A characterization of the scientific impact of Brazilian institutions

Aristoklis D. Anastasiadis
Electrical and Computer Engineering Department, University of Patras, Rio, Achaia 26500, Greece and
Centro Brasileiro de Pesquisas Fisicas, Rua Xavier Sigaud 150 22290-180 Rio de Janeiro Brazil

Marcelo P. de Albuquerque and Marcio P. de Albuquerque∗
Centro Brasileiro de Pesquisas Fisicas, Rua Xavier Sigaud 150 22290-180 Rio de Janeiro Brazil

(Received on 23 January, 2009)

In this paper we studied the research activity of Brazilian Institutions for all sciences and also their performance in the area of physics between 1945 and December 2008. All the data come from the Web of Science database for this period. The analysis of the experimental data shows that, within a nonextensive thermostatistical formalism, the Tsallis q-exponential distribution $N(c) \propto 1/[1 + (q - 1) c/T]^{1/q}$ with $q \approx 4/3$ for all the available citations $c$, $T$ being an “effective temperature”. The present analysis ultimately suggests that via the “effective temperature” $T$, we can provide a new performance metric for the impact level of the research activity in Brazil, taking into account the number of the publications and their citations. This new performance metric takes the “quantity” (number of publications) and the “quality” (number of citations) for different Brazilian Institutions into account. In addition we analyzed the research performance in Brazil to show how the scientific research activity changes with time, for instance between 1945 to 1985, then during the period 1986-1990, 1991-1995, and so on until the present. Finally, this work intends to show a new methodology that can be used to analyze and compare institutions within a given country.

Keywords: Citation analysis, Nonextensive statistical mechanics, Web of science, Tsallis $q$-exponential distribution

1. INTRODUCTION

The analysis of the citations of scientific papers is an important issue that can enable a better understanding of the research activity of the authors and the institutions [1, 2]. The evaluation of the productivity of individual scientists has traditionally relied on the number of papers they have published. It is becoming popular to use citation analysis as a bibliometric tool for the evaluation of the scientific and academic performance for individual researchers [1], journals [3, 4], universities [2, 5] even entire countries [6]. Nowadays, with the easy access to the Internet and to large databases, including the Web of Science [7], the comparison of the impact of scientific contributions is a much easier and more rapid process.

Research productivity is usually measured by taking two different variables into account, namely the number of total publications and their citations. The first measure reflects research quantity and the other reflects research impact. The degree to which published works are cited by other authors is generally considered as a reflection of the quality of those works [8]. Prior citation works have analyzed a wide variety of factors such as the distribution of citation rates [9, 11, 12]. Prior citation works have analyzed a wide variety of factors such as the distribution of citation rates [9–12].

A stretched exponential fitting was applied to model citation distributions based on multiplicative processes [13]. Lehmann [11] attempted to fit both a power law and stretched exponential to the citation distribution of 281 717 papers in the SPIRES database and showed it is impossible to discriminate between the two models. Redner analyzed the ISI and Physical Review databases [9]. In Redner’s work the applied fitting distribution had only partial success while the same numerical data for large citation count $c$ showed that it can be fitted quite satisfactorily with a single curve by using nonextensive thermostatistical formalism [12]. Another fitting distribution that was applied was the lognormal distribution, which was used in order to measure the research activity [14]. A recent characterization of scientific impact has been conducted using Tsallis $q$-exponential distribution [6]. In that work the scientific research activity was considered in terms of the number of publications and number of citations using data from Thomson ISI Web of Science database [7] for many different countries in Latin America, Europe and South Africa. That study showed that the data for all the tested countries can be satisfactorily fitted with a single curve, which naturally emerges within the Tsallis theory [15].

In this work further study has been done for the Brazilian scientific community. Traditionally, researchers and institutions have been evaluated by peer review, which is the main mechanism for merit assessment for funding, appointment, and promotion decisions. Currently, there is also a global trend towards developing and broadening the use of bibliometric indicators to help making these decisions [16]. The experimental data shows that each year there is an increase in Brazilian contribution to international science (this is obtained by the total number of publications). The number of Brazilian authors and the number of Brazilian publications in the international scientific literature have grown substantially during the last decades [7]. Many studies have been done to analyze the Brazilian scientific activity further and also provide a performance metric for the Brazilian Institutions [16–18].

This manuscript provides an analysis of the scientific citations of the Brazilians institutions and their impact within a nonextensive thermostatistical formalism, the Tsallis $q$-exponential distribution $N(c) \propto 1/[1 + (q - 1) c/T]^{1/q}$.
with \( q \approx 4/3 \) for all the available citations \( c \), \( T \) being an “effective temperature”. Emphasis is also given on the performance of the Brazilian Institutions of Physics and Physics departments of Brazil’s universities. The outputs of this study could be useful for the national Brazilian agencies, such as CAPES (Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior) and other research support agencies, which are responsible for creating and assessing programs and projects. Finally, the “effective temperature” will be a scientific metric for the Brazilian institutions’ growing performance and will help Brazilian agencies in the evaluation process of the research programs.

2. NONEXTENSIVE STATISTICAL MECHANICS AND TSALLIS \( q \)-EXPONENTIAL DISTRIBUTION

Nowadays, the idea of nonextensivity has been used in many applications. Nonextensive statistical mechanics has been applied successfully in physics (astrophysics, astronomy, cosmology, nonlinear dynamics) [19], biology [20], economics [21], human and computer sciences [22–24] and provide interesting insights into a variety of physical systems, and among others [25].

Nonextensive statistical mechanics is based on Tsallis entropy. Tsallis statistics [15] is currently considered useful in describing the thermostatistical properties of nonextensive systems; it is based on the generalized entropic form [26]:

\[
S_q \equiv k \frac{1 - \sum_{i=1}^{W} p_i^q}{q - 1} \quad (q \in \mathbb{R}),
\]

where \( W \) is the total number of microscopic configurations, whose probabilities are \( \{p_i\} \), and \( k \) is a conventional positive constant. When \( q = 1 \) it reproduces the Boltzmann-Gibbs entropic form \( S_{BG} = -k \sum_{i=1}^{W} p_i \ln p_i \). The nonextensive entropy \( S_q \) achieves its extreme value at the equiprobability \( p_i = 1/W, \forall i \), and this value equals \( S_q = k \frac{W^{1-q} - 1}{1-q} \) \( (S_1 = S_{BG} = k \ln W) \) [26, 27]. The Tsallis entropy is nonadditive in such a way that, for statistical independent systems \( A \) and \( B \), the entropy satisfies the following property:

\[
\frac{S_q(A + B)}{k} = \frac{S_q(A)}{k} + \frac{S_q(B)}{k} + (1-q) \frac{S_q(A) S_q(B)}{k}. \tag{2}
\]

It is subadditive for \( q > 1 \), superadditive for \( q < 1 \), and, for \( q = 1 \), it recovers the BG entropy, which is additive [27]. The Boltzmann factor is generalized into a power-law. The mathematical basis for Tsallis statistics includes \( q \)-generalized expressions for the logarithm and the exponential functions which are the \( q \)-logarithm and the \( q \)-exponential functions. The \( q \)-exponential function, which reduces to \( \exp(x) \) in the limit \( q \to 1 \), is defined as follows

\[
e_q^x \equiv \left[ 1 + (1-q)x \right]^{1/(1-q)} = \frac{1}{\left[ 1 - (1-q)x \right]^{1/(1-q)}} \quad (e_q^1 = e^x). \tag{3}
\]

Its inverse is the \( q \)-logarithm function and it is equal to

\[
\ln_q(x) \equiv \left[ \frac{x}{1-q} \right]^{1/(1-q)} \quad (x > 0). \]

We shall from now on refer to these two functions as the \( q \)-exponential and the \( q \)-logarithm respectively. We remind that extremizing entropy \( S_q \) under appropriate constraints we obtain a probability distribution, which is proportional to \( q \)-exponential function.

In this work, we focus on the analysis of the distribution of citations of scientific publications, more precisely those that have been catalogued by the Institute for Scientific Information (ISI) for the Brazilian Institutions and for the whole of Brazil. The proposed fitting distributions follow from the nonextensive formalism as \( N(x) \approx 1/[1 + (q - 1) c/T]^{1/q} \). In this study we adopt the following expression:

\[
N(c) = N(2) \exp_{q}^{-c^2} \tag{4}
\]

where \( N(2) \) is the number of papers with two citations, and, as already mentioned, \( T \) plays the role of an effective temperature.

3. THOMSON ISI WEB OF KNOWLEDGE- DATA ACQUISITION

Traditionally, the most commonly used source of bibliometric data is Thomson ISI Web of Knowledge, in particular the (Social) Science Citation Index and the Journal Citation Reports (JCR), which provide the yearly Impact Factors (IF) [7]. The subject categories and terminology provided by ISI are widely recognized by many researchers and scientometricians in their studies and are relatively simple to use [6, 14]. The Institute for Scientific Information has made an industry of providing citation data to libraries since the mid-1960s; the products are currently available as part of Thomson/ISI. Although the ISI database has a few shortcomings, overall it gives a wide coverage of most research fields [28]. Therefore in our survey we utilize Thomson/ISI Web of Science database to study the distribution of the citation within a variety of countries.

To obtain all the necessary data we developed a program which automatically downloads the ISI bibliographic information. We take all the document types into account, e.g. articles, proceeding papers, meeting abstracts, etc, for all the available subject areas, for instance neurosciences, mathematics, chemistry etc, to select all the data for the Brazil and then the same procedure for the Brazilian Institutions and the departments and institutes of physics that we are interested in.

The program is written in Delphi 7 and uses the TWebBrowser component. This component provides access to the Web browser functionality and saves all the “html” pages. When the page is completely downloaded, an OnDownloadComplet event is generated and we automatically go to the next “html” page. When all pages are downloaded we process each “html” page to obtain the specific information that we are interested in using the TPerlRegEx component from the open source PCRE library [http://www.pcre.org/]. In this case, we have gathered the number of citations for each publication and the total number of the published papers, for each Institution. We applied filters to take all these data sorted by the times they are cited, using the Citation databases namely Science Citation Index Expanded (SCI—EXPANDED 1945–present), Social Sciences Citation Index (SSCI 1956–present), and Arts and humanities Citation Index (AandHCI 1975–present). All the data were captured in December 2008.
4. PRESENTATION OF RESULTS

Firstly, we are going to present the data for captured until December 2008 and then describe the procedure that we follow to conduct the final citations fitting. All the papers included in the Web of Science and having at least one author with at least one affiliation address in Brazil have been collected. This means that the work includes all the documents with at least one Brazilian address with citations till December 2008. Research done by Brazilians abroad, i.e with only foreign addresses, is disregarded in the considered database. Note that the data and results are presented on a log-log scale. Initially we evaluate the values of $q$ in order to find its optimal value, and then, with this value, we move to the final fitting procedure.

Probability distribution for citations in Brazil (down)

$$T_{optimal} = \frac{1}{q}$$

FIG. 1: Process to obtain best $q$ and $T$ (up): $q$ is evaluated in order to find its optimal value, and then a data fit determines the effective temperature $T$. Probability distribution for citations in Brazil (down)

Next we investigate how the temperature changes during the years. As the temperature is a characterization of the scientific impact its evolution over the years can offer a deeper understanding of how the Brazilian research activity evolved.

Next, Table II presents the Brazilian Institutions in the tested Brazilian society of Physics. Note that the results for “Brazil” do not represent the average of the particular Brazilian institutions that we are considering in the Tables but all Brazilian institutions. This happens because these results are taken by placing “Brazil” in the address field. It should also be clear that when we refer to “Brazil Physics”, it is the average research performance for all the Brazilian institutions in the area of physics and not only the tested Brazilian institutes, i.e in this case we apply the word “Brazil”, and Physics (“Fis”) in the address field to obtain these results. Finally, in the tables II and III we study the institutions with temperature greater or equal to the whole Brazilian temperature, i.e $T \geq 4.0$.

Table I presents the total number of publications, and the percentage of zero, and papers cited once for the tested Institutions. University of Sao Paulo (USP) achieves the highest publication productivity with 66 404 published papers. Then University Estadual Campinas (UNICAMP) and Federal University of Rio de Janeiro (UFRJ) have published 24 209 and 21 656 research papers respectively. The rest of the tested Brazilian Institutions attain a significantly lower rate of published papers, i.e Federal University of Pernambuco (UFPE), Federal University of Rio Grande do Sul (UFRGS), and Federal University of Fluminense (UFF) have published 5 540, 5 318, and 5 318 papers respectively. Finally, the Federal University of Minas Gerais (UFMG) has 1 887 publications.

Table I: Number of total publications, and the percentage of zero, and papers cited once for the tested Institutions

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Total # Papers</th>
<th># Zero citations</th>
<th># One citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>USP</td>
<td>66 404</td>
<td>24 197 (36.4%)</td>
<td>7 041 (10.6%)</td>
</tr>
<tr>
<td>UNICAMP</td>
<td>24 209</td>
<td>8 215 (33.9%)</td>
<td>2 771 (11.5%)</td>
</tr>
<tr>
<td>UFRJ</td>
<td>21 656</td>
<td>7 498 (34.6%)</td>
<td>2 591 (12.0%)</td>
</tr>
<tr>
<td>UFPE</td>
<td>6 052</td>
<td>2 067 (34.3%)</td>
<td>794 (13.2%)</td>
</tr>
<tr>
<td>UFRGS</td>
<td>5 540</td>
<td>2 868 (51.6%)</td>
<td>695 (12.5%)</td>
</tr>
<tr>
<td>UFF</td>
<td>5 318</td>
<td>1 919 (36.1%)</td>
<td>668 (12.6%)</td>
</tr>
<tr>
<td>UFMG</td>
<td>1 887</td>
<td>680 (36.0%)</td>
<td>286 (15.2%)</td>
</tr>
<tr>
<td>Brazil</td>
<td>285 570</td>
<td>108 984 (38.2%)</td>
<td>33 428 (11.7%)</td>
</tr>
</tbody>
</table>

lights how the scientific research activity changes with time. It is remarkable how effective temperature is as a reliable performance metric for the research activity in Brazil. This part of the analysis uses the entire available year publication window for all disciplines for papers published between 1945 to December 2008. Note that for the last periods from 2001 to 2004 and 2005 to 2008 there has not been enough time for the publications to become widely known to the scientific community so the number of their citations is small. Thus the overall temperature is smaller as there is this delay. Also Figure 2 (right) illustrates the performance of Brazil in Physics domain. 39 617 papers (8 688 zero citations, (21.9%)) have been published in Physics until January 2009 giving $T=4.44$, which characterizes the overall research performance of our tested Brazilian society of Physics.

Next, Table II presents the Brazilian Institutions in the ranking based on the temperature that we obtain through the nonextensive distribution fitting. Notice that this ranking differs from the one presented in Table I, where the total amount of the published papers (quantity ranking) is shown. The effective temperature $T$ characterizes the scientific impact of the tested Institutions. As we can perceive from Table II, in
almost all cases the range value of the entropic index $q$ is around $q = 4/3$. The linear regression coefficient $R^2$ is also indicated in each case. As we can see comparing Tables I and II, the rankings are quite different. Let us check UFRJ, for instance. Although it has a relatively smaller number of papers compared to UNICAMP, its effective temperature is higher $T = 4.55$.

Table III presents the best fitting values of $q$ and the effective temperature $T$, which characterizes the research impact of the Brazilian Institutions with emphasis on Physics. In this analysis UFMG was not included as the available publications in the Web of Science database [7] are few (not enough to have a good statistical analysis). In this survey the Centro Brasileiro de Pesquisas Físicas (CBPF) is also included. It becomes evident from Table III that CBPF, USP and UNICAMP achieved the highest temperature in research activity in Physics by applying the new metric ($T$). It is also worth mentioning that the Institutes/Departments of Physics of the Universities have the responsibility of both undergraduate/graduate students and are administratively located at the Ministry of Education, whereas CBPF has the responsibility of only graduate students and is administratively located at the Ministry of Science and Technology. This is possibly one of the reasons that can help this institute to achieve a higher temperature. Moreover it is important to mention at this point the performance of the UFPE and UFRJ. While the UFPE significantly increases its Temperature on the domain of Physics the UFRJ has lower $T = 4.10$ compared to the overall research impact value ($T = 4.55$) in all sciences.

Figures 3 and 4 illustrate the fitting of different Brazilian Institutions using the nonextensive distribution $N(c)$. Figure 3 left side shows publications of all sciences and right side demonstrates the research activity in physics domain. As we can observe the general tendency for physics science has a higher research impact than the overall university activity. Finally, Figure 4 presents the CBPF and UFPE fitting curves by applying the new characterization of citations impact. CBPF achieves the highest performance with $T=5.32$ and $q=1.336$. UFPE physics domain attains $T=4.76$ while the whole UFPE’s university citations impact metric is 4.08.

From all the above experimental results, we obtain a value of $q$ close to $4/3$.

5. CONCLUSIONS

Nowadays the number of citations is among the most widely used measures of academic performance. Extended study of citation distributions helps to better understand the mechanics behind citations and can objectively establish a comparative measure for scientific performance. Citations of scientific papers constitute in fact a connection network consisting of authors (nodes) and directed links (citations) among them. Recently, connection networks have been described, studied, characterized and represented by parameters using typical concepts in the area of Complex Systems.
The entropic index \( q \) in Tsallis entropy is usually interpreted as a quantity characterizing the degree of nonextensivity of a system. An appropriate choice of the entropic index \( q \) to nonextensive physical systems still remains an open field of study. In some cases, the physical meaning of the index \( q \) is unknown; it provides nevertheless new possibilities of comparison between theoretical approaches and experimental data. Other cases are better understood, and then \( q \) has a clear physical meaning, either at a microscopic or at a mesoscopic level, or both.

In this paper we characterize the citations impact of the Brazilian institutions using the Tsallis q-exponential distribution. We also show how the scientific research activity changes with time, between six periods from 1945 to 2008. The present study provides a new performance metric based on Nonextensive Statistical Mechanics for ranking and evaluating institutions’ research production. The proposed Tsallis q-exponential distribution satisfactorily describes Institute of Scientific Information citations for Brazilian institutions and Brazilian physics departments between 1945 and December 2008.

Our study provides evidence that the citation distribution for all tested cases within this period could be the Tsallis q-exponential distribution. Our findings in this work gives an evidence for the effectiveness of \( T \), and the ranking that we proposed based on the Temperature. Figure 5 illustrates the \( q \)-logarithmic number of publications \( \ln_q[N(c)/N(1)] \) versus the \( (c-1) \) number of citations for three different Brazilian universities (UFF, UNICAMP, USP). USP has the highest citation impact, the UNICAMP an intermediate \( T \) and UFF a
temperature lower than the average. It is important to notice that \((-1/T)\) corresponds to the average slope associated with each university. It also gives an explanation for the meaning of \(T\), and the ranking that we proposed based on the new performance metric \(T\).

It is remarkable how the proposed nonextensive distribution satisfactorily fits all cited papers for all the institutions. This part of the analysis uses the entire available-year publication window for all disciplines for papers published between 1945 to December 2008. The present article also focuses on the performance of the Brazilian Institutions and their activities in physics science. In the present study we used a single database for the extraction of the articles, and their number of citations. The ISI/Web of Science was chosen because it is one of the main databases providing information on citations. Although our strategy might have left publications out of the analysis, we believe that the sample of articles was representative of the core international scientific production of the Brazilian Institutions. The new performance metric of citations impact is a balanced combination of “quantity” (number of publications) and “quality” (number of citations). These are the main factors of this performance metric. Keeping in mind that citation rate reflects the use and impact of scientific information, not necessarily expressing quality.

This work intends to show how the new methodology can be used to analyze and compare institutions within a given country. A case study of certain Brazilian institutions and their physics departments is used to investigate the effectiveness of the new characterization of citations’ impact. Future work can address other scientific fields in these important Brazilian universities or universities of other countries and how they evolved observing the same analyzed period of time. It is also important to study cases of universities, countries or other scientific institutions with extremely high number of zero or one citations and observe the impact of their research activity. The extent to which this number of citations affects the proposed performance metric will be a field of further study.

**Acknowledgements** The authors thank Professor C.Tsallis for helpful discussions, D. B. Mussi for the program development, the support from the National Council for Scientific and Technological Development (CNPq) of the Brazilian Ministry of Science and Technology, the State of Rio de Janeiro Research Foundation (FAPERJ) and the Brazilian Coordination Office for the Improvement of Staff with Higher Education (CAPES).

----

2. TIMES higher education World University Rankings 2008: http://www.timeshighereducation.co.uk/

**TABLE II: Best fitting values of \(q\) and effective temperature \(T\). Note that tested Institutions are ranked according to \(T\)**

<table>
<thead>
<tr>
<th>Institutions</th>
<th>Entropic index (q)</th>
<th>Linear regression coefficient (R^2)</th>
<th>Temperature (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USP</td>
<td>1.339</td>
<td>0.99</td>
<td>4.75</td>
</tr>
<tr>
<td>UFRJ</td>
<td>1.300</td>
<td>0.99</td>
<td>4.55</td>
</tr>
<tr>
<td>UNICAMP</td>
<td>1.330</td>
<td>0.99</td>
<td>4.35</td>
</tr>
<tr>
<td>UFPE</td>
<td>1.336</td>
<td>0.99</td>
<td>4.08</td>
</tr>
<tr>
<td>UFF</td>
<td>1.335</td>
<td>0.99</td>
<td>4.00</td>
</tr>
<tr>
<td>Brazil</td>
<td>1.339</td>
<td>0.99</td>
<td>4.00</td>
</tr>
</tbody>
</table>

**TABLE III: Best fitting values of \(q\) and effective temperature \(T\). Note that tested Institutions are ranked according to \(T\)**

<table>
<thead>
<tr>
<th>Physics</th>
<th>Total # Papers (\sum_{0}^{\infty}N(c))</th>
<th># Zero citations (N(0)) (%)</th>
<th>Entropic index (q)</th>
<th>Linear regression coefficient (R^2)</th>
<th>Temperature (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBPF</td>
<td>3680</td>
<td>658 (17.9%)</td>
<td>1.336</td>
<td>0.99</td>
<td>5.32</td>
</tr>
<tr>
<td>USP</td>
<td>8781</td>
<td>1776 (20.2%)</td>
<td>1.320</td>
<td>0.99</td>
<td>5.13</td>
</tr>
<tr>
<td>UNICAMP</td>
<td>3992</td>
<td>809 (20.3%)</td>
<td>1.330</td>
<td>0.99</td>
<td>5.0</td>
</tr>
<tr>
<td>UFPE</td>
<td>1685</td>
<td>311 (18.5%)</td>
<td>1.336</td>
<td>0.99</td>
<td>4.76</td>
</tr>
<tr>
<td>UFRJ</td>
<td>5089</td>
<td>1646 (32.3%)</td>
<td>1.336</td>
<td>0.99</td>
<td>4.10</td>
</tr>
<tr>
<td>UFF</td>
<td>1512</td>
<td>309 (20.4%)</td>
<td>1.332</td>
<td>0.99</td>
<td>4.08</td>
</tr>
<tr>
<td>Brazil Physics</td>
<td>39617</td>
<td>8688 (21.9%)</td>
<td>1.332</td>
<td>0.99</td>
<td>4.44</td>
</tr>
</tbody>
</table>
FIG. 4: Probability distribution for citations of CBPF (left) and UFPE (right) up to January 2009

FIG. 5: \( \ln\left[\frac{N(c)}{N(1)}\right] \) versus \((c - 1)\) for 3 Brazilian Universities (UFF, UNICAMP, USP). The figure is a zoom with an upper limit of 100 citations.


