SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
STATISTICAL PHYSICS, THERMODYNAMICS AND NONLINEAR DYNAMICS
APPLIED PHYSICS
STATISTICS
THEORY

Received
18 October 2011

Accepted
14 November 2011

Published
5 December 2011

Correspondence and requests for materials should be addressed to A.M.P. (petersen.xander@gmail.com)

Statistical regularities in the rank-citation profile of scientists

Alexander M. Petersen[1,2], H. Eugene Stanley[2] & Sauro Succi[3,4]

[1]IMT Lucca Institute for Advanced Studies, 55100 Lucca, Italy, [2]Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA, [3]Istituto Applicazioni Calcolo C.N.R., Rome, IT, [4]Freiburg Institute for Advanced Studies, Albertstrasse, 19, D-79104, Freiburg, Germany.

Recent science of science research shows that scientific impact measures for journals and individual articles have quantifiable regularities across both time and discipline. However, little is known about the scientific impact distribution at the scale of an individual scientist. We analyze the aggregate production and impact using the rank-citation profile $c_i(r)$ of 200 distinguished professors and 100 assistant professors. For the entire range of paper rank r, we fit each $c_i(r)$ to a common distribution function. Since two scientists with equivalent Hirsch h-index can have significantly different $c_i(r)$ profiles, our results demonstrate the utility of the β_i scaling parameter in conjunction with h_i for quantifying individual publication impact. We show that the total number of citations C_i tallied from a scientist's N_i papers scales as $C_i \sim h_i^{1+\beta_i}$. Such statistical regularities in the input-output patterns of scientists can be used as benchmarks for theoretical models of career progress.

A scientist's career path is subject to a myriad of decisions and unforeseen events, such as Nobel Prize worthy discoveries[1], that can significantly alter an individual's career trajectory. As a result, the career path can be difficult to analyze since there are potentially many factors (individual, mentor-apprentice, institutional, coauthorship, field)[2–9] to account for in the statistical analysis of scientific panel data.

The rank-citation profile, $c_i(r)$, represents the number of citations of individual i to his/her paper r, ranked in decreasing order $c_i(1) \geq c_i(2) \geq \ldots c_i(N)$, and provides a quantitative synopsis of a given scientist's publication career. Here, we analyze the rank-ordered citation distribution $c_i(r)$ for 300 scientists in order to better understand patterns of success and to characterize scientific production at the individual scale using a common framework. The review of scientific achievement for post-doctoral selection, tenure review, award and academy selection, at all stages of the career is becoming largely based on quantitative publication impact measures. Hence, understanding quantitative patterns in production are important for developing a transparent and unbiased review system. Interestingly, we observe statistical regularities in $c_i(r)$ that are remarkably robust despite the idiosyncratic details of scientific achievement and career evolution. Furthermore, empirical regularities in scientific achievement suggest that there are fundamental social forces governing career progress[10–13].

We group the 300 scientists that we analyze into three sets of 100, referred to as datasets A, B and C, so that we can analyze and compare the complete publication careers of each individual, as well as across the three groups:

- [A] 100 highly-profile scientists with average h-index $\langle h \rangle = 61 \pm 21$. These scientists were selected using the citation shares metric[9] to quantify cumulative career impact in the journal *Physical Review Letters* (PRL).
- [B] 100 additional "control" scientists with average h-index $\langle h \rangle = 44 \pm 15$.
- [C] 100 current Assistant professors with average h-index $\langle h \rangle = 14 \pm 7$. We selected two scientists from each of the top-50 US physics departments (departments ranked according to the magazine *U.S. News*).

In the methods section we describe in detail the selection procedure for datasets A, B, and C and in tables S1-S6 we provide summary statistics for each career.

There are many conceivable ways to quantify the impact of a scientist's N_i publications. The h-index[14] is a widely acknowledged single-number measure that serves as a proxy for production and impact simultaneously. The h-index h_i of scientist i is defined by a single point on the rank-citation profile $c_i(r)$ satisfying the condition

$$c_i(h_i) = h_i. \tag{1}$$

To address the shortcomings of the h-index, numerous remedies have been proposed in the bibliometric sciences[15]. For example, Egghe proposed the g-index, where the most cited g papers cumulate g^2 citations overall[16], and Zhang proposed the e-index which complements the h and g indices quantitatively[17].

To justify the importance of analyzing the entire profile $c_i(r)$, consider a scientist $i = 1$ with rank-citation profile $c_1(r) \equiv [100, 50, 33, 25, 20, 16, 14, 12, 11, 10, 9\ldots]$ and a scientist $i = 2$ with $c_2(r) \equiv [10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 9\ldots]$. Both scientists have the same h-index value $h = 10$, although $c_1(r)$ tallies 2.9 times as many citations as $c_2(r)$ from his/her most-cited 10 papers. Hence, an additional parameter β_i is necessary in order to distinguish these two example careers. Specifically, the β_i parameter quantifies the scaling slope in $c_i(r)$ for the high-rank papers corresponding to small r values. In this simple illustration, $\beta_1 \approx 1$ while $\beta_2 \approx 0$.

In Fig. 1 we plot $c_i(r)$ for 5 extremely high-impact scientists. The individuals EW, ACG, MLC, and PWA are physicists with the largest h_i values in our data set; BV is a prolific molecular biologists who we include in this graphical illustration in order to demonstrate the generality of the statistical regularity we find, which likely exists across discipline. However, citation and h-index metrics should not be compared across discipline since baseline publication and citation rates can vary significantly between research fields Refs[8, 9]. To demonstrate how the singe point $c_i(h_i)$ is an arbitrary point along the $c_i(r)$ curve, we also plot the lines $H_p(r) \equiv p\,r$ for 5 values of $p = \{1, 2, 5, 20, 80\}$. The value $p \equiv 1$ recovers the h-index $h_1 = h$

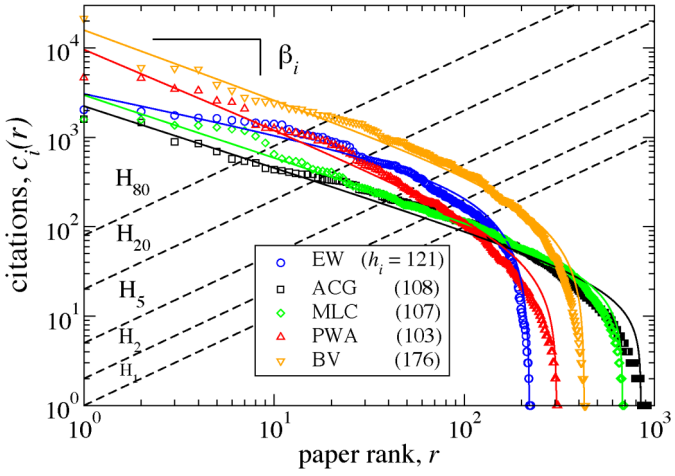

Figure 1 | The citation distribution of individual scientists is heavy-tailed. We show 5 empirical rank-citation $c_i(r)$ profiles , each belonging to an extremely high-impact scientists (E. Witten, A. C. Gossard, M. L. Cohen, P. W. Anderson and B. Vogelstein) whose initials and h-index as of Jan. 2010 are listed in the figure legend. The hierarchical scaling pattern in $c_i(r)$ for small r values indicate that the pillar contributions of top scientists are "off-the-charts" since they have no characteristic scale. Put in the framework of the citation distribution, consider the probability distribution $P_i(c)$ of the citation impact c calculated for an individual's N_i papers. If $P_i(c)$ is heavy tailed with asymptotic power-law scaling $P_i(c) \sim c^{-\zeta_i}$, then $\zeta_i = 1 + 1/\beta_i$. Ref. [9] calculates $\zeta \approx 3$, corresponding to $\beta = 1/2$, using the entire set of citations for papers from six individual journals. Hence, the citation impact of stellar scientists can be significantly more skewed than the aggregate population. This statistical regularity demonstrates the utility of the β_i scaling exponent in characterizing the highly cited papers of a given scientist i. Interestingly, each scientist has coauthored a significant number of papers that are significantly lower impact than their $c_i(1)$ pillar paper. The $c_i(r)$ distributions show significant variability in both the high-rank (β) and low-rank (γ) regimes. Moreover, for $c_i(r)$ with similar h values, the h-index (a single point on each curve) is insufficient to adequately distinguish career profiles. The solid curves are the best-fit DGBD functions (see Eq. 3) for each corresponding $c_i(r)$ over the entire rank range in each case. The intersection of $c_i(r)$ with the line $H_p(r)$ corresponds to the generalized h-index h_p, which together uniquely quantify the $c_i(r)$ profile. Five $H_p(r)$ lines are provided for reference, with $p = \{1, 2, 5, 20, 80\}$.

proposed by Hirsch. The intersection of any given line $H_p(r)$ with $c_i(r)$ corresponds to the "generalized h-index" h_p,

$$c(h_p) = p h_p, \tag{2}$$

proposed in[18] and further analyzed in[19], with the relation $h_p \leq h_q$ for $p > q$. Since the value $p \equiv 1$ is chosen somewhat arbitrarily, we take an alternative approach which is to quantify the entire $c_i(r)$ profile at once (which is also equivalent to knowing the entire h_p spectrum). Surprisingly, because we find regularity in the functional form $c_i(r)$ for all 300 scientists analyzed, we can relate the relative impact of a scientist's publication career using the small set of parameters that specify the $c_i(r)$ profile for the entire set of papers ranging from rank $r = 1\ldots N_i$. Using a much smaller parameter space than the h_p spectrum, we can begin to analyze the statistical regularities in the career accomplishments of scientists.

The aim of this analysis is not to add another level of scrutiny to the review of scientific careers, but rather, to highlight the regularities across careers and to seed further exploration into the mechanisms that underlie career success. The aim of this brand of quantitative social science is to utilize the vast amount of information available to develop an academic framework that is sustainable, efficient and fruitful. Young scientific careers are like "startup" companies that need appropriate venture funding to support the career trajectory through lows as well as highs[13].

Results

A Quantitative Model for $c_i(r)$. For each scientist i, we find that $c_i(r)$ can be approximated by a scaling regime for small r values, followed by a truncated scaling regime for large r values. Recently a novel distribution, the discrete generalized beta distribution (DGBD)

$$c_i(r) \equiv A_i r^{-\beta_i} (N_i + 1 - r)^{\gamma_i} \tag{3}$$

has been proposed as a model for rank profiles in the social and natural sciences that exhibit such truncated scaling behavior[20,21]. The parameters A_i, β_i, γ_i and N_i are each defined for a given $c_i(r)$ corresponding to an individual scientists i, however we suppress the index i in some equations to keep the notation concise. We estimate the two scaling parameters β_i and γ_i using *Mathematica* software to perform a multiple linear regression of $\ln c_i(r) = \ln A_i - \beta_i \ln r + \gamma_i \ln(N_i + 1 - r)$ in the base functions $\ln r$ and $\ln(N_i + 1 - r)$. In our fitting procedure we replace N with r_1, the largest value of r for which $c(r) \geq 1$ (we find that $r_1/N_i \approx 0.84 \pm 0.01$ for careers in datasets A and B). Figs. 1 and 2 demonstrate the utility of the DGBD to represent $c_i(r)$, for both large and small r. The regression correlation coefficient $R_i > 0.97$ for all $\ln c_i(r)$ profiles analyzed.

The DGBD proposed in[20] is an improvement over the Zipf law (also called the generalized power-law or Lotka-law[22]) model and the stretched exponential model[14] since it reproduces the varying curvature in $c_i(r)$ for both small and large r. Typically, an exponential cutoff is imposed in the power-law model, and justified as a finite-size effect. The DGBD does not require this assumption, but rather, introduces a second scaling exponent γ_i which controls the curvature in $c_i(r)$ for large r values. The DGBD has been successfully used to model numerous rank-ordering profiles analyzed in[20,21] which arise in the natural and socio-economic sciences. The relative values of the β_i and γ_i exponents are thought to capture two distinct mechanisms that contribute to the evolution of $c_i(r)$[20,21]. Due to the data limitations in this study, we are not able to study the dynamics in $c_i(r)$ through time. Each $c_i(r)$ is a "snapshot" in time, and so we can only conjecture on the evolution of $c_i(r)$ throughout the career. Nevertheless, we believe that there is likely a positive feedback effect between the "heavy-weight" papers and "newborn" papers, whereby the reputation of the "heavy-weight" papers can increase the exposure and impact the perceived significance of "newborn" papers during their infant phase. Moreover, the 2-regime power-law

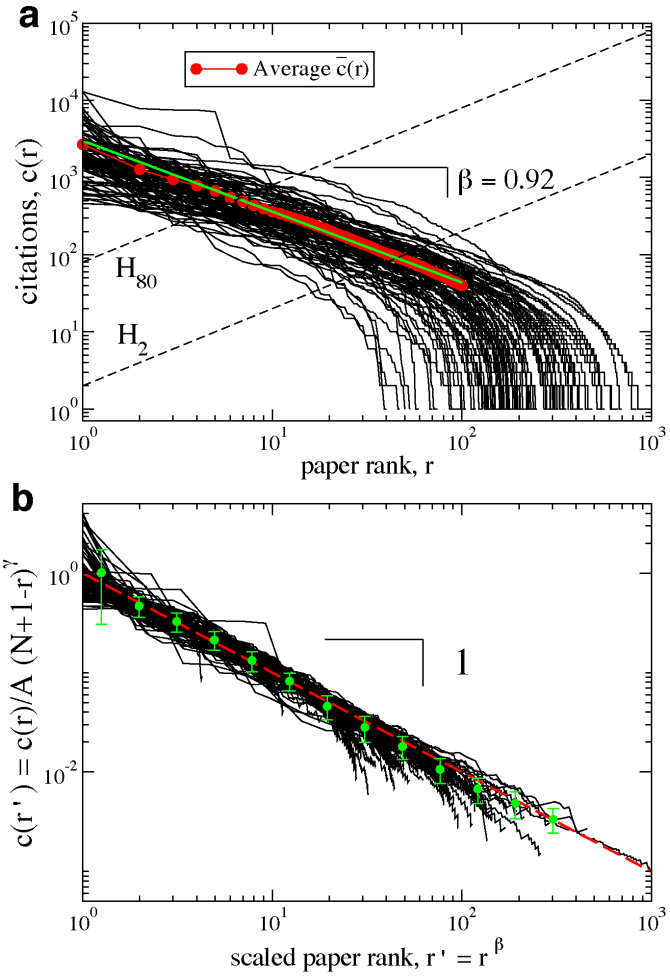

Figure 2 | Data collapse of each $c_i(r)$ along a universal curve. A comparison of 100 rank-citation profiles $c_i(r)$ demonstrates the statistical regularity in career publication output. Each scientist produces a cascade of papers of varying impact between the $c_i(1)$ pillar paper down to the least-known paper $c_i(N_i)$. (a) Zipf rank-citation profiles $c_i(r)$ for 100 scientists listed in dataset [A]. For reference, we plot the average $\bar{c}(r)$ of these 100 curves and find $\bar{c}(r) \sim r^{-\beta}$ with $\beta = 0.92 \pm 0.01$. The solid green line is a least-squares fit to $\bar{c}(r)$ over the range $1 \leq r \leq 100$. We also plot the $H_2(r)$ and $H_{80}(r)$ lines for reference. (b) We re-scale the curves in panel (a), plotting $c_i(r') \equiv c_i(r)/A(r_1 + 1 - r)^{\gamma}$, where we use the best-fit γ_i and A_i parameter values for each individual $c_i(r)$ profile. Using the rescaled rank value $r' \equiv r^{\beta_i}$, we show excellent data collapse onto the expected curve $c(r') = 1/r'$. (see Figs. S1 and S2 for analogous plots for dataset [B] and [C] scientists). Green data points correspond to the average $c(r')$ value with 1σ error bars calculated using all 100 $c_i(r')$ curves separated into logarithmically spaced bins.

behavior of $c_i(r)$ suggests that the reinforcement dynamics can be quantified by the scale-free parameters β and γ.

The β_i value determines the relative change in the $c_i(r)$ values for the high-rank papers, and thus it can be used to further distinguish the careers of two scientists with the same h-index. In particular, smaller β values characterize flat profiles with relatively low contrast between the high and low-rank regions of any given profile, while larger β values indicate a sharper separation between the two regions.

In Fig. 2(a) we plot $c_i(r)$ for each scientist from dataset [A] as well as the average of the 100 individual curves $\bar{c}(r) \equiv \frac{1}{100}\sum_{i=1}^{100} c_i(r)$ (see Figs. S1 and S2 for analogous plots for datasets [B] and [C]). We find robust power-law scaling

$$\bar{c}(r) \sim r^{-\beta} \; [\beta \approx 0.92 \pm 0.01] \qquad (4)$$

for $10^0 \leq r \leq 10^2$. The scaling value calculated for other rank-size (Zipf) distributions in the social and economic sciences is typically around unity, $\beta \approx 1$, for example in studies of word frequency[23] and city size[20,21,24]. Here we calculate β_i for each individual author and observe a distribution which is centered around characteristic values $\langle\beta\rangle = 0.83 \pm 0.23$ [A], $\langle\beta\rangle = 0.70 \pm 0.16$ [B], $\langle\beta\rangle = 0.79 \pm 0.38$ [C].

We calculate each β_i value using a multilinear least-squares regression of $\ln c_i(r)$ for $1 \leq r \leq r_1$ using the DGBD model defined in Eq. [3]. To properly weight the data points for better regression fit over the entire range, we use only 20 values of $c_i(r)$ data points that are equally spaced on the logarithmic scale in the range $r \in [1, r_1]$. We elaborate the details of this fitting technique in the methods section. We plot five empirical $c_i(r)$ along with their corresponding best-fit DGBD functions in Fig. 1 to demonstrate the goodness of fit for the entire range of r.

In order to demonstrate the common functional form of the DGBD model, we collapse each $c_i(r)$ along a universal scaling function $c(r') = 1/r'$, by using the rescaled rank values $r' \equiv r^{\beta_i}$ defined for each curve. In Figs. 2(b), S1(b) and S2(b), we plot the quantity $c_i(r') \equiv c_i(r)/A(r_1 + 1 - r)^{\gamma}$, using the best-fit γ_i and A_i parameter values for each individual $c_i(r)$ profile. While the curves in Fig. 2(a) are jumbled and distributed over a large range of $c(r)$ values, the rescaled $c_i(r)$ curves in Fig. 2(b) all lie approximately along the predicted curve $c(r') = 1/r'$.

Using $c_i(r)$ to quantify career production and impact. A main advantage of the h-index is the simplicity in which it is calculated, e.g. *ISI Web of Knowledge*[25] readily provides this quantity online for distinct authors. Another strength of the h-index is its stable growth with respect to changes in $c_i(r)$ due to time and information-dependent factors[26]. Indeed, the h-index is a "fixed-point" of the citation profile. This time stability is evident in the observed growth rates of h for scientists. Average growth rates, calculated here as h/L, where L is the duration in years between a given author's first and most recent paper, typically lie in the range of one to three units per year (this annual growth rate corresponds to the quantity m introduced by Hirsch[14]). Annual growth rates $h/L \approx 3$ correspond to exceptional scientists (for the histogram of $P(h/L)$ see Fig. S3 and for h/L values see the SI text (Tables S1–S6)). As a result, h/L is a good predictor for future achievement along with h[27].

It is truly remarkable how a single number, h_i, correlates with other measures of impact. Understandably, being just a single number, the h-index cannot fully account for other factors, such as variations in citation standards and coauthorship patterns across discipline[28–30], nor can h_i incorporate the full information contained in the entire $c_i(r)$ profile. As a result, it is widely appreciated that the h-index can underrate the value of the best-cited papers, since once a paper transitions into the region $r \leq h_i$, its citation record is discounted, until other less-cited papers with $r > h_i$ eventually overcome the rank "barrier" $r = h_i$. Moreover, as noted in[14], the papers for which $r > h_i$ do not contribute any additional credit.

Instead of choosing an arbitrary h_p as an productivity-impact indicator, we use the analytic properties of the DGBD to calculate a crossover value r_i^*. In the methods section, we derive an exact expression for r_i^* which highlights the distinguished papers of a given author. To calculate r_i^*, we use the logarithmic derivative $\chi(r) \equiv d \ln c(r)/dr$ to quantify the relative change in $c_i(r)$ with increasing r. We defined papers as "distinguished" if they satisfy the inequality $c_i(r)/c_i(r+1) > \exp(\overline{\chi})$, where $\overline{\chi}$ is the average value of $\chi(r)$ over the entire range of r values. This inequality selects the peak papers which are significantly more cited than their neighbors. The peak region $r \in [1, r_i^*]$ corresponds to a "knee" in $c_i(r)$ when plotted on log-linear axes. The dependence of $\overline{\chi}$ and r_i^* on the three DGBD parameters β_i, γ_i and N_i are provided in the methods section.

The advantage of r_i^* is that this characteristic rank value is a comprehensive representation of the stellar papers in the high-rank

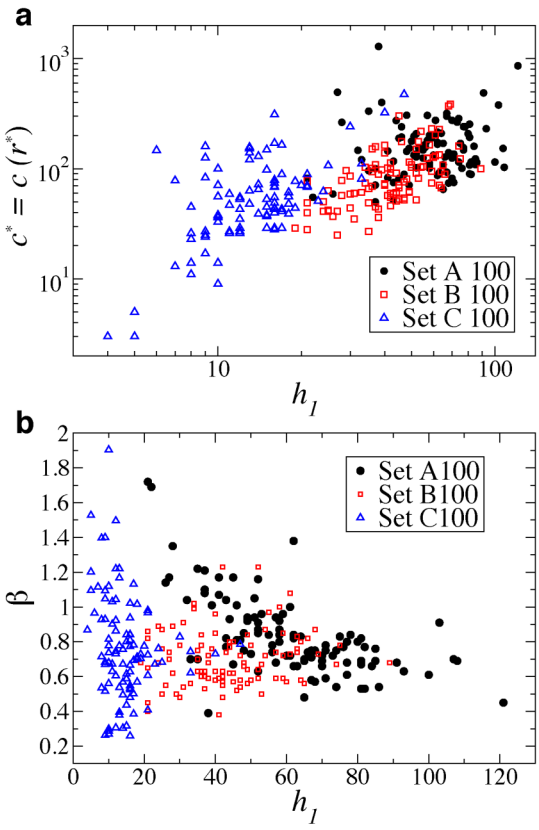

Figure 3 | Limitations to the use of the *h*-index alone. The *h*-index can be insufficient in comprehensively representing $c_i(r)$. (a) The *h*-index does not contain any information about $c_i(r)$ for $r < h_i$, and can shield a scientist's most successful accomplishments which are the basis for much of a scientist's reputation. This is evident in the cases where $c(r_i^*) \gg h_i$, in which case the *h*-index cannot account for the stellar impact of the papers. (b) For a given h_i value, prolific careers are characterized by a large β_i value, as it is harder to maintain large β_i values for large h_i. As a result, the β_i vs h_i parameter space can be used to identify anomalous careers and to better compare two scientists with similar h_i indices. We find that a third career metric C_i, the total number of citations to the papers of author *i*, can be calculated with high accuracy by the scaling relation $C_i \sim h_i^{1+\beta_i}$, which we illustrate in Fig. 4(b).

scaling regime since it depends on the DGBD parameter values β_i, γ_i and N_i, and thus probes the entire citation profile. Fig. 3 shows a scatter plot of the "*c*-star" $c_i^* \equiv c_i(r_i^*)$ and h_i values calculated for each scientist and demonstrates that there is a non-trivial relation between these two single-value indices. It also shows that for scientists within a small range of *c** there is a large variation in the corresponding *h* values, in some cases straddling across all three sets of scientists. Also, there are several c_i^* values which significantly deviate from the trend in Fig. 3, which is plotted on log-log axes. These results reflect the fact that the *h*-index cannot completely incorporate the entire $c_i(r)$ profile. We plot the histogram of c_i^* and r_i^* values in Figs. S4 and S5, respectively.

To further contrast the values of c_i^* and the *h*-index, we propose the "peak indicator" ratio $\Lambda_i \equiv c_i^*/h_i$, which corrects specifically for the *h*-index penalty on the stellar papers in the peak region of $c_i(r)$. Thus, all papers in the peak region of $c_i(r)$ satisfy the condition $c_i(r) \geq h_i\Lambda_i$. In an extreme example, R. P. Feynman has a peak value $\Lambda \approx 36$, indicating that his best papers are monumental pillars with respect to his other papers which contribute to his *h*-index. Fig. S6 shows the histogram of Λ_i values, with typical values for dataset [A] scientists $\langle\Lambda\rangle \approx 3.4 \pm 3.9$, and for dataset [B] scientists $\langle\Lambda\rangle \approx 2.2 \pm 1.1$. This indicator can only be used to compare scientists with similar *h* values, since a small h_i can result in a large Λ_i.

An alternative "single number" indicator is C_i, an author's total number of citations

$$C_i = \sum_{r=1}^{N} c_i(r), \tag{5}$$

which incorporates the entire $c_i(r)$ profile. However, it has been shown that $\sqrt{C_i}$ correlates well with h_i[31], a result which we will demonstrate in Eq. [6] to follow directly from a $c_i(r)$ with $\beta_i \approx 1$.

We test the aggregate properties of $c_i(r)$ by calculating the aggregate number of citations $C_{\beta,h}$ for a given profile,

$$C_{\beta,h} \equiv \sum_{r=1}^{N} Ar^{-\beta} \approx h^{1+\beta} \sum_{r=1}^{N'} r^{-\beta} = h^{1+\beta} H_{N',\beta} \sim h^{1+\beta} \tag{6}$$

where $H_{N',\beta}$ is the *generalized harmonic number* and is of order $O(1)$ for $\beta \approx 1$. We neglect the γ_i scaling regime since the low-rank papers do not significantly contribute to an author's C_i tally. We approximate the coefficient *A* in Eq. [6] using the definition $c(h) \equiv h$, which implies that $A/h^\beta \approx h$. We use the value $N' \equiv 3\ h$, so that $C_{\beta,h}$ can be approximated by only the two parameters h_i and β_i for any given author. We justify this choice of N' by examining the rescaled $c_i(r/h)$, which we consider to be negligible beyond rank $r = 3\ h_i$ for most

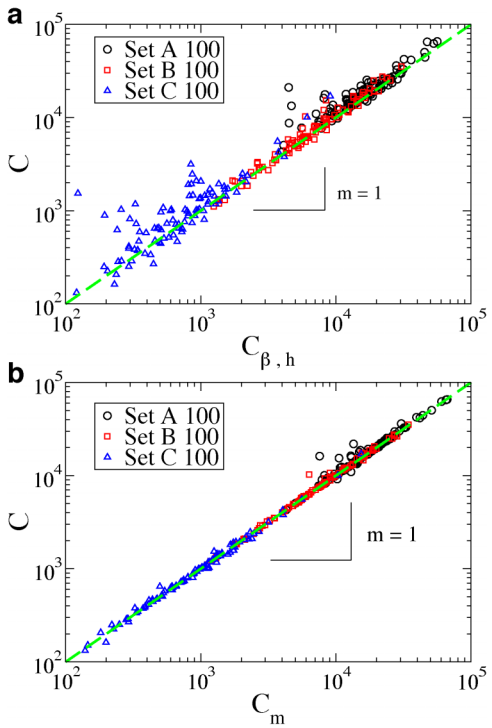

Figure 4 | Aggregate publication impact *C*. The total number of citations C_i is also comprehensive productivity-impact measure. For most best-fit DGBD model curves, the C_i value is preserved with high precision. This shows that the difference between a given $c_i(r)$ and the corresponding best-fit DGBD model function are negligible on the macroscopic scale. (a) The exact aggregate number of citations C_i, calculated from $c_i(r)$ using Eq. [5], can be analytically approximated by $C_{\beta,h} \sim h_i^{1+\beta_i}$ using Eq. [6] which depends only on the scientist's β_i and h_i values. (b) We justify the use of the DGBD model defined in Eq. [3] for the approximation of $c_i(r)$ by comparing the aggregate citations C_i with the expected aggregate citations $C_m = \sum_{r=1}^{r_1} c_m(r)$ calculated from the best-fit DGBD model $c_m(r)$. Including the extra scaling-parameter, as in the DGBD model, improves the agreement between the theoretical and empirical C_i values in (a) and (b). We plot the line $y = x$ (dashed-green line) for visual reference.

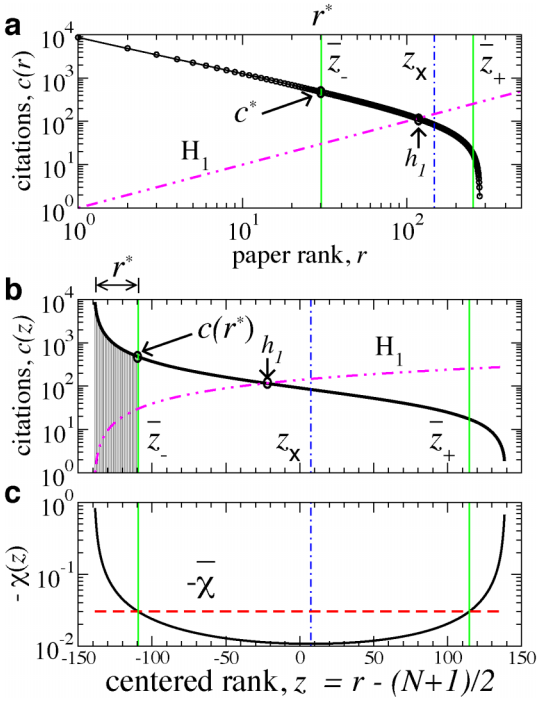

Figure 5 | Characteristic properties of the DGBD. We graphically illustrate the derivation of the characteristic $c_i(r)$ crossover values that locate the two tail regimes of $c_i(r)$, in particular, the distinguished "peak" paper regime corresponding to paper ranks $r \leq r^*$ (shaded region). The crossover between two scaling regimes suggests a complex reinforcement relation between the impact of a scientist's most famous papers and the impact of his/her other papers. (a) The $c_i(r)$ plotted on log-log axes with $N = 278$, $\beta = 0.83$ and $\gamma = 0.67$, corresponding to the average values of the Dataset [A] scientists. The hatched magenta curve is the $H_1(z)$ line on the log-linear scale with corresponding h-index value $h = 104$. The r^* value for $c_i(r)$ is not visibly obvious. (b) We plot on log-linear axes the centered citation profile $c_i(z)$ (solid black curve) given by the symmetric rank transformation $z = r - z_0$ in Eq. [7]. This representation better highlights the peak paper regime, but fails to highlight the power-law β scaling. (c) We plot the corresponding logarithmic derivative $\chi(z)$ of $c(z)$ (solid black curve), which represents the relative change in $c(z)$. The dashed red line corresponds to $-\bar{\chi}$, where $\bar{\chi}$ is the average value of $\chi(z)$ given by Eq. [12]. The values of $\bar{z}_{\pm}$, indicated by the solid vertical green lines, are defined as the intersection of $\bar{\chi}$ with $\chi(z)$ given by Eq. [13]. The regime $z < \bar{z}_-$ corresponds to the best papers of a given author. The hatched blue line corresponds to $z_\times$ which marks the crossover between the β and γ scaling regimes.

scientists. In Fig. 4(a), we plot for each scientist the predicted $C_{\beta,h}$ value versus the empirical C_i value, and we find excellent agreement with our theoretical prediction $C_i \sim h_i^{1+\beta_i}$ given by Eq. [6]. In Fig. 4(b), we plot for each scientist the total number of citations $C_m = \sum_{r=1}^{r_1} c_m(r)$ using the best-fit DGBD model $c_m(r) \equiv c_i(r; \beta_i, \gamma_i, A_i, r_1)$ to approximate $c_i(r)$. The excellent agreement demonstrates that the fluctuations in the residual difference $c_m(r) - c_i(r)$ cancel out on the aggregate level. Furthermore, a comparison of the quality of agreement between the theoretical C_i values and the empirical C_i values in Fig. 4(a) and (b) shows the importance of the additional γ_i scaling regime in the DGBD model.

Discussion

We use the DGBD model to provide an analytic description of $c_i(r)$ over the entire range of r, and provide a deeper quantitative understanding of scientific impact arising from an author's career publication works. The DGBD model exhibits scaling behavior for both large and small r, where the scaling for small r is quantified by the exponent β_i, which for many scientists analyzed, can be approximated using only two values of the generalized h-index h_p (see SI text). In particular, we show that for a given h-value, a larger β_i value corresponds to a more prolific publication career, since $C_i \sim h_i^{1+\beta_i}$.

Many studies analyze only the high rank values of generic Zipf ranking profiles $c(r)$, e.g. computing the scaling regime for $r < r_c$ below some some rank cutoff r_c. However, these studies cannot quantitatively relate the large observations to the small observations within the system of interest. To account for this shortcoming, our method for calculating the crossover values $r_i^* \equiv \bar{r}_-$, $r_\times$, and $\bar{r}_+$, which we elaborate in the methods section, can be used in general to quantitatively distinguish relatively large observations and relatively small observations within the entire set of observations. Moreover, the DGBD model has been shown to have wide application in quantifying the Zipf rank profiles in various phenomena[21].

To measure the upward mobility of a scientist's career, in the SI text we address the question: given that a scientist has index h, what is her/his most likely h-index value Δt years in the future? In consideration of the bulk of $c_i(r)$, and following from the regularity of $c_i(r)$ for $r \approx h$, we propose a model-free gap-index $G(\Delta h)$ as both an estimate and a target for future achievement which can be used in the review of career advancement. The gap index $G(\Delta h)$, defined as a proxy for the total number of citations a scientist needs to reach a target value $h+\Delta h$, can detect the potential for fast h-index growth by quantifying $c_i(r)$ around h. This estimator differs from other estimators for the time-dependent h-index[33–35] in that $G(\Delta h)$ is model independent.

Even though the productivity of scientists can vary substantially[9,36–39], and despite the complexity of success in academia, we find remarkable statistical regularity in the functional form of $c_i(r)$ for the scientists analyzed here from the physics community. Recent work in[8,9,40] calculates the citation distributions of papers from various disciplines and shows that proper normalization of impact measures can allow for comparison across time and discipline. Hence, it is likely that the publication careers of productive scientists in many disciplines obey the statistical regularities observed here for the set of 300 physicists. Towards developing a model for career evolution, it is still unclear how the relative strengths of two contributing factors (i) the extrinsic cumulative advantage effect[2,3,9] versus (ii) the intrinsic role of the "sacred spark" in combination with intellectual genius[37] manifest in the parameters of the DGBD model.

With little calculation, the β_i metric developed here, used in conjunction with the h_i, can better answer the question, "How popular are your papers?"[41]. Since the cumulative impact and productivity of individual scientists are also found to obey statistical laws[9,11], it is possible that the competitive nature of scientific advancement can be quantified and utilized in order to monitor career progress. Interestingly, there is strong evidence for a governing mechanism of career progress based on cumulative advantage[9,11,42] coupled with the the inherent talent of an individual, which results in statistical regularities in the career achievements of scientists as well as professional athletes[11,43,44]. Hence, whenever data are available[45,46], finding statistical regularities emerging from human endeavors is a first step towards better understanding the dynamics of human productivity.

Methods

Selection of scientists and data collection. We use disambiguated "distinct author" data from *ISI Web of Knowledge*. This online database is host to comprehensive data that is well-suited for developing testable models for scientific impact[9,32,40] and career progress[11]. In order to approximately control for discipline-specific publication and citation factors, we analyze 300 scientists from the field of physics.

We aggregate all authors who published in *Physical Review Letters* (PRL) over the 50-year period 1958–2008 into a common dataset. From this dataset, we rank the scientists using the citations shares metric defined in[9]. This citation shares metric divides equally the total number of citations a paper receives among the n coauthors, and also normalizes the total number of citations by a time-dependent factor to account for citation variations across time and discipline.

Hence, for each scientist in the PRL database, we calculate a cumulative number of citation shares received from only their PRL publications. This tally serves as a proxy for his/her scientific impact in all journals. The top 100 scientists according to this citation shares metric comprise dataset [A]. As a control, we also choose 100 other dataset [B] scientists, approximately randomly, from our ranked PRL list. The selection criteria for the control dataset [B] group are that an author must have published between 10 and 50 papers in PRL. This likely ensures that the total publication history, in all journals, be on the order of 100 articles for each author selected. We compare the tenured scientists in datasets A and B with 100 relatively young assistant professors in dataset [C]. To select dataset [C] scientists, we chose two assistant professors from the top 50 U.S. physics and astronomy departments (ranked according to the magazine *U.S. News*).

For privacy reasons, we provide in the SI tables only the abbreviated initials for each scientist's name (last name initial, first and middle name initial, e.g. L, FM). Upon request we can provide full names.

We downloaded datasets A and B from ISI Web of Science in Jan. 2010 and dataset C from ISI Web of Science in Oct. 2010. We used the "Distinct Author Sets" function provided by ISI in order to increase the likelihood that only papers published by each given author are analyzed. On a case by case basis, we performed further author disambiguation for each author.

Statistical significance tests for the $c(r)$ DGBD model. We test the statistical significance of the DGBD model fit using the χ^2 test between the 3-parameter best-fit DGBD $c_m(r)$ and the empirical $c_i(r)$. We calculate the p-value for the χ^2 distribution with $r_1 - 3$ degrees of freedom and find, for each data set, the number $N_{>P_c}$ of $c_i(r)$ with p-value $> p_c$: $N_{>P_c} = 4$ [A], 19 [B], 22 [C] for $p_c = 0.05$, and 8 [A], 22 [B], 37 [C] for $p_c = 0.01$.

The significant number of $c_i(r)$ which do not pass the χ^2 test for $P_c = 0.05$, results from the fact that the DGBD is a scaling function over several orders of magnitude in both r and $c_i(r)$ values, and so the residual differences $[c_i(r) - c_m(r)]$ are not expected to be normally distributed since there is no characteristic scale for scaling functions such as the DGBD. Nevertheless, the fact that so many $c_i(r)$ do pass the χ^2 test at such a high significance level, provides evidence for the quality-of-fit of the DGBD model. For comparison, none of the $c_i(r)$ pass the χ^2 test using the power-law model at the $P_c = 0.05$ significance level. In the next section, we will also compare the macroscopic agreement in the total number of citations for each scientist and the total number of citations predicted by the DGBD model for each scientist, and find excellent agreement.

Derivation of the characteristic DGBD r values. Here we use the analytic properties of the DGBD defined in Eq. [3] to calculate the special r values from the parameters β, γ and N which locate the two tail regimes of $c(z)$, and in particular, the distinguished paper regime. The scaling features of the DGBD do not readily convey any characteristic scales which distinguish the two scaling regimes. Instead, we use the properties of $\ln\, c_i(r)$ to characterize the crossover between the high-rank and the low-rank regimes of $c_i(r)$.

We begin by considering $c_i(r)$ under the centered rank transformation $z = r - z_0$, where $z_0 = (N + 1)/2$, then

$$c(z) = A\frac{(z_0 - z)^{\gamma}}{(z_0 + z)^{\beta}}, \tag{7}$$

in the domain $z \in [-(z_0 - 1), (z_0 - 1)]$. The logarithmic derivative of $c(z)$ expresses the relative change in $c(z)$,

$$\begin{aligned}\chi(z) &\equiv \frac{d\ln c(z)}{dz} = \frac{dc(z)/dz}{c(z)} \\ &= -\left(\frac{\gamma}{z_0 - z} + \frac{\beta}{z_0 + z}\right) = -m\left(\frac{1+\theta x}{1-x^2}\right),\end{aligned} \tag{8}$$

where $x = z/z_0$, $\theta = \frac{\gamma-\beta}{\gamma+\beta}$, and $m = \left(\frac{\gamma+\beta}{z_0}\right)$. The extreme values of $\frac{d\ln c(z)}{dz}$ for $z_0 \gg 1$ are given by

$$\frac{d\ln c(z)}{dz}\Big|_{z=-(z_0-1)} \approx -\beta \tag{9}$$

$$\frac{d\ln c(z)}{dz}\Big|_{z=z_0-1} \approx -\gamma \tag{10}$$

and the average value $\overline{\chi}$ is calculated by,

$$\begin{aligned}\overline{\chi} &\equiv \left\langle \frac{d\ \ln\ c(z)}{dz} \right\rangle \\ &= \frac{-m}{(1-1/z_0)-(1/z_0-1)}\int_{-(1-1/z_0)}^{(1-1/z_0)} dx \frac{(1+\theta x)}{1-x^2} \\ &= \frac{-m}{2}\ln\ N\end{aligned} \tag{11}$$

The function $\chi(z)$ takes on the value of $\overline{\chi}$ twice at the values $\overline{z}_{\pm} = z_0\overline{x}_{\pm}$ corresponding to the solutions to the quadratic equation,

$$\overline{\chi} = -m\left(\frac{1+\theta x}{1-x^2}\right), \tag{12}$$

which has the solution

$$\begin{aligned}\overline{x}_{\pm} &= -\frac{\theta}{\ln\ N} \pm \frac{\sqrt{(\ln\ N)^2 - 2\ \ln\ N + \theta^2}}{\ln\ N} \\ &\approx -\frac{\theta}{\ln\ N} \pm \sqrt{1 - 2/\ln\ N}\end{aligned} \tag{13}$$

for $\theta^2/\ln^2\ N \ll 1$. Converting back to rank, then

$$\overline{r}_{\pm} \approx \left(\frac{N}{2}\right)\left(1 - \frac{\theta}{\ln\ N} \pm \sqrt{1 - 2/\ln\ N}\right), \tag{14}$$

and so the value $r^* \equiv \overline{r}_-$ is the special rank value which distinguishes the set of excellent papers of each given author. The c-star value $c_i(r^*)$ is thus a characteristic value arising from the special analytic properties of $c_i(r)$. This method for determining the crossover value r^* can be applied to any general rank order profile which can be modeled by the DGBD.

Furthermore, the crossover z_x between the β scaling regime and the γ scaling regime is calculated from the inflection points of $\ln\, c(z)$,

$$0 = \frac{d^2\ \ln\ c(z)}{dz^2}\Big|_{z=z_x} = \frac{-\gamma}{(z_0 - z_x)^2} + \frac{\beta}{(z_0 + z_x)^2} \tag{15}$$

which has 2 solutions $z_x^{\pm} = z_0\left(\frac{1\pm\zeta}{1\mp\zeta}\right)$, where $\zeta \equiv \sqrt{\gamma/\beta}$. only $|z_x^-| < z_0$ is a physical solution. Transforming back to rank values, we find $r_x = z_0 + z_x^- = z_0\frac{2}{1+\zeta} = \frac{N+1}{1+\zeta}$. We illustrate these special z values in Fig. 5.

1. Mazloumian, A., Eom, Y.-H., Helbing, D., Lozano, S., Fortunato, S. How citation boosts promote scientific paradigm shifts and Nobel prizes. *PLoS ONE* **6**(5), e18975 (2011).
2. Merton, R. K. The Matthew effect in science. *Science* **159**, 56–63 (1968).
3. Merton, R. K. The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *ISIS* **79**, 606–623 (1988).
4. Cole, J. R. *Social Stratification in Science* (Chicago, Illinois, The University of Chicago Press, 1981).
5. Guimera, R., Uzzi, B., Spiro, J., Amaral, L. A. N. Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308**, 697–702 (2005).
6. Malmgren, R. D., Ottino, J. M., Amaral, L. A. N. The role of mentorship in protégé performance. *Nature* **463**, 622–626 (2010).
7. Azoulay, P., Zivin, J. S. G., & Wang, J. Superstar Extinction. *Q. J. of Econ.* **125** (2), 549–589 (2010).
8. Radicchi, F., Fortunato, S. & Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl. Acad. Sci. USA* **105**, 17268–17272 (2008).
9. Petersen, A. M., Wang, F., Stanley, H. E. Methods for measuring the citations and productivity of scientists across time and discipline. *Phys. Rev. E* **81**, 036114 (2010).
10. Simonton, D. K. Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychol. Rev.* **104**, 66–89 (1997).
11. Petersen, A. M., Jung, W.–S., Yang, J.–S. & Petersen, A. M., Jung, W.–S., Yang, J.–S. & Stanley, H. E. Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proc. Natl. Acad. Sci. USA* **108**, 18–23 (2011).
12. Wu, J., Lozano, S., Helbing, D. Empirical study of the growth dynamics in real career h-index sequences. *J. Informetrics* **5**, 489–497 (2011). (In press)
13. Petersen, A. M., Riccaboni, M., Stanley, H. E., Pammolli, F. *Persistency and Uncertainty in the Academic Career.* (2011). In preparation.
14. Hirsch, J. E. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA* **102**, 16569–16572 (2005).
15. Bornmann, L., Mutz, R., Daniel, H.–J. Are there better indices for evaluation purposes than the h Index? A comparison of nine different variants of the h Index using data from biomedicine. *JASIST* **59**, 001–008 (2008).
16. Egghe, L. Theory and practise of the g-index. *Scientometrics* **69**, 131–152 (2006).
17. Zhang, C–T. Relationship of the h-index, g-index, and e-index. *JASIST* **62**, 625–628 (2010).
18. van Eck, J. N., Waltman, L. Generalizing the h- and g-indices. *J. Informetrics* **2**, 263–271 (2008).
19. Wu, Q. The w-index: A measure to assess scientific impact focusing on widely cited papers. *JASIST* **61**, 609–614 (2010).
20. Naumis, G. G., Cocho, G. Tail universalities in rank distributions as an algebraic problem: The beta-like function. *Physica A* **387**, 84–96 (2008).
21. Martinez-Mekler, G., Martinez, R. A., del Rio, M. B., Mansilla, R., Miramontes, P., Cocho, G. Universality of rank-ordering distributions in the arts and sciences. *PLoS ONE* **4**, e4791 (2009).
22. Egghe, L., Rousseau, R. An informetric model for the Hirsch-index. *Scientometrics* **69**, 121–129 (2006).
23. Zipf, G. *Human Behavior and the principle of least effort* (Cambridge, MA, Addison-Wesley, 1949).
24. Gabaix, X. Zipf's law for cities: An explanation. *Q. J. of Econ.* **114** (3), 739–767 (1999).
25. ISI Web of Knowledge: www. isiknowledge.com/

26. Henzinger, M., Sunol, J., Weber, I. The stability of the h-index. *Scientometrics* **84**, 465–479 (2010).
27. Hirsch, J. E. Does the h index have predictive power. *Proc. Natl. Acad. Sci. USA* **104**, 19193–19198 (2008).
28. Batista, P. D., Campiteli, M. G., Martinez, A. S. Is it possible to compare researchers with different scientific interests? *Scientometrics* **68**, 179–189 (2006).
29. Iglesias, J. E., Pecharromán, C. Scaling the h-index for different scientific ISI fields. *Scientometrics* **73**, 303–320 (2007).
30. Bornmann, L., Daniel, H.–J. What do we know about the h index? *JASIST* **58**, 1381–1385 (2007).
31. Redner, S. On the meaning of the h-index. *J. Stat. Mech.* **2010**, L03005 (2010).
32. Radicchi, F., Fortunato, S., Markines, B., Vespignani, A. Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E* **80**, 056103 (2009).
33. Egghe, L. Dynamic h-Index: the Hirsch index in function of time. *JASIST* **58**, 452–454 (2006).
34. Burrell, Q. L. Hirsch's h-index: A stochastic model. *J. Informetrics* **1**, 16–25 (2007).
35. Guns, R., Rousseau, R. Simulating growth of the h-index. *JASIST* **60**, 410–417 (2009).
36. Shockley, W. On the statistics of individual variations of productivity in research laboratories. *Proc. of the IRE* **45**, 279–290 (1957).
37. Allison, A. D., Stewart, J. A. Productivity differences among scientists: Evidence for accumulative advantage. *Amer. Soc. Rev.* **39**(4), 596–606 (1974).
38. Huber, J. C. Inventive productivity and the statistics of exceedances. *Scientometrics* **45**, 33–53 (1998).
39. Peterson, G. J., Presse, S., Dill, K. A. Nonuniversal power law scaling in the probability distribution of scientific citations. *Proc. Natl. Acad. Sci. USA* **107**, 16023–16027 (2010).
40. Radicchi, F., Castellano, C. Rescaling citations of publications in Physics. *Phys. Rev. E* **83**, 046116 (2011).
41. Redner, S. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys J. B* **4**, 131–134 (1998).
42. De Solla Price, D. A general theory of bibliometric and other cumulative advantage processes. *JASIST* **27**, 292–306 (1976).
43. Petersen, A. M., Jung, W.-S. & Stanley, H. E. On the distribution of career longevity and the evolution of home-run prowess in professional baseball. *EPL* **83**, 50010 (2008).
44. Petersen, A. M., Penner, O. & Stanley, H. E. Methods for detrending success metrics to account for inflationary and deflationary factors. *Eur. Phys. J. B* **79**, 67–78 (2011).
45. Lazer, D., *et al.* Computational social science. *Science* **323**, 721–723 (2009).
46. Castellano, C., Fortunato, S., Loreto, V. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
47. Redner, S. Citation statistics from 110 years of Physical Review. *Phys. Today.* **58**, 49–54 (2005).

Acknowledgments

We thank J. E. Hirsch and J. Tenenbaum for helpful suggestions.

Author contributions

A. M. P., H. E. S., & S. S. designed research, performed research, wrote, reviewed and approved the manuscript. A. M. P. performed the numerical and statistical analysis of the data.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Petersen, A.M., Stanley, H.E. & Succi, S. Statistical regularities in the rank-citation profile of scientists. *Sci. Rep.* **1**, 181; DOI:10.1038/srep00181 (2011).