Journal of Informetrics 10 (2016) 1–8

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

The advantage of simple paper abstracts

Adrian Letchford*, Tobias Preis, Helen Susannah Moat

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

ARTICLE INFO

Article history:
Received 16 June 2015
Received in revised form 1 November 2015
Accepted 1 November 2015

Keywords:
Citation analysis
Scientific writing
Computational social science
Science of science

ABSTRACT

Each year, researchers publish an immense number of scientific papers. While some receive many citations, others receive none. Here we investigate whether any of this variance can be explained by the choice of words in a paper's abstract. We find that doubling the word frequency of an average abstract increases citations by 0.70%. We also find that journals which publish papers whose abstracts are shorter and contain more frequently used words receive slightly more citations per paper. Specifically, adding a 5 letter word to an abstract decreases the number of citations by 0.02%. These results are consistent with the hypothesis that the style in which a paper's abstract is written bears some relation to its scientific impact.

1. Introduction

Written communication is now being recorded online on a colossal scale (Conte et al., 2012; King, 2011; Lazer et al., 2009; Michel et al., 2011; Moat, Preis, Olivola, Liu, & Chater, 2014; Petersen, Tenenbaum, Havlin, & Stanley, 2012; Petersen, Tenenbaum, Havlin, Stanley, & Perc, 2012; Watts, 2007). The datasets left behind offer unprecedented insight into how information flows between humans. Recent studies have attempted to link data on information exchanged online to data on actions taken in the real world, considering sources such as *Google* (Curme, Preis, Stanley, & Moat, 2014; Preis & Moat, 2014; Preis, Moat, & Stanley, 2013; Preis, Moat, Stanley, & Bishop, 2012), *Wikipedia* (Kristoufek, 2013; Mestyán, Yasseri, & Kertész, 2013; Moat et al., 2013; Yasseri, Sumi, Rung, Kornai, & Kertész, 2012), online news (Alanyali, Moat, & Preis, 2013), and *Twitter* (Bollen, Mao, & Zeng, 2011; Ciulla et al., 2012; Gonçalves, Perra, & Vespignani, 2011; Mocanu et al., 2013). In this paper, we focus on the communication of scientific findings in journals and papers. We investigate whether the way in which these findings are communicated in the abstract bears any relationship to the number of times other scientists cite the paper.

Online services such as *Web of Science* provide access to vast collections of scientific papers. These services also track the number of times each paper is cited as a measure of impact. Here, we define a successful paper as one that has received a greater number of citations. Recently, advances have been made in quantifying scientific output based on publication statistics, offering remarkable insight into academic conversation (Hartley, 2005, 2007; Laurance, Useche, Laurance, & Bradshaw, 2013; Letchford, Moat, & Preis, 2015; Lewison & Hartley, 2005; Penner, Pan, Petersen, Kaski, & Fortunato, 2013; Petersen & Penner, 2014; Petersen, Stanley, & Succi, 2011; Petersen & Succi, 2013; Petersen, Wang, & Stanley, 2010; Soler, 2007; van Dijk, Manor, & Carey, 2014; Yogatama et al., 2011). For example, recent analyses have indicated that a paper's success can be partially predicted by its early success (Acuna, Allesina, & Kording, 2012; Hirsch, 2007; Wang, Song, & Barabasi, 2013) as well as the reputation of the authors (Petersen et al., 2014).

* Corresponding author.
E-mail address: Adrian.Letchford@wbs.ac.uk (A. Letchford).

http://dx.doi.org/10.1016/j.joi.2015.11.001
1751-1577/

Here, we focus on whether or not the style in which a paper is written may relate to its success. For example, some paper titles include a question mark (Jamali & Nikzad, 2011) or other non-alphanumeric characters (Buter & van Raan, 2011) where inclusion of these characters has been linked to fewer citations. Not all of these results are reproducible across different samples of papers. For example, one analysis suggested that papers with a colon in their title tend to receive fewer citations (Jamali & Nikzad, 2011) and another analysis concludes the opposite (Jacques & Sebire, 2009). These studies use samples of 2,172 and 50 papers respectively. While we have previously analysed the relationship between paper title characteristics and citation rates (Letchford et al., 2015), here we focus on characteristics of a paper's abstract.

In this paper, we use a sample of 300,000 papers across all disciplines to investigate whether or not the style in which a paper's abstract is written may relate to its success. A previous analysis of a sample of 196 papers concluded that the length of an abstract is not indicative of the number of times the paper is cited (Falagas, Zarkali, Karageorgopoulos, Bardakas, & Mavros, 2013). A recent larger scale study however provided contradictory evidence, suggesting that papers with longer abstracts attract more citations (Weinberger, Evans, & Allesina, 2015). This same study also found that using simpler words in an abstract resulted in fewer citations for the paper.

However, in the light of evidence from psychological experiments, this finding might be considered surprising. In an experiment called the lexical decision task, participants attempt to classify sequences of characters as real or nonsense words. Studies have found that the frequency of words is a very influential factor in determining participant's response speed (Whaley, 1978). In this same task, longer words have also been shown to lead to longer response times (New, Ferrand, Pallier, & Brysbaert, 2006). These experiments suggest that the length and frequency of words may provide at least a crude measure of how easy they are to understand. We hypothesise that if comprehending a paper's abstract requires a higher cognitive load, due to uncommon words and lengthy prose, then the paper may not receive as many citations.

2. Methods

We obtained bibliometric data from *Web of Science* (http://webofknowledge.com). Between 20th November 2014 and 23rd November 2014 we retrieved the 30,000 most highly cited papers per year between 1999 and 2008 for a total of 300,000 papers. This represents the most frequently cited 1.04% of papers in 1999 to 0.73% in 2008. By searching for all papers in a specific year and sorting the results by the number of time each paper has been cited to date, we were able to download the records in batches of 500 papers. After approximately 30,000 papers, the *Web of Science* online interface is unable to continue fulfilling requests. For each paper, the dataset includes the title, the abstract, the publishing journal's title and serial number (ISSN) as well as the number of times the paper has been cited to date, including self-citations.

We clean our bibliometric dataset by first removing all records with a missing title, abstract, journal name, or ISSN. We then identify all journals which have 10 or fewer papers in a given year in our sample and remove the papers in such journals for that year. Our analysis in the *Results* section utilises maximum likelihood estimation which is unable to converge on a solution if we retain papers in journals with fewer than 11 papers in a given year. After cleaning, we have 216,280 papers in our dataset. The basic characteristics of this dataset before and after cleaning are displayed in Figure S11.

For each paper, we calculate the number of sentences in the abstract by counting the occurrences of a full stop followed by a space: ". ". We increment this number if the abstract also ends with a full stop.

We obtain yearly word counts from all the English books indexed by *Google* from *Google's Ngram* project.[1] We used the entire 1-gram dataset excluding counts for punctuation, parts of speech and non English characters. We divide each word's annual word count by the total number of words in that year. The basic characteristics of this dataset are displayed in Figure S12.

For each paper, we determine the median word frequency of the abstract using the *Google Ngram* counts from the year during which the paper was published. Any word that does not appear in the books that *Google* has indexed is considered to have a frequency of zero. The distribution of median word frequencies is shown in Figure S12D.

3. Results

Here, we quantify the relationship between the length of a papers abstract and the number of citations it receives. We also study how the frequency of words in a papers abstract relates to citations received.

3.1. Abstract length

Papers published in earlier years have had more time to attract citations. To normalise their citation counts and remove this effect, we convert the number of citations into percentiles as follows. For each year in our dataset, we rank all of the papers in terms of the number of citations received and transform these ranks into percentiles. For each journal, we then calculate the mean percentile of this citation count distribution. We also calculate the median number of characters in the abstracts of papers published in each journal, including all letters, punctuations, spaces and special characters. We find that,

[1] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

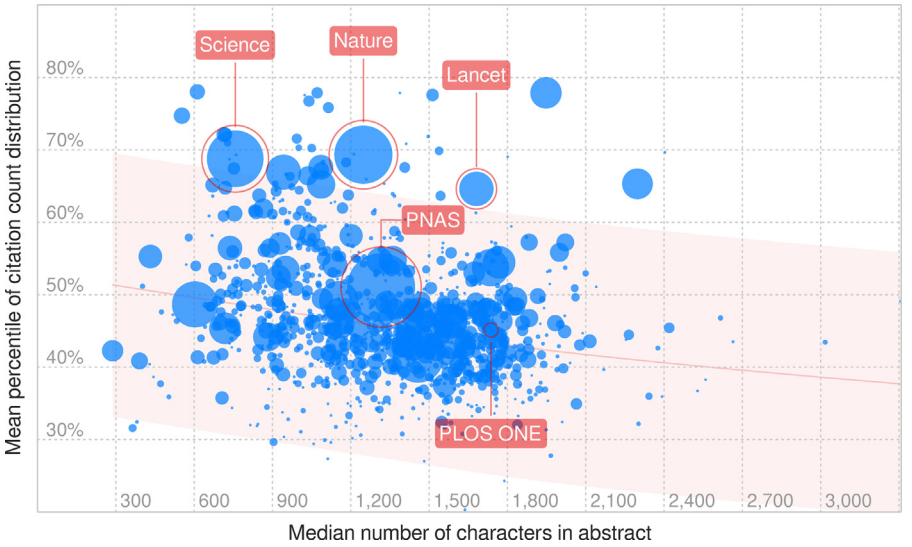

Fig. 1. Journals which publish papers with shorter abstracts receive more citations per paper. Papers published in earlier years have had more time to attract citations. To normalise the citation counts against this effect we convert the number of citations into percentiles as follows. For each year in our dataset, we rank all of the papers in terms of the number of citations received and transform these ranks into percentiles. For each journal, we then calculate the mean percentile of the citation distribution. We also calculate for each journal the median number of characters in the abstracts of papers published in that journal. Here, each blue circle represents a journal and the size of each circle represents the number of papers in our dataset published in that journal. We find that, in our sample, journals which publish papers with shorter abstracts tend to receive more citations per paper (Kendall's $\tau = -0.18$, $N = 955$, $p < 0.001$). (For interpretation of reference to color in this figure legend, the reader is referred to the web version of this article.)

in our sample, journals which publish papers with shorter abstracts tend to receive more citations per paper (Kendall's $\tau = -0.18$, $N = 955$, $p < 0.001$; Fig. 1).

We check whether this relationship holds for yearly subsets of our dataset. For each year in our sample, we calculate the median number of citations and median number of characters in the abstracts of papers published by each journal. We find that, in our sample, journals that publish papers with shorter abstracts tend to receive slightly more citations per paper (for all years, all Kendall's τs < -0.10, all Ns ≥ 489, all ps < 0.001, FDR corrected; Fig. S1).

The number of times a paper is cited is influenced in part by the field of research and the impact factor of the journal that published it. Some fields also have their own writing style which may be apparent in their abstracts. For example, some medical journals use structured abstracts containing headings and summaries of the background, methods, results and conclusions. As both the field of research and the impact factor are largely journal level characteristics, we analyse the performance of papers in comparison to others published in the same journal, in order to control for these variables.

For each year and journal in our dataset, we rank all of the papers in terms of the number of citations received and transform these ranks into percentiles. We also convert the length of each abstract into percentiles in the same fashion. Relative to papers published by the same journal, we find a weak relationship where papers with longer abstracts tend to receive slightly more citations (Kendall's $\tau = 0.01$, $N = 216,279$, $p < 0.001$). However, when we repeat this analysis for each year, we find that the relationship disappears (for all years, all Kendall's $|\tau|$s ≤ 0.01, all Ns $\geq 21{,}419$, all ps > 0.05, FDR corrected).

Our findings do not qualitatively change when measuring the length of each abstract in terms of the number of sentences or number of words (see Supplementary Information).

3.2. Word frequency

In the previous section, we examined the relationship between the length of a paper's abstract and the number of citations it receives. Inspired by the idea that more frequently used words might require a lower cognitive load to understand, we investigate whether papers with more common words in their abstract receive more citations.

We use the database of English language books indexed by *Google's Ngram* project (Google Ngram Viewer., 2012) to calculate word frequencies. For each word in a papers abstract, we determine how frequently the word occurred within English language books that were published during the same year as the paper, and divide this count by the total word count for this corpus.

As before, we rank all of the papers in terms of the number of citations received and transform these ranks into percentiles. For each journal, we then calculate the mean percentile of the citation count distribution. We also calculate the median word frequency across all the abstracts published in each journal. We find a weak positive relationship between the median word frequency of the abstracts of papers published in each journal and the mean of their citation count distribution (Kendall's

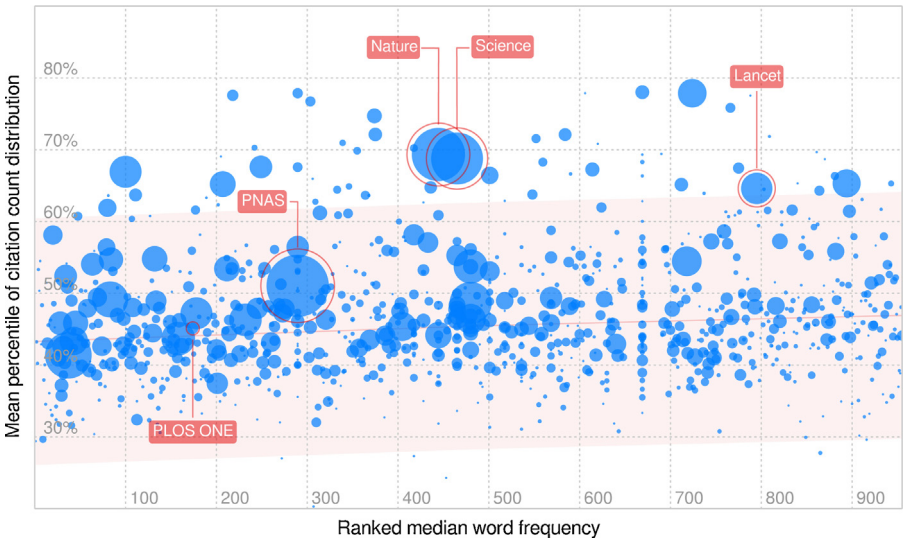

Fig. 2. Journals which publish papers with more frequently used words in their abstracts receive slightly more citations per paper. We define the frequency of a word as the percentage of times it appears in the English language books indexed by *Google*. For each journal, we calculate the median word frequency across all the abstracts in our dataset published by that journal. As in Fig. 1, for each year, we rank the papers by their citations and transform these ranks into percentiles. For each journal, we then calculate the mean percentile of the citation count distribution. Here, each blue circle represents a journal and the size of each circle represents the number of papers in our dataset published in that journal. We find a weak positive relationship between the median word frequency of the abstracts of papers published in each journal and the mean of their citation count distribution (Kendall's $\tau = 0.09$, $N = 955$, $p < 0.001$). Our finding suggests that, in our sample, journals which publish papers with more frequently used words in their abstracts tend to receive more citations. (For interpretation of reference to color in this figure legend, the reader is referred to the web version of this article.)

$\tau = 0.09$, $N = 955$, $p < 0.001$; Fig. 2). This finding suggests that, in our sample, journals which publish abstracts with more frequently used words tend to receive slightly more citations per paper.

We check whether this relationship holds for yearly subsets of our datasets. For each year in our sample, we calculate the median number of citations and median word frequency across the abstracts of papers published by each journal. We find that in most years, journals which publish papers with more frequently used words in their abstract tend to receive slightly more citations per paper (for all years all τs >0.03, all Ns $\geq$489, for all years except 1999, 2001 and 2004 all ps <0.05, FDR corrected Kendall's Tau, Fig. S6).

We conduct a similar correlation analysis to see if papers that have more frequently used words in their abstract tend to receive more citations, irrespective of the journal in which they are published. Again, we group the papers by year and journal before transforming the citations and median word frequencies to percentiles. This removes any temporal effects, and journal driven effects such as field and impact factor. Relative to papers published by the same journal during the same year, we find that papers whose abstracts contain more frequently used words tend to receive slightly more citations (Kendall's $\tau = 0.03$, $N = 216{,}279$, $p < 0.001$, Fig. S7). We find that this relationship holds for individual years within our dataset (for all years, all Kendall's τs > 0.01, all Ns $\geq$ 21,419, all ps < 0.01, FDR corrected, Fig. S8).

3.3. Mixed effects model parallel analysis

Our correlation analysis in the previous two sections revealed that papers whose abstracts contain more frequently used words tend to receive slightly more citations. The analysis also revealed that journals which publish papers whose abstracts are shorter and contain more frequently used words tend to receive more citations. We complement this analysis by building mixed effects models of the relationship between abstract characteristics and the median number of citations for papers published in each journal, as well as the number of citations for each individual paper.

3.3.1. Journal regression model

We start by fitting a linear regression model to the median number of citations received by papers published in each journal. The independent variables are the median abstract length and median word frequency across all papers published by each journal. We define our model as:

$$\log_{10}(c_{y,j}) = I + I_y + Ll_{y,j} + Ff_{y,j} + \epsilon_{y,j} \quad (1)$$

where $c_{y,j}$ is the median number of citations received by papers published in journal j and year y. The grand intercept is I while I_y is an intercept for each year. There is a slope L for the median number of characters in abstract $l_{y,j}$ published in journal j and during year y. There is also a slope F for the median word frequency $f_{y,j}$ of all the abstracts of papers published in journal j and during year y. We assume that the error term $\epsilon_{y,j}$ is Gaussian.

We first fit only the intercepts as a base model. We then include the two slopes for a full model. We perform an analysis of variance between these two models and find that the median abstract length and word frequency significantly contribute to the model (F statistic = 145.14, df = 2, $p < 0.001$). We find that journals which publish papers with shorter abstracts tend to receive more citations ($t = -16.27$, df = 5120, $p < 0.001$, t-test of the slope L). We also find that journals which publish papers with more frequently used words tend to receive more citations ($t = 5.63$, df = 5120, $p < 0.001$, t-test of the slope F).

We verify whether these findings hold for annual subsets of our data. For each year, we fit the following simple linear regression model:

$$\log_{10}(c_j) = I + Ll_j + Ff_j + \epsilon_j \tag{2}$$

We find that journals which publish papers with shorter abstracts tend to receive more citations (for all years, all ts $<= -4.23$, all df ≥ 486, all ps < 0.001, FDR corrected, t-test on slopes L; Figs. S9A and S9B). However, for half of the ten years, we do not find evidence that journals which publish papers with more frequently used words tend to receive more citations per paper (for 1999, 2002, 2003, 2005, 2008 all ts $>= 2.10$, all df ≥ 504, all ps < 0.05; for all other years all ts < 2, all df ≥ 486, all ps > 0.05; FDR corrected, t-test on slope F; Figs. S9C and S9D).

3.3.2. Paper mixed effects model

We fit a mixed effects model to the number of citations a paper receives, controlling for the journal in which each paper is published. The independent variables are both the total number of characters and median frequency of words in each paper's abstract. We define our model as:

$$\log_{10}(c_{y,j,p}) = I + I_y + I_{y,j} + Aa_p + (L + L_j)l_{y,j,p} + (F + F_j)f_{y,j,p} + \epsilon_{y,j,p} \tag{3}$$

where $c_{y,j,p}$ is the number of citations received by paper p published in journal j and during year y. The grand intercept is I while I_y is a random intercept for each year and $I_{y,j}$ is a random intercept for each journal nested in each year. The random intercepts are modelled as Gaussian random variables with a mean of zero.

There is a fixed slope L for the number of characters in the abstract and a fixed slope F for the median frequency of words in the abstract. To control for any effects of the field or impact factor of each journal, we consider the journal as a random effect which may affect the slope for abstract length L_j and the slope for the word frequency F_j. We also include a fixed slope A for the number of authors of each paper a_p. We assume the error term $\epsilon_{y,j,p}$ is Gaussian and that the random effects slopes are Gaussian with a mean of 0. We fit the model using maximum likelihood.

We first fit only the intercepts and the number of authors as a base model. We then include the length and word frequency slopes for a full model. We perform an analysis of variance between these two models and find that the abstract length and word frequency significantly contribute to the model ($\chi^2 = 1476.4$, df = 5, $p < 0.001$). We find that papers with shorter abstracts tend to receive more citations ($L = -0.0071$; $t = -5.45$, df = 692, $p < 0.001$, t-test of slope L, Satterthwaite approximation to degrees of freedom). We also find that papers with more frequently used words tend to receive more citations ($F = 0.008$; $t = 9.184$, df = 427, $p < 0.001$, t-test of slope F, Satterthwaite approximation to degrees of freedom).

We find that the average abstract contains words that occur 4.5 times per million words in the *Google Ngram* dataset. According to our model, doubling the median word frequency of an average abstract to 9 times per million words will increase the number of times it is cited by approximately 0.74%. If the average English word is approximately 5 letters long, then removing a word from an abstract increases the number of citations by 0.02% according to our model.

We verify whether these findings hold for annual subsets of our data. For each year, we fit the following model:

$$\log_{10}(c_{j,p}) = I + I_j + Aa_p + (L + L_j)l_{j,p} + (F + F_j)f_{j,p} + \epsilon_{j,p} \tag{4}$$

We do not find a relationship between the length of a paper's abstract and the number of citations it received (for all years except 2000, all ts, $-3 < t < 0$, all df > 204, all ps > 0.05, for 2000 $t = -3.62$, $p < 0.001$, FDR corrected, t-tests of abstract length slopes, Satterthwaite approximation to degrees of freedom; Figs. S10A and S10B). We do find that papers whose abstracts contain more frequently used words tend to receive more citations (all years except 2001 and 2004, all ts $>= 3.39$, all df > 121, all ps < 0.01, FDR corrected, t-tests of word frequency slopes, Satterthwaite approximation to degrees of freedom; Figs. S10C and S10D). We depict the variation of these effects across years in Fig. 3.

This complementary analysis using mixed effects models confirms the results found using the correlation analysis. Both analyses find that papers whose abstracts are shorter and contain more frequently used words tend to receive slightly more citations, although the relationship between abstract length and citations received is not found when evaluating yearly subsets of the data. Similarly, journals which publish papers whose abstracts are shorter and contain more frequently used words tend to receive more citations.

4. Discussion

In this paper, we investigate whether the length of paper abstracts and the frequency of the words contained in the abstracts might explain some of the variance in the number of times a paper has been cited. Our analysis considers the 30,000 most highly cited papers in each of the years 1999 to 2008, representing a sample size between 1.04% in 1999 and 0.73% in 2008.

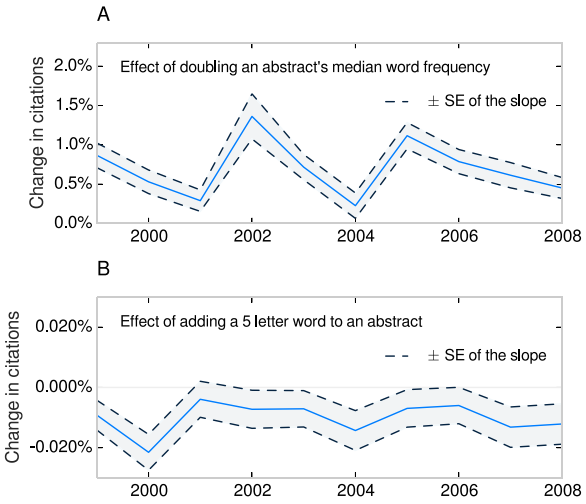

Fig. 3. Effect of the median word frequency and length of a paper's abstract on the number of times it is cited. (A) We fit a mixed effects model to estimate the number of times a paper will been cited given its abstract's median word frequency and length. Here, we show the expected increase in citations if we double the median word frequency of an average abstract. When fitting a model across all years, we find that the number of citations increase by 0.70%. (B) We also calculate the change in citations if 6 letters are added to an abstract. We find that the number of citations decreases by 0.02% for each 6 letters added to an abstract.

We find a relationship between the frequency of words found in the paper's abstract and the number of times a paper has been cited. In our sample, journals which publish papers whose abstracts contain more frequently used words tend to receive more citations per paper. In addition, papers whose abstracts use more frequently used words tend to receive more citations relative to the other papers published in the same journal. Our model indicates that doubling the median word frequency of an average abstract would result in 0.74% more citations.

There is also evidence of a relationship between the length of a paper's abstract and the number of times a paper has been cited. Our analyses suggest that journals which publish papers with shorter abstracts tend to receive slightly more citations per paper. Papers with shorter abstracts also tend to receive more citations relative to other papers published in the same journal, where adding a 5 letter word decreases the number of citations by 0.02%. However, a mixed effects analysis fails to find evidence of this relationship when evaluating yearly subsets of the data.

Papers in some fields tend to receive more citations than papers in other fields. For example, papers in cell biology tend to receive many more citations than papers in mathematics. At the same time, different research fields tend to have their own writing conventions. For example, many journals in medicine ask authors to provide abstracts with headings signposting their structure, such as "Method", "Results" and "Conclusions". As such, relationships between features of an abstract and citation counts at a journal level may be driven by differences in practices between fields. However, when analysing the performance of individual papers, our regression model factors out the effect of individual journals, such that any effect of differences between fields is also accounted for.

We propose three possible explanations of why papers with shorter abstracts or more frequently used words may gain more citations. High impact journals might restrict the length of their papers' abstracts and require writing suitable for a wider audience. For example, abstracts in *Science* are restricted to 125 words. Similarly, papers reporting greater scientific advances might be written with shorter abstracts and contain less technical language. A third potential explanation is that shorter abstracts with more commonly used words may be easier to read and hence attract more citations.

A study by Weinberger et al. (2015) also considered the relationship between the frequency of words in an abstract and the number of times a paper was cited. Contrary to our results, Weinberger et al. (2015) found that abstracts which used less frequent words were cited more often. In this study, Weinberger et al. (2015) evaluated how simple a word was by determining whether or not it was included in a list of 2,954 words known as the Dale-Chall list of Easy Words. In our study, we calculate the frequency of a word using *Google's Ngram* database, which provides information on how frequently words appeared in a very large corpus of English language books. This allows us to account for almost every single word used. For example, the Dale-Chall list contains the word "honeymoon", but not the word "linear". Data from *Google Ngram* however suggests that the word "linear" is used up to 20 times more than the word "honeymoon". We also calculate the frequency of each word in an abstract using *Googles* data from the same year in which the paper was published. This allows us to account for changes in language through time.

Weinberger et al. (2015) also find that papers with longer abstracts are cited more frequently, in contrast to our results. One possible explanation for the difference in our findings may be that our dataset contains the 30,000 most frequently cited papers in each year across all disciplines across a 10 year period, whereas Weinberger et al. (2015) analyse all articles published within only eight specific disciplines, between 1996 and 2012. Further analyses could investigate whether the relationship between the number of citations received and abstract length may differ for papers with few citations.

Overall, our results suggest that the style in which a paper's abstract is written may relate to the number of times the paper is cited. Future analyses may investigate how the findings we report here may fit into a broader model of the relationship between citation counts and various characteristics of scientific articles, such as title length, article length, use of non-alphanumeric characters, the number of keywords, the number of references in the paper, the number of affiliated institutions and the reputation of the authors who write the papers (Buter & van Raan, 2011; Falagas et al., 2013; Letchford et al., 2015; Petersen et al., 2014; Weinberger et al., 2015).

Competing interests

The authors declare no competing financial interests.

Author contributions

Conceived and designed the analysis: Adrian Letchford, Tobias Preis, Helen Susannah Moat.
Collected the data: Adrian Letchford.
Contributed data or analysis tools: Adrian Letchford, Tobias Preis, Helen Susannah Moat.
Performed the analysis: Adrian Letchford, Tobias Preis, Helen Susannah Moat.
Wrote the paper: Adrian Letchford, Tobias Preis, Helen Susannah Moat.

Acknowledgements

The authors acknowledge the support of Research Councils UK Digital Economy via grant EP/K039830/1.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joi.2015.11.001.

References

Google Ngram Viewer. URL: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html. 2012.
Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Future impact: Predicting scientific success. *Nature*, *489*, 201–202.
Alanyali, M., Moat, H. S., & Preis, T. (2013). Quantifying the relationship between financial news and the stock market. *Scientific Reports*, *3*, 3578.
Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.
Buter, R. K., & van Raan, A. F. J. (2011). Non-alphanumeric characters in titles of scientific publications: An analysis of their occurrence and correlation with citation impact. *Journal of Informetrics*, *5*(4), 608–617.
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. *European Physical Journal Data Science*, *1*, 8.
Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertész, J., Loreto, V., Moat, S., Nadal, J. P., Sanchez, A., Nowak, A., Flache, A., San Miguel, M., & Helbing, D. (2012). Manifesto of computational social science. *European Physical Journal Special Topics*, *214*(1), 325–346.
Curme, C., Preis, T., Stanley, H. E., & Moat, H. S. (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(32), 11600–11605.
van Dijk, D., Manor, O., & Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology*, *24*(11), R516–R517.
Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., & Mavros, M. N. (2013). The Impact of Article Length on the Number of Future Citations: A Bibliometric Analysis of General Medicine Journals. *PLOS ONE*, *8*(2), e49476.
Gonçalves, B., Perra, N., & Vespignani, A. (2011). Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PLoS ONE*, *6*(8), e22656.
Hartley, J. (2005). To attract or to inform: what are titles for? *Journal of Technical Writing and Communication*, *35*(2), 203–123
Hartley, J. (2007). Planning that title: Practices and preferences for titles with colons in academic articles. *Library & Information Science Research*, *29*(4), 553–568.
Hirsch, J. E. (2007). Does the H index have predictive power? *Proceedings of the National Academy of Sciences of the United States of America*, *104*(49), 19193–19198.
Jacques, T. S., & Sebire, N. J. (2009). The impact of article titles on citation hits: an analysis of general and specialist medical journals. *Journal of the Royal Society Medicine Short Reports*, *1*(2), 1–5.
Jamali, H. R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics*, *88*(2), 653–661.
King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science*, *331*(6018), 719–721.
Kristoufek, L. (2013). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, *3*, 3415.
Laurance, W. F., Useche, D. C., Laurance, S. G., & Bradshaw, C. J. A. (2013). Predicting Publication Success for Biologists. *BioScience*, *63*(10), 817–823.
Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Alstyne, M. V. (2009). Computational Social Science. *Science*, *323*(5915), 721–723.
Letchford, A., Moat, H. S., & Preis, T. (2015). The advantage of short paper titles. *Royal Society Open Science*, *2*(8), 150266.
Lewison, G., & Hartley, J. (2005). What's in a title? Numbers of words and the presence of colons. *Scientometrics*, *63*(2), 341–356.
Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLOS ONE*, *8*(8), e71226.
Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182.
Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports*, *3*, 1801.
Moat, H. S., Preis, T., Olivola, C. Y., Liu, C., & Chater, N. (2014). Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences*, *37*(1), 92–93.

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*(4), e61981.
New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*(1), 45–52.
Penner, O., Pan, R. K., Petersen, A. M., Kaski, K., & Fortunato, S. (2013). On the predictability of future impact in science. *Scientific Reports*, *3*, 3052.
Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H. E., & Pammolli, F. (2014). Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(43), 15316–15321.
Petersen, A. M., & Penner, O. (2014). Inequality and cumulative advantage in science careers: a case study of high-impact journals. *European Physical Journal Data Science*, *3*(1), 24.
Petersen, A. M., Stanley, H. E., & Succi, S. (2011). Statistical regularities in the rank-citation profile of scientists. *Scientific Reports*, *1*, 181.
Petersen, A. M., & Succi, S. (2013). The Z-index: A geometric representation of productivity and impact which accounts for information in the entire rank-citation profile. *Journal of Informetrics*, *7*(4), 823–832.
Petersen, A. M., Tenenbaum, J., Havlin, S., & Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*, *2*, 313.
Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, H. E., & Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports*, *2*, 943.
Petersen, A. M., Wang, F., & Stanley, H. E. (2010). Methods for measuring the citations and productivity of scientists across time and discipline. *Physical Review E*, *81*(3), 036114.
Preis, T., & Moat, H. S. (2014). Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*, *1*, 140095.
Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google trends. *Scientific Reports*, *3*, 1684.
Preis, T., Moat, H. S., Stanley, H. E., & Bishop, S. R. (2012). Quantifying the advantage of looking forward. *Scientific Reports*, *2*, 350.
Soler, V. (2007). Writing titles in science: An exploratory study. *English for Specific Purposes*, *26*(1), 90–102.
Wang, D., Song, C., & Barabasi, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*, 127–132.
Watts, D. J. (2007). A twenty-first century science. *Nature*, *445*, 489.
Weinberger, C. J., Evans, J. A., & Allesina, S. (2015). Ten simple (empirical) rules for writing science. *PLOS Computational Biology*, *11*(4), e1004205.
Whaley, C. P. (1978). Wordnonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*(2), 143–154.
Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in wikipedia. *PLoS ONE*, *7*(6), e38869.
Yogatama, D., Heilman, M., O'Connor, B., Dyer, C., Routledge, B. R., & Smith, N. A. (2011). Predicting a scientific community's response to an article. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 594–604).