Scientometrics (2021) 126:1173–1186
https://doi.org/10.1007/s11192-020-03789-8

# Machine learning misclassification of academic publications reveals non-trivial interdependencies of scientific disciplines

**Alexey Lyutov**[1] · **Yilmaz Uygun**[1] · **Marc-Thorsten Hütt**[2]

Received: 7 April 2020 / Accepted: 10 November 2020 / Published online: 27 November 2020


## Abstract

Exploring the production of knowledge with quantitative methods is the foundation of scientometrics. In an application of machine learning to scientometrics, we here consider the classification problem of the mapping of academic publications to the subcategories of a multidisciplinary journal—and hence to scientific disciplines—based on the information contained in the abstract. In contrast to standard classification tasks, we are not interested in maximizing the accuracy, but rather we ask, whether the *failures* of an automatic classification are systematic and contain information about the system under investigation. These failures can be represented as a 'misclassification network' inter-relating scientific disciplines. Here we show that this misclassification network (1) gives a markedly different pattern of interdependencies among scientific disciplines than common 'maps of science', (2) reveals a statistical association between misclassification and citation frequencies, and (3) allows disciplines to be classified as 'method lenders' and 'content explorers', based on their in-degree out-degree asymmetry. On a more general level, in a wide range of machine learning applications misclassification networks have the potential of extracting systemic information from the failed classifications, thus allowing to visualize and quantitatively assess those aspects of a complex system, which are not machine learnable.

**Keywords** Machine learning · Scientometrics · Maps of science · Classification algorithms · Interdisciplinary research

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s11192-020-03789-8) contains supplementary material, which is available to authorized users.

✉ Alexey Lyutov
a.lyutov@jacobs-university.de

[1] Department of Mathematics and Logistics, Jacobs University, Campus Ring 1, 28759 Bremen, Germany

[2] Department of Life Sciences and Chemistry, Jacobs University, Campus Ring 1, 28759 Bremen, Germany

## Introduction

The rich research landscape exploring the possibility of constructing 'maps of science', allowing for a locally and globally accurate representation of the relationships, distances, and proximities of scientific disciplines (the 'scientific landscape') is one of the cornerstones of scientometrics. This field of research provides quantitative analyses of the mechanisms, prerequisites, and predictors of academic success and the creation of meaningful representations of the interdependencies among scientific disciplines, as a basis for strategic decisions (Boyack et al. 2005; Leydesdorff 2001).

Starting from the first networks of scientific publications Price (1965), which can be seen as the initiation of scientometrics, and the diverse approaches of constructing 'maps of science' (Boyack et al. 2005; Small 2010; Leydesdorff and Rafols 2009; Enders et al. 2018), the field of scientometrics (Leydesdorff 2001; Mingers and Leydesdorff 2015) has led to some remarkable results about the mechanisms underlying the production of knowledge. In Guimera et al. (2005) the optimal composition of a research team (as well as other types of creative teams) is studied using both, a data-driven approach and a minimal model. The formal definition and predictability of 'sleeping beauties'—papers, which receive the bulk of their scientific attention only years after their publication—has been discussed in Ke et al. (2015). The increasing rate of publication and its consequences for progress in science has been studied in Shiffrin et al. (2018). How past academic training and current working environment (in particular, the academic status of the host institution) affect the productivity and success of researchers has been quantitatively explored in Way et al. (2019). In Ma and Uzzi (2018) the association between scientific awards and academic success is investigated and put in the context of collaboration networks and academic student-advisor relationships. Evidence for a reduction of career lengths in science as a consequence of higher competition for academic positions, more diverse career paths and an imbalance between temporary and permanent positions is accumulated in Milojević et al. (2018). The statistical analysis in Wuchty et al. (2007) has revealed a drift towards larger team sizes in scientific work, even for high-impact publications. Using novel metrics assessing a publication's impact, a more recent study emphasized that disruptive publications are rather associated with smaller team sizes (Wu et al. 2019). Other findings at the interface of scientometrics and the 'science of team science' (SciTS) are summarized in Börner et al. (2010).

An even broader perspective than in scientometrics is adopted under the label 'science of science' (Fortunato et al. 2018), a research field built on the insight that the interplay of social, economic and content-based drivers requires us to think about the production of knowledge as a complex system. Some prominent observations in this complex system view of the production of knowledge are the power law of citation frequencies (Price 1965, 1976; Redner 1998; Tsallis and de Albuquerque 2000), the basic organizational laws of successful scientific work Guimera et al. (2005), as well as the scale-free nature of citation networks (Price 1965; Bilke and Peterson 2001), the observation that topological features of co-authorship networks may serve as predictors of citation frequencies (which can be seen as a proxy of academic success) (Krumov et al. 2011; Klosik et al. 2014) and a multitude of examples summarized in Fortunato et al. (2018). The most relevant data resources for maps of science are academic publications. Their distribution across academic journals, their success (in terms of citations) and the resulting networks derived, e.g., from co-citations or co-authorships have for a long time been instrumental in the construction of maps of science Boyack et al. (2005). Methods from machine learning and artificial intelligence

are currently in the process of triggering disruptive changes in many aspects of societal organization. Their explosion in the last years enables us to draw a map of science from a novel perspective, namely by assessing how machines classify abstracts of scientific publications into the categories of a journal, compared to humans.

The part of the literature methodologically closest to our investigation focuses on creating maps of science by correlating subjects using machine learning. Most of such articles are similar in using the LDA method (Blei et al. 2003) or its modifications but differ in the investigated problems and datasets. Blei and Lafferty (2007) plot the map of science using the Science articles from the JSTOR archive. Yau et al. (2014) solve a similar problem using bibliographical data from the Web of Science, as well as articles from EI Compendex. Suominen and Toivanen (2015) apply the LDA to plot the map of Finnish science. Velden et al. (2017) perform an elaborate comparison of various unsupervised clustering methods, in the problem of topic extraction from bibliographic data of scientific publications. The main dataset used in their research is the Astro Data Set, astrophysical journals indexed by the Web of Science. Zong et al. (2013) design a simple word frequency algorithm that detects the co-occurrence of words in Doctoral dissertations from Chinese universities. Based on the algorithm, they plot a topic closeness map. Zhang et al. (2017) go in a different, yet interesting direction of plotting the evolution of science by identifying birth, change, and death of scientific disciplines. To do so, they apply a self-designed learning algorithm to the US NSF award data. Being similar by using the machine learning methods, these articles however, differ from the current study as they omit the available ground truth, i.e. the original disciplines of the documents. Whereas our method of looking at the data centers on the knowledge of the original category of a document and, more importantly, the wrong identification of those.

Here we analyze all papers published in PNAS between 2004 and 2017, a total of 48,000 papers. Each paper is represented by freely available data at the PNAS online archive: authors, title, keywords, abstract, significance statement, and journal section and subsection. In addition, directly after the initial data acquisition procedure, the corresponding citations count was obtained for each paper using the Scopus database (Scopus https://dev.elsevier.com/).

Using the very basic classification task of sorting abstracts of scientific publications into predefined categories – the topical subsections of a journal—we show that such links between content and categories can be successfully learned by data-centric algorithms.

Taking the human classification (i.e. the actual section and subsection assignment of the paper) as the 'ground truth', the probabilities that a machine learning device falsely classifies an abstract from one category into other categories define a directed, weighted graph—the misclassification network, which interlinks scientific disciplines represented by the categories of the journal at hand. This research is an attempt to use the limits of learnability to identify boundaries of scientific disciplines and to draw a network of disciplinary interdependencies. This network is fundamentally different from approaches based on co-citations (Boyack et al. 2005), co-authorships (Krumov et al. 2011), or surveys among scientists (Enders et al. 2018). The application of machine learning techniques to the classification of scientific abstracts is an indirect way of measuring the information content in a scientific text segment.

We subsequently interpret these findings from a scientometric perspective and discuss four main results:

1. The view of the classification capabilities via a machine learning as an indirect measure of information content offers the possibility to study how this quantity changes along an abstract. To investigate this, we parametrize abstracts with a normalized coordinate, where the beginning of the text is 0 and the full abstract is 1 and plot the classification accuracy as a function of the position of a small text window along this coordinate. We find that these curves are indicative of an 'hourglass' structure Derntl (2014).
2. The misclassification network reveals strikingly different features than other 'maps of science'. In particular, we can quantitatively assess that it follows an organizational principle distinct from a section co-occurrence network derived from all papers with multiple section labels.
3. We find that some disciplines have a much higher out-degree than in-degree, suggesting that their disciplinary boundaries are formally less well defined. More generally, the asymmetry of in- and out-degrees in the misclassification network can be used to classify disciplines into 'methods lenders' and 'methods receivers' (or 'content explorers').
4. By analyzing the citation frequencies of correctly and incorrectly classified publications we find a tendency of higher citations in the incorrectly classified publications. This finding may be indicative of a higher impact of—and relatively a higher interest in—interdisciplinary articles located at the boundaries of scientific disciplines.

## Methods

### Machine learning devices

Text classification is performed in the Python environment using the NLTK library (Bird et al. 2009) to process text data and the SCIKIT-learn library (Pedregosa et al. 2012) for machine learning tools. The following machine learning methods have been used: Naive Bayes, Maximum Entropy, Multi-Layer Perceptron, and Multi-Voting classifier. Each of the first three represents a different algorithm with a proven performance. The fourth one, Multi-Voting classifier, combines these three methods into a single one by comparing their answers and returning the most popular one. In all experiments where a single classifier is needed the Multi-Voting one has been used.

For most of the classification algorithms, default parameters have been used, except for the Maximum Entropy classifier where the default solver has been changed to the 'lbfgs' which supports a multi-class output and the maximum number of iterations has been increased from 100 to 300 to handle convergence problems in boundary cases. A similar tuning has been done to the Multi-Layer Perceptron Classifier where the maximum number of iterations has been increased from 200 to 1000.

### Publication data

We use the PNAS online archive with 52000 scientific articles from categories "Biological Sciences", "Physical Sciences", and "Social Sciences" for years 2004–2017. Each article contains a title, an abstract, an assigned section and subsection, a list of authors, and a publication date. 43,000 of these articles belong to a single section and subsection, while the remaining 4500 have multiple sections or subsections, thus appearing in the database more than once, leading to 48,000 unique articles.

A single-label classification method based on abstract texts has been used in most tests of this research. Some additional tests (reported in the Supplementary Information) are performed with subsets of the data: 37,,500 of the 43,000 single-section articles have keywords and 12,500 of 37,500 also have significance statements. The significance statement is a simplified version of an abstract designed for a general educated audience without specific knowledge in the given domain.

In addition to the general data available from the PNAS, we have used the Scopus database to acquire each article's total citation count. The citation count, as opposed to the general article data, is a temporal variable that implicitly depends on other parameters (e.g. journal prestige, disciplines popularity, etc.). This might lead to potential confounders when performing citations-related experiments. In order to control them, we make sure that the citations are gathered in a short time window (within one week in June 2019), the range of years used for the research is relatively small (2004–2017), and all articles are published in the same journal, thus having a somewhat equal initial probability to become successful.

## Statistical features of the data

The available data set has a high disproportion among classes on both the section (Table 1) and the subsection (Table S1) levels, which is well-known in the machine learning community Chawla et al. (2004); He and Garcia (2009) as an imbalanced data set problem. In the current manuscript, the simplest approach of under-sampling is used to overcome the imbalance. In this approach, the training set size is limited by $n$ articles per each label, $n = min(|l|, \theta)/2$, where $|l|$ is the size of the article set with label $l$ and $\theta$ is the threshold which serves as an upper limit of article sets. For example, when training a section classifier with the threshold $\theta = 1000$, 1000 articles are taken from bigger sections (Biological Sciences and Physical Sciences) while only a half (462) is taken from the smaller, Social Sciences, section. The test set, in this case, consists of all the remaining articles that are not taken into the training set. In the case of the smallest category, the ratio between training and test set sizes would be at least 1:1, while in other cases it can reach 375:1.

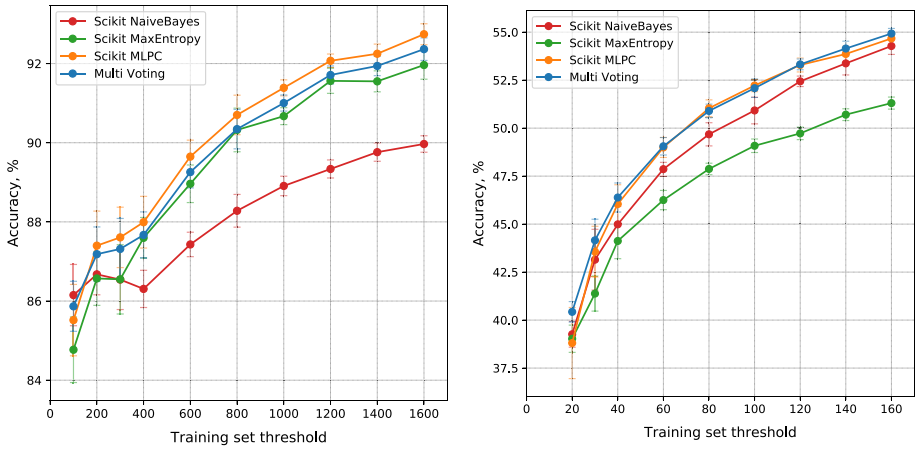
Such an approach might lead to an information loss based on the frequency distribution of the classes. However, this is an acceptable drawback, as the high accuracy of classification is not the primary goal in the current research.

## Classification accuracy

The accuracy of all classification algorithms has been tested against an increasing training set size. In this test, the threshold varied from 100 to 1600 during the section and from 20 to 160 during the subsection classification. For each threshold, 10 random training sets have been picked to reduce the effect of random sampling. The overall accuracy for each threshold has been taken as an average accuracy in those 10 tests. Figure 1a and b show the

**Table 1** Distribution of single-section PNAS articles by section

| Section | N articles |
|---|---|
| Biological Sciences | 37,650 |
| Physical Sciences | 4591 |
| Social Sciences | 925 |




**(a)** section classification **(b)** subsection classification

**Fig. 1** Mean and standard deviation of abstracts classification accuracy by their section and subsection name

achieved accuracy in this test. The maximum value is 92.7% in the case of section classification and 54.9% in the case of subsection classification. The accuracy of a dummy classifier in these cases would be 33.3% and 2.7%, respectively (42% and 3.4% with the prior knowledge of class proportions in the training sets). Another important characteristic of a classifier, a receiver operating characteristic (ROC) curve, is shown in the supplements, Fig. S1.

## Results

Figure 1 provides a general orientation of the performance for the classification of publications into the sections and subsections of the journal based on the abstracts.

As expected, with the increasing amounts of training data the classification accuracy increases, both at the section and the subsection levels. Furthermore, different machine learning algorithms produce qualitatively similar results. It should be noted that diverse technical details affect aspects of this general behavior, e.g., the exact implementation and parameter choices of the machine learning methods, the precise processing of the word list derived from the abstract, as well as some features of the split between training and test data.

In this investigation, however, our aim is *not* to maximize the accuracy of the classification or prediction task. It is rather to probe the systematic features of misclassification for information on the interdependences of scientific disciplines. We checked that the results discussed in the following are not depending strongly on these technical details.

In the beginning, we analyze the variation of information content along an abstract. Scientific abstracts usually have a typical structure, as described, for example, in the "Nature guide to authors" (Nature https://www.nature.com/documents/nature-summary-paragraph.pdf) or in the traditional hourglass model (Derntl 2014). The pattern can be seen in all the disciplines: An abstract starts with a general introduction to the field, then it narrows down

to a specific problem or method and broadens again at the end to describe the benefit to the field in general.

To detect how the information content follows this pattern along an abstract using machine learning, we define a window of size $\alpha \in (0, 1)$ as $\alpha \cdot N$ consecutive words from an abstract of size $N$ and slide this window across the abstract, varying its position from $x_0 = \alpha/2$ to $x_k = 1 - \alpha/2$. For each position, training and testing is performed using only the words from this window. Thus, the information content of abstracts is represented as the accuracy of identifying the correct field of an article based on some part of the abstract.

The results of this investigation with the training set threshold = 800 and different window sizes $\alpha$ from 0.4 to 0.1 are given in Fig. 2. For each window size, there is a typical signal of higher accuracy at the beginning of abstracts, then reaching its minimum at the middle, and rising again at the end of the abstracts. This finding clearly shows that most scientists have adopted this pattern of abstracts writing. Classification accuracy as a function of position, therefore, offers a quantitative confirmation of the 'hourglass model'.

When analyzing the classification results, it is possible to construct a confusion matrix that shows how often an item from class $A$ is misclassified as class $B$. An example of such a matrix for the case of section classification is given in Table S2, as well as in Tables S3–S5 for the subsection level. The confusion matrix reflects how strongly the contents of class $A$ are entangled with class $B$. Such a matrix can be considered as a weighted adjacency matrix of a directed graph. This graph represents a network of scientific disciplines with links reflecting the proximity of two disciplines.

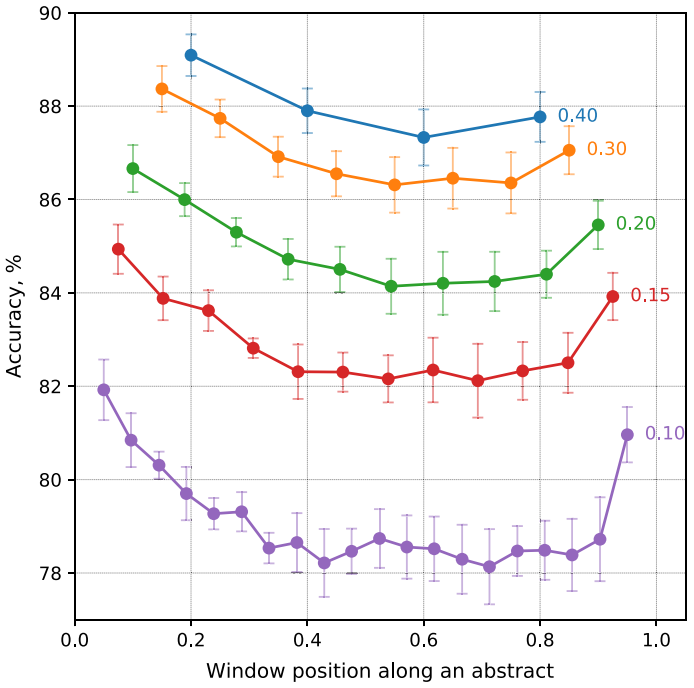



**Fig. 2** The accuracy of section classification using the words from a window sliding along an abstract; training set threshold = 800. Different lines represent different window sizes varied from 0.4 to 0.1 of the whole abstract (shown to the right of each line)

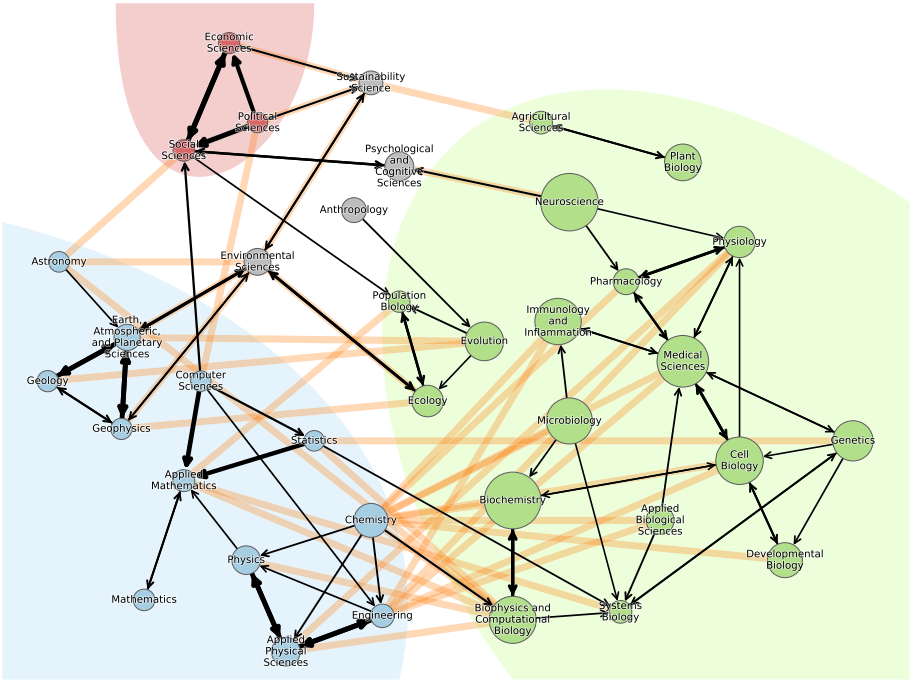



**Fig. 3** Two different maps of sciences with PNAS subsections as nodes. Node sizes are proportional to the number of articles in the corresponding subsection. (1) Black, directed—misclassification network based on the confusion matrix from subsection classification with threshold = 40. Links represent articles from starting subsection that are misclassified as an ending subsection, the threshold for drawing a link is 6% of starting subsection size. Widths of links are proportional to the percentage of misclassified articles. (2) Orange, undirected—multi-section network based on articles that were assigned to multiple subsections by PNAS. Links represent articles that belong to both connected subsections at the same time, the threshold for drawing a link is 20% of smallest connected subsection size

Figure 3 (black directed links) shows a network based on the confusion matrix from the subsection classification. From this figure, it can be seen how a wrong classification caused by content closeness creates a systematic and informative pattern of links among disciplines. To ensure the robustness of this network, we perform 10 classification runs for different training and test datasets and plot only those links which are systematically repeated in each run. Each network link is normalized by the size of the starting subsection. If there were no systematic connections between the subsections, links between categories would appear at random and will be filtered out during the different runs.

In the network, some subsections have a systematically higher in-degree than out-degree (e.g. ”Systems Biology” or ”Applied Mathematics”) or out-degree than in-degree (e.g. ”Neuroscience” or ”Chemistry”). Based on the meaning of the wrong classification the following node degree interpretation can be formulated. Qualitatively speaking, based on the asymmetry of in- and out-degree in the misclassification network we can thus classify disciplines into ’methods lenders’ and ’content explorers’.

Discipline with a high **out-degree** and a small in-degree is characterized by articles systematically misclassified into a range of other disciplines. This is indicative of a methods-oriented discipline with a highly successful method export history.

On the other hand, a discipline with a high **in-degree** and a comparatively low out-degree has a substantial set of articles being misclassified as belonging to this discipline, even though in the journal these articles have appeared in another section (i.e., assigned to another discipline). This is indicative of a discipline with well-defined content that is investigated using resources and methods from a wide range of other disciplines.

Disciplines with a less pronounced degree asymmetry are uniting aspects of both types. These preliminary categories are rising from the topological properties of our misclassification network require validation with other datasets and may serve as a starting point for subsequent research with more content-driven methods.

It should be noted that the source data and the methods used for drawing the networks shown in Leydesdorff and Rafols (2009) and Boyack et al. (2005) are fundamentally different from the one we are using here.

To illustrate how qualitatively different this network is from more typical 'maps of science' in the literature, we also depict a subsection co-occurrence network. To do so, we take the 4500 articles from the original dataset that are assigned to more than one section and construct a multi-section network (orange undirected links in Fig. 3). A link between two disciplines here is established when an article is assigned to both disciplines at the same time. Link weights are then normalized by the minimum size of the connected nodes. This network, as opposed to the classification-based network, is closer to the conventional methods employed in scientometrics (Leydesdorff and Rafols 2009; Klavans et al. 2007; Boyack et al. 2005; Small 2010).

Both network construction approaches evaluate the proximity of disciplines (journal subsections) suggesting structural similarities of the two networks in Fig. 3. Visual comparison, however, reveals strong qualitative differences. For example, there is only a single link in the multi-section network that connects two nodes from the same section (mixed type: "sustainability science" and "environmental sciences"), while in case of the misclassification network 51 link connects same sections and 30 links connect different ones, showing a slight preference towards connecting the disciplines of the same field (subsection belonging to the same sections).

To support the visual observations of a stronger-than-random difference between the two networks we have calculated the spectral distance (Ipsen and Mikhailov 2002) between these two networks and between their switch-randomized (Maslov and Sneppen 2002) versions (Fig. S13).

Thus, the difference between multi-section and misclassification networks reflects the different features of those networks. While the misclassification highlights the closeness of disciplines regardless of their original field (section), the multi-section network captures the closeness across the fields. It should be stated that the main source of difference between the misclassification and the multisection networks will be due to the PNAS tradition that a paper typically receives multiple subcategory labels from different major categories.

In the following, we explore the relationship between the two network-generating quantities—misclassification and subsection co-occurrence, with the impact of a publication. Citation frequencies will be used as a proxy of an academic impact.

Some evidences suggest that interdisciplinary articles receive a higher citation impact (Yegros-Yegros et al. 2015). Our analysis of citation frequencies of multi-section articles shows some support for this. We compare citation counts of single- and the multi-section articles. The comparison is complicated by the fact that the number of single-section articles in the available data is almost 10 times higher than the number of multi-section ones. Their distribution over time, as can be seen in Fig. S6, is also substantially

different from a marked growth of multi-section articles number from 2004 until 2012. Therefore, to compare these two datasets we calculate their yearly frequency distributions for each separate year from 2004 to 2017, normalize them by the number of articles in each set from that year, and summarize the normalized distributions across all years. As can be seen in Fig. 4, the multi-section curve is systematically shifted towards the higher citation counts compared to the single-section curve. This observation is also supported by a higher time-averaged mean, 64.3 for multi-section vs. 54.2 for single-section. The citation data covers a range of years with some of the articles not reaching their steady state as discussed by Stringer et al. (2008). To overcome this issue, a year-by-year comparison of correctly and incorrectly classified abstracts is given in Fig. S12.

Multi-section articles (i.e., the ones behind the orange links depicted in Fig. 3) thus tend to be more frequently cited than their single-section counterparts (see Fig. S10a). It is, therefore, an intriguing question, whether the single-section articles misclassified by ML approaches (which hence can be thought of to lie at the 'boundaries' of subsections) also have a tendency towards higher citation frequencies. Figure S10b confirms this visible, though weak, effect.

To test this hypothesis, an abstracts-based classification of sections is performed. This test is performed with only sinlge-section articles to exclude any potential biases from citations shift in multisection data. During the test phase we split the citations count of the classified articles into two sets: with correctly and incorrectly assigned sections. This procedure is repeated 10 times with different training and test sets. Here, the yearly-averaging procedure is not performed, because the number of correctly and incorrectly classified articles is uniform over the years. Figure 5 shows that there is a consistent increase of citation count for systematically misclassified articles, which is comparable with the one

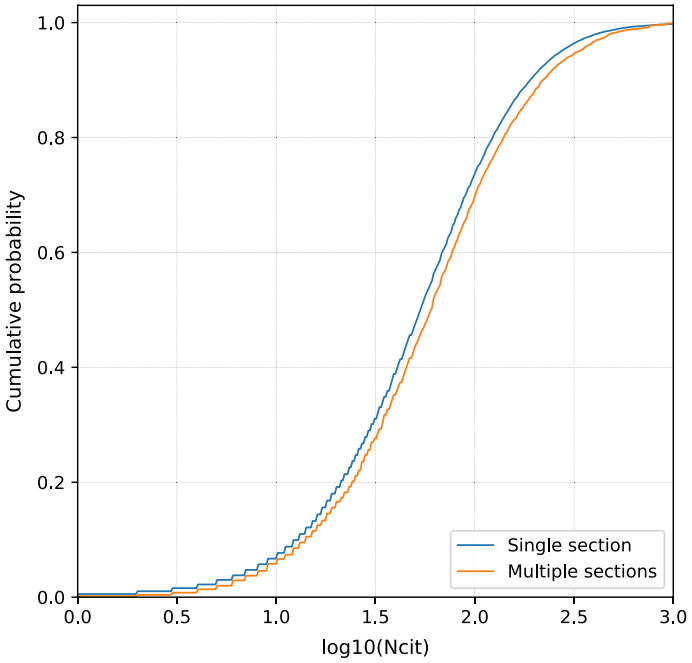



**Fig. 4** Cumulative frequency distribution of single- and multi-section articles by their citations count

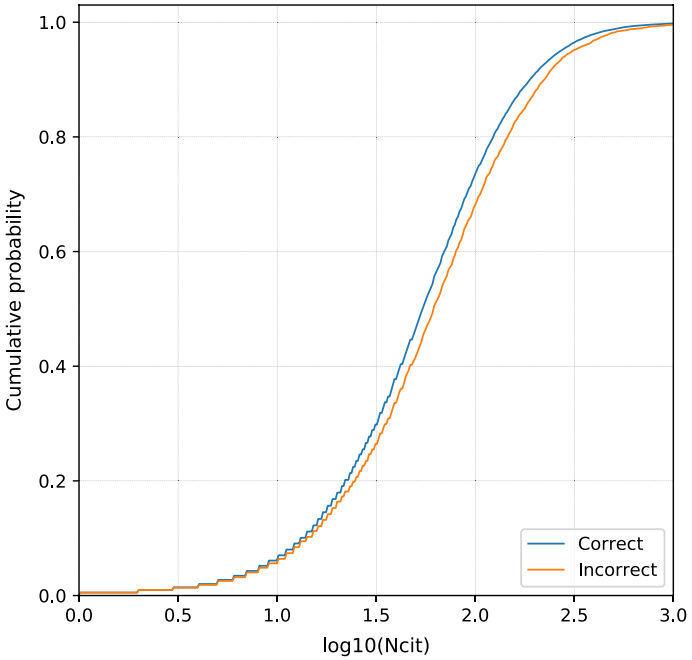



**Fig. 5** Cumulative frequency distribution of correctly and incorrectly classified articles during the abstract-based section classification with threshold = 1200

for multi-section articles. This slight trend is indicative of a potential interest created by a more fuzzy wording and could point to a fundamental difference between a human and a machine reading a scientific abstract. As it is shown in the supplements (Fig. S8, S9), the shift in the citations count for the incorrectly classified abstracts is not explained by an abstract size effect. Another set of experiments there (Fig. S10, S11) aims to show that this weak signal exists and is not a random effect. Apparently, publications that challenge the ML approaches, tend to receive a (slightly) higher attention by the scientific community.

Interestingly, a similar experiment with subsections classification gives two distributions for correct and incorrect datasets that match well, indicating that there is no citation shift for misclassified subsections. We believe that this happens because misclassification on the subsection level does not reveal the interdisciplinarity of an article as the disciplines on this level are a lot closer to each other.

## Conclusion

Here we have applied machine learning tools to problems in scientometrics. Instead of the conventional approach of maximizing the efficiency of those algorithms, we question if there is anything systematic in the patterns of wrong answers and, more importantly, do those patterns reveal any additional information about the underlying system.

The first result of our investigation is the validation of the standard scheme used by scientists to write scientific abstracts. The variation of machine learning classification accuracy along the abstract shows that the beginnings and the ends of abstracts provide

systematically higher topic-related information content, thus correlating well with the standard ”hourglass model”.

The second result is that the wrong classifications of an algorithm can be used to search for dependencies among scientific fields, allowing to create networks of sciences and draw conclusions about relations among disciplines. Applying a similar analysis over a longer timeframe (e.g. decades of publication history) can demonstrate an evolution of scientific disciplines, revealing, for example, how the status of a discipline drifts from a content-oriented, borrowing methods from other disciplines, to a mature, providing its own methods to other fields.

The third result is an evidence of a relationship between fuzzy wording (leading to machine learning misclassification), multidisciplinarity (as provided by multi-section articles), and impact (as measured by citation frequencies).

In practice, human readers will often base their decision, whether to read an academic publication (i.e. whether the published work is related to their own field of interest) on the abstract. Therefore, the classification based on the abstract – and in particular the misclassification network derived from it – are of particular interest. Nevertheless, a natural next step of this investigation could be to perform a full-text classification and see (i) whether the accuracy is substantially increased, (ii) whether the misclassification network changes strongly in this case and (iii) whether the relationship between misclassification and citation frequencies remains intact (see also the discussion of this point in the Supplementary Information around Fig. S2).

We see two principles emerging from our study which would allow doing so. (1) The notion of misclassification networks derived from machine learning applications to a field where a reliable ground truth is available. We envision applications to all levels of societal organization. One example of such an application could be a job market, where a CV serves as the input text and companies or departments are the classes. Another example is product specifications as input texts and business units in a company as the classes (Lyutov et al. 2019). (2) The generalization of our findings from a single academic journal to a broad range of publication outlets. Here normalizations of the citation frequencies to the median of the journal, as well as other measures of academic success of a publication (e.g., the ’disruptiveness’ discussed in Wu et al. (2019)), will require reflection.

Lastly, an avenue of research prompted by our findings is the better characterization of interdisciplinary science, e.g., by comparison with the approaches outlined in Trujillo and Long (2018) and Basurto-Flores et al. (2018).

**Funding** Open Access funding enabled and organized by Projekt DEAL.

## References


Basurto-Flores, R., Guzmán-Vargas, L., Velasco, S., Medina, A., & Hernandez, A. C. (2018). On entropy research analysis: Cross-disciplinary knowledge transfer. *Scientometrics*, *117*(1), 123–139.

Bilke, S., & Peterson, C. (2001). Topological properties of citation and metabolic networks. *Physical Review E*, *64*(3), 036106.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol: O'Reilly Media, Inc.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35. https://doi.org/10.1214/07-AOAS114.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation david. *Journal of Machine Learning Research*, *3*, 993–1022.

Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S. M., Hall, K. L., Keyton, J., et al. (2010). A multi-level systems perspective for the science of team science. *Science Translational Medicine*, *2*(49), 49cm24–49cm24.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 6. https://doi.org/10.1145/1007730.1007733.

Derntl, M. (2014). Basics of research paper writing and publishing. *International Journal of Technology Enhanced Learning*, *6*(2), 105–123. https://doi.org/10.1504/IJTEL.2014.066856.

Enders, M., Hütt, M. T., & Jeschke, J. M. (2018). Drawing a map of invasion biology based on a network of hypotheses. *Ecosphere*, *9*(3), e02146.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Petersen, A. M., et al. (2018). Science of science. *Science*,. https://doi.org/10.1126/science.aao0185.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, *308*(5722), 697–702.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.2174/156802608786786589.

How to construct a Nature summary paragraph. Nature https://www.nature.com/documents/nature-summary-paragraph.pdf.

Ipsen, M., & Mikhailov, A. (2002). Evolutionary reconstruction of networks. *Physical Review E*, *66*, 1–4.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, *112*(24), 7426–7431.

Klavans, R., & Boyack, K. W. Is there a convergent structure of science? A comparison of maps using the ISI and scopus databases. In Proceedings of ISSI 2007: 11th International Conference of the International Society for Scientometrics and Informetrics (June), 437–448 (2007).

Klosik, D. F., Bornholdt, S., & Hütt, M. T. (2014). Motif-based success scores in coauthorship networks are highly sensitive to author name disambiguation. *Physical Review E*, *90*(3), 032811.

Krumov, L., Fretter, C., Müller-Hannemann, M., Weihe, K., & Hütt, M. T. (2011). Motifs in co-authorship networks and their relation to the impact of scientific publications. *The European Physical Journal B*, *84*(4), 535–540.

Leydesdorff, L. (2001). *The challenge of scientometrics: The development, measurement, and self-organization of scientific communications*. Irvine: Universal-Publishers.

Leydesdorff, L., & Rafols, I. (2009). A Global Map of Science Based on the ISI Subject Categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362. https://doi.org/10.1002/asi.20967.

Lyutov, A., Uygun, Y., & Hütt, M. T. (2019). Managing workflow of customer requirements using machine learning. *Computers in Industry*, *109*, 215–225. https://doi.org/10.1016/j.compind.2019.04.010.

Maslov, S., & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, *296*(5569), 910–913.

Ma, Y., & Uzzi, B. (2018). Scientific prize network predicts who pushes the boundaries of science. *Proceedings of the National Academy of Sciences*, *115*(50), 12608–12615.

Milojević, S., Radicchi, F., & Walsh, J. P. (2018). Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences*, *115*(50), 12616–12623.

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*(1), 1–19. https://doi.org/10.1016/j.ejor.2015.04.002.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2.

Price, D. J. D. S. (1965). Networks of Scientific Papers. *Science*, *149*(3683), 510–515. https://doi.org/10.1126/science.149.3683.510.

Price, D. J. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, *27*(5), 292–306.

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

Scopus API documentation. Scopus https://dev.elsevier.com/.

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, *115*(11), 2632–2639.

Small, H. (2010). Maps of science as interdisciplinary discourse: Co-citation contexts and the role of analogy. *Scientometrics*, *83*(3), 835–849. https://doi.org/10.1007/s11192-009-0121-z.

Stringer, M. J., Sales-Pardo, M., & Nunes Amaral, L. A. (2008). Effectiveness of Journal Ranking Schemes as a Tool for Locating Information. *PLoS ONE*, *3*(2), e1683. https://doi.org/10.1371/journal.pone.0001683.

Suominen, A., & Toivanen, H. (2015). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464–2476. https://doi.org/10.1002/asi.23596.

Trujillo, C. M., & Long, T. M. (2018). Document co-citation analysis to enhance transdisciplinary research. *Science advances*, *4*(1), e1701130.

Tsallis, C., & de Albuquerque, M. P. (2000). Are citations of scientific papers a case of nonextensivity? *The European Physical Journal B-Condensed Matter and Complex Systems*, *13*(4), 777–780.

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. *Scientometrics*, *111*(2), 1169–1221. https://doi.org/10.1007/s11192-017-2306-1.

Way, S. F., Morgan, A. C., Larremore, D. B., & Clauset, A. (2019). Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences*, *116*(22), 10729–10733.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767–786. https://doi.org/10.1007/s11192-014-1321-8.

Yegros-Yegros, A., Rafols, I., & D'Este, P. (2015). Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity. *PLoS ONE*, *10*(8), 1–21. https://doi.org/10.1371/journal.pone.0135095.

Zhang, Y., Zhang, G., Zhu, D., & Lu, J. (2017). Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics. *Journal of the Association for Information Science and Technology*, *68*(8), 1925–1939. https://doi.org/10.1002/asi.23814.

Zong, Q. J., Shen, H. Z., Yuan, Q. J., Hu, X. W., Hou, Z. P., & Deng, S. G. (2013). Doctoral dissertations of Library and Information Science in China: A co-word analysis. *Scientometrics*, *94*(2), 781–799. https://doi.org/10.1007/s11192-012-0799-1.