ISSI 2023 Proceedings, v2, pp 441–447
DOI: 10.5281/zenodo.8429001 https://cns-iu.github.io/workshops/2023-07-02_issi/

Mind the gap: Nowcasting the citation impact of research institutions

Stephan Stahlschmidt[1] and Dimity Stephen[2]

[1]stahlschmidt@dzhw.eu, [2]stephen@dzhw.eu
German Centre for Higher Education Research and Science Studies, Schützenstraße 6a,
10117 Berlin, (Germany)

Abstract

Citation-based impact indicators are the foundation of bibliometric evaluation, but the impact reporting is substantially delayed by the required citation window of at least three years. This lag considerably hinders the application of bibliometrics for timely decision making. We propose to adopt a nowcasting approach to overcome this drawback and compute estimates of research institutions current and most recent impact before they become available at the end of the citation window. By uncovering the underlying dynamic communication structure via a temporal Bayesian network, we observe an informative relation between very early news mentions and subsequent Tweets and Mendeley readerships with the – delayed available – citation impact of the corresponding publications. These relations are applied to produce early nowcasts of the citation impact before the end of the citation window to facilitate timely bibliometric reporting. The current nowcast implementation of this work-in-progress reduces the error rate of the citation impact estimate by more than 25% if compared to baseline extrapolation model.

Introduction

Citation-based indicators are the foundation of quantitative bibliometric indicators. However, a substantial period of time is required for a publication to be ingested by the academic system and citing publications in turn to be produced before it accrues a level of citations representative of its long-term impact. Accounting for the need for up-to-date data, empirical studies suggest a period of three to five years is necessary (Wang, 2013). While essential for the reliability and validity of citation-based bibliometric indicators, this lag greatly delays impact assessments.

The introduction of altmetrics in recent years presents a potential opportunity to reduce this delay. Unlike citations, the majority of a publication's altmetric attention can occur within days or weeks of publication (e.g. Eysenbach, 2011). Harnessing this speed, it may be feasible to use attention on the diverse altmetrics channels to overcome the requisite citation window and provide up-to-date impact assessments, provided that the attention observed by altmetrics results informative for the citation impact.. Studies examining the correlations between citations and various altmetric sources have found stronger associations of up to $r = 0.76$, arguably because the users bookmarking or downloading academic papers strongly overlap with those citing them in later publications.

Previous bibliometric studies have sought to predict a publication's future number of citations using either univariate time series or multivariate models. Time series models predict future citations based on the citations already received at time t following publication, thus continuing the univariate time series (Wang et al., 2013), while multivariate models model a relationship between potential explanatory variables and future citations (Brody et al. 2006; Shibata et al., 2007; Fu and Aliferis, 2008; Eysenbach, 2011; Bornmann et al., 2014).

In the current study, we translate the technique of nowcasting established in macroeconomics and epidemiology (Evans, 2005) into bibliometrics. In contrast to forward-looking forecasting, nowcasting uses statistical models to generate estimates for the most recent past and the current status of a time series. Here, these estimates are based on correlations between the established but delayed citation indicators and currently available altmetrics data and used to nowcast the impact of German research institutions.

In contrast to previous studies mentioned above, we are not interested in individual publications, but in the publications of an aggregated unit, i.e. all publications of a research institution in a certain year. With this change from the micro level to the meso level perspective, any random inaccuracies on the micro publication level lose their importance. Secondly, we do not aim to estimate the publications' total number of citations. Instead, we attempt to bridge the 3-5 year citation window. This limited time constraint greatly reduces the complexity of the estimation problem, because estimates over a short time horizon, due to their proximity to the known present, leave less room for uncertainty than estimates for the distant future.

Thus, the aim of this study is to create a nowcasting model based on the complex communication model linking citations with altmetrics data to address the research question: can altmetrics data bridge the citation window to produce sufficiently accurate citation impact estimates for a timely bibliometric reporting on research institutions?

Data

In order to account for the dynamic in citation and altmetrics events accumulation, we collect time series data for more than 200 German research institutions from 2011 to 2019. Publications and citation data are sourced from the German Kompetenznetzwerk Bibliometrie in-house Web of Science database, which matches WoS-indexed publications to German research institutions (Rimmert et al., 2017). The corresponding Altmetric data, sourced from Altmetrics LLP, is identified via a DOI match and includes yearly counts of mentions on Twitter, in News, on blogs, on Facebook, on Wikipedia and readership counts on the Mendeley platform.

We compute the 10% excellence rate (PPtop10) indicator for German institutions with at least 100 publications/year incorporating a three-year citation window. As we like to bridge this window, we divide the data into overlapping time slices covering each three years, i.e. the publication years 2011 to 2017 and for each publication year the subsequent two years, as in the third year after the publication year the actual excellence rate can be computed. Obviously nowcasts are required before the third year.

In Figure 1 these time slices are normalized on the publication year t0={2011,…, 2017}, the year directly after the publication year (t1) and the second year after the publication year (t2) to observe the time dynamics irrespective of the actual publication year. Mendeley data is only available for the publication years 2015 to 2017.

Observing the Figure 1 three distinct patterns can be identified. Mendeley counts in the top left panel arise instantly in the publication year and surpass the publication counts by 80% on average, i.e. the number of Mendeley counts on the institution's publications is on average 80% higher than the publication counts themselves. However, the underlying distribution is skewed as the German research institutions find on average 35% of their publications without any readership indication on Mendeley in the citation window. In t1 the ratio raises on average to 10, i.e. there are 9 times more Mendeley counts than publications for an institution. This ratio decreases slightly to more than 8.5 on average in t2. As these counts are yearly counts, the accumulated surplus at t2 amounts to 0.8+9+7.5 = 17.3 more Mendeley mentions than publications on average if the publications are not removed from the Mendeley profiles.

In contrast the Wikipedia counts in the lower right corner show neither a noticeable uptake, nor any dynamics. On average for every 100 publications, we observed one mention on Wikipedia within the publication year. This low number does not vary over time and aligns with former observations on the particular nature of Wikipedia as an online encyclopaedia with an extensive selection discourse among editors (Schmidt et al., 2021).

Mentions in the Altmetric channels Twitter, News, Facebook and Blog exhibit the third pattern, where mentions primarily arise in the publication year and drop down rapidly in t1 before decreasing slightly more in t2. Twitter can be differentiated by having the most profound reaction, that is on average we count double the number of tweets compared to the number of publications of the German research institutions. However, the attention on Twitter decreases sharply afterwards with only 25 (12) additional tweets on average for every 100 publications in t1 (t2). As with Mendeley the attention on Twitter is not uniformly distributed on the publications, but the institutions find on average more than 60% of their publications not mentioned at all on Twitter in the citation window. In the more recent time frame 2015-2018 this share is reduced to around 50% showing a generally increasing uptake of Twitter.

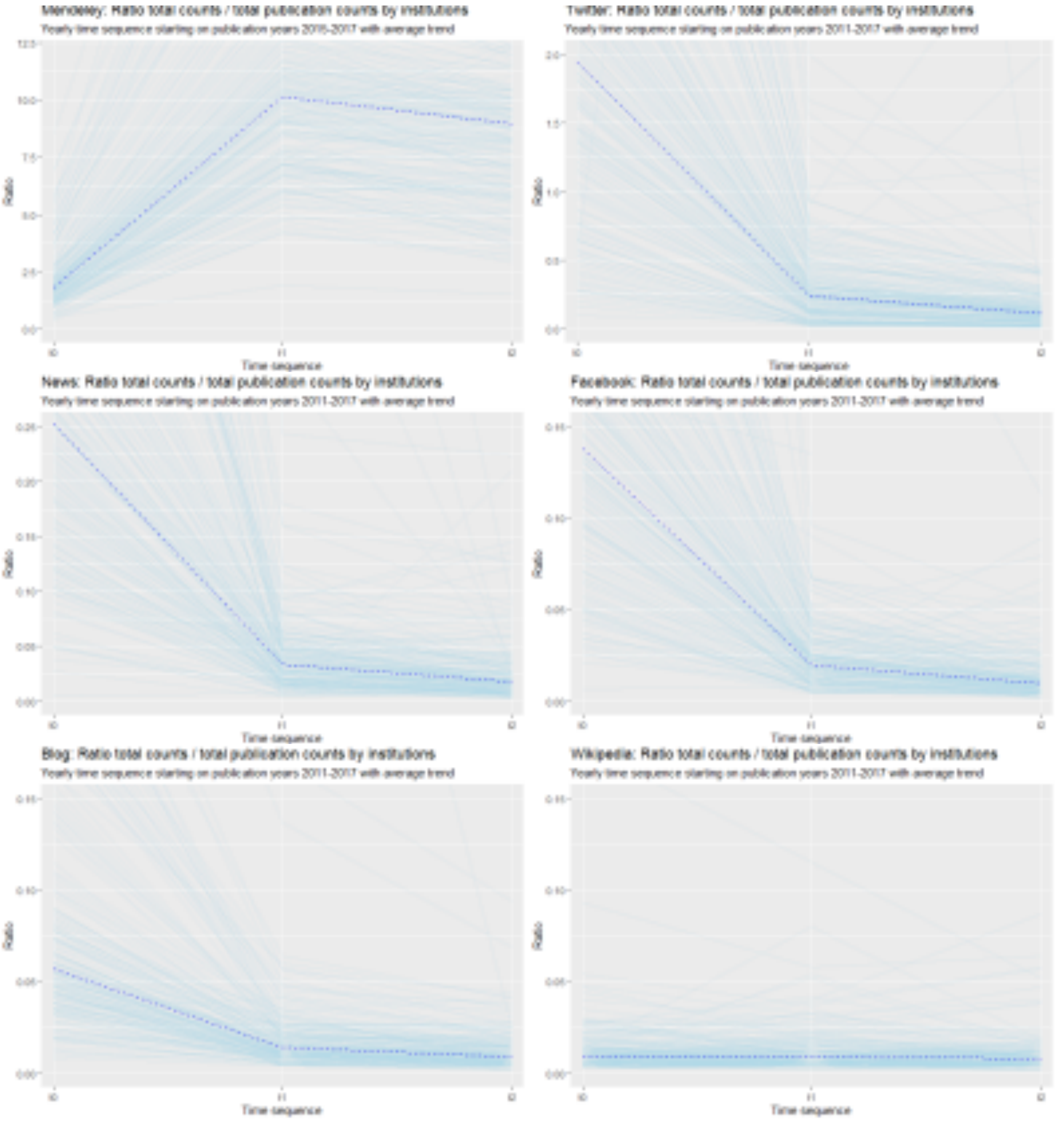

Figure 1. Time dynamics of Altmetrics channels in the publication year (t0) and the subsequent two years (t1, t2) represented as the ratio of total counts of mentions in Altmetrics channels on institution's publications divided by the number of publications at each point in time. Every solid line denotes a German institution, the dashed lines represent the average dynamics. The scale of the y-axis varies for visual readability.

Mentions in traditional news follow on average a similar pattern although on a much lower level. The institutions obtain on average 25 news mentions for every 100 publications in the publication year. This ratio drops sharply to less than 4 in t1 and below 2 at t2. Obviously, publications are newsworthy shortly after publishing and according to the data lose their appeal to the general news quickly afterwards.

Also mentions on Facebook show the same overall patterns as mentions on Twitter, but on a lower level. For every 100 publications more than 12 Facebook mentions are observed in the publication year. In t1 this falls to around 2 and less than 2 at t2. Mentions on blogs obtain even lower numbers in the publication year (around 5 blog mentions for 100 publications), but the decrease in t1 (more than 1 blog mention for 100 publications) is noticeable less pronounced than for Twitter, Facebook or News. Relative to the other short-winded channels, blog mentions seem less hyped, but miss the coverage as the average German institutions see less than 5% of their publications covered by blogs in the citation window. Given the low coverage of mentions on Facebook, Blogs and Wikipedia, we restrict our nowcasting model to Mendeley, Twitter and News.

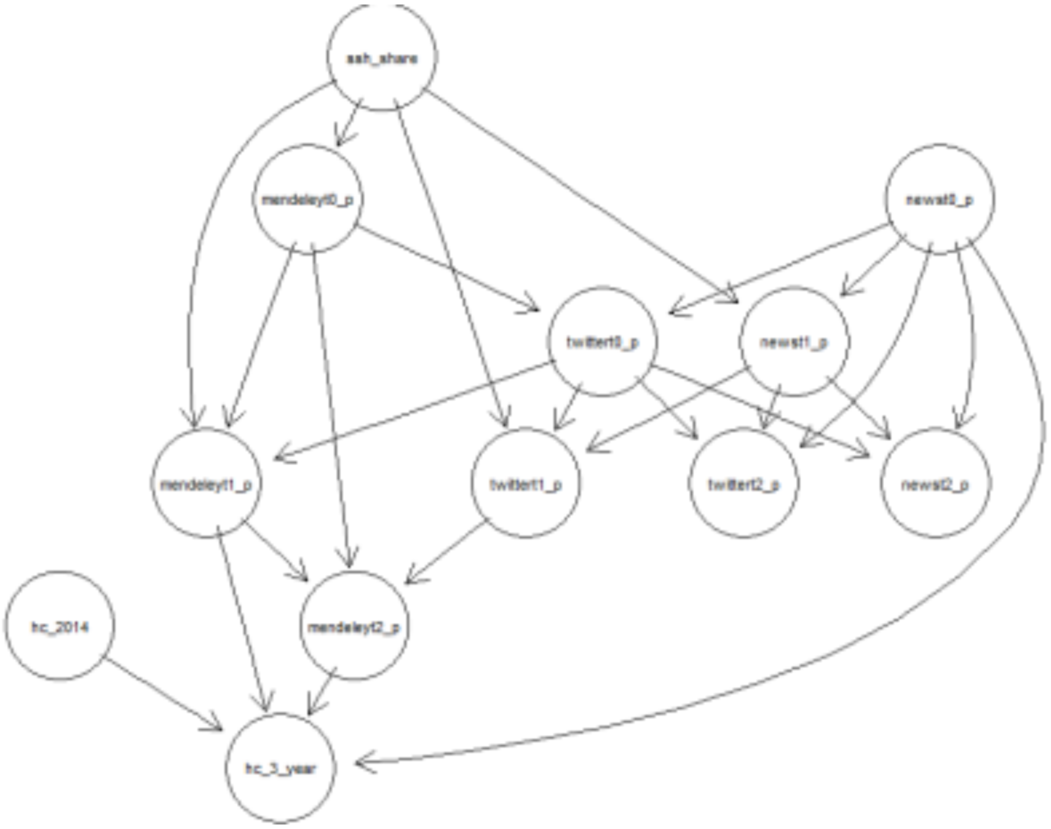

Figure 2. Bayesian network for German universities of the 10% excellence rate of 2014 (hc_2014) and 2015 (hc_3_year), respectively the time series of the Altmetric channels in the corresponding citation window 2015-2017.

Apart from the Altmetric variables we also include the share of publications in the SSH by institution in our model to account for variation in Altmetric usage and publication/citation culture. Finally, the former values of the 10% excellence rate for every German institution are computed and supplied to the model to incorporate path dependencies arising from relatively stable research staff and research infrastructure.

Methods

To observe the structure among the diverse Altmetrics channels and their relationship with the excellence rate, we generate a temporal Bayesian network (BN) among the variables. A BN consists of a directed acyclic graph (DAG) which mirrors a factorisation of a probability distribution over several variables by including a directed edge between two dependent variables. Conditional independence among certain variables leads to a sparse graph, in which the nodes, representing variables, can be endowed with conditional probabilities. This combination of (in-)dependence statements and conditional probabilities describes the possibly causal relations among all factors of a specific domain and facilitates thereby statistical inference (Pearl, 2000). BNs offer several advantages, as they describe the structure of a domain, here the complex communication structure of altmetrics and citations. Hence generating the BN mainly by data may be used for learning the unknown structure in this domain. Furthermore, BNs may also be employed for prediction, as information on the variables, e.g. the number of news mentions in the publication year t0, can be entered into the network and this information is propagated via the edges to the other nodes, e.g. the belatedly available excellence rate of t0 publications. Hence, any Altmetrics information becoming available during the citation window can instantly be employed in the BN to update the estimate/nowcast of the true excellence rate unknown during the citation window.

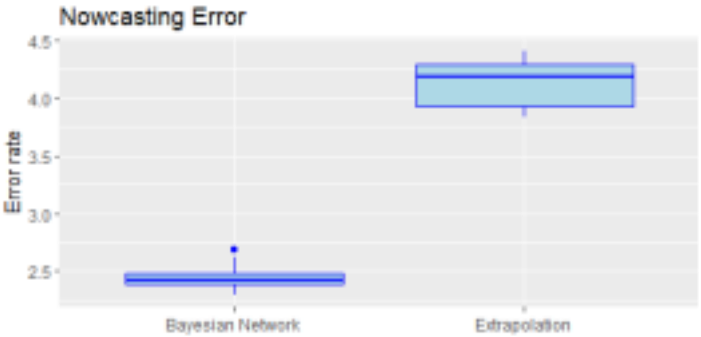

Figure 3. Mean Square Error of BN and extrapolation approach to nowcast the 2015 excellence rate of German universities from 2017 information

We learn the BN via a score-based algorithms, which maximises the Bayesian Information Criterion (BIC). The plain Hill Climbing Greedy Search is employed and evaluates by which action (adding or deleting an edge or changing an edge's direction) the BIC improves most, conducts this step and reiterates until convergence. 500 random restarts are completed to avoid any local optima of the BIC in this domain. Before we inform this structure learning algorithm on the temporal sequence of the time series described above.

Results

Communication structure

Figure 2 depicts the Bayesian network learned from the data of German universities for their excellence rate in 2015 (to be nowcasted), the explanatory Altmetric variables for their 2015 publications as well as the 2014 excellence rate. According to the edges, the excellence rate is explained by the 2014 excellence rate, the news at t0 and the Mendeley readerships at t1, respectively t2. That is the first important signal for the excellence rate arises already in the initial publication year in form of mentions in traditional news. Afterwards the Mendeley readerships in the subsequent two years inform the probability distribution of the excellence rate, as scientists list the papers to be cited in their Mendeley accounts. Also, the 2014 excellence rate results informative for the 2015 excellence rate, which seem reasonable given the pronounced path dependency on the institutional level. For the Altmetrics channels News and Mendeley we can observe the autocorrelation of Figure 1, whereas for Twitter counts at t0 with their high spike informs both the Twitter count at t1 and t2. The share of SSH publications affects all three Altmetrics time series at t1, highlighting the uniqueness of SSH disciplines.

Nowcasting

We now employ the learned BN to compute nowcasts for the 2015 excellence rate of German universities on their 2015 publications from information available in 2017, i.e. one year before the actual 2015 excellence rates becomes available in 2018. As proposed by the BN pictured in Figure 2 we include the values of the 2014 excellence rate (initially available in 2017), the Mendeley counts in 2016 and 2017 and news mentions in 2015 in the respective nodes and have this information propagated via the edges to compute a nowcast of the 2015 excellence rate. By comparing these estimates to the true value via the mean square error (MSE) we get informed about the precision of this particular nowcasting model while accounting for positive or negative deviations from the true values. Accordingly, the left boxplot in Figure 3 presents the MSE arising from a 10-fold cross validation on parameter learning to avoid overfitting. We compare this approach with a simple baseline model, which extrapolates the 2014 excellence rate to the 2015 excellence rate, i.e., omit all Altmetric nodes in Figure 2. The error rate of this baseline model is depicted in the right boxplot in Figure 3.

It may be observed that the error rate is substantially reduced by the BN approach. While the extrapolation approach incorporates a median error of 4.2, the BN approach with its additional information from the Altmetrics channels lowers the median error rate to 2.4. Taking the root to align the error scale with the scale of the excellence rate, we observe a 25% reduction in the error and an average BN error of 1.5 percentage points. Given the excellence rate range of the 72 German universities starting at 4.5% and topping at 24.6%, the proposed BN nowcasting approach seems informative for many reporting use cases.

In this ongoing work we will continue to fine-tune the model, as well as include the extra university research institutions in Germany. Furthermore, we plan to extend the analysis from its current focus on 2015 publications and incorporate also the publications years 2016 and 2017 to present a more general model at the ISSI2023 conference.

Acknowledgments

This study was funded by the Federal Ministry of Education and Research of Germany (BMBF) under grant number 16PU17024. WoS data has been obtained via the BMBF funded Kompetenznetzwerk Bibliometrie (16WIK2101A). This study and the simultaneously submitted study “Confounding cross-correlation: A meta-analysis of the association between citations and altmetrics”, are studies in the same project. The authors like to thank Rodrigo Costas and CWTS for providing the Mendeley data.

References

Bornmann, L., Leydesdorff, L. and Wang, J. (2014). How to improve the prediction based on citation impact percentiles for years shortly after the publication date? *Journal of Informetrics, 8*, 175-180. DOI: 10.1016/j.joi.2013.11.005.

Evans, M. D. D. (2005). Where are we now? Real-time estimates of the macroeconomy. *International Journal of Central Banking, 1*, 127-175.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional netrics of scientific impact. *Journal of Medical Internet Research, 13*:e123. DOI: 10.2196/jmir.2012.

Fu, L. und Aliferis, C. (2008). *Models for predicting and explaining citation count of biomedical articles.* AMIA Annual Symposium Proceedings 2008. Maryland: AMIA.

Haustein, S., Peters, I., Sugimoto, C., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology, 65*(4), 656-669, DOI: 10.1002/asi.23101. Pearl, J. (2000). Causality. New York: Cambridge University Press.

Rimmert C., Schwechheimer H., Winterhager M. (2017). Disambiguation of author addresses in bibliometric databases – technical report. Bielefeld: Universität Bielefeld, Institute for Interdisciplinary Studies of Science (I²SoS).

Schmidt, M., Kircheis, W., Simons, A., Potthast, M., & Stein, B. (2021). Does Wikipedia Cover the Relevant Literature on Major Innovations Timely? An Exploratory Case Study of CRISPR/Cas9. In W. Glänzel, S. Heeffer, P.-S. Chi, & R. Rousseau (Hrsg.), *18th International Conference on Scientometrics & Informetrics – Proceedings* (S. 1021-1026). Leuven: International Society for Scientometrics and Informetrics (I.S.S.I.).

Shibata, N., Kajikawa, Y. und Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. *Journal of the Association for Information Science and Technology, 58*(6), 872-882.

Wang, D., Song, C. und Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science, 342*, 127-132. DOI: 10.1126/science.1237825.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*, 851-872. DOI: 10.1007/s11192-012-0775-9.