 0.0025 r)$), finally for p = 0.53 we have a Bradford curve with a so-called Groos droop.

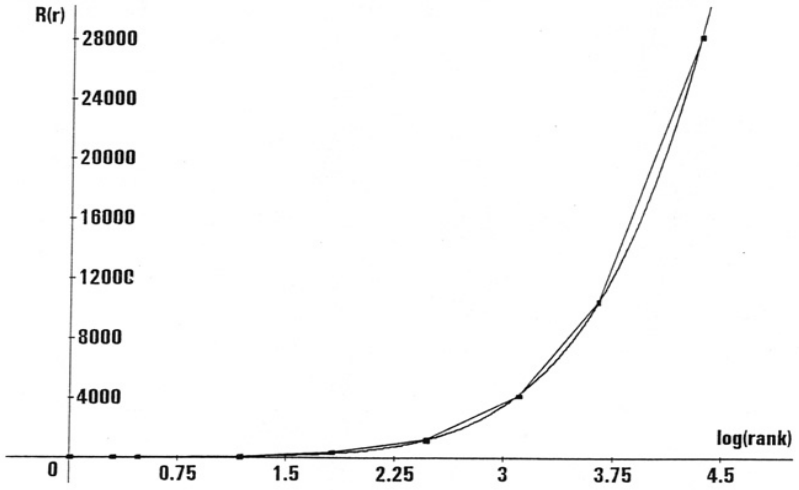

Fig.2 Data and best fitting Bradford curve for p = 0.10

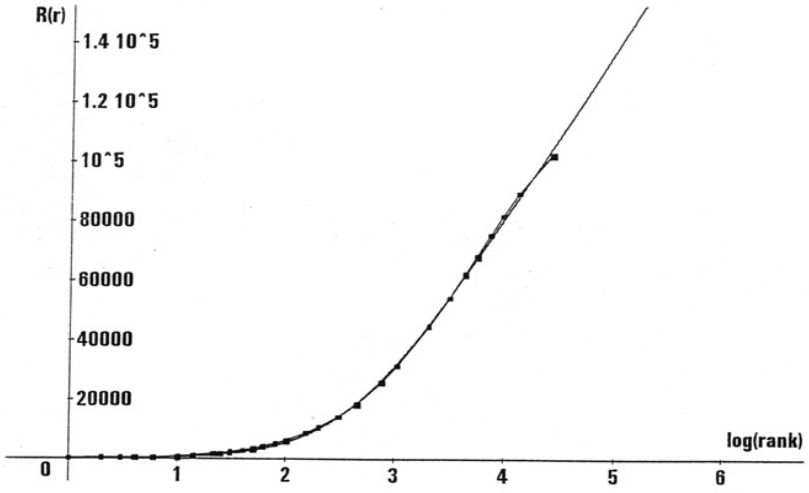

Fig. 3 Data and best fitting Bradford curve for p = 0.40

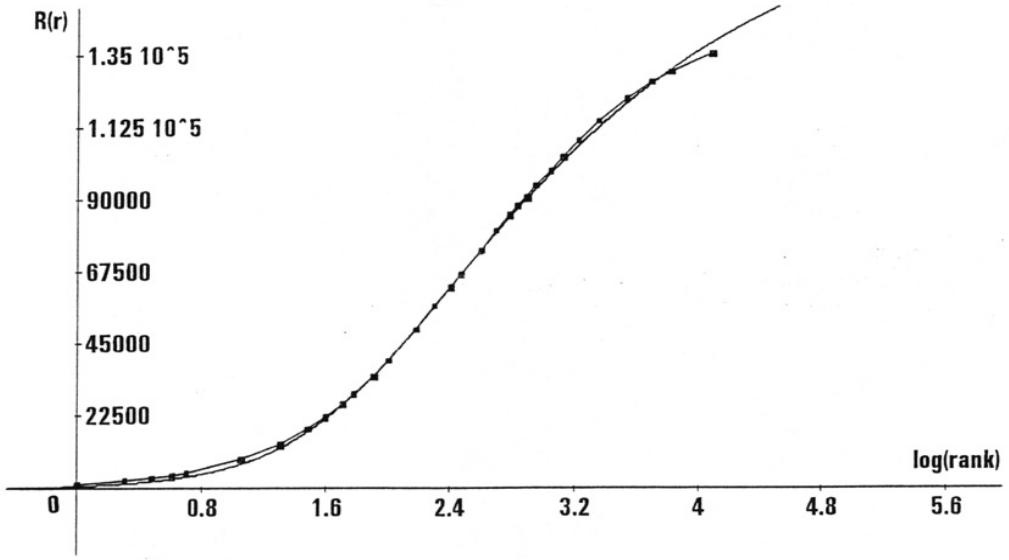

Fig.4 Data and best fitting Bradford curve for p = 0.53

Clearly something is happening around the critical value p_c = 0.5928. Next, we will use another technique in an attempt to make this visible. For small p values we have a lot of small clusters, while for p values larger than the critical value there is one big cluster and some smaller ones. It seemed to us that a concentration or a diversity measure, as used in informetric and ecological studies [23], should be able to measure this difference, and perhaps even show the 'phase transition'. We choose the Shannon-Wiener or entropy measure because of its links with physics. Recall that the entropy diversity measure H' is defined as:

$$H' = -\sum_{i=1}^{N} a_i \ln(a_i)$$

In this formula N denotes the number of clusters and a_i denotes the relative contribution of the i-th cluster. Concretely, if the i-th cluster consists of q_i lattice cells and there are S cells occupied, then $a_i = q_i/S$. Fig.5 shows the amazing result: around p = 59% the value of the entropy index suddenly changes. Note that in physical texts this effect is usually demonstrated by plotting only the size of the largest cluster.

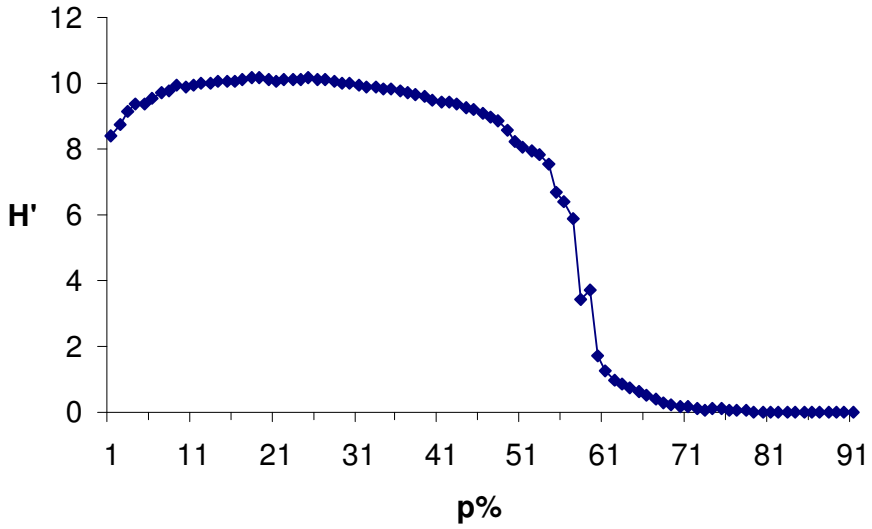

Fig.5 Entropy values showing a sudden decrease around the critical value $p_c = 0.5928$

4. Discussion

The outstanding results of our fitting exercise are not a coincidence. On the contrary, we will show that percolation is a concrete way to simulate the success-breeds-success (SBS) principle [24-29]. Hence, it is no surprise to obtain the results of the SBS-principle or the cumulative advantage effect (we use both terms as synonyms). Using Bradford's terminology [3], every lattice cell can be thought of as an article about a specific topic written (or not written) by a scientist (or group of scientists). In this way, every potential source (article) has a chance, given by p, to become active. Note that this aspect is usually not modelled in the SBS-model. If a source becomes active it can be isolated, corresponding to publishing in a new journal, or it can form a cluster with other, already existing, articles. This corresponds to an article being published in a journal that has already published articles on this particular topic. The larger the patch (journal) the larger the probability that the new article can join it and make it even larger (important). This is the success-breeds-success or cumulative advantage aspect of the procedure. Basically, a larger percolation cluster has a larger perimeter and therefore grows faster.

If p is small only relative few lattice cells are occupied. These cells moreover, have only a small chance to merge into larger patches. This situation corresponds to the emergence of a new field of science. The larger p the more cells are occupied, having also a larger probability to cluster. This corresponds to a more active and older field of science.

In this way percolation yields an interesting simulation of the activity of a science field. The larger p, the more mature the simulated subdomain of science (or studied

topic). The results of our fitting exercise clearly show that larger p values correspond to smaller alpha values (cf. Fig.6). Alpha-values smaller than 2 give cumulative rank-frequency distributions with a so-called 'Groos droop' [4]. Simulations using increasing p-values hence correspond to studying the evolution of a field [30,31].

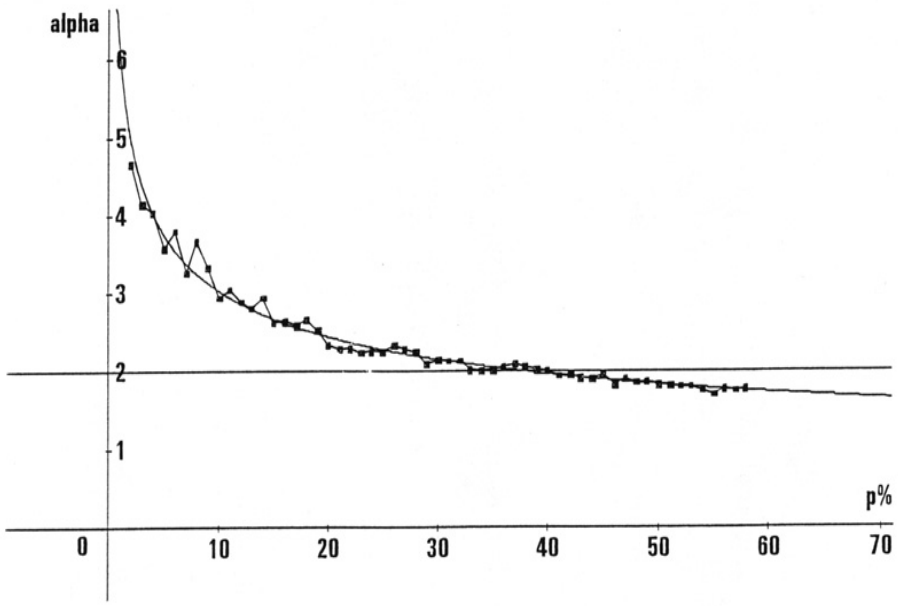

Fig.6 The relation between the parameter α and a lattice cell's probability to be occupied

We further note that the power law

$$\alpha = \frac{6.184}{(100p)^{0.311}}$$

yields an excellent fit to the data of Fig. 6. From an ecological perspective we may say that this monotone decreasing (p,α)-trend curve (Fig.6) enables the

characterisation of the fragmentation phenomenon by one single characteristic value, namely α.

Maps with $p > 0.60$ can not be described by the classical informetric distributions. This should not be interpreted as a disadvantage: landscapes represented by these maps can be considered as the result of a (very) low degree of fragmentation. Usually, studies are directed towards landscapes with an intermediate or high degree of fragmentation for which an adequate conservation policy can compensate for these negative consequences of the fragmentation process.

Finally, we have shown how an inequality measure (in this case the entropy diversity measure) is able to demonstrate the phase transition in percolation maps.

5. Conclusion

Percolation seems to be an interesting technique to simulate generalised bibliographies (IPPs), such as publication lists of scientific institutes, and their time evolution. This can best be appreciated by comparing our Figs 2,3,4 with semi-log plots of existing bibliographies as in 'Bradford curves'[2]. Besides this promising application in informetrics, we note a new interpretation in ecology. Indeed, the α-value can be used as a quantitative indicator of habitat fragmentation in the sense that high values of this indicator are related to high p-values, i.e. highly fragmented landscapes. This article shows that it is possible to give knowledge and techniques used in scientometrics and informetrics a new interpretation in other fields of science.

Acknowledgement

Jan Bogaert is Research Assistant of the Fund for Scientific Research - Flanders (F.W.O.). We thank the referees for a number of pertinent observations leading to a, hopefully, better article.

References

1. R. ROUSSEAU, S. ROUSSEAU, Informetric distributions: a tutorial review. *Canadian Journal of Information and Library Science*, 18 (1993) 51-63.
2. R. ROUSSEAU, Bradford curves. *Information Processing & Management*, 30 (1994) 267-277.
3. S.C. BRADFORD, Sources of information on specific subjects. *Engineering*, 137 (1934) 85-86. Reprinted in: *Journal of Information Sciences* 10 (1985) 176-180.
4. R. ROUSSEAU, Lotka's law and its Leimkuhler representation. *Library Science with a Slant to Documentation and Information Studies*, 25 (1988) 150-178.
5. R. ROUSSEAU, A weak goodness-of-fit test for rank-frequency distributions. In: (C. Macias-Chapula, ed.), *Proceedings of the Seventh Conference of the International Society for Scientometrics and Informetrics,* Universidad de Colima (Mexico), 1999, 421- 430.
6. A. BUNDE, S. HAVLIN, A brief introduction to fractal geometry, in: A. BUNDE, S. HAVLIN (Eds.), *Fractals in Science*, Springer-Verlag, Berlin, 1995, 1-25.
7. D. STAUFFER, *Introduction to percolation theory*. Taylor & Francis, London, UK, 1985.
8. G. GRIMMETT, *Percolation*. New York (etc.): Springer-Verlag (1989).

9. J. BOGAERT, I. IMPENS, Generating random percolation clusters. *Applied Mathematics and Computation*, 91 (1998) 197-208.
10. J. BOGAERT, P. VAN HECKE, R. MOERMANS, I. IMPENS, Twist number statistics as an additional measure of habitat perimeter irregularity. *Environmental and Ecological Statistics*, 9 (1999) 275-290.
11. R.H. GARDNER, B.T. MILNE, M.G. TURNER, R.V. O'NEILL, Neutral models for the analysis of broad-scale landscape pattern. *Landscape Ecology*, 1 (1987) 19-28.
12. R.H. GARDNER, M.G. TURNER, V.H. DALE, R.V. O'NEILL, A percolation model of ecological flows, in: A.J. HANSEN, F. DI CASTRI (Eds.) *Landscape boundaries. Consequences for biotic diversity and ecological flows*, Springer-Verlag, New York, USA, 1992, 259-269.
13. K.A. WITH, A.W. KING, The use and misuse of neutral landscape models in ecology. *Oikos,* 79 (1997) 219-299.
14. R.T.T. FORMAN, *Land mosaics. The ecology of landscapes and regions*, Cambridge University Press, Cambridge, UK. , 1997, p.406.
15. D.J. BENDER, T.A. CONTRERAS, L. FAHRIG, Habitat loss and population decline: a meta-analysis of the patch size effect. *Ecology*, 79 (1998) 517-533.
16. M.J. GROOM, N. SCHUMACKER, Evaluating landscape change: patterns of worldwide deforestation and local fragmentation, in: P.M. KAREIVA, J.G. KINGSOLVER & R.B. HUEY (Eds.), *Biotic interactions and global change*, Sinauer Associates Inc., Sunderland, MA, USA, 1993, 24-44.
17. J.M. LORD, D.A. NORTON, Scale and the spatial concept of fragmentation. *Conservation Biology*, 4 (1990) 197-202.

18. D.A. SAUNDERS, R.J. HOBBS, C.R. MARGULES, Biological consequences of ecosystem fragmentation: a review. *Conservation Biology*, 5 (1991) 18-32.

19. A. FARINA, *Principles and methods in landscape ecology*, Chapman & Hall, London, UK, 1998, p.58.

20 L. EGGHE, Bridging the gaps - conceptual discussions on informetrics, Scientometrics , 30 (1994) 35-47.

21. B.F.J. MANLY, Randomization and Monte Carlo methods in biology, Chapman & Hall, London, UK., 1991, p.38-42.

22. USA-CERL, *GRASS 4.1 User's reference manual*, United States Army Corps of Engineers. Construction Engineering Research Laboratories, Champaign, Illinois, USA, 1993.

23. D. NIJSSEN, R. ROUSSEAU, P. VAN HECKE, The Lorenz curve: a graphical representation of evenness, *Coenoses* 13 (1998) 33-38.

24. H.A. SIMON, On a class of skew distributions*, Information and Control* 52 (1955) 425-440.

25. D. DE SOLLA PRICE, A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27.

26. J. TAGUE, The success-breeds-success phenomenon and bibliometric processes. *Journal of the American Society for Information Science*, 32 (1981) 280-286.

27. W. GLÄNZEL, A. SCHUBERT, The cumulative advantage function. A mathematical formulation based on conditional expectations and its application to scientometric distributions, in: (L. EGGHE, R. ROUSSEAU, eds), *Informetrics 89/90,* Amsterdam, Elsevier, 1990, 139-147.

28. Y.-S. CHEN, P.P. CHONG, M.Y.TONG, Dynamic behavior of Bradford's law. *Journal of the American Society for Information Science*, 46 (1995) 370-383.
29. L. EGGHE, R. ROUSSEAU, Generalized success-breeds-success principle leading to time-dependent informetric distributions. *Journal of the American Society for Information Science,* 46 (1995) 426-445.
30. G.M. BRAGA, Some aspects of the Bradford's distribution. *Proceedings of the American Society for Information Science Annual Meeting*, 29 (1978) 51-54.
31. V. OLUIĆ-VUKOVIĆ, Journal productivity distribution: quantitative study of dynamic behavior. *Journal of the American Society for Information Science*, 43 (1992) 412-421.

Appendix: Results of fitting equation (1) to percolation data: parameter values and R^2

p%	α	K	M	R^2
2	4.67	10045	3.67	1
3	4.15	13833	3.56	1
4	4.04	17522	3.72	0.99999
5	3.59	15813	4.17	0.99999
6	3.79	19629	4.86	0.99999
7	3.26	20275	4.02	0.99999
8	3.67	26071	5.90	0.99994
9	3.33	24645	4.71	0.99998
10	2.93	23893	4.56	0.99998
11	3.03	25523	5.06	0.99998
12	2.88	26773	4.92	0.99995
13	2.80	26544	5.44	0.99997
14	2.92	29324	8.17	0.99990
15	2.63	26437	7.03	0.99997
16	2.64	28669	7.29	0.9999
17	2.58	28690	8.24	0.9999
18	2.65	31622	8.71	0.9998
19	2.52	30352	8.65	0.9998
20	2.32	27244	8.97	0.9999
21	2.27	26428	9.85	0.9998
22	2.30	28419	10.94	0.9996
23	2.22	27315	10.83	0.9997
24	2.25	29047	11.37	0.9996
25	2.24	29971	11.62	0.9997
26	2.33	34196	15.14	0.9992
27	2.27	32657	15.57	0.9992
28	2.24	32368	16.03	0.9991
29	2.08	28076	15.98	0.9994
30	2.14	30232	19.12	0.9991
31	2.12	30155	22.34	0.9990
32	2.12	29871	27.22	0.9990
33	2.01	26026	22.83	0.9991
34	2.01	26265	26.77	0.9986
35	2.01	27309	26.22	0.9988
36	2.05	26910	36.41	0.9992
37	2.08	29186	39.78	0.9983
38	2.05	27955	44.59	0.9988
39	2.02	26663	48.94	0.9992
40	2.00	25119	62.18	0.9993
41	1.94	22226	58.84	0.9992
42	1.95	22372	65.48	0.9994
43	1.91	21993	63.26	0.9986
44	1.89	21158	71.98	0.9982
45	1.95	23427	91.51	0.9979
46	1.82	17537	87.85	0.9981

47	1.90	20595	123.5	0.9972
48	1.85	17043	139.2	0.9987
49	1.86	17041	207.1	0.9982
50	1.82	12654	354.5	0.9992
51	1.81	11321	481.3	0.9994
52	1.80	10845	530.2	0.9994
53	1.80	10164	734.5	0.9995
54	1.76	7965	758.7	0.9992
55	1.68	4278	2113	0.9987
56	1.78	5425	5669	0.9997
57	1.74	3744	8888	0.9989
58	1.77	3092	41540	0.9944
59	1.76	1944	213129	0.9837