The advantage of short paper titles

Adrian Letchford, Helen Susannah Moat and Tobias Preis

Data Science Lab, Behavioural Science, Warwick Business School, University of Warwick, Coventry CV4 7AL, UK

Vast numbers of scientific articles are published each year, some of which attract considerable attention, and some of which go almost unnoticed. Here, we investigate whether any of this variance can be explained by a simple metric of one aspect of the paper's presentation: the length of its title. Our analysis provides evidence that journals which publish papers with shorter titles receive more citations per paper. These results are consistent with the intriguing hypothesis that papers with shorter titles may be easier to understand, and hence attract more citations.

1. Introduction

Written communication is now being recorded online on a massive scale [1–5]. Colossal amounts of data on collective information gathering and distribution via online services such as *Twitter* [6–9], *Wikipedia* [10–13], *Google* [14–17], news services [18] and even large digitized collections of books [19–21] can now be analysed, widening our understanding of economic decision-making [11,14,16], human conflict [7,12] and natural disasters [22,23].

Scientific endeavours also generate extensive written communication, in the form of papers. We define a paper to be more successful than others if it has received a greater number of citations. The online database *Scopus* contains citation records of papers, offering remarkable insights into academic conversation. Recently, advances have been made in quantifying scientific output based on publication statistics [24–28]. A number of studies have provided evidence that the long-term success of scientists depends on their early publications [29,30]. Further analyses have indicated that a paper's success can be partially predicted by its early success [31–33] as well as the reputation of the authors [34]. In addition, papers in particular academic domains gain more citations than others [35].

Here, we consider whether we can find any evidence that the style in which a paper is written may relate to its success. Specifically, we consider the length of the article title chosen by the authors and investigate whether the length bears any relation to the number of citations. Previous studies have explored different

characteristics of scientific paper titles [36–41]. A subset of these studies have focused on identifying stylistic attributes of academic writing and the use of a colon or question in a paper's title [36–39]. Those which have investigated the relationship between the length of an article's title and the number of citations it receives have been limited to relatively small samples, up to a maximum of 2200 papers [40,41]. These analyses have reported conflicting results, with one study suggesting that papers with longer titles might receive more citations [41] and another finding no evidence of a relationship [40]. Here, we exploit data on a much larger sample of 140 000 papers in order to investigate whether a paper's title length bears any relation to the number of citations it receives.

2. Results

We analyse data provided by *Scopus*, one of the leading bibliometric platforms. A *Scopus* user can search and export data on journal articles in batches of 20 000 records, including data on how often each article has been cited since publication. We download data on the 20 000 most cited papers in each year between 2007 and 2013.

We determine the number of characters in each paper's title, including spaces and punctuation. Using the year 2010 as an example, we rank the papers' title length and citations (figure 1*a*). Upon visual inspection, there appears to be a high concentration of papers with short titles and many citations, as well as a high concentration of papers with long titles and few citations. We find that for the top 20 000 most highly cited papers published in 2010, papers with shorter titles receive more citations (Kendall's $\tau = -0.07$, $N = 15\,395$, $p < 0.001$). We apply the same analysis to each year in our sample and find that papers from all years exhibit this relationship between their title length and citations (figure 1*b*; all τs < -0.042, all ps < 0.001, $\alpha = 0.05$, Kendall's τ correlation with false discovery rate (FDR) correction).

Some journals may attract a greater number of citations for their papers owing to their reputation. To remove any potential influence of the journal in which a paper is published on the relationship between citations received and paper title length, we rank all of the papers in terms of the number of citations received and transform these ranks into percentiles. We calculate percentiles in terms of the length of papers' titles in the same fashion. In this transformed dataset, for papers published in 2010, we find that papers with shorter titles receive more citations (figure 1*c*; $\tau = -0.020$, $N = 15\,395$, $p < 0.001$, Kendall's τ correlation). Again, we run parallel analyses for the 20 000 most cited papers in each year between 2007 and 2013. For years 2007–2010, we find that papers with shorter titles receive more citations, whereas papers published during 2011–2013 do not (figure 1*d*; for years 2007–2010, all τs < -0.016, all Ns $> 14\,791$, all ps < 0.01; for years 2011–2013, all $|\tau|$s < 0.01, all Ns $> 15\,396$, all ps > 0.05; Kendall's τ with FDR correction). These smaller τs suggest that the journal in which a paper is published may help explain the relationship between paper title length and the number of citations the paper receives.

To investigate this hypothesis further, we group papers by their journal. Again, using 2010 as an example, we calculate the median number of citations and median title length for each journal. We find that journals which published papers with shorter titles also tend to receive more citations per paper (figure 2*a*; Kendall's $\tau = -0.19$, $N = 361$, $p < 0.001$). Parallel analyses for papers published in each year between 2007 and 2013 show that this relationship holds for papers published in all 7 years in our sample (figure 2*b*; 2012: $\tau = -0.1$, $N = 320$, $p < 0.05$; 2013: $\tau = -0.11$, $N = 352$, $p < 0.01$; all other years: all τs ≤ -0.14, all ps < 0.001, $\alpha = 0.05$; Kendall's τ correlation with FDR correction). Finally, we carry out a complementary aggregated analysis across all years of data in our sample. We rank all papers published in a given year by citations received and by title length, and transform these ranks into percentiles for that year. Again, we find that journals which publish papers with shorter titles also tend to receive more citations per paper (figure 3; $\tau = -0.19$, $N = 625$, $p < 0.001$, Kendall's τ correlation).

Our primary analysis is based on rank-based statistics. To complement our analysis, we fit a mixed-effects model to the log of the number of citations a paper receives as a function of its title length controlling for the journal in which each paper is published. A mixed-effects models allows us to control for the journal in which each paper is published. We define our model as

$$\log_{10}(c_{j,p}) = I + I_j + (L + L_j)l_{j,p} + \epsilon_{j,p}, \tag{2.1}$$

where $c_{j,p}$ is the number of citations received by paper p published in journal j. The distribution of citations received by a paper is highly positively skewed. For this reason, we log these citation counts, so that the distribution of the residuals of our model, ϵ, is closer to a Gaussian distribution. The grand intercept is I, whereas I_j is an intercept for each journal. There is a fixed slope L for the number of characters in the title $l_{j,p}$ for paper p published in journal j. There is also a journal-level random effects
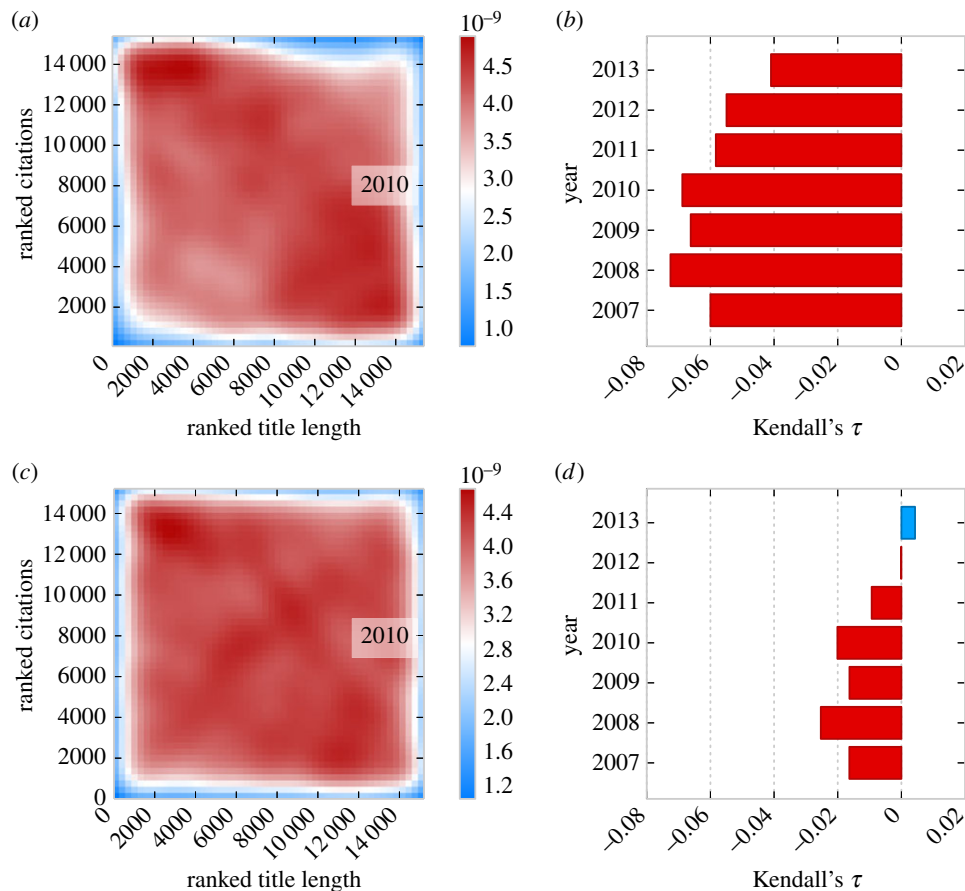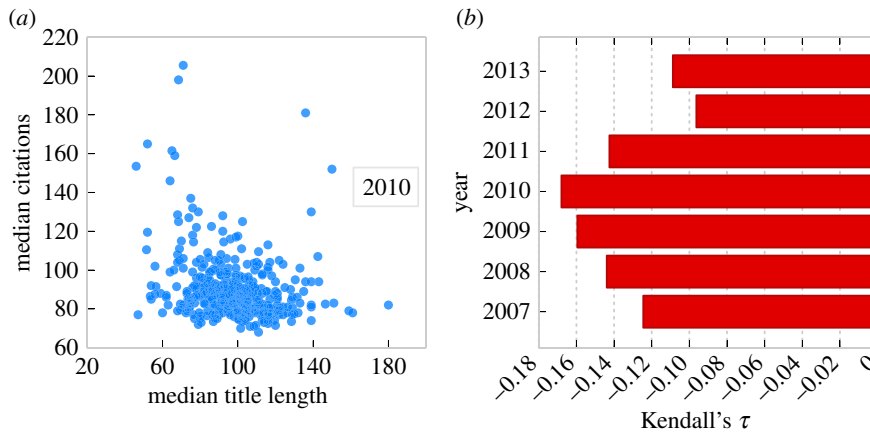
Figure 1. Paper title length and citations received. (*a*) We consider the 20 000 most cited papers in 2010. We rank the papers both in terms of citations received and title length. A density plot of the ranked citations and title length reveals that papers with shorter titles receive more citations (Kendall's $\tau = -0.07$, $N = 15\,395$, $p < 0.001$). (*b*) We run parallel analyses for the 20 000 most cited papers in each year between 2007 and 2013. For each of these years, we find that papers with shorter titles receive more citations (all τs < -0.042, all ps < 0.001, $\alpha = 0.05$, Kendall's τ correlation with FDR correction). (*c*) To remove any potential influence of the journal in which a paper is published on this relationship, we rank all of the papers in terms of the number of citations received and transform these ranks into percentiles. We calculate percentiles in terms of the length of papers' titles in the same fashion. In this transformed data, we find that papers with shorter titles receive more citations in 2010 (Kendall's $\tau = -0.020$, $N = 15\,395$, $p < 0.001$). (*d*) We run parallel analyses for the 20 000 most cited papers in each year between 2007 and 2013. For years 2007–2010, we find that papers with shorter titles receive more citations, whereas papers published during 2011–2013 do not (for years 2007–2010, all τs < -0.016, all Ns $> 14\,791$, all ps < 0.01; for years 2011–2013, all $|\tau|$s < 0.01, all Ns $> 15\,396$, all ps > 0.05; Kendall's tau with FDR correction). The smaller τs in (*d*) suggest that the journal in which a paper is published may help explain the relationship between paper title length and the number of citations the paper receives.

slope for the title length L_j. We fit the model for each year using maximum likelihood. We find that papers published during 2007–2011 with shorter titles tend to receive more citations while those published during 2012 and 2013 do not (for years 2007–2010: all ts < -3.832, all ps < 0.001; 2011: $t = -3.314$, $N = 345$ $p < 0.01$; 2012–2013: both ts < -0.251, both ps > 0.05; t-test on slope L with FDR correction). The values of the slope L are given for all years in table 1.

Again, we investigate if this relationship exists when aggregating papers by the journal in which they are published. We fit a linear regression model to the median number of citations papers receive per journal as a function of the median title length. We define our model as

$$\log_{10}(c_j) = I + Ll_j + \epsilon_j, \tag{2.2}$$

where c_j is the median number of citations received by papers published in journal j. The intercept is I, and there is a slope L for the median number of characters in the titles of papers l_j published in journal j. Again, we log the citation counts, so that the distribution of the residuals of our model, ϵ, is closer to a

Figure 2. Paper title length and citations received, analysed at journal level. (*a*) For each journal in 2010, we plot the median citations for a paper against the median title length. We find that journals which publish papers with shorter titles receive more citations per paper (Kendall's $\tau = -0.19$, $p < 0.001$, $N = 361$). (*b*) Parallel analyses of the data for each year between 2007 and 2013 confirm that this relationship holds across all 7 years of data (2012: $\tau = -0.1$, $N = 320$, $p < 0.05$; 2013: $\tau = -0.11$, $N = 352$, $p < 0.01$; all other years: all τs ≤ -0.14, all ps < 0.001, $\alpha = 0.05$; Kendall's τ correlation with FDR correction).

Table 1. Mixed effects model of the relationship between paper title length and citations received. Our primary analysis in figures 1 and 2 are based on rank statistics. To complement this analysis, we fit linear models to the data. We fit a mixed-effects model to the log of the number of citations a paper receives as a function of its title length (equation (2.1)). The model includes a fixed slope L for the number of characters in the length of a paper's title. We fit this model for each year in our dataset and display the slopes here under 'for individual papers'. We find that, for each year, the slope is negative. We also investigate if this relationship exists when aggregating papers by the journal in which they are published. We fit a linear regression model to the log of the median number of citations papers receive per journal as a function of the median title length (equation (2.2)). There is a slope L for the median number of characters in the titles of papers published in each journal. We fit this model for each year in our dataset and display the slopes here under 'for individual journals'. Again, we find that for each year, the slope is negative.

	slope of length (*L*)	
year	for individual papers	for individual journals
2007	−0.0118***	−0.0147***
2008	−0.0118***	−0.0208***
2009	−0.0093***	−0.0190***
2010	−0.0099***	−0.0174***
2011	−0.0078**	−0.0183***
2012	−0.0040	−0.0080*
2013	−0.0005	−0.0116**

Asterisks represent FDR-corrected *p*-values for a *t*-test of the slope. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Gaussian distribution. We fit the model for each year. We find that journals which publish papers with shorter titles also tend to receive more citations per paper (for years 2007–2011: all ts < -4.215, all ps < 0.001; 2012–2013: both ts < -2.022, both ps < 0.05; t-test of slope L with FDR correction). The values of the slope L are given for all years in table 1.

3. Discussion

In this study, we investigate whether the length of a scientific paper's title is related to the number of citations it receives. We analyse the 20 000 most highly cited papers for the years 2007–2013, representing a sample size between 1.12% and 1.53% of all papers published in each of these years. Previous studies analysing much smaller sets of papers have reported conflicting evidence, suggesting either that the
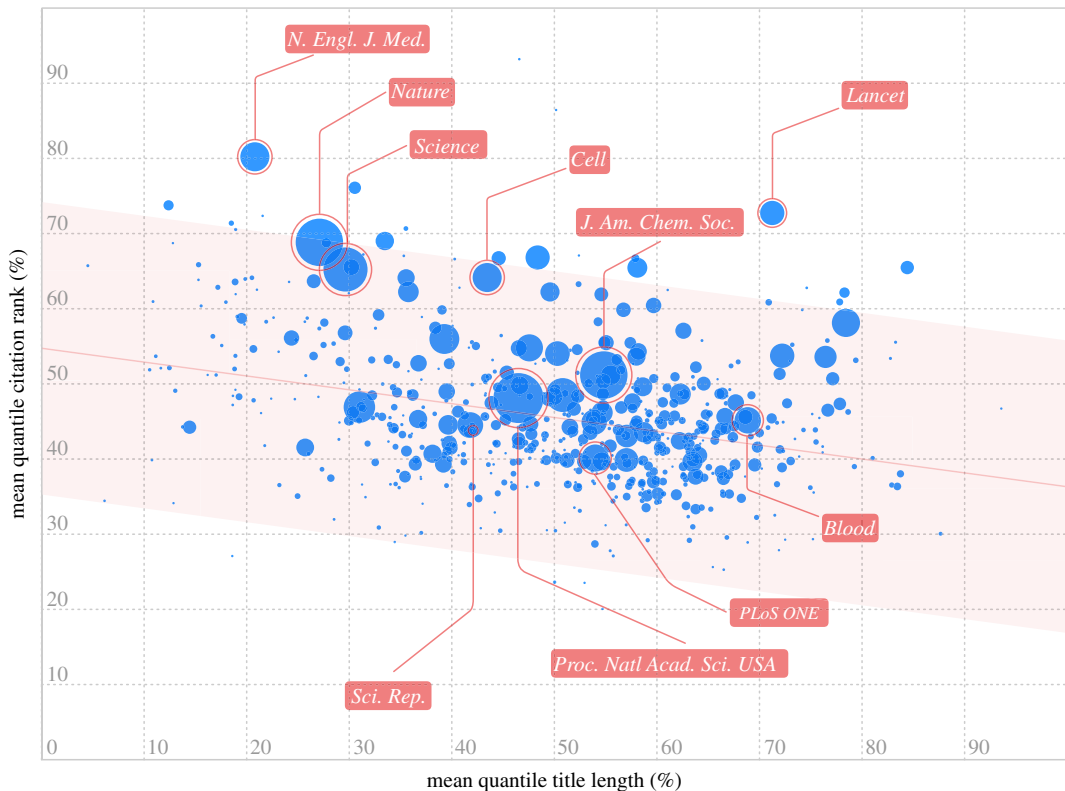
Figure 3. Journals which publish papers with shorter titles receive more citations per paper. For each year in our dataset, we rank all of the papers in terms of the number of citations received and in terms of the length of the titles, and transform these ranks into percentiles for a given year. For each journal, we then calculate the average quantile of the citations and of the title lengths, across papers and across years. Here, each blue circle represents a journal, the size of each circle represents the number of papers in our sample for that journal. Again, we find that journals that publish papers with shorter titles also tend to receive more citations per paper (Kendall's $\tau = -0.19$, $N = 625, p < 0.001$).

length of a paper's title bears no relation to its scientific impact [40], or that longer titles can be linked to greater citation counts [41].

Our analysis suggests that papers with shorter titles do receive greater numbers of citations. However, it is well known that papers published in certain journals attract more citations than papers published in others. When citation counts are adjusted for the journal in which the paper is published, we find that the strength of the evidence for the relationship between title length and citations received is reduced. Our results do however reveal that journals which publish papers with shorter titles tend to receive more citations per paper.

We propose three possible explanations for these results. One potential explanation is that high-impact journals might restrict the length of their papers' titles. Similarly, incremental research might be published under longer titles in less prestigious journals. A third possible explanation is that shorter titles may be easier to understand, enabling wider readership and increasing the influence of a paper.

Our findings provide evidence that elements of the style in which a paper is written may relate to the number of times it is cited. Future analysis will investigate whether further stylistic attributes of the language used in a paper can be related to the number of citations it receives.

4. Methods

We retrieve bibliometric data from *Scopus* (http://www.scopus.com) between 21 October 2014 and 14 November 2014. To obtain data on the 20 000 most cited papers published in each of the 7 years from 2007 to 2013, we search for any papers that are marked by *Scopus* as an 'article' with the following search query:

```
DOCTYPE(ar) AND PUBYEAR = {year},
```

where {year} is replaced by each of the years 2007–2013. In total, we retrieve 140 000 records. In 2007, *Scopus* reports 1 302 973 published papers which increases to 1 788 065 papers in 2013. The top 20 000 most cited papers published in each year represent a sample of 1.53% in 2007, decreasing to 1.12% in 2013.

Some journals are referred to with multiple variations of their name (for example, '*Analyst*' and '*The Analyst*'). For this reason, we clean the dataset from *Scopus* by deleting leading 'The's from each journal's title, and converting the title to lower case. We also identify all journals which have fewer than 10 papers in the most cited 20 000 papers for a given year, and remove the papers in such journals for that year. The basic characteristics of our dataset before and after cleaning are depicted in the electronic supplementary material, figure S1.

References

1. Conte R *et al.* 2012 Manifesto of computational social science. *Eur. Phys. J. ST.* **214**, 325–346. (doi:10.1140/epjst/e2012-01697-8)

2. King G. 2011 Ensuring the data-rich future of the social sciences. *Science* **331**, 719–721. (doi:10.1126/science.1197872)

3. Lazer D *et al.* 2009 Computational social science. *Science* **323**, 721–723. (doi:10.1126/science.1167742)

4. Moat HS, Preis T, Olivola CY, Liu C, Chater N. 2014 Using big data to predict collective behavior in the real world. *Behav. Brain Sci.* **37**, 92–93. (doi:10.1017/S0140525X13001817)

5. Watts DJ. 2007 A twenty-first century science. *Nature* **445**, 489. (doi:10.1038/445489a)

6. Bollen J, Mao H, Zeng X. 2011 Twitter mood predicts the stock market. *J. Comput. Sci.* **2**, 1–8. (doi:10.1016/j.jocs.2010.12.007)

7. Ciulla F, Mocanu D, Baronchelli A, Gonçalves B, Perra N, Vespignani A. 2012 Beating the news using social media: the case study of American Idol. *EPJ Data Sci.* **1**, 1–11. (doi:10.1140/epjds1)

8. Gonçalves B, Perra N, Vespignani A. 2011 Modeling users' activity on Twitter networks: validation of Dunbar's number. *PLoS ONE* **6**, e22656. (doi:10.1371/journal.pone.0022656)

9. Mocanu D, Baronchelli A, Perra N, Gonçalves B, Zhang Q, Vespignani A. 2013 The Twitter of Babel: mapping world languages through microblogging platforms. *PLoS ONE* **8**, e61981. (doi:10.1371/journal.pone.0061981)

10. Mestyán M, Yasseri T, Kertész J. 2013 Early prediction of movie box office success based on Wikipedia activity big data. *PLoS ONE* **8**, e71226. (doi:10.1371/journal.pone.0071226)

11. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T. 2013 Quantifying *Wikipedia* usage patterns before stock market moves. *Sci. Rep.* **3**, 1801. (doi:10.1038/srep01801)

12. Yasseri T, Sumi R, Rung A, Kornai A, Kertész J. 2012 Dynamics of conflicts in wikipedia. *PLoS ONE.* **7**, e38869. (doi:10.1371/journal.pone.0038869)

13. Kristoufek L. 2013 *BitCoin* meets *Google Trends* and *Wikipedia*: quantifying the relationship between phenomena of the Internet era. *Sci. Rep.* **3**, 3415. (doi:10.1038/srep03415)

14. Curme C, Preis T, Stanley HE, Moat HS. 2014 Quantifying the semantics of search behavior before stock market moves. *Proc. Natl Acad. Sci. USA* **111**, 11 600–11 605. (doi:10.1073/pnas.1324054111)

15. Preis T, Moat HS, Stanley HE, Bishop SR. 2012 Quantifying the advantage of looking forward. *Sci. Rep.* **2**, 350. (doi:10.1038/srep00350)

16. Preis T, Moat HS, Stanley HE. 2013 Quantifying trading behavior in financial markets using *Google Trends*. *Sci. Rep.* **3**, 1684. (doi:10.1038/srep01684)

17. Preis T, Moat HS. 2014 Adaptive nowcasting of influenza outbreaks using Google searches. *R. Soc. Open Sci.* **1**, 140095. (doi:10.1098/rsos.140095)

18. Alanyali M, Moat HS, Preis T. 2013 Quantifying the relationship between financial news and the stock market. *Sci. Rep.* **3**, 3578. (doi:10.1038/srep03578)

19. Michel JB *et al.* 2011 Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182. (doi:10.1126/science.1199644)

20. Petersen AM, Tenenbaum J, Havlin S, Stanley HE. 2012 Statistical laws governing fluctuations in word use from word birth to word death. *Sci. Rep.* **2**, 313. (doi:10.1038/srep00313)

21. Petersen AM, Tenenbaum JN, Havlin S, Stanley HE, Perc M. 2012 Languages cool as they expand: allometric scaling and the decreasing need for new words. *Sci. Rep.* **2**, 943. (doi:10.1038/srep00943)

22. Sakaki T, Okazaki M, Matsuo Y. 2010 Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW '10, 26–30 April 2010, Raleigh, NC.*, pp. 851–860. New York, NY: ACM.

23. Preis T, Moat HS, Bishop SR, Treleaven P, Stanley HE. 2013 Quantifying the digital traces of hurricane sandy on Flickr. *Sci. Rep.* **3**, 3141. (doi:10.1038/srep03141)

24. Petersen AM, Stanley HE, Succi S. 2011 Statistical regularities in the rank-citation profile of scientists. *Sci. Rep.* **1**, 181. (doi:10.1038/srep00181)

25. Petersen AM, Wang F, Stanley HE. 2010 Methods for measuring the citations and productivity of scientists across time and discipline. *Phys. Rev. E* **81**, 036114. (doi:10.1103/PhysRevE.81.036114)

26. Penner O, Pan RK, Petersen AM, Kaski K, Fortunato S. 2013 On the predictability of future impact in science. *Sci. Rep.* **3**, 3052. (doi:10.1038/srep03052)

27. Petersen AM, Penner O. 2014 Inequality and cumulative advantage in science careers: a case study of high-impact journals. *Eur. Phys. J. Data Sci.* **3**, 24. (doi:10.1140/epjds/s13688-014-0024-y)

28. Petersen AM, Succi S. 2013 The Z-index: a geometric representation of productivity and impact which accounts for information in the entire rank-citation profile. *J. Informetr.* **7**, 823–832. (doi:10.1016/j.joi.2013.07.003)

29. van Dijk D, Manor O, Carey LB. 2014 Publication metrics and success on the academic job market. *Curr. Biol.* **24**, R516–R517. (doi:10.1016/j.cub.2014.04.039)

30. Laurance WF, Useche DC, Laurance SG, Bradshaw CJA. 2013 Predicting publication success for biologists. *BioScience* **63**, 817–823. (doi:10.1525/bio.2013.63.10.9)

31. Acuna DE, Allesina S, Kording KP. 2012 Predicting scientific success. *Nature* **489**, 201–202. (doi:10.1038/489201a)

32. Hirsch JE. 2007 Does the H index have predictive power? *Proc. Natl Acad. Sci. USA* **104**, 19 193–19 198. (doi:10.1073/pnas.0707962104)

33. Wang D, Song C, Barabasi AL. 2013 Quantifying long-term scientific impact. *Science* **342**, 127–132. (doi:10.1126/science.1237825)

34. Petersen AM, Fortunato S, Pan RK, Kaski K, Penner O, Rungi A, Riccaboni M, Eugene Stanleye H, Pammollia F. 2014 Reputation and impact in academic careers. *Proc. Natl Acad. Sci. USA* **111**, 15 316–15 321. (doi:10.1073/pnas.1323111111)

35. Yogatama D, Heilman M, O'Connor B, Dyer C, Routledge BR, Smith NA. 2011 Predicting a scientific community's response to an article. In *Proc. EMNLP, 27–31 July 2011, Edinburgh, UK.*, pp. 594–604. Stroudsburg, PA: Association for Computational Linguistics.

36. Soler V. 2007 Writing titles in science: an exploratory study. *ESP* **26**, 90–102. (doi:10.1016/j.esp.2006.08.001)

37. Lewison G, Hartley J. 2005 What's in a title? Numbers of words and the presence of colons. *Scientometrics* **63**, 341–356. (doi:10.1007/s11192-005-0216-0)

38. Hartley J. 2005 To attract or to inform: what are titles for? *JTWC* **35**, 203–123. (doi:10.2190/NV6E-FN3N-7NGN-TWQT)

39. Hartley J. 2007 Planning that title: practices and preferences for titles with colons in academic articles. *Libr. Inf. Sci. Res.* **29**, 553–568. (doi:10.1016/j.lisr.2007.05.002)

40. Jamali HR, Nikzad M. 2011 Article title type and its relation with the number of downloads and citations. *Scientometrics* **88**, 653–661. (doi:10.1007/s11192-011-0412-z)

41. Jacques TS, Sebire NJ. 2009 The impact of article titles on citation hits: an analysis of general and specialist medical journals. *J. R. Soc. Med. Short Rep.* **1**, 1–5. (doi:10.1258/shorts.2009.100020)