SpringerBriefs in Complexity

SpringerBriefs in Complexity are a series of slim high-quality publications encompassing the entire spectrum of complex systems science and technology. Featuring compact volumes of 50 to 125 pages (approximately 20,000–45,000 words), Briefs are shorter than a conventional book but longer than a journal article. Thus Briefs serve as timely, concise tools for students, researchers, and professionals.

Typical texts for publication might include:

- A snapshot review of the current state of a hot or emerging field
- A concise introduction to core concepts that students must understand in order to make independent contributions
- An extended research report giving more details and discussion than is possible in a conventional journal article,
- A manual describing underlying principles and best practices for an experimental or computational technique
- An essay exploring new ideas broader topics such as science and society

Briefs allow authors to present their ideas and readers to absorb them with minimal time investment. Briefs are published as part of Springer's eBook collection, with millions of users worldwide. In addition, Briefs are available, just like books, for individual print and electronic purchase. Briefs are characterized by fast, global electronic dissemination, straightforward publishing agreements, easy-to-use manuscript preparation and formatting guidelines, and expedited production schedules. We aim for publication 8-12 weeks after acceptance.

SpringerBriefs in Complexity are an integral part of the Springer Complexity publishing program. Proposals should be sent to the responsible Springer editors or to a member of the Springer Complexity editorial and program advisory board (springer.com/complexity).

More information about this series at http://www.springer.com/series/8907

Michael Golosovsky

Citation Analysis and Dynamics of Citation Networks

Springer

Michael Golosovsky
Racah Institute of Physics
Hebrew University of Jerusalem
Jerusalem, Israel

ISSN 2191-5326 ISSN 2191-5334 (electronic)
SpringerBriefs in Complexity
ISBN 978-3-030-28168-7 ISBN 978-3-030-28169-4 (eBook)
https://doi.org/10.1007/978-3-030-28169-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book belongs to the science of science. The idea of this book appeared in 2007–2010 when I attended the interdisciplinary seminar of Prof. Sorin Solomon in the Racah Institute of Physics, Hebrew University of Jerusalem. The purpose of the seminar was to construct physical models of social phenomena. My long experience with Web of Science suggested me to look for citations to scientific papers and to try to model their dynamics as physicists do. The modeling of citation dynamics has been popular among physicists, and almost all such models were built by theoreticians. These models were quite general and mathematically rigorous but lacked proper calibration, namely, comparison to measurements was insufficient. I set my goal to build a fully calibrated and validated model of citation dynamics. To achieve this goal, I hoped to use my background and experience in experimental solid-state physics which should help me to design special measurements for model validation.

This book presents a stochastic model of citation dynamics which is based on the well-known copying or redirection mechanism and which was built using methods of network science. The combination of modeling and measurement revealed that citation dynamics of scientific papers is nonlinear. This nonlinearity has far-reaching consequences including nonstationary citation distributions, diverging citation trajectories of similar papers, and runaways or "immortal papers" with an infinite citation life-span.

This book presents a fully calibrated and validated model of citation dynamics. It can serve as a practical tool for quantitative analysis and forecasting of citations and impact factors. This book appeals to students and researchers in network science, citation analysis, and bibliometrics.

I am indebted to Sorin Solomon who introduced me into the wonderful world of complexity, supported me through all stages of this research, and induced me to write this book. I am grateful to Sidney Redner and Peter Richmond for their encouragement and advices.

Jerusalem, Israel
December 2018

Michael Golosovsky

Abstract

We consider network of citations of scientific papers and use a combination of theoretical and experimental tools to uncover microscopic details of its growth. Namely, we develop a stochastic model of citation dynamics based on copying/redirection/triadic closure mechanism. In a complementary and coherent way, the model accounts both for statistics of references of scientific papers and for their citation dynamics. Originating in empirical measurements, the model is cast in such a way that it can be verified quantitatively in every aspect. Such verification is performed by measuring citation dynamics of Physics papers. The measurements revealed nonlinear citation dynamics, the nonlinearity being intricately related to network topology. The nonlinearity has far-reaching consequences including non-stationary citation distributions, diverging citation trajectory of similar papers, runaways or "immortal papers" with infinite citation lifetime etc. Nonlinearity in complex network growth is our most important finding. In a more specific context, our results can be a basis for quantitative probabilistic prediction of citation dynamics of individual papers and of the journal impact factor.

Contents

1 Introduction 1
1.1 The Place of Citation Analysis in Science 1
1.1.1 Bibliometrics 1
1.1.2 Discrete Power-Law Statistical Distributions 3
1.1.3 Complex Networks 4
1.2 The Purpose of the Book: A Quantitative Microscopic Model of Citation Dynamics 4

2 Complex Network of Scientific Papers 7
2.1 Statistics of References and Citations 7
2.2 Expanding Science 10
2.3 Temporal Aspect of the Citation Network 12
2.3.1 Age Distribution of References 12
2.3.2 Age Distribution of Citations 13
2.3.3 Reference-Citation Duality 15

3 Stochastic Modeling of References and Citations 19
3.1 Copying/Recursive Search Model of the Citation Process 19
3.1.1 Scenario: The Author's Strategy to Compose the Reference List of a Paper 19
3.1.2 Recursive Search Model: Mathematical Formalism 21
3.1.3 Recursive Search Algorithm 24
3.2 Modeling Age Distribution of References in the Reference Lists of Scientific Papers 24
3.2.1 A Mean-Field Model for References 24
3.2.2 Model Calibration: Direct and Indirect References 26
3.3 Stochastic Model of Citation Dynamics of Individual Papers 28
3.4 Continuous Approximation of the Model 31
3.4.1 Relation to the Bass Model 31
3.4.2 Analytic Solution 32

4 Citation Dynamics of Individual Papers: Model Calibration ... 35
4.1 Measurements ... 35
4.1.1 Methodology ... 35
4.1.2 Direct Citations ... 37
4.1.3 Indirect Citations ... 38
4.2 Degree-Degree Correlations in the Citation Network ... 39
4.2.1 Statistics of the Second-Generation Citing Papers ... 39
4.2.2 Probability of Indirect Citation ... 41

5 Model Validation ... 45
5.1 Numerical Simulation of Stochastic Model ... 45
5.1.1 Methodology ... 46
5.1.2 Citation Distributions ... 47
5.1.3 Citation Trajectories ... 48
5.2 Stochastic Component of Citation Dynamics ... 49
5.2.1 Methodology ... 49
5.2.2 Statistical Distribution of Additional Citations ... 50
5.2.3 Uncited Papers ... 55

6 Comparison of Citation Dynamics for Different Disciplines ... 57
6.1 Extension of the Model to Other Disciplines ... 57
6.1.1 Mean Number of Citations ... 58
6.1.2 Citation Distributions and the Aging Function for Citations ... 59
6.1.3 Indirect Citations ... 61
6.2 Citation Lifetime ... 63
6.3 Universality of Citation Distributions ... 64
6.4 Uncited Papers ... 66
6.5 Summary ... 68

7 Prediction of Citation Dynamics of Individual Papers ... 69
7.1 Introduction ... 69
7.1.1 Our Goal ... 70
7.2 Probabilistic Character of the Citation Process and Its Implications with Respect to Predictability of Future Citations ... 71
7.2.1 Divergence of Citation Dynamics of Similar Papers: Measurements ... 71
7.2.2 Divergence of Citation Dynamics of the Papers with the Same Fitness: Numerical Simulation ... 72
7.2.3 Fitness Estimation ... 74
7.3 Fitness Estimation Basing on Paper's Content ... 75
7.4 Timeliness of Results ... 76
7.5 Summary ... 78

8 Power-Law Citation Distributions are Not Scale-Free 81
8.1 Introduction 81
8.1.1 Power-Law Distributions 81
8.1.2 Experimental Assessment of the Fat Tailed Distributions 83
8.2 Empirical Characterization of Citation Distributions 84
8.3 Recursive Search Model Explains the Shape of Citation Distributions 86
8.3.1 Citation Distributions 86
8.3.2 Citation Lifetime 88
8.4 Discussion 90

9 Comparison to Existing Models 93
9.1 Preferential Attachment Mechanism 93
9.1.1 Theoretical Model 93
9.1.2 Model Validation by Measurements 94
9.2 Fitness-Based Preferential Attachment 96
9.2.1 Multiplicative Fitness 96
9.2.2 Additive Fitness 97
9.3 Fitness-Only Models 98
9.4 Explanatory Models 98
9.5 Equivalence Between Preferential Attachment and Fitness Models 100
9.6 The Genuine Preferential Attachment Exists and is Related to Nonlinear Citation Dynamics 104

A Details of Numerical Simulation 107

References 109

Index 119

List of Symbols

$A(t)$	Aging function for references
$\tilde{A}(t)$	Aging function for citations
a	Prefactor in the expression for the probability of indirect citations
$\tilde{a}$	Parameter of the $s(K)$ dependence
a_{mn}	Adjacency matrix
$B(t)$	Generalized aging function
b	Coefficient in the expression for the probability of indirect citations
$\tilde{b}$	Parameter of the $s(K)$ dependence
C	Total number of citations in a citation network
$c_{t,t-1}$	Pearson autocorrelation coefficient for additional citations
f_l	Fraction of the second-generation citing papers connected to their progenitor by l two-hop paths
$f_{uncited}(t)$	Fraction of uncited papers
G	The slope of the $\Gamma(K)$ dependence
$K(t)$	Cumulative number of citations of a paper after t years
K^{∞}	Longtime limit of citations of an individual paper
K_r	Onset of the runaway behavior
K_0	Initial attractivity in the preferential attachment mechanism
$k(t)$	Annual citation rate of an individual paper at year t
$k^{nn}(t)$	Mean annual number of the second-generation citations per one first-generation citing paper (the nearest-neighbor connectivity)
$M(t)$	Mean number of cumulative citations after t years
$m(t)$	Mean annual citation rate at year t
$m_{dir}(t)$	Mean annual direct citation rate
$m_{indir}(t)$	Mean annual indirect citation rate
$N(t)$	Annual number of publications in one discipline
N^{nn}	Mean number of the second-generation citing papers per one first-generation citing paper
$n^{nn}(t)$	Mean annual number of the second-generation citing papers per one first-generation citing paper
$P(t)$	Probability of indirect citation through a paper t years old

P_0	Probability amplitude of indirect citation
Q	Community size
q	Reduced probability of indirect citation
R_{0i}	Reference list length of paper i
$R_0(t_0)$	Average reference list length of the papers published in year t_0
$R(t)$	Age distribution of references
$R_{dir}(t)$	Age distribution of direct references
$R_{indir}(t)$	Age distribution of indirect references
$r(t)$	Reduced age distribution of references
$r_{dir}(t)$	Reduced age distribution of direct references
$r_{indir}(t)$	Reduced age distribution of indirect references
s	The number of two-hop paths connecting a second-generation citing paper to its progenitor
s_0	Intercept of the $P_0(s)$ dependence
$T(t)$	Memory function for indirect citations
t_0	Publication year of a paper
α	Exponent characterizing the growth of the number of publications
β	Exponent characterizing the growth of the reference list length
$\gamma, \tilde{\gamma}$	Exponent of the memory function
Γ	Obsolescence rate
Γ_0	Obsolescence rate of the low-cited papers
δ	Exponent of the aging function for direct citations
ϵ, ζ	Exponent of the preferential attachment
η	Paper's fitness
Θ_{ikj}	Probability of indirect citation: paper i cites paper j through the intermediate paper k
λ	Probabilistic citation rate
λ_{dir}	Probabilistic direct citation rate
λ_{indir}	Probabilistic indirect citation rate
μ	Mean of the log-normal distribution
ν	Exponent of the power-law distribution
Π_{ij}	Probability of citation of paper j by paper i
Π_{ij}^{dir}	Probability of direct citation of paper j by paper i
Π_{ij}^{indir}	Probability of indirect citation of paper j by paper i
π_l	Probability of indirect citation in l-multiplet
$\rho(\eta)$	Fitness distribution
σ	Standard deviation of the log-normal distribution
τ_0	Citation lifetime