ELSEVIER
ACADEMIC
PRESS

Power Laws in the Information Production Process: Lotkaian Informetrics

Leo Egghe

Power Laws in the Information Production Process: Lotkaian Informetrics

Leo Egghe

2005

ELSEVIER
ACADEMIC
PRESS

Amsterdam – Boston – Heidelberg – London – New York – Oxford – Paris
San Diego – San Francisco – Singapore – Sydney – Tokyo

ELSEVIER B.V.
Radarweg 29
P.O.Box 211, 1000 AE
Amsterdam, The Netherlands

ELSEVIER Inc.
525 B Street, Suite 1900
San Diego, CA 92101-4495
USA

ELSEVIER Ltd
The Boulevard, Langford Lane
Kidlington, Oxford OX5 1GB
UK

ELSEVIER Ltd
84 Theobalds Road
London WC1X 8RR
UK

© 2005 Elsevier Ltd. All rights reserved.

This work is protected under copyright by Elsevier Ltd, and the following terms and conditions apply to its use:

Photocopying
Single photocopies of single chapters may be made for personal use as allowed by national copyright laws. Permission of the Publisher and payment of a fee is required for all other photocopying, including multiple or systematic copying, copying for advertising or promotional purposes, resale, and all forms of document delivery. Special rates are available for educational institutions that wish to make photocopies for non-profit educational classroom use.

Permissions may be sought directly from Elsevier's Rights Department in Oxford, UK: phone (+44) 1865 843830, fax (+44) 1865 853333, e-mail: permissions@elsevier.com. Requests may also be completed on-line via the Elsevier homepage (http://www.elsevier.com/locate/permissions).

In the USA, users may clear permissions and make payments through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA; phone: (+1) (978) 7508400, fax: (+1) (978) 7504744, and in the UK through the Copyright Licensing Agency Rapid Clearance Service (CLARCS), 90 Tottenham Court Road, London W1P 0LP, UK; phone: (+44) 20 7631 5555; fax: (+44) 20 7631 5500. Other countries may have a local reprographic rights agency for payments.

Derivative Works
Tables of contents may be reproduced for internal circulation, but permission of the Publisher is required for external resale or distribution of such material. Permission of the Publisher is required for all other derivative works, including compilations and translations.

Electronic Storage or Usage
Permission of the Publisher is required to store or use electronically any material contained in this work, including any chapter or part of a chapter.

Except as outlined above, no part of this work may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission of the Publisher.
Address permissions requests to: Elsevier's Rights Department, at the fax and e-mail addresses noted above.

Notice
No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

First edition 2005

ISBN: 0 12 088753 3

∞ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER BOOK AID International Sabre Foundation

Power Laws in the Information Production Process: Lotkaian Informetrics

Library and Information Science

Series Editor: *Bert R. Boyce*
School of Library & Information Science
Louisiana State University, Baton Rouge

Recent and Forthcoming Volumes

Mike Thelwall
An Information Science Approach

Donald Case
Looking for Information

Matthew Locke Saxton and John V. Richardson
Understanding Reference Transactions: Turning Art Into a Science

Robert M. Hayes
Models for Library Management, Decision-Making, and Planning

Charles T. Meadow, Bert R. Boyce, and Donald H. Kraft
Text Information Retrieval Systems, Second Edition

Charles T. Meadow
Text Information Retrieval Systems

A. J. Meadows
Communicating Research

V. Frants, J. Shapiro, & V. Votskunskii
Automated Information Retrieval: Theory and Methods

Harold Sackman
Biomedical Information Technology: Global Social Responsibilities for the Democratic Age

Peter Clayton
Implementation of Organizational Innovation: Studies of Academic and Research Libraries

Bryce L. Allen
Information Tasks: Toward a User-Centered Approach to Information Systems

试读结束：需要全本请在线购买： www.ertongbook.com

My wife is thanking me for the
many quiet evenings at the time
of the writing of this book.

To Ute, because she asked for it.

PREFACE

Explaining and hence understanding is one of the key characteristics of human beings. Explaining is making logical, mathematical deductions based on a minimum of unexplained properties, called axioms. Indeed, without axioms, one is not able to make deductions. How these axioms are selected is the only non-explainable part of the theory. In fact, different choices are possible leading to different, in themselves consistent, theories which conflict when considered together. A typical example is the construction of the different types of geometries, Euclidean and non-Euclidean geometries which are in contradiction when considered together but which have their own applications.

In this book the object of study is a two-dimensional information production process, i.e. where one has sources (e.g. journals, authors, words, ...) which produce (or have) items (e.g. respectively articles, publications, words occurring in texts, ...) and in which one considers different functions describing quantitatively the production quantities of the different sources. All functions can be reduced to one type of function, namely the size-frequency function f: such a function gives, for every n = 1, 2, 3, ..., the number f(n), being the number of sources with n items. This is the framework of study and in this framework we want to explain as many regularities that one encounters in the literature as possible, using the size-frequency function f.

As explained above, we need at least one axiom. The only axiom used in this book is that the size-frequency function f is Lotkaian, i.e. a power function of the form $f(n)=C/n^{\alpha}$, where $C>0$ and $\alpha^{3}\ 0$, hence also implying that the function f decreases. The name comes from its introduction into the literature by Alfred Lotka in 1926, see Lotka (1926). Based on this one assumption, one can be surprised about the enormous amounts of regularities that can be explained. In all cases the parameter α turns out to be crucial and is capable of, dependent on the different values that α can take, explaining different shapes of one phenomenon.

In this book we encounter explanations in the following directions: other informetric functions that are equivalent with Lotka's law (e.g. Zipf's law), concentration theory

(theory of inequality), fractal theory, modelling systems in which items can have multiple sources (as is the case in the system: articles written by several authors), modelling citation distributions. These models are developed from Chapter II on.

Although the law of Lotka is an axiom in this book, in Chapter I we investigate what other models are capable of "explaining" Lotka's law. The only purpose for these developments is to understand why Lotka's law is choosen in this book (and not another type of size-frequency function such as, e.g., an exponential function) and hence to make the choice acceptable (although this is not needed, strictly speaking). For the same reason we also give an overview of situations where Lotka's law is encountered, being the majority of the situations (including the "new" situations as networks, including the Internet).

The author is basing himself on the publications that he has written the past 20 years on the subject but also benifits from the work of many others. To them my sincerest thanks. The author is especially grateful to Professor Ronald Rousseau (co-winner of the 2001 Derek De Solla Price Award) with whom he co-authored several papers but with whom he also had numerous long discussions (often by phone) on the different topics described in this book.

We strongly hope this book will serve the informetrics community in the sense that it shows the logical links between many (at first sight unrelated) informetrics aspects. The informetrician, having read this book, can use it each time he/she encounters a new informetric phenomenon (often in the form of a data set) in the sense that one can investigate if the phenomenon shows regularities that are (or can be) explained using the arguments given in this book. The mathematical knowledge required is limited to elementary mathematics such as first-year calculus. Other, more advanced topics, are introduced in this book.

The author is indebted to the Limburgs Universitair Centrum (LUC) and the University of Antwerp (UA) for their support in doing informetric research: in LUC, the author is chief librarian and coordinator of the research project "bibliometrics" while in UA, he is professor in the School of Library and Information Science, where he teaches the courses on informetrics and on information retrieval. The author thanks

Mr. M. Pannekoeke (LUC) for the excellent typing and organization of this manuscript.

Leo Egghe
Diepenbeek, Belgium
Summer 2004

TABLE OF CONTENTS

Preface vii

Table of contents xi

Introduction 1

Chapter I Lotkaian Informetrics: An Introduction 7

I.1 Informetrics 7

I.2 What is Lotkaian informetrics ? 14

I.2.1 The law of Lotka 14

I.2.2 Other laws that are valid in Lotkaian informetrics 19

I.3 Why Lotkaian informetrics ? 25

I.3.1 Elementary general observations 26

I.3.2 The scale-free property of the size-frequency function f 27

I.3.3 Power functions versus exponential functions for the size-frequency function f 32

I.3.4 Proof of Lotka's law based on exponential growth or based on exponential obsolescence 34

I.3.4.1 Proof of Lotka's law based on exponential growth: the Naranan model 34

I.3.4.2 Proof of Lotka's law based on exponential obsolescence: solution of a problem of Buckland 40

I.3.5 Derivation of Mandelbrot's law for random texts 42

I.3.6 "Success Breeds Success" 45

I.3.6.1 The urn model 46

I.3.6.2 General definition of SBS in general IPPs 49

I.3.6.3 Approximate solutions of the general SBS 52

I.3.6.4 Exact results on the general SBS and explanation of its real nature 55

I.3.7 Entropy aspects 65

I.3.7.1 Entropy: definition and properties 66

I.3.7.2 The Principle of Least Effort (PLE) and its relation with the law of Lotka 70

I.3.7.3 The Maximum Entropy Principle (MEP) 76

I.3.7.4 The exact relation between (PLE) and (MEP) 78

I.4 Practical examples of Lotkaian informetrics 85

I.4.1 Important remark 85

I.4.2 Lotka's law in the informetrics and linguistics literature 86

I.4.3 Lotka's law in networks 87

I.4.4 Lotka's law and the number of authors per paper 90

I.4.5 Time dependence and Lotka's law 92

I.4.6 Miscellaneous examples of Lotkaian informetrics 94

I.4.7 Observations of the scale-free property of the size-frequency function f 98

Chapter II Basic Theory of Lotkaian Informetrics 101

II.1 General informetrics theory 101

II.1.1 Generalized bibliographies: Information Production Processes (IPPs) 101

II.1.2 General informetric functions in an IPP 104

II.1.3 General existence theory of the size-frequency function 110

II.2 Theory of Lotkaian informetrics 114

II.2.1 Lotkaian function existence theory 114

II.2.1.1 The case $\rho_m = \infty$ 114

II.2.1.2 The general case $\rho_m < \infty$ 116

II.2.2 The informetric functions that are equivalent with a Lotkaian size-frequency function f 121

试读结束：需要全本请在线购买： www.ertongbook.com

II.3 Extension of the general informetrics theory: the dual size-frequency function h 144

II.4 The place of the law of Zipf in Lotkaian informetrics 150

II.4.1 Definition and existence 150

II.4.2 Functions that are equivalent with Zipf's law 152

Chapter III Three-dimensional Lotkaian Informetrics 157

III.1 Three-dimensional informetrics 157

III.1.1 The case of two source sets and one item set 158

III.1.2 The case of one source set and two item sets 159

III.1.3 The third case: linear three-dimensional informetrics 161

III.1.3.1 Positive reinforcement 163

III.1.3.2 Type/Token-Taken informetrics 168

III.1.4 General notes 172

III.2 Linear three-dimensional Lotkaian informetrics 175

III.2.1 Positive reinforcement in Lotkaian informetrics 175

III.2.2 Lotkaian Type/Token-Taken informetrics 177

Chapter IV Lotkaian Concentration Theory 187

IV.1 Introduction 187

IV.2 Discrete concentration theory 192

IV.3 Continuous concentration theory 196

IV.3.1 General theory 196

IV.3.2 Lotkaian continuous concentration theory 199

IV.3.2.1 Lorenz curves for power laws 199

IV.3.2.2 Concentration measures for power laws 205

IV.3.3 A characterization of Price's law of concentration in terms of Lotka's law and of Zipf's law 214

IV.4 Concentration theory of linear three-dimensional informetrics 218

IV.4.1 The concentration of positively reinforced IPPs 219

IV.4.2 Concentration properties of Type/Token-Taken informetrics 226

Chapter V Lotkaian Fractal Complexity Theory 231

V.1 Introduction 231

V.2 Elements of fractal theory 232

V.2.1 Fractal aspects of a line segment, a rectangle and a parallelepiped 233

V.2.2 The triadic von Koch curve and its fractal properties. Extension to general self-similar fractals 234

V.2.3 Two general ways of expressing fractal dimensions 236

V.2.3.1 The Hausdorff-Besicovitch dimension 236

V.2.3.2 The box-counting dimension 239

V.3 Interpretation of Lotkaian IPPs as self-similar fractals 242

Chapter VI Lotkaian Informetrics of Systems in which Items can have Multiple Sources 247

VI.1 Introduction 247

VI.2 Crediting systems and counting procedures for sources and "super sources" in IPPs where items can have multiple sources 253

VI.2.1 Overview of crediting systems for sources 254

VI.2.1.1 First or senior author count 254

VI.2.1.2 Total author count 254

VI.2.1.3 Fractional author count 255

VI.2.1.4 Proportional author count 255

VI.2.1.5 Pure geometric author count 255

VI.2.1.6 Noblesse Oblige 256

VI.2.2 Crediting systems for super sources 256

VI.2.3 Counting procedures for super sources in an IPP 256

VI.2.3.1 Total counting 257

VI.2.3.2 Fractional counting 258

VI.2.3.3 Proportional counting 258

VI.2.4 Inequalities between $Q_T(c)$ and $Q_F(c)$ and consequences for the comparison of $Q_T(c)$, $Q_F(c)$ and $Q_P(c)$ 261

VI.2.5 Solutions to the anomalies 266

VI.2.5.1 Partial solutions 267

VI.2.5.2 Complete solution to the encountered anomalies 269

VI.2.6 Conditional expectation results on $Q_T(c)$, $Q_F(c)$ and $Q_P(c)$ 270

VI.3 Construction of fractional size-frequency functions based on two dual Lotka laws 276

VI.3.1 Introduction 276

VI.3.2 A continuous attempt: $z \in \mathbb{R}^+$ 278

VI.3.3 A rational attempt: $q \in \mathbb{Q}^+$ 282

Chapter VII Further Applications in Lotkaian Informetrics 295

VII.1 Introduction 295

VII.2 Explaining "regularities" 297

VII.2.1 The arcs at the end of a Leimkuhler curve 297

VII.2.2 A "type/token-identity" of Chen and Leimkuhler 298

VII.3 Probabilistic explanation of the relationship between citation age and journal productivity 300

VII.4 General and Lotkaian theory of the distribution of author ranks in multi-authored papers 304

VII.4.1 General theory 304

VII.4.2 Modelling the author rank distribution using seeds 308

VII.4.3 Finding a seed based on alphabetical ranking of authors 310

VII.5 The first-citation distribution in Lotkaian informetrics 313

VII.5.1 Introduction 313

VII.5.2 Derivation of the model 317

VII.5.3 Testing of the model 320

VII.5.3.1 First example: Motylev (1981) data 320

VII.5.3.2 Second example: JACS to JACS data of Rousseau 322

VII.5.4 Extensions of the first-citation model 323

VII.6 Zipfian theory of N-grams and of N-word phrases: the Cartesian product of IPPs 326

VII.6.1 N-grams and N-word phrases 326

VII.6.2 Extension of the argument of Mandelbrot to 2-word phrases 329

VII.6.3 The rank-frequency function of N-grams and N-word phrases based on Zipf's law for N = 1 333

VII.6.4 The size-frequency function of N-grams and N-word phrases derived from Subsection VII.6.3 347

VII.6.5 Type/Token averages μ_N and Type/Token-Taken averages μ_N^* for N-grams and N-word phrases 352

Appendix 365

Appendix I 365

Appendix II 370

Appendix III Statistical determination of the parameters in the law of Lotka 372

A.III.1 Statement of the problem 372

A.III.2 The problem of incomplete data (samples) and Lotkaian informetrics 373

A.III.3 The difference between the continuous Lotka function and the discrete Lotka function 378

A.III.4 Statistical determination of the parameters K, a, n_{max} in the discrete Lotka function K/n^a, $n = 1,...,n_{max}$ 386

A.III.4.1 Quick and Dirty methods 387

A.III.4.2 Linear Least Squares method 388

A.III.4.3 Maximum Likelihood Estimating method 390

A.III.5 General remarks 393

A.III.5.1 Fitting Zipf's function 393

A.III.5.2 The estimation of ρ_m and n_{max} 394

A.III.5.3 Fitting derived functions such as Price's law 394

A.III.5.4 Goodness-of-fit tests 395

Bibliography 397

Subject Index 423

INTRODUCTION

The most facinating aspect of informetrics is the study of what we could call two-dimensional informetrics. In this discipline one considers sources (e.g. journals, authors, ...) and items (being produced by a source - e.g. articles) and their interrelations. By this we mean the description of the link that exists between sources and items. Without the description of this link we would have two times a one dimensional informetrics study, one for the sources and one for the items. Essentially in two-dimensional informetrics the link between sources and items is described by two possible functions: a size-frequency function f and a rank-frequency function g. Although one function can be derived from the other, they are different descriptions of two-dimensionality. A size-frequency function f describes the number f(n) of sources with n = 1, 2, 3, ... items while a rank-frequency function g describes the number g(r) of items in the source on rank r = 1, 2, 3, ... (where the sources are ranked in decreasing order of the number of items they contain (or produce)). So, in essence, in f and g, the role of sources and items are interchanged. This is called duality, hence f and g can be considered as dual functions.

Rank-frequency functions are well-known in the literature, especially in the economics and linguistics literature, where one usually considers Pareto's law and Zipf's law, respectively, being power laws. Less encountered in the literature (except in information sciences) is the size-frequency function. If studied, one supposes in most cases also a power law for such a function, i.e. a function of the type f(n)=C / n^{α} with $\alpha \geq 0$. Such a function is then called the law of Lotka referring to its introduction in the informetrics literature in 1926, see Lotka (1926). The law of Lotka gives rise to a variety of derived results in informetrics, the description of them being the subject of this book. That we choose a size-frequency function as the main study-object is explained e.g. by its simplicity in formulation (in the discrete setting simpler than a rank-frequency function since the latter uses ranks which have been derived from the "sizes" n but also in the continuous setting, where sizes and ranks are taken in an interval, the formulation of the size-frequency function is more appealing and direct). A size-frequency function also allows for a study of fractional quantities (see Chapter VI), needed e.g. in the description of two-dimensional informetrics in which

试读结束：需要全本请在线购买： www.ertongbook.com