Journal of Informetrics 10 (2016) 471–484

Contents lists available at ScienceDirect

Journal of Informetrics

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

A data analytic approach to quantifying scientific impact

Xuanyu Cao*, Yan Chen, K.J. Ray Liu

Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 9 September 2015
Received in revised form 22 February 2016
Accepted 22 February 2016

Keywords:
Citation prediction
Data analytic approach

ABSTRACT

Citation is perhaps the mostly used metric to evaluate the scientific impact of papers. Various measures of the scientific impact of researchers and journals rely heavily on the citations of papers. Furthermore, in many practical applications, people may need to know not only the current citations of a paper, but also a prediction of its future citations. However, the complex heterogeneous temporal patterns of the citation dynamics make the predictions of future citations rather difficult. The existing state-of-the-art approaches used parametric methods that require long period of data and have poor performance on some scientific disciplines. In this paper, we present a simple yet effective and robust data analytic method to predict future citations of papers from a variety of disciplines. With rather short-term (e.g., 3 years after the paper is published) citation data, the proposed approach can give accurate estimate of future citations, outperforming state-of-the-art prediction methods significantly. Extensive experiments confirm the robustness of the proposed approach across various journals of different disciplines.

1. Introduction

Citation is frequently used as a performance metric for quantifying the scientific impact of papers. In many applications, we need to know not only the current citations of papers, but also predictions of their future citations. In this section, we first motivate the study of citation prediction problem and briefly review the existing literature on citation prediction and citation distributions. Drawbacks and limitations of existing citation prediction approaches are discussed. Given these drawbacks, the principle of a novel data analytic citation prediction approach is introduced. Based on this principle, the proposed two data analytic prediction methods are epitomized.

1.1. Motivation and related works

Assessing the impact of a paper is a very important issue in academia. Besides the traditional paper awards and other subjective recognitions, an objective measure of the impact of a publication is highly desirable. It has been a trend that citation is perhaps the most often used metric to assess the scientific impact of a paper. A common argument is that the citations of a highly cited paper reflects its influence and contributions to the development of scientific advances. In fact, in addition to individual publications, citations have also been popularly used to assess the scientific impact of researchers and journals. Many popular impact measures, e.g., h-index and impact factor, rely directly on the citations of publications.

* Corresponding author.
E-mail address: apogne@umd.edu (X. Cao).

http://dx.doi.org/10.1016/j.joi.2016.02.006
1751-1577/© 2016 Elsevier Ltd. All rights reserved.

Despite its wide usage, the citation can only measure the current and past scientific impact of the papers, while in many scenarios people want to go beyond that to foresee the future scientific impact. Consequently, we need not only the current citations of a paper, but also a prediction of its future citations, which can reflect its future scientific impact. Unfortunately, the complex heterogeneous temporal patterns of citation dynamics make the prediction of future citations rather difficult. To make things even worse, in most of the meaningful practical applications, the task is often to predict the citations of those recently published papers, meaning that we need to make predictions based on a very short-term observation of the citation dynamics. In fact, people have already used such kind of citation prediction in practice. But their predictions are based on human heuristics (Waltman & Costas, 2014), which could be quite subjective and unreliable. All of these motivate us to study a critical yet challenging problem: can one predict the future citations of a paper based on a short-term observation of its citation dynamics?

Hitherto, many works have been devoted to the characterization of the citation distributions and fair comparison between papers from different disciplines (Eom & Fortunato, 2011; Haunschild & Bornmann, 2016; Peterson, Pressé, & Dill, 2010; Radicchi, Fortunato, & Castellano, 2008; Redner, 1998, 2005; Rodrí guez-Navarro, 2011; Schubert & Braun, 1996; Smolinsky, 2016; Stringer, Sales-Pardo, & Amaral, 2008; van Leeuwen & Moed, 2005; Zhang, 2013), yet few have considered the prediction of the future citations of individual papers. Bornmann, Leydesdorff, and Wang (2013, 2014) and Wang (2013) studied the correlation between the citation percentile of early years and that in the future and found a pessimistic result: the correlation is low. Hence, future citation prediction seems to be challenging.

In the literature, researchers have done works related to the citation prediction problem. Acuna, Allesina, and Kording (2012) predicted the future h-index of neuroscientists based on a variety of factors including number of articles written, current h-index, years since publishing the first article, number of distinct journals published in, and number of articles in several top journals. The method combined these factors with a linear regression model to predict future h-index. Furthermore, Ajiferuke and Famoye (2015) systematically studied the relations between the count response variables (e.g., numbers of citations, authors, references, views, downloads) by using several statistical models such as linear regression, lognormal regression, negative binomial regression. Hirsch (2007) compared several indicators of individual scientific achievement (e.g., h-index, total citations, citations per paper) on the task of predicting future scientific achievements. He found that h-index was the best indicator in predicting future achievements of individuals. However, Schreiber (2013) discovered that h-index was an inert indicator since it often severely depended on the growth of the citations of very old publications. This suggested that the real predictive power of h-index was limited. Petersen et al. (2014) measured the impact of authors' reputations on the future success of papers. Based on empirical observations, they argued that when a paper's current citation is low (e.g., at the early stage), the reputations of the authors are important in determining the future citations of the paper. However, if a paper's current citation is higher than a certain threshold, then the reputations of the authors are not important any more in determining the future success of the paper. Moreover, Breitzman and Thomas (2015) proposed to use the size of the inventor team to predict future citation of patents while Havemann and Larsen (2015) compared different bibliometric indicators' predictive powers on future success of young astrophysicists.

All the aforementioned existing literatures are related to the paper citation prediction problem, but none of them tackle it directly. Recently, Stegehuis, Litvak, and Waltman (2015) used the impact factor of the publishing journal and the first year citation count to predict the probability distribution of future citation of papers. However, the usage of only one single year's citation count would inevitably limit the accuracy of the prediction. Bornmann et al. (2014) made use of several other relevant factors (e.g., numbers of authors, pages, references) to predict the long-term citation percentile of papers. Yu, Yu, Li, and Wang (2014) exploited various features of papers (e.g., journal features, author features etc.) to predict the future citations with parametric regression models. But the experiments are confined to papers in the field of information science and library science. Wang, Song, and Barabási (2013) proposed a universal parametric model (hereafter the WSB model) for the temporal citation dynamics and used it to predict the future citations. The WSB model uses three parameters to characterize the citation dynamics as a function of time and explains the underlying mechanism dominating the citation process. The authors claimed that for any paper, by tuning these three parameters, the WSB model can always fit the citation dynamics well. When making predictions, given a period of citation dynamics data of a paper, the authors used it to estimate the three parameters and afterwards employed the trained WSB model to predict future citations.

However, this method has several limitations. First, since the model is parametric, the parameters need to be accurately estimated in order to make accurate predictions. To do so, they usually need a relatively long-term (usually at least 5 years, and the longer the better) observation of the citation dynamics to make meaningful predictions. If only a short-term observation (e.g., 3 years) is provided, their method does not work well, as will be shown in the later experimental results. However, as we previously mentioned, in many scenarios, the observation can be rather short-term. Hence, the usage of WSB model is limited in practice, as pointed out by Van Noorden (2013). Second, only limited experiments based on observations from high impact factor journals (e.g., *Science*, *Nature*) of fundamental sciences (e.g., chemistry, physics and biology) are conducted by Wang et al. (2013). Yet little is known about the performance on other journals such as engineering journals. Actually, according to our experiments, the WSB model performs much worse on papers in *IEEE*, which constitutes a popular journal database for electrical engineering and computer science research. Hence, the WSB model, though claimed to be universal, is not reliable for papers from different disciplines. Third, as pointed out by Wang, Mei, and Hicks (2014) and admitted by Wang, Song, Shen, and Barabási (2014), the WSB model may perform poorly on a few outliers due to severe overfitting, even with some regularization methods (Shen, Wang, Song, & Barabási, 2014). Though the outliers are minority and do not hurt the effectiveness of the WSB model too much, they somehow reduce the reliability of the prediction.

In all, in spite of being strongly desired by research community, there still lacks a reliable and robust method that can predict the future citations of individual papers accurately by using short-term citation data. This motivates us to study the challenging problem of predicting future citations from a completely different perspective in this paper.

1.2. Principle of data analytic approach

The drawbacks of the WSB model mainly come from the parametric nature of the model. It is very hard, if not unrealistic, to use a simple model with several parameters to characterize the complicated citation dynamics of all papers. In addition, with short-term citation data, even if the model fits the real data well, one may not estimate the parameters in the model accurately due to lack of statistics, leading to poor prediction of future citations. In this paper, we take an alternative data analytic approach. We observe that, given a short-term citation dynamics of a testing paper (paper A), if we find several existing papers whose early citation dynamics are similar to that of paper A, then their future citations match that of paper A well. If we build a database consisting of existing papers and use the short-term citation dynamics of a testing paper to match the (large number of) papers in the database, then we could predict the future citations of the testing paper by using the citation dynamics of those matched papers.

This suggests a data analytic citation prediction framework. The underlying principle is that, instead of using data from time domain to collect statistics which necessitates a long-term observation of the citation dynamics, we use a large number of existing papers to make predictions accurately. We conjecture that the complex patterns of the citation dynamics, which are hard to characterize using a simple parametric model, lie in the citation data of existing papers as long as there are sufficient enough of data – a basic principle of big data analytics. Actually, our data-analytic approach can also be motivated by the WSB model. The WSB model discovers that there is some universal temporal pattern of the citation dynamics satisfied by all papers. This suggests the usage of other papers' citation histories to predict the current paper's future citations.

To verify and validate the above principle, in this paper, two methods are proposed to make predictions based on those matched papers. The first one is to simply take the mean (average method) of the citation dynamics of the matched papers as the predictor of the future citations of the testing paper. Extensive experiments on real-world datasets from *Science*, *Nature*, *NEJM*, *PNAS*, *PRL* and *IEEE* show that this method is very effective, outperforming state-of-the-art WSB method (Wang et al., 2013). Specifically, with short-term (e.g., three years) citation dynamics data, the method is able to predict the future citations accurately. In addition, the method is also robust on datasets from various journals of different disciplines. In contrast, the WSB method fails on engineering journals like *IEEE*. Furthermore, the proposed method is also free of severe overfitting problem incurred by parametric methods such as the WSB method.

The second method is a model approach to further cluster the citation dynamics of the matched papers into several (e.g., 3) clusters using a Gaussian mixture model (GMM). Then, we use the centroids (i.e., the mean of each Gaussian component) as the predictors. Thus, given a testing paper, we can recommend several possible trends of the future citations along with their probabilities. The recommended trend closest to the ground-truth is generally very near to the ground-truth and, without surprise, it outperforms the average method since we are recommending more than one possible trends.

Overall, the proposed data analytic approach for citation prediction is simple but effective (high accuracy), robust (on any tested journal dataset) and timely (only needs short-term observation of the citation dynamics).

2. Data analytic approach

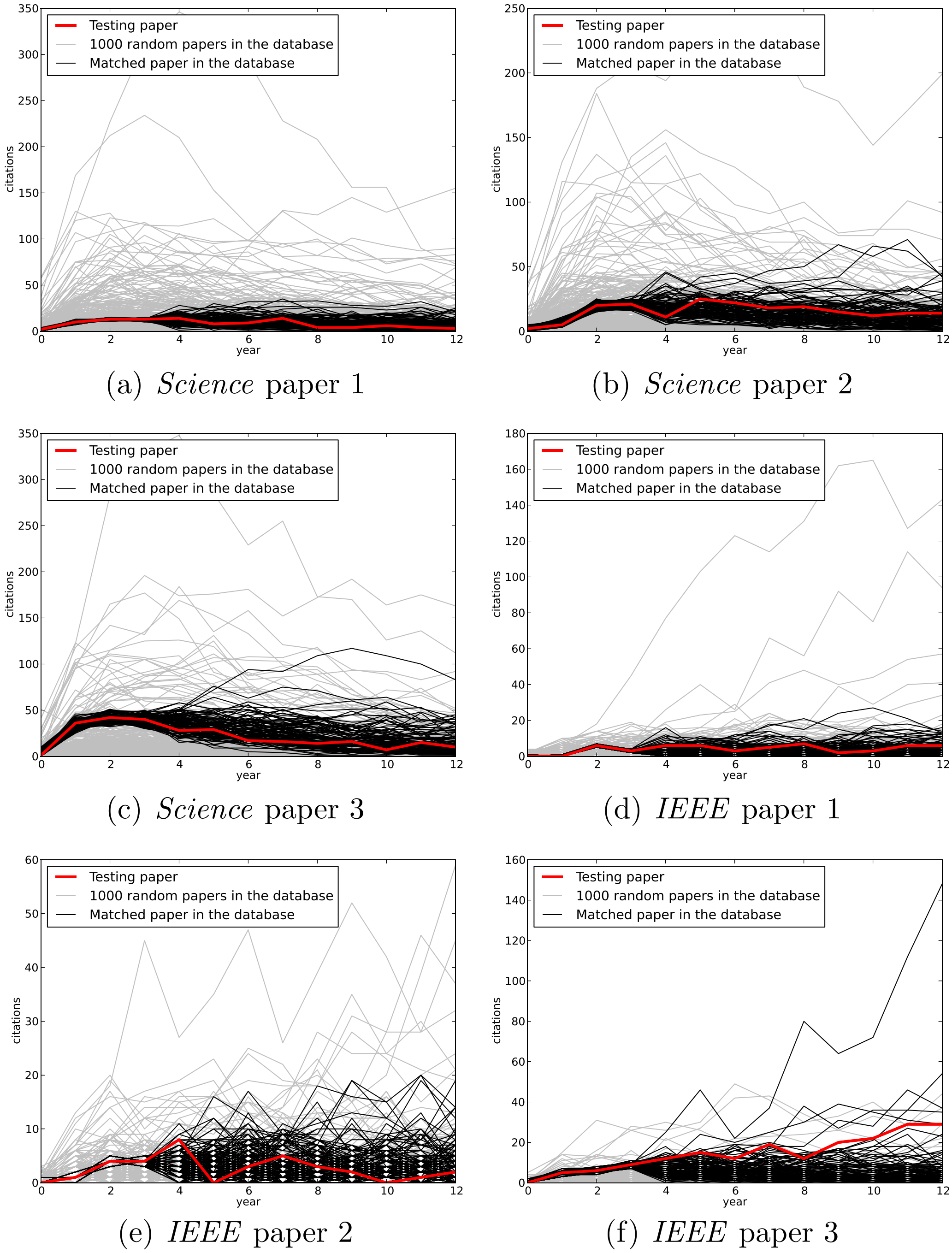
In this section, we first introduce the database used in this paper. Then, several diverting phenomena are observed from real-world citation data, which lead to a fundamental hypothesis of this paper. Based on this hypothesis, two data analytic approaches, namely AVR and GMM, are proposed and elaborated.

2.1. The database

The establishment of a large database is the most essential step of the proposed methodology. In this study, the citation data from the Web of Science of Thomson Reuters are chosen. The Web of Science database collects citations from a variety of sources with comprehensive coverage. Our dataset encompasses various kinds of journals, including general science journals, like *Science*, *Nature* and *Proceedings of the National Academy of Sciences (PNAS)*; journals from certain scientific disciplines, like *Cell*, *New England Journal of Medicine (NEJM)*, *Physical Review Letters (PRL)* and the entire *Physical Review* corpus; and journals from electrical engineering and computer science, like *Institute of Electrical and Electronics Engineers (IEEE)*. We download the citation data of all papers published by these journals from 1981 to 2001. The citation data of each individual paper consists of its citation counts in every year from its publication year to the year 2013.

2.2. Some observations

Let us start by asking a question: given a short period of citation dynamics, what is the uncertainty or predictability of the future citation dynamics? To answer this question, we randomly pick one paper (paper A) published in 2001. Given its citation dynamics from 2001 to 2004, we wish to predict its future citations from year 2005 to year 2013. From the dataset consisting of all the papers published in the same journal between 1982 and 1991, we select 100 papers whose first 4-year



(a) *Science* paper 1

(b) *Science* paper 2

(c) *Science* paper 3

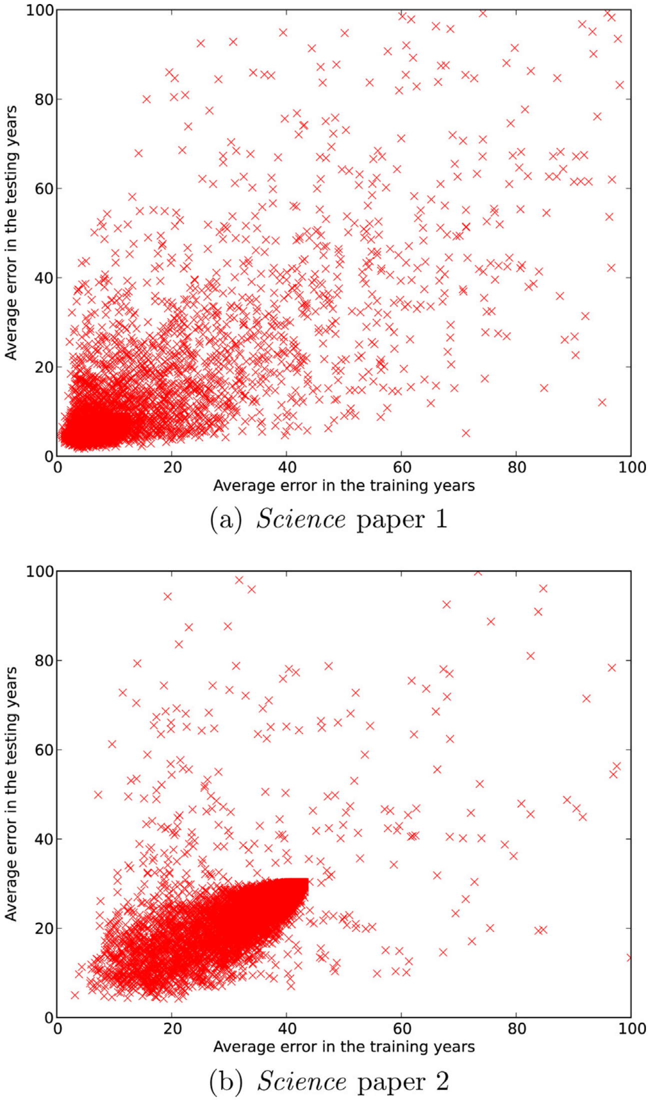
(d) *IEEE* paper 1

(e) *IEEE* paper 2

(f) *IEEE* paper 3

Fig. 1. Testing papers and their corresponding most matched 100 papers in the database. Three papers are published in *Science* 2001 and the other three are published in *IEEE* 2001. The citation dynamics of the matched papers are highly concentrated in a small region, in which the future citation of the testing paper, i.e., the ground-truth, generally lies in. This indicates the predictability of citations based on a data analytic approach by taking the citation histories of past papers into account.

(including the year that it is published) citation dynamics are closest to that of paper A. We choose six such paper A's, three from *Science* and three from *IEEE*, to conduct the above procedure. We plot the citation dynamics of those matched papers as the black curves and the citation dynamics of paper A as the red curve in Fig. 1. Without surprise, the first 4-year citation dynamics of the matched papers and paper A almost overlap, since we are matching these 4-year dynamics. But, more importantly, the following 9-year dynamics of the matched papers and paper A are also very close in most cases except an outlier in Fig. 1(f). This suggests strong predictability of future citations based on short-term observations of the citation dynamics and existing citation histories of the papers in the database.



(a) *Science* paper 1

(b) *Science* paper 2

Fig. 2. The predictive correlation between the citation dynamics during early years and those during future years. The larger the error during the training years, the larger the error during the testing years. This suggests us to match the early year citation dynamics of a testing paper with the papers in the database and then use the future citation dynamics of the matched papers to make predictions for the testing paper's future citations.

In the database, most of the papers do not match paper A well in the first 4-year dynamics. Another important question is: will the dynamics of these large number of papers coincide with that of A in the following 9-year? To see this, we randomly select 1000 papers in the corresponding datasets and plot their citation dynamics in gray in Fig. 1. We observe that these dynamics are quite diverse and do not match the dynamics of paper A well, which highlights the importance of the matching in the first 4-year.

2.3. *The hypothesis*

The proposed citation prediction method in this paper is based on the hypothesis that the future citation dynamics are correlated with the past citation dynamics. We validate this hypothesis in Fig. 2. For each subfigure we pick one paper, paper A, from *Science* 2001. Then for every *Science* paper published between 1981 and 1991, we calculate its average root mean square error (RMSE) deviation compared to paper A during the training years (the first 4 years) and the testing years (the later 9 years), respectively. The x-axis and y-axis correspond to the RMSE during the training years and the testing years respectively. Each point in the figure corresponds to one paper. As illustrated in the figure, the larger the error during the training years, the larger the error during the testing years. This suggests us to match the early year citation dynamics of a testing paper with the papers in the database and then use the mean behavior of the future citation dynamics of the matched papers to make predictions. In this way, we are essentially collecting accurate citation statistics from the database of a large number of past papers. Therefore, as we will demonstrate in the sequel, compared to the parametric WSB model (Wang et al., 2013), we can reduce the length of the training period to give a good prediction of the future citations.

2.4. The proposed methods

In this subsection, we describe the proposed two data analytic methods in detail. Suppose we are given the first N-year citation dynamics of a paper (paper A), x_1, x_2, . . ., x_N, where x_n is the citation counts during the year that is $n-1$ years after paper A is published. Note that x_1, the citation counts at the year that paper A got published, may be skewed by the publication month of the paper. We wish to predict the future M years' citation dynamics of paper A, i.e., to predict x_{N+1}, . . ., x_{N+M} by using the citation dynamics up to $N-1$ years after the publish year.

Assume that we have a database of past papers $\mathcal{D}$. Then, for every paper $y \in \mathcal{D}$, we calculate the matching error of the first N years citation dynamics between x and y as $e_x(y) = \sum_{n=1}^{N}(y_n - x_n)^2$. Evidently, the smaller the $e_x(y)$, the more similar dynamics between x and y. If the dynamics of x and y are very similar in the first N years, we expect the citation dynamics of y in the future years can be used to predict that of x.

Suppose we find L papers, $y^{(1)}$, $y^{(2)}$, . . ., $y^{(L)}$, from the database $\mathcal{D}$ with smallest matching error $e_x(y)$. Now we want to predict the citations of x based on the citation data of those matched papers $y^{(l)}$. To this end, we propose two methods of prediction.

1. *Average (*AVR*) method*: The first method is simply taking the average of the citation dynamics of the matched papers y_i as the predictor for the future citations of x. In other words, the prediction is $\hat{x}_p = (1/L)\sum_{l=1}^{L} y_p^{(l)}$, where $p = N+1, \ldots, N+M$.
2. *Gaussian mixture model (*GMM*) method*: Sometimes we wish to predict several possible trends of the future citations of a paper. To this end, we cluster the L matched citation dynamics $y^{(l)}$, $l = 1, 2, \ldots, L$, into K clusters by fitting a Gaussian mixture model (GMM) with K Gaussian components. Then, we predict the K Gaussian means of the K Gaussian components as the K possible trends of the paper's future citations. The weights of each Gaussian component can be regarded as the probability of that corresponding trend. In our experiment, we generally set $K = 3$ to make the predictions meaningful.

2.5. Relation with existing methods

Overall, our data analytic method first finds the most matched papers in the database based on the available citation observations and then makes predictions based on some mean behaviors of those matched papers. In fact, this framework is commonly used in the literature of non-parametric machine learning and various domain sciences. For instance, in the nearest-neighbor classification, for a testing data point, several training points (the matched examples) nearest to it are found and the corresponding labels are used to classify the given testing data point (Bishop, 2006). In computer vision, given an input image with a missing region, the known region is used to find matching scenes from a large-scale image database to complete the missing region of the input image (Hays & Efros, 2007). In the literature of computer-aided diagnose, a patient's medical images such as brain CT (computed tomography) or MR (magnetic resonance) are used to match the instances in a database consisting of many medical images with known diagnose information to guide diagnosis (Huang, Nielsen, Nelson, & Liu, 2005; Yuan, Tian, Zou, Bai, & You, 2011).

In fact, our approach falls into the general category of non-parametric prediction methods for time series (Ferraty & Vieu, 2006; Vilar-Fernández & Cao, 2007). In these methods, given a piece of time series, people use the average behavior of the past similar time series to predict the future dynamics of the given time series. In the literature, people propose different distance metrics to quantify the notion of similarity and use different weighting factors for averaging. Non-parametric forecasting has been successfully applied to several fields. Nikolov (2012) used non-parametric method to detect trending topics on Twitter in the early stage of the potentially popular Tweets. Scholz, Nielsen, and Sperlich (2012) exploited non-parametric methods to predict the stock prices.

3. Experiments

In this section, we conduct experiments to evaluate the performance of the proposed data analytic citation prediction methods. We first present the definitions of the two performance metrics used in the experiments as well as the experiment setup. Then, we show experimental results to confirm the advantages of the proposed methods over the WSB model (Wang et al., 2013), a state-of-the-art citation prediction method. Finally, we investigate the performance of the proposed methods in various scenarios.

3.1. Performance metric and experiment setup

For a paper x, suppose we want to use $x_1, x_2, \ldots, x_N$ (x_i denotes the citations during the $(i-1)$th year after being published) to predict the future M years' citation dynamics, i.e., $x_{N+1}, \ldots, x_{N+M}$. Denote the corresponding predictions as $\hat{x}_k$.

- Performance metric 1 (PM1): The relative prediction error of the total citations in the testing years. In other words,

$$PM1 = \left| 1 - \frac{\sum_{k=N+1}^{N+M} \hat{x}_k}{\sum_{k=N+1}^{N+M} x_k} \right| . \tag{1}$$

- Performance metric 2 (PM2): The average of the relative prediction error of the citations in each testing year. In other words,

$$PM2 = \frac{1}{M} \sum_{k=N+1}^{N+M} \left| 1 - \frac{\hat{x}_k}{x_k} \right| . \tag{2}$$

In the performance metric, if the denominator is zero, then we treat it as one. Basically, PM1 reflects the prediction error of the total citations while PM2 indicates the prediction error of each individual year. For instance, if the true total citation in the testing years is 100 and PM1 = 0.3, then the predicted total citation is 70 or 130. Furthermore, suppose PM2 = 0.4, then on average if the citation in one year is 10, the predicted citation in that year will be 6 or 14. Note that in general, we have PM1 < PM2 since PM2 takes the noisy vibration of the dynamics into account. When we evaluate the performance of the GMM method, we select one of the $K = 3$ predicted dynamics that is closest to the ground-truth to be the predictor. We also evaluate the performance of the WSB model (Wang et al., 2013) to serve as benchmark. In our experiment, the testing papers are all the papers published in year 2001 in a certain journal with total citations up to year 2013 at least 100. For each testing paper, we use its citation data up to 3 (5) years after published, i.e., $N = 4$ ($N = 6$), for prediction. The corresponding database consists of papers published in the same journal from year 1981 to year 1991 (for $N = 4$) or year 1993 (for $N = 6$). The reason of the chosen years is as follows. Suppose it is year 2004 now and we try to predict the future citations up to year 2014 of a paper published in 2001. Thus, we are predicting the future citations of a paper 13 years after it is published. The last year that we have citation data for 13 years after published is year 1991. So, we can only use citation data of papers published in or before 1991, which is the ending year of the database. As for the starting year of the papers in the database, generally a good choice is 10 years before the ending year, which is a good tradeoff between using too few data and using too out-dated data.

3.2. Comparison with the WSB model

In this subsection, we show experimental results to see the advantages of the proposed data analytic citation prediction methods over the WSB model. Before any experiment, we note that in practice, the available citation data, e.g., those on Google scholar and Web of Science, are yearly citation counts rather than the timestamps of every citation, which is a significantly larger amount of data. It is very costly, if not prohibitive, for individuals to obtain the detailed timestamps of every citation of their papers. Consequently, the input data to the citation prediction system is chosen to be the yearly citation counts of each paper. However, we notice that the input to the original WSB method (Wang et al., 2013) should be timestamps of every citation. Hence, we adapt the WSB method to the yearly citation count data by fitting the WSB model with the yearly citation dynamics, which is still referred to as the WSB method in this paper. But, one interesting issue is that whether we can improve the performance of the WSB method by feeding it with detailed timestamp of every citation? To answer this question, we compare the original WSB method based on detailed timestamp of every citation with our implementation of the WSB method based on yearly citation counts on papers published in *Nature* 1990. The results are shown in Fig. 3, from which we observe that the performances of the two versions of the WSB method are basically the same. Thereby, requiring detailed timestamps of every citation does not improve much, if any, over just using the yearly citation counts.

We first conduct experiments on papers in *Science* and *IEEE* with $N = 4$, i.e., using the citation data up to three years after publish. The distribution of the relative errors, i.e., the PM1 and PM2, of each paper is shown in Fig. 4. The relative errors are sorted in decreasing order. The inset of Fig. 4(a) is the 10% papers in *Science* with the worst performance, i.e., biggest relative errors. Similarly, the inset of Fig. 4(b) is the 25% papers in *IEEE* with the worst performance. As shown in the figure, when we use the WSB model for prediction, a minority of papers suffer from severe overfitting and thus have very bad performance, i.e., meaningless predictions with relative error much larger than 1. Actually, even when the training year is longer, this overfitting problem, though alleviated, still exists for a minority of papers. On the other hand, with our proposed AVR method, even those worst-behaved papers have reasonable performance. So, our method is more reliable. The above phenomenon also holds in other journals: about 10% papers with worst performance have relative error much larger than 1 when we use the WSB model for prediction. Consequently, in the rest of this paper, when evaluating the average relative errors, we always exclude the 10% papers with the worst performance in order to make the results of the WSB model meaningful.

Next, we conduct extensive experiments on various journals and the results are shown in Fig. 5. We note that the proposed data analytic approach always outperforms the parametric WSB model significantly. Particularly, when the observation of citation data is short-term ($N = 4$, i.e., using citation data up to 3 years after published), our AVR method already has PM1 near to 0.3, suggesting that the predicted total citations is accurate. In contrast, the PM1 of the WSB model is around 0.65, which

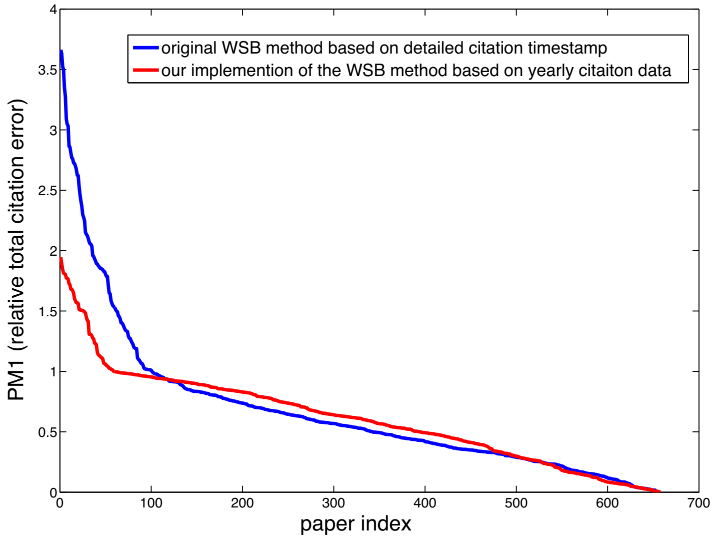


Fig. 3. Comparison between the original WSB method using the detailed timestamp of every citation and the our implementation of the WSB method using the yearly citation counts. The performances of the two versions of the WSB are basically the same.

is more than twice the errors of the AVR. In addition, the performance of the GMM method is better than the AVR method, indicating that if we are asked to recommend 3 possible trends of a paper's future citation dynamics, then at least one of our recommended trends will match the ground truth accurately (even for $N = 4$). If the length of the training years is longer, all the methods give better results and the comparisons above still hold. Among the journals considered in Fig. 5, *IEEE* is the unique engineering journal while all others are journals on fundamental science. For *IEEE*, the results of the WSB model are not shown when $N = 4$ since they are very bad (with performance metrics larger than 2, which can be regarded as meaningless predictions). Even when $N = 6$, the performance of the WSB model on *IEEE* papers is not good, as can be seen in Fig. 5(b). The reason may be that the 3-parameter WSB model cannot fit the citation dynamics of *IEEE* well. In other words, the WSB model, though effective in most of the journals on fundamental science like *Science*, does not work well for engineering journals like *IEEE*. As for our proposed GMM method and AVR method, the performance does become worse (the performance metrics increase 0.1 roughly) on *IEEE* papers but the prediction is still relatively accurate. This indicates the robustness of our proposed data analytic approach across different journals. Shen et al. (2014) proposed to add a Bayesian prior to the WSB model in order to enhance the performance by about ten percent on average. This performance improvement is much smaller than that of the proposed approach here, which is more than a half.

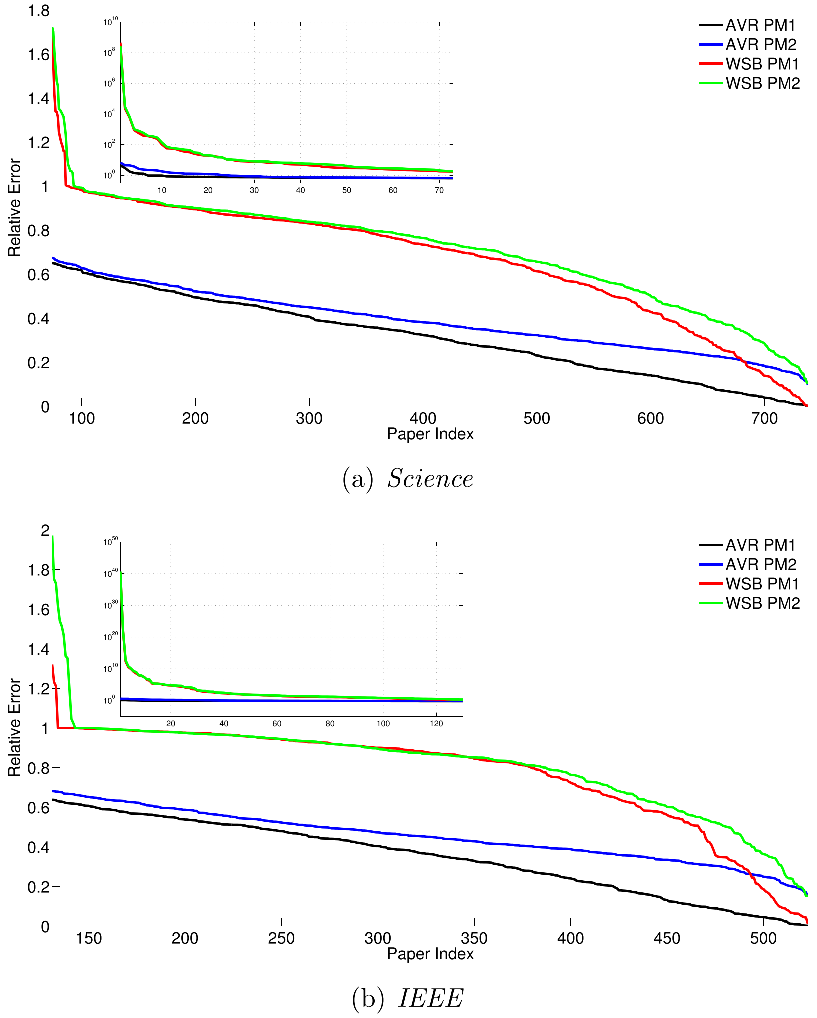
We also consider the scenario of predicting a paper's very long-term citations, e.g., more than 20 years after the paper is published. For this, with different length of training years, we predict the citations (up to year 2013) of the papers published in the entire *PR* corpus in the year 1990. The corresponding database is the *PR* corpus from year 1960 to year 1970 ($N = 4$), 1972 ($N = 6$) or 1977 ($N = 11$). The results are shown in Fig. 6. We do not consider PM2 here since the citations of most papers become near to zero after ten years since published, which makes the metric PM2 not meaningful. Additionally, we do not show the results of the WSB model when $N = 4$ because the performance is very bad and thus meaningless (with performance metrics much larger than 1). We observe that the proposed method again outperforms the WSB model a lot even in terms of long-term citation predictions.

3.3. Performance in various scenarios

In this section, we investigate the performance of the proposed citation prediction methods in various scenarios.

First, we evaluate the impact of the length of the training years on the performance of the proposed methods. For this end, we vary the length of training years from 2 years to 11 years, i.e., from $N = 2$ to $N = 11$. Note that when $N = 2$, we are essentially only using citation data up to 1 year after the paper got published. The performance of the proposed AVR method on papers published in *Science* 2001 is shown in Fig. 7. As the length of the training years increases, the performance gets better without surprise. The results for the GMM method and other journals are similar. An impressive observation is that even with $N = 2$, the PM1 of the proposed AVR method is about 0.38, which is already reasonably accurate.

We note that in the proposed citation prediction method, when matching papers in the database, the starting point of our matching is just the starting point of the citation dynamics, i.e., the publication year of the paper. For citation prediction, this matching methodology is reasonable and gives us accurate results. How about the starting point of the matching locates at some middle point of the dynamics? To test if the proposed method is still effective in such a circumstance, we perform another experiment, where the starting point of the matching is no longer constrained to be the publication year of the papers but can be any feasible middle point of the citation dynamics. The result is shown in Fig. 8. We observe that although



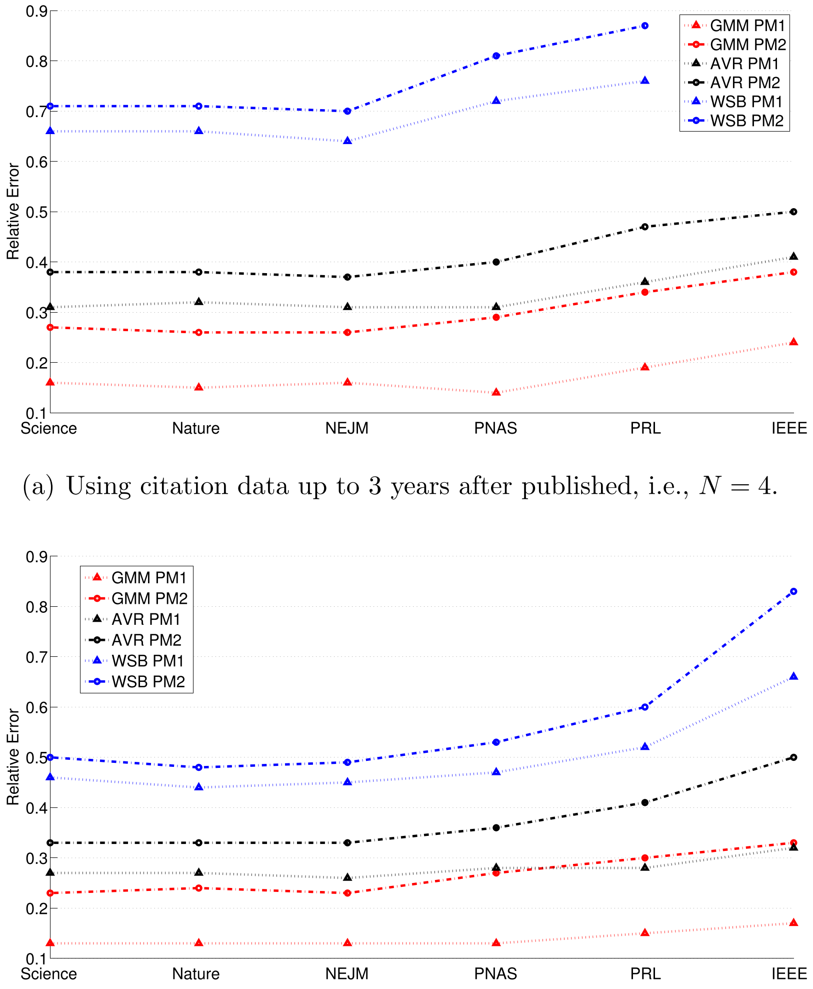
(a) *Science*

(b) *IEEE*

Fig. 4. The relative errors of the AVR method and the WSB model on papers in *Science* and *IEEE* when $N = 4$, using citation data up to three years after publish for prediction. The WSB model incurs severe overfitting problem on a minority of papers, which have very bad performance and meaningless predictions. Our proposed AVR method is free of such problem and thus more reliable.

the performance of this middle-point-matching version of the AVR method somewhat degrades compared to the originally proposed start-point-matching AVR method, it is still reasonably good, much better than the WSB model. So far, to predict the citations of papers of a journal, we always use the database consisting of papers from the same journal. Sometimes, we may run into difficulty because the number of papers published in that journal is too small. An example is the journal *Cell*. The number of papers published in *Cell* in each year of the 1990s is around 500, which may not be enough to construct a database for prediction. To resolve this issue, we use the database of *Science* and *Nature* to make predictions of papers published in *Cell* in 2001. The result is shown in Fig. 9. The result is very good (even better than that of *Science* and *Nature* and the reason may be that the citation dynamics of *Cell* are more regular and predictable) and the observations mentioned earlier still hold. This inspires us to see if making the database and testing papers to lie in the same journal is essential to our proposed data analytic approach. To this end, we use *Science* database to predict the citations of *Nature* papers and vice versa. The result is about the same as using *Nature* database to predict the citations of *Nature* papers, indicating the robustness of our approach. But, if we use *IEEE* database to predict the citations of *Science* papers, the performance becomes worse, and conversely, if we use *Science* database to predict the citations of *IEEE* papers, the performance also becomes a little worse. The reason may be that the temporal patterns of the citation dynamics of papers in *IEEE* are very different from that of *Science* and *Nature*. However, we notice that with such a performance degradation, the prediction is still accurate, indicating the robustness of our method when matching papers with very different citation dynamic patterns.

In practice, we may want to know the most influential papers in advance. We note that the task of identifying influential papers is difficult since, compared to normal papers, many highly influential papers follow quite different, even abnormal, temporal patterns in citation dynamics. We predict the top-50 papers published in year 2001 in each journal. Specifically, we predict the top-50 highly cited papers for total citations until year 2013, based on short-term observations, i.e., $N = 4$ (use



(a) Using citation data up to 3 years after published, i.e., $N = 4$.

(b) Using citation data up to 5 years after published, i.e., $N = 6$.

Fig. 5. The performance of the proposed data analytic approach on various journal datasets. The performance of the WSB model (Wang et al., 2013) serves as benchmark. Our approach always outperforms the WSB model. For the *IEEE* dataset, the WSB model fails and we do not plot its performance on *IEEE* when $N = 4$ because the performance is too bad (PM1 = 2.84, PM2 = 3.54) and thus out of the scale of this figure.

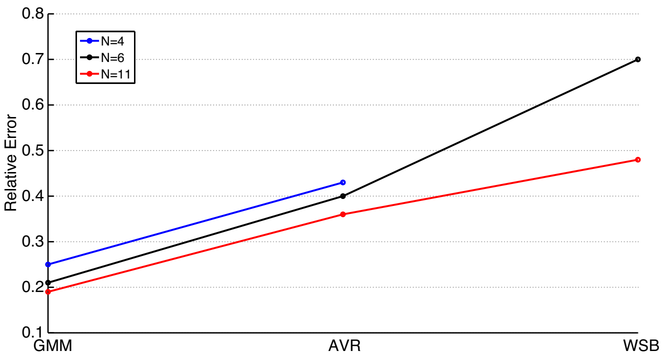


Fig. 6. Performance evaluation for papers in *PR* in 1990. We predict their citations up to year 2013, i.e., 23 years after they are published.

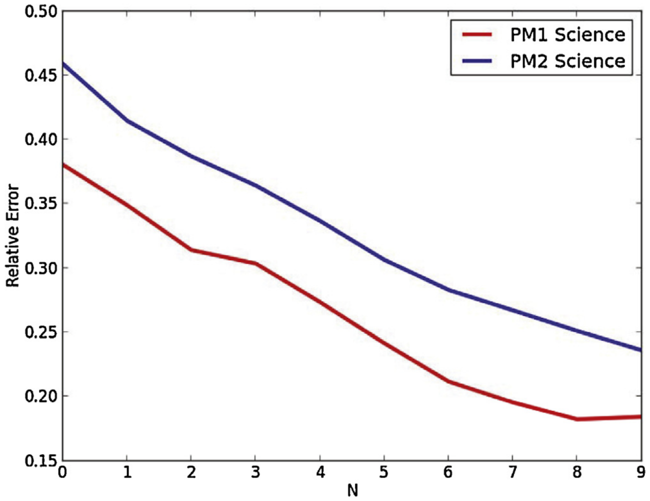


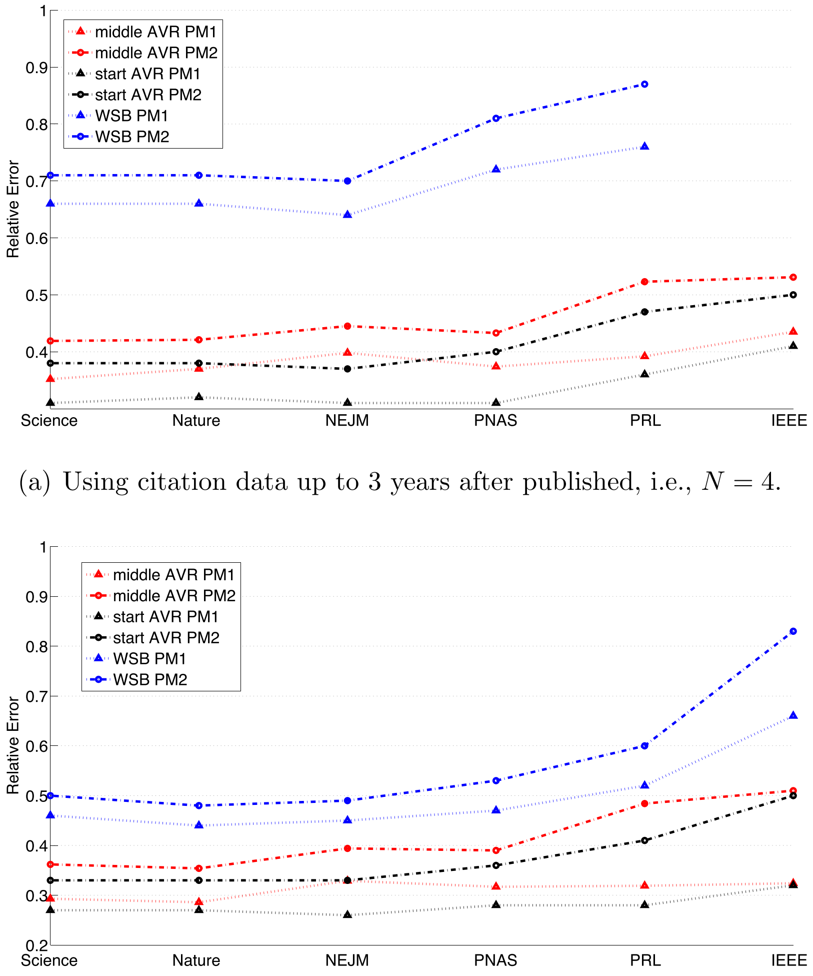
Fig. 7. Impact of the length of the training years on the performance of our approach. N is the length of the training years: $N=2$ means we are using citation data up to 1 year after the publish year.

citation data up to 3 years after published) and $N=6$ (use citation data up to 5 years after published). We predict 50 papers as the future top-50 highly cited papers and compute the proportion among the true top-50 papers that is predicted as top-50. The results, as shown in Fig. 10, are very good. Again, the *IEEE* papers are the most difficult ones to predict, in accordance with the previous experiments, due to the lack of regularity of the citation dynamics of *IEEE* papers.

4. Discussion

Through the previous experiments, the proposed data analytic citation prediction methods are validated to be effective. Several comments are given below.

1. Note that the proposed data analytic approach only uses two pieces of information of a paper to predict its future citations: (i) the early-year citation dynamics, (ii) the journal that the paper is published in. We use (ii) to select the correct database for matching and use (i) to match the papers in the database. The excellent performance of our method indicates that the early-year (e.g., 3 years after published) citation dynamics and the journal title already gives enough information to make accurate predictions. In other words, the future citations are largely determined once (i) and (ii) are given. This is also illustrated in Fig. 1, where the future citations of the matched papers are concentrated with low variance. If one takes the topic, author or other side information into account, instead of using an average of the future citations of the matched papers as predictor, one may give different weights to these matched papers by using the side information, which may improve the prediction accuracy. However, here we observe that even without these side information, the prediction is already reasonably accurate.
2. Our method uses paper database consisting of papers published in the past years (in most of the experiments, 10–20 years before the publication of the testing papers). The citations of the old papers are traditionally considered not good indicators of the citations of the new papers because the community size, citation customs and research directions have changed a lot. On the contrary, the excellent performance of our approach indicates that the time variance of the temporal patterns of the citation dynamics does not change too much. This means if two papers, published in 1985 and 2000 respectively, have similar citation dynamics within 3 years after being published, then with high probability, their future citation dynamics are also similar. However, we still note that time variance does exist. The reason is that if we extend the start year of the database to be even earlier, then the performance of our methods will begin degrading. This prevents us from extending the size of the database wildly.
3. Different journals may have similar or quite different citation dynamics patterns. Specifically, top scientific journals with high impact factors such as *Science*, *Nature* and *Cell* share similar citation dynamics patterns. The evidence is that if we use papers from one journal as database to predict the citations of papers from another journal, the performance is quite good. However, if we use papers from *IEEE* as database to predict the citations of *Science* papers, the performance degrades. This indicates that *IEEE*, an engineering journal, has quite different citation dynamics patterns.



(a) Using citation data up to 3 years after published, i.e., $N = 4$.

(b) Using citation data up to 5 years after published, i.e., $N = 6$.

Fig. 8. Confirming the effectiveness of using the middle piece of the citation dynamics for matching, which is highlighted by the red curve. We observe that though using the middle piece for matching degrades the performance of the proposed AVR method compared to the initial one (the black curve), it still appears to be effective with a much better performance than that of the WSB model (the blue curve). This indicates the potential application of our method to fields other than citation prediction, such as stock price prediction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

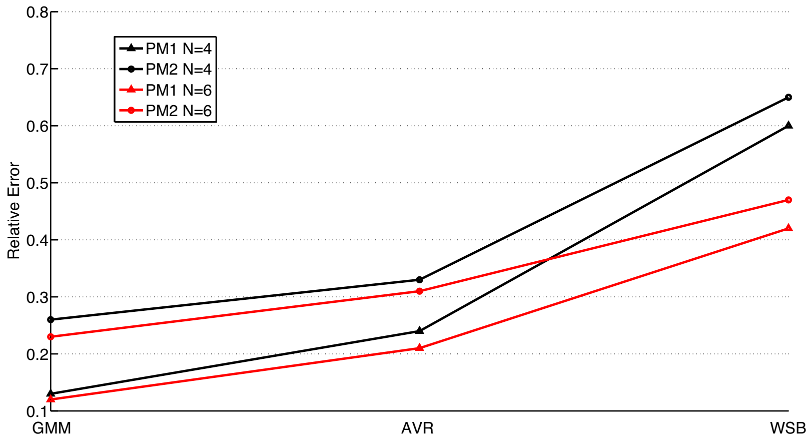


Fig. 9. Performance evaluation for papers in *Cell*. We use papers in *Science* and *Nature* as database to predict citations of papers in *Cell*.

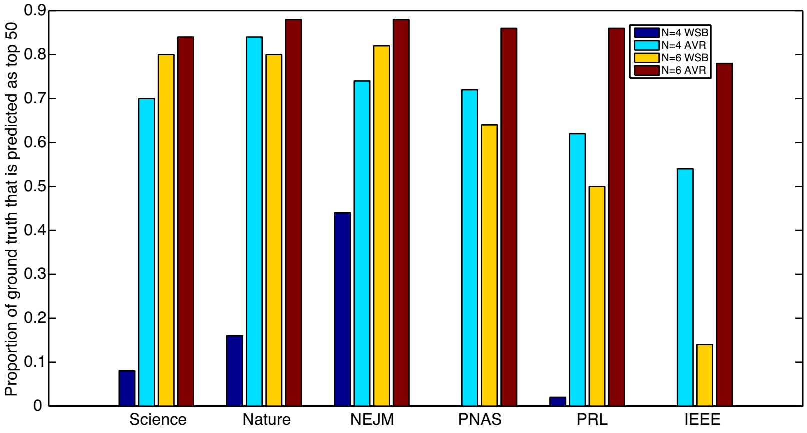


Fig. 10. Identifying top 50 highly cited papers in advance. *N* is the length of the training year. In the experiment, we predict 50 papers as the future top-50 highly cited papers based on short-term observations of their citation data. The vertical axis is the proportion among the true top-50 papers that is predicted as top-50. Two blue bars are missing because the corresponding quantities are zero.

5. Conclusion

In this paper, we propose a simple data analytic approach to predict the future citations of individual papers. The approach is able to predict citations based on short-term observations of the early citation data. The approach is accurate, robust and reliable across journals from different disciplines and outperforms the state-of-the-art WSB model in all experiments.

Author contributions

Conceived and designed the analysis: Xuanyu Cao; Yan Chen; K.J. Ray Liu.
Collected the data: Xuanyu Cao.
Contributed data or analysis tools: Xuanyu Cao.
Performed the analysis: Xuanyu Cao; Yan Chen; K.J. Ray Liu.
Wrote the paper: Xuanyu Cao; Yan Chen; K.J. Ray Liu.

References

Acuna, D. E., Allesina, S., & Kording, K. P. (2012). Future impact: Predicting scientific success? *Nature*, *489*(7415), 201–202.

Ajiferuke, I., & Famoye, F. (2015). Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models. *Journal of Informetrics*, *9*(3), 499–513.

Bishop, C. M. (2006). *Pattern recognition and machine learning.*

Bornmann, L., Leydesdorff, L., & Wang, J. (2013). Which percentile-based approach should be preferred for calculating normalized citation impact values? An empirical comparison of five approaches including a newly developed citation-rank approach (p100). *Journal of Informetrics*, *7*(4), 933–944.

Bornmann, L., Leydesdorff, L., & Wang, J. (2014). How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics*, *8*(1), 175–180.

Breitzman, A., & Thomas, P. (2015). Inventor team size as a predictor of the future citation impact of patents? *Scientometrics*, *103*(2), 631–647.

Eom, Y.-H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE*, *6*(9), e24926.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice*. Springer Science & Business Media.

Haunschild, R., & Bornmann, L. (2016). Normalization of Mendeley reader counts for impact assessment? *Journal of Informetrics*, *10*(1), 62–73.

Havemann, F., & Larsen, B. (2015). Bibliometric indicators of young authors in astrophysics: Can later stars be predicted? *Scientometrics*, *102*(2), 1413–1434.

Hays, J., & Efros, A. A. (2007). Scene completion using millions of photographs. *ACM Transactions on Graphics (TOG)*, *26*(3), 4.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 19193–19198.

Huang, H., Nielsen, J., Nelson, M. D., & Liu, L. (2005). Image-matching as a medical diagnostic support tool (DST) for brain diseases in children? *Computerized Medical Imaging and Graphics*, *29*(2), 195–202.

Nikolov, S. (2012). *Trend or no trend: A novel nonparametric method for classifying time series*. Twitter Inc (PhD thesis).

Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., et al. (2014). Reputation and impact in academic careers? *Proceedings of the National Academy of Sciences*, *111*(43), 15316–15321.

Peterson, G. J., Pressé, S., & Dill, K. A. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations? *Proceedings of the National Academy of Sciences*, *107*(37), 16023–16027.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact? *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B – Condensed Matter and Complex Systems*, *4*(2), 131–134.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*.

Rodrí guez-Navarro, A. (2011). A simple index for the high-citation tail of citation distribution to quantify research performance in countries and institutions. *PLoS ONE*, *6*(5), e20510.

Scholz, M., Nielsen, J. P., & Sperlich, S. (2012). *Nonparametric prediction of stock returns guided by prior knowledge. Technical report*. University of Graz, Department of Economics.

Schreiber, M. (2013). How relevant is the predictive power of the h-index? A case study of the time-dependent Hirsch index. *Journal of Informetrics*, *7*(2), 325–329.

Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators? *Scientometrics*, *36*(3), 311–324.

Shen, H.-W., Wang, D., Song, C., & Barabási, A.-L. (2014). *Modeling and predicting popularity dynamics via reinforced Poisson processes.* , arXiv preprint arXiv:1401.0778.

Smolinsky, L. (2016). Expected number of citations and the crown indicator? *Journal of Informetrics*, *10*(1), 43–47.

Stegehuis, C., Litvak, N., & Waltman, L. (2015). Predicting the long-term citation impact of recent publications? *Journal of Informetrics*, *9*(3), 642–657.

Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2008). Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*, *3*(2), e1683.

van Leeuwen, T. N., & Moed, H. F. (2005). Characteristics of journal impact factors: The effects of uncitedness and citation distribution on the understanding of journal impact factors? *Scientometrics*, *63*(2), 357–371.

Van Noorden, R. (2013). Formula predicts research papers' future citations. *News, Nature*, *3*.

Vilar-Fernández, J. M., & Cao, R. (2007). Nonparametric forecasting in time series – A comparative study? *Communications in Statistics – Simulation and Computation*, *36*(2), 311–334.

Waltman, L., & Costas, R. (2014). F1000 recommendations as a potential new data source for research evaluation: A comparison with citations? *Journal of the Association for Information Science and Technology*, *65*(3), 433–445.

Wang, J. (2013). Citation time window choice for research impact evaluation? *Scientometrics*, *94*(3), 851–872.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact? *Science*, *342*(6154), 127–132.

Wang, J., Mei, Y., & Hicks, D. (2014). Comment on "Quantifying long-term scientific impact"? *Science*, *345*(6193), 149.

Wang, D., Song, C., Shen, H.-W., & Barabási, A.-L. (2014). Response to comment on "Quantifying long-term scientific impact"? *Science*, *345*(6193), 149.

Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis? *Scientometrics*, *101*(2), 1233–1252.

Yuan, K., Tian, Z., Zou, J., Bai, Y., & You, Q. (2011). Brain ct image database building for computer-aided diagnosis using content-based image retrieval? *Information Processing & Management*, *47*(2), 176–185.

Zhang, C.-T. (2013). The h′-index, effectively improving the h-index based on the citation distribution. *PLOS ONE*, *8*(4), e59912.