ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Dias L, Gerlach M, Scharloth J, Altmann EG. 2018 Using text analysis to quantify the similarity and evolution of scientific disciplines. *R. Soc. open sci.* **5**: 171545. http://dx.doi.org/10.1098/rsos.171545

Received: 12 October 2017
Accepted: 29 November 2017

Subject Category:
Computer science

Subject Areas:
complexity

Keywords:
science of science, information theory, dissimilarity measures

Author for correspondence:
Eduardo G. Altmann
e-mail: eduardo.altmann@sydney.edu.au

THE ROYAL SOCIETY
PUBLISHING

Using text analysis to quantify the similarity and evolution of scientific disciplines

Laércio Dias[1], Martin Gerlach[1,2], Joachim Scharloth[3] and Eduardo G. Altmann[1,4]

[1]Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany
[2]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA
[3]Department of German, TU Dresden, Applied Linguistics, 01062 Dresden, Germany
[4]School of Mathematics and Statistics, University of Sydney, Sydney 2006, New South Wales, Australia

EA, 0000-0002-1932-3710; MG, 0000-0002-0879-7865

We use an information-theoretic measure of linguistic similarity to investigate the organization and evolution of scientific fields. An analysis of almost 20 M papers from the past three decades reveals that the linguistic similarity is related but different from experts and citation-based classifications, leading to an improved view on the organization of science. A temporal analysis of the similarity of fields shows that some fields (e.g. computer science) are becoming increasingly central, but that on average the similarity between pairs of disciplines has not changed in the last decades. This suggests that tendencies of convergence (e.g. multi-disciplinarity) and divergence (e.g. specialization) of disciplines are in balance.

1. Introduction

The digitization of scientific production opens new possibilities for quantitative studies on scientometrics and science of science [1], bringing new insights into questions such as how knowledge is organized (maps of science) [2–6], how impact evolves over time (bibliometrics) [7,8], or how to measure the degree of interdisciplinarity [9,10]. At the heart of these questions lies the problems of identifying scientific fields and how they relate to each other. The difficulty of these problems, and the inadequacy of a purely essentialist approach, were clear already to Popper [11]: ‘The belief that there is such a thing as physics, or biology, or archaeology, and that these "studies" or "disciplines" are distinguishable by the subject matter which they investigate, appears to me to be a residue from the time when one believed that a theory had to proceed from a definition of its own subject matter. But subject matter, or kinds of things, do not, I hold, constitute a basis for distinguishing disciplines.' [11, pp. 124–125]. Instead, he argued that disciplines have a cognitive and a social dimension [12], i.e. they 'are distinguished partly for historical reasons and reasons of administrative convenience (such as the organization of teaching and of appointments), and partly because the theories which we construct to solve our problems have a tendency to grow into unified systems.' [11, p. 125].

On the one hand, the social dimension of scientific fields can be defined in terms of different institutions establishing stable recurring patterns of behaviour [13]: producing and reproducing institutions such as research institutes and universities, communicative institutions such as scientific societies, journals or conferences, collecting institutions (journals, libraries), as well as directing institutions (ministries, scientific advisory boards), etc. All these institutions contribute to the formation, stabilization and reproduction of a discipline as well as its distinction from others. On the other hand, the cognitive dimension has been specified in [13] as a number of fundamental invariants in the procedural knowledge, which lead to the categorical construction of scientific knowledge. If this process causes a change in the cognitive realm for an object of knowledge, it constitutes a certain discipline.

The brief discussion above is sufficient to show that both the definition and relation between scientific fields depend on multiple dimensions (e.g. essentialist, social and cognitive). There is no correct or intrinsically better classification or organization of scientific fields. The consequence of this view to the problem of identifying relationships between scientific fields from empirical data is that there is no *ground truth* against which results (e.g. clusters) can be compared [14]. While previous quantitative analyses often aimed at obtaining measures which lead to *better* classification methods [15], the approach we pursue here is to design methods with a clear interpretation and with known statistical properties in order to quantify and understand the different aspects captured by different dimensions. This approach is in line with very recent work in scientometrics [6].

Traditional (expert) classifications are mostly motivated by the '*subject matters*' under investigation and can be associated with an essentialist view. The empirical analysis of citation networks, an approach with a long tradition in scientometrics [16,17], can be regarded as capturing the social dimension (i.e. collecting institutions in the form of journals). While citations offer valuable insights into the structure and dynamics of science, they thus reflect only one particular dimension of the relationship between publications (or scientists) largely ignoring the actual content of the scientific articles. By contrast, the cognitive dimension can be operationalized with the help of linguistic features (e.g. keywords as indicators for conceptual imprints of disciplines). The increasing availability of full text of scientific articles (e.g. of Open Access journals) provides new opportunities to study the latter aspect in the form of written language. Examples include: (i) the tracking of the spread of ideas [18], individual words (memes) [19], or scientific concepts [20]; (ii) quantifying differences in the scientific discourse between subdomains in biomedical literature [21] or 'hard' and 'soft' science [22]; and (iii) efforts to combine citation and textual information [2,23–26].

In this work, we advance the idea that the organization and evolution of science should be studied through different, complementary dimensions. We add a new methodology that provides a meaningful, language-based, organization of scientific disciplines based on written text. We perform a systematic analysis how this compares with the organization obtained from experts and citations and, furthermore, trace the temporal evolution in the relation between different scientific disciplines.

More specifically, we introduce an unsupervised methodology to analyse the text of scientific articles. Our methodology is based on an information-theoretic dissimilarity measure we proposed recently [27] (more technically, it is a generalized and normalized Jensen–Shannon divergence between two corpora). The main advantage of this measure is that it has an absolute meaning (i.e. it is not based on relative comparisons) and it is statistically more robust than traditional approaches [27,28], e.g. with respect to the detection of spurious trends owing to rare words and increasing corpus sizes. We measure the similarity between scientific fields based on $\approx 10^7$ abstracts from the last three decades (Web of Science database). Comparing our language analysis to a citation analysis and an experts classification, we find that the language and citation are more similar to each other but the language is even more distinct from the experts than the citation analysis. Following the relation between scientific fields over time, our language analysis reveals the scientific fields that are becoming more central in science. However, overall (averaged over all pairs of disciplines) we find that the similarity between the language of different fields is not increasing.

2. Dissimilarity of scientific fields

We are interested in the general problem [2,4] of quantifying the relationship between two scientific fields i, j through the computation of a dissimilarity measure $D(i, j)$, i.e. a quantification of how different i and j are. Dissimilarity measures are symmetric $D(i, j) = D(j, i)$, non-negative $D(i, j) \geq 0$ and $D(i, i) = 0$ [29]. Each scientific field is defined by (at least hundreds of) papers classified by Web of Science as belonging to the same category (see Material and methods §5.1 for details on the data). We consider dissimilarities computed from the following three different types of information (expert, citation and language).

2.1. Experts

The classification of disciplines by their relationship is as old as science itself. The most used structure is a strict hierarchical tree, as seen in the traditional departmental division of universities. The collection of papers used here, provided by ISI Web of Science [30], provides a classification of papers according to the Organisation for Economic Cooperation and Development (OECD) classification of fields of science and technology [31]. This scheme is a hierarchical tree with scientific fields defined at three levels (domains, disciplines and specialities). For instance, *Applied Mathematics* (a speciality) is part of *Mathematics* (a discipline) which is part of *Natural Sciences* (a domain). The natural dissimilarity measure $D_{\text{exp}}(i, j)$ between two fields in this structure is the number of links needed to reach a common ancestor of i and j. For instance, considering i, j at the specialty level, D_{exp} can assume three different values: $D_{\text{exp}} = 1$ for specialties belonging to the same discipline (e.g. *Applied Mathematics* and *Statistics & Probability*), $D_{\text{exp}} = 2$ for specialties belonging to the same domain (e.g. *Applied Mathematics* and *Condensed Matter Physics*) and $D_{\text{exp}} = 3$ for the other pairs of specialties (e.g. *Applied Mathematics* and *Linguistics*). While researchers have pointed out potential issues with classification into categories of ISI Web of Science [4], it offers the most extensively available classification and remains widely used to relate articles and journals to disciplines [9,32].

2.2. Citations

Another popular approach is to consider that fields i and j are more similar if there are citations from (to) papers in i to (from) papers in j [4,16,17]. Here, we consider a dissimilarity measure $D_{\text{cite}}(i, j)$ which decreases for every citation between papers in i and j and increases with every citation from i that is not to j (and vice versa), but that remains unchanged by the number of citations that do not involve either i or j. These requirements are achieved using (for $i \neq j$) a symmetrized Jaccard-like dissimilarity [29,33]:

$$D_{\text{cite}}(i, j) = \frac{1}{2}\left(\frac{C_{i,\bar{j}} + C_{\bar{i},j}}{c_{i,j} + C_{i,\bar{j}} + C_{\bar{i},j}} + \frac{C_{j,\bar{i}} + C_{\bar{j},i}}{c_{j,i} + C_{j,\bar{i}} + C_{\bar{j},i}}\right), \tag{2.1}$$

where $c_{i,j}$ are the number of citations from i to j, $C_{a,\bar{b}} = \sum_{t=1,t\neq b}^{N} c_{a,t}$ and $C_{\bar{a},b} = \sum_{t=1,t\neq a}^{N} c_{t,b}$[1].

2.3. Language

We compare the language of fields i and j based on the frequency of words in each field using methods from information theory. Measuring the frequency $p(w)$ of word w, for each field i we obtain a vector of frequencies $\mathbf{p}_i \equiv p_i(w)$ for $w = 1, \ldots, V$, where V is the size of the vocabulary (i.e. number of different words). From this, following [27], the dissimilarity between two fields i and j is

$$D_{\text{lang}}(i, j) = \frac{2H_2(\mathbf{p}_i + \mathbf{p}_j)/2 - H_2(\mathbf{p}_i) - H_2(\mathbf{p}_j)}{(1/2)(2 - H_2(\mathbf{p}_i) - H_2(\mathbf{p}_j))}, \tag{2.2}$$

where $H_2(\mathbf{p}_i) = 1 - \sum_w p_i(w)^2$ is the generalized entropy of order 2 and the denominator ensures normalization (i.e. $0 \leq D_{\text{lang}}(i, j) \leq 1$). To increase the discrimination power and to avoid statistical biases in our estimation, we removed a list of stop words and included only the $V = 20\,000$ most frequent words (see Material and methods §5.3 for a justification). The dissimilarity (2.2) corresponds to a generalized (and normalized) Jensen–Shannon divergence which yields statistically robust estimations in texts [27,28] (for details and motivation, see Material and methods §5.4).

In contrast with most previously proposed methods, equation (2.2) has two critical properties that are essential in order to obtain the interpretable results mentioned in the Introduction. On the one hand, it

[1] Each of the two terms in equation (2.1) can be interpreted as a directed Jaccard distance $i \rightarrow j$ ($j \rightarrow i$) in the sense that we divide the number of edges that are out-links of field i (j) *and* in-links of field j (i) by the number of edges that are out-links of field i (j) *or* in-links of field j (i).

is well founded in information theory and its statistical properties (in terms of systematic and statistical errors) are well understood [27,34], distinguishing it from other heuristic approaches. On the other hand, it has convenient properties: $D_{\rm lang}(i,j)$ depends only on the papers contained in fields i and j and it is normalized $0 \leq D_{\rm lang}(i,j) \leq 1$. As a result, the measured distance between two fields, $D_{\rm lang}(i,j)$, has an absolute meaning. This is in contrast with alternative similarity measures [2,4], including machine-learning approaches (e.g. topic models [15,35]) based on (un-) supervised classification of documents into coherent subgroups. Here, the main limitations stem from the fact that either (i) the division into subgroups is typically based on statistically significant differences in the usage of words between the different subgroups independent of the actual effect size, or (ii) the resulting distance between two fields depends also on all other fields (e.g. the distance between '*Physics*' and '*Chemistry*' depends on whether one includes articles about '*Anthropology*' in the classification).

3. Results

We now present and interpret results obtained computing the three dissimilarity measures ($D_{\rm exp}, D_{\rm cite}$ and $D_{\rm lang}$) reported above for scientific fields i, j defined by papers published in different time intervals and categorized (by Web of Science) as belonging to the same specialty (e.g. *Applied Mathematics*), discipline, (e.g. *Mathematics*) or domain (e.g. *Natural Sciences*).

3.1. Comparison of dissimilarity measures

Figure 1 shows the three $D(i,j)$ at the level of specialties (i,j) for the complete time interval 1991–2014. The concentration of low $D(i,j)$ close to the diagonal shows that both the citations and language of scientific papers partially reflect the disciplinary classification done by the experts. However, visual inspection already reveals that citations and our language analysis show relationships not present in the expert classification, e.g. the low dissimilarity between *Engineering* and *Natural Sciences* (most clearly between *Electrical Engineering* and *Physical Sciences*) and between the disciplines inside the *Agriculture* domain and *Biological Sciences*.

We start by quantifying the relationship between the three different dissimilarity measures, i.e. ($D_{\rm exp}, D_{\rm cite}$ and $D_{\rm lang}$), across all pairs of specialties (i,j). In table 1, we report the rank correlation between the three measures, which we obtain from ranking for each dissimilarity the pairs of (i,j) according to $D(i,j)$. The choice of this non-parametric correlation is motivated by the fact that the range of the three measures differs dramatically (e.g. $D_{\rm exp} \in \{0,1,2,3\}$ and $D_{\rm lang} \in [0,1]$). The positive statistically significant correlation between all pairs of $D(i,j)$'s confirms the visual impression described above. The correlation between citations and language is higher than the correlation with the experts classification. Remarkably, language and citations show a very similar correlation with experts, but language is systematically less correlated than citations (p-value $= 1.8 \times 10^{-5}$ for Spearman-ρ and p-value $= 2.2 \times 10^{-5}$ for Kendall-τ[2]). We conclude that the language dissimilarity $D_{\rm lang}$ introduced here is able to retrieve the well-known relationships between disciplines to a similar extent with that of the (well-studied) citation analysis.

We now explore how the relationship between the different dimensions depends on the different scientific fields. The results in figure 2 confirm the conclusions of the aggregated analysis but show further interesting features. First, the correlation in $(D_{\rm exp}, D_{\rm lang})$ is smaller than $(D_{\rm exp}, D_{\rm cite})$ mainly in the natural sciences. Second, while the correlation between citations and language remains largely constant, large fluctuations in the correlations between expert and citations (as well as expert and language) exist. This is seen both as the strong downward spikes and also in the manifested dependence on disciplines and domains. The titles of the specialties at the low peaks already suggest that these are specialties with interdisciplinary connections. For instance, *Chemistry, Medicinal* is a specialty that (according to the experts classification) belongs to the discipline *Basic Medicine* and to the domain *Medical Science*. Therefore, $D_{\rm exp} = 3$ between *Chemistry, Medicinal* and all specialties of the *Natural Sciences* (in particular, for all specialties from the discipline *Chemical Sciences*). Instead, the dissimilarity measured by citations $D_{\rm cite}$ and language $D_{\rm lang}$ yields much smaller values revealing the proximity of *Chemistry, Medicinal* to the *Natural Sciences*, thus explaining the low correlation in $(D_{\rm exp}, D_{\rm cite})$ and in $(D_{\rm exp}, D_{\rm lang})$. The other downward spikes seen in the figure can also be understood as a consequence of the central role played by the *Natural Sciences* in other disciplines. To confirm this, we computed for a list of selected specialties

[2] Obtained from 10^3 bootstrapping samples of each joint distribution $P(D_{\rm exp}, D_{\rm cite})$ and $P(D_{\rm exp}, D_{\rm lang})$, i.e. comparison of 10^6 pairs of correlation values.

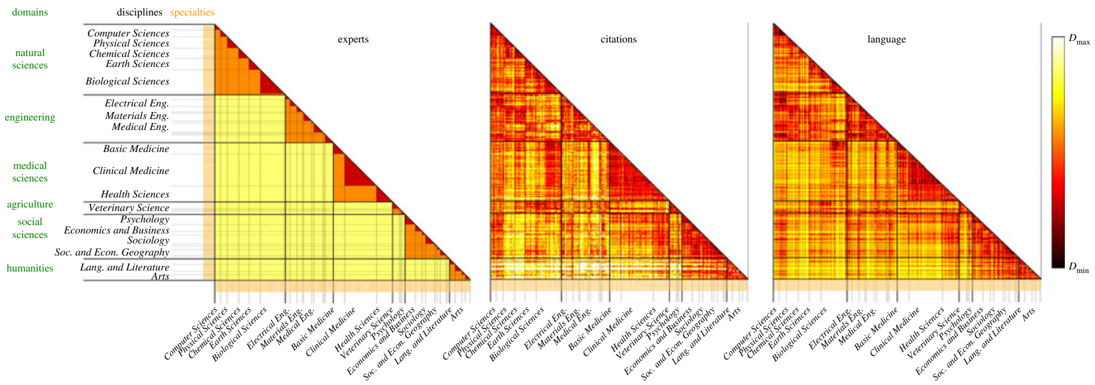

Figure 1. Dissimilarity between specialties measured in three different dimensions: (*a*) D_{exp} based on experts classification [31], where $D_{min} = 0$ and $D_{max} = 4$; (*b*) citations dissimilarity D_{cite} (2.1), where $D_{min} = 0$ and $D_{max} = 7.5$; (*c*) language dissimilarity D_{lang} (2.2), where $D_{min} = 0$ and $D_{max} = 1$. $N = 225$ specialties of the OECD classification scheme are considered. Results based on ≈ 21 M papers from 1991 to 2014 (see §5.1 for details).

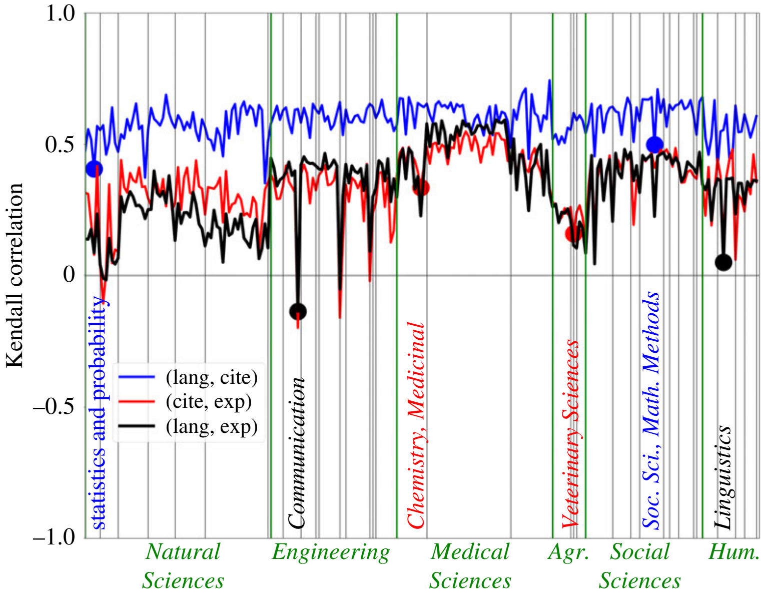

Figure 2. Correlation between the different dissimilarity measures varies across fields. The Kendall correlation $\tau(x, y)$ (shown in the vertical axis) for two measures x and y is computed between $D_x(i, j)$ and $D_y(i, j)$ over all specialties j for a fixed specialty i (shown in the horizontal axis). The three possible comparisons (x, y) are indicated in the caption. Six specialties (one from each domain) with low correlation are highlighted.

Table 1. Rank correlation between the dissimilarities $D_x(i, j)$ obtained from different dimensions $x \in \{\text{exp (experts)}, \text{cite (citations)}, \text{lang (language)}\}$ computed over all specialty pairs (i, j). (All values are significantly different from zero (p-values $< 10^{-5}$). The two values in each cell denote the Kendall-τ and Spearman-ρ (in parenthesis). Qualitatively equivalent results are obtained in three different time intervals (indicated in the left row).)

time	lang-cite	lang-exp	cite-exp
all, 1991–2014	0.57 (0.76)	0.32 (0.39)	0.34 (0.42)
first half, 1991–2002	0.60 (0.80)	0.34 (0.41)	0.37 (0.46)
second half, 2003–2014	0.64 (0.84)	0.35 (0.43)	0.38 (0.47)

$i = i^{\text{spikes}}$ the pairs (i, j) which suffered the largest change in ranks (comparing D_{exp} with D_{lang}), finding that nine from the top 10 specialties were from the domain *Natural Sciences* (five of them from the discipline *Chemical sciences*, including the top two specialties).

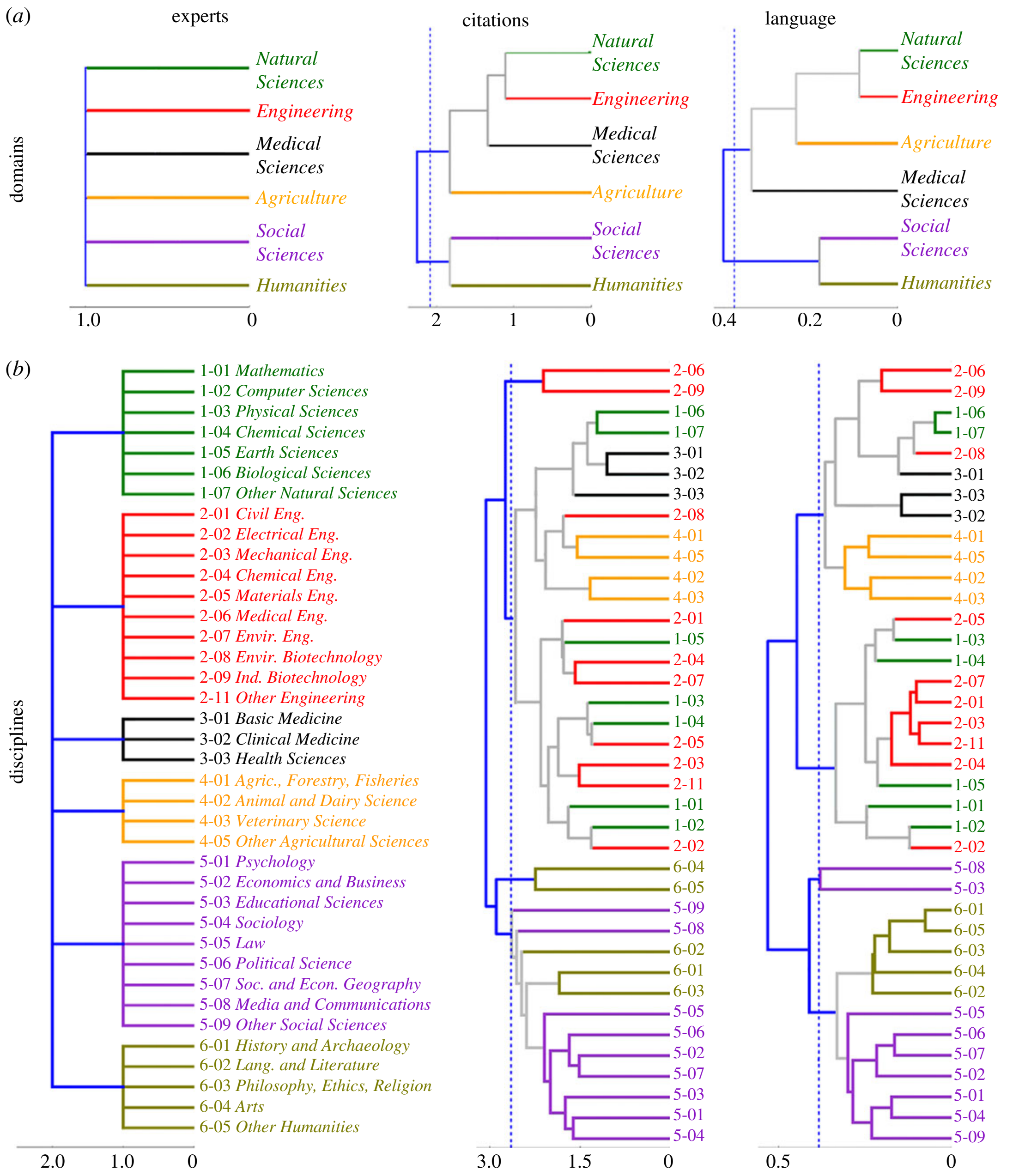

Figure 3. Hierarchical clusterings at the level of domains (*a*) and disciplines (*b*). Results for citations (language) were obtained by agglomerative hierarchical clustering, applying the group average method [36] to $D_{\text{cite}}(i,j)$ ($D_{\text{lang}}(i,j)$). The x-axis shows the clustering dissimilarity (i.e. the dissimilarity of two clusterings that are merged). The dashed line corresponds to a clustering dissimilarity equal to the percentile 0.92 of the values of all cluster dissimilarities at each measure (citations/language).

3.2. Hierarchical clustering

A strict hierarchical classification of scientific fields is both aesthetically appealing and of practical use in bibliographical and document classification tasks. It also allows us to further highlight the differences in the relationship between scientific fields revealed by the different dissimilarity measures (in particular by D_{lang}). While D_{exp} is precisely based on one such hierarchical classification, D_{cite} and D_{lang} are not. In figure 3, we show the hierarchical classifications induced by D_{cite} and D_{lang} through the computation of a simple clustering method at the level of domains and disciplines.

At the top level of the six domains (figure 3*a*), the clustering obtained from citations and from language are very similar. In particular, both identify *Engineering-Natural Sciences* and *Humanities-Social Science* as clusters that separate from the other domains in a similar fashion. The only difference is that,

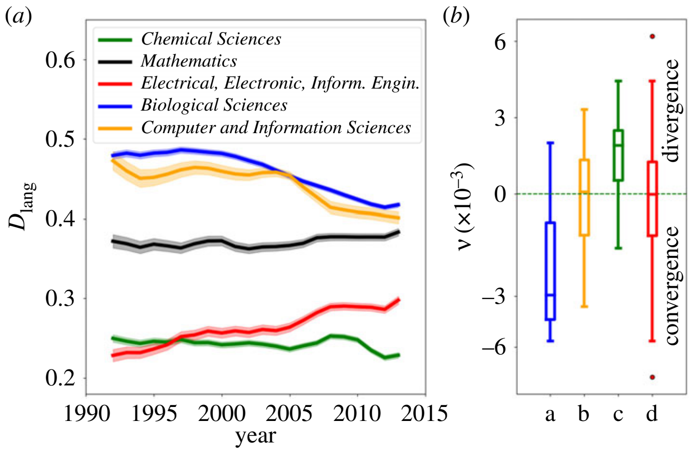

Figure 4. Evolution of the similarity between disciplines in the last three decades. (*a*) Distance $D_{\text{lang}}(i, j)$ between *Physical Sciences* (*i*) and other five selected disciplines (*j*) (3-year moving averages). (*b*) Total variation ν—defined in equation (3.1)—of the distance for pairs of disciplines with histories longer than 12 years. Each boxplot corresponds to the distribution of ν for pairs of disciplines where we fixed one of the disciplines. At position (*a*) we fixed *Computer and Information Sciences*, at (*b*) *Chemical sciences*, at (*c*) *Psychology* and at (*d*) we used all pairs of disciplines.

Table 2. Contribution of individual words *w* to the dissimilarity between *Physical Sciences* (*p*) and *Computer and Information Sciences* (*c*) in years 1991 (left) and 2014 (right). (The 10 words *w* that contributed most to D_{lang} in each year are shown. The relative contribution *C* is measured as the fraction of the total D_{lang} owing to word *w*. It is computed dividing the absolute contribution of word *w* to D_{lang}—obtained from equation (2.2) by using the single *w* in the numerator and all *w*'s in the denominator (the normalization factor)—by the total D_{lang} leading to $C = 0.5(p_c(w) - p_p(w))^2/(2H_2((\mathbf{p}_p + \mathbf{p}_c)/2) - H_2(\mathbf{p}_p) - H_2(\mathbf{p}_c))$. The last column shows the ratio of frequencies of word *w* in the two disciplines.)

year = 1991, $D^{(1991)}_{\text{lang}}(p, c) = 0.49$			year = 2014, $D^{(2014)}_{\text{lang}}(p, c) = 0.40$		
word *w*	*C*	$p_c(w)/p_p(w)$	word *w*	*C*	$p_c(w)/p_p(w)$
system	9.9%	3.981	algorithm	6.4%	14.134
algorithm	6.7%	30.196	propose	5.9%	5.091
problem	4.4%	10.355	problem	3.5%	11.388
paper	3.5%	7.782	paper	3.3%	5.107
temperature	2.3%	0.015	method	2.6%	2.246
language	2.2%	488.726	approach	2.6%	4.377
program	2.0%	34.313	data	2.5%	2.934
energy	1.9%	0.025	network	2.3%	11.313
field	1.5%	0.141	system	2.0%	2.079
set	1.3%	6.738	temperature	1.6%	0.015

based on citations, *Agriculture* appears more isolated, while based on language this happens for *Medical Science*. A more detailed picture of the differences between language and citation is revealed at the level of disciplines (figure 3*b*). While at the first division, both citations and language create a cluster in which all disciplines of the domains *Humanities* and *Social Sciences* appear, further divisions show more subtle differences between the two dissimilarity measures.

Remarkably, the hierarchy obtained from language creates a cluster containing all and only *Humanities* disciplines. By contrast, the hierarchy based on citations creates one clustering with three of the five *Humanities* disciplines (*Lang. and Literature*, *Arts* and *Other Humanities* while the two remaining ones (*History & Archaeology* and *Philosophy, ethics, religion*) are clustered together in the middle of a cluster of disciplines in *Social Science*. Another interesting difference between the clusterings is revealed looking at three disciplines of the domain *Medicine*: In the analysis based on citations, the minimum cluster that includes the three disciplines includes *Biological Sciences* and *Other Natural Sciences*, while in the language analysis, this cluster includes additionally three related *Engineering* disciplines (*Medical Eng.*, *Ind. Biotechnology* and *Environ. Biotechnology*).

Probably the most remarkable feature of the clustering obtained by, both, citations and language is that it repeatedly clusters together related disciplines from *Natural Sciences* with disciplines from *Engineering* and *Medicine* (e.g. *Chemical Sciences* and *Materials Science*). This clustering, not present in the experts classification, suggests that the distinction between fundamental and applied sciences present in the expert classification has no strong effect on citations and the language of the publications. Instead, in this specific case, the citation and language analysis seem to be capturing a connection between 'subject matters' that was necessarily absent from the strict hierarchical expert classification.

3.3. Temporal evolution

While in the previous sections we looked at a static snapshot of the relation between disciplines, here we are interested in how the linguistic relationship $D_{\text{lang}}(i,j)$ between pairs (i,j) of disciplines evolved over the last three decades.[3] In figure 4, we show the temporal evolution for five out of 703 pairs (i,j), with focus on the discipline *Physical Sciences*, illustrating different types of dynamic patterns. On the one hand, the dissimilarity to *Chemical Sciences* (its most similar discipline) and *Mathematics* stays roughly constant over time. On the other hand, we also observe systematic trends of disciplines becoming more or less similar over time. While the proximity to *Biological Sciences* and *Computer and Information Science* has steadily increased (decreased dissimilarity $D_{\text{lang}}(i,j)$) after the year 2000, the opposite trend is seen for *Electrical, Electronical and Information Engineering*. These observations are consistent with the increasing number of biological and computational-related publications in *Physics*, and with a departure from the historical connections to *Engineering*.

The observations reported above raise the question as to whether scientific disciplines are showing an overall tendency to become more similar to each other. In a more general context, this amounts to the question of whether the purported increase in *interdisciplinarity* leads to a larger overlap in the language used by different disciplines. We address this question by computing, for each pair of disciplines, the mean yearly variation

$$\nu(i,j) = \frac{1}{\Delta t}\sum_{t\in\Delta t} D^{(t)}_{\text{lang}}(i,j) - D^{(t-1)}_{\text{lang}}(i,j) \tag{3.1}$$

$$= \frac{1}{\Delta t}(D^{(t_f)}_{\text{lang}}(i,j) - D^{(t_0)}_{\text{lang}}(i,j)), \tag{3.2}$$

where the time interval $\Delta t \equiv t_f - t_0$ was usually from $t_0 = 1991$ to $t_f = 2014$. The distribution of values of ν for all discipline pairs (i,j) is shown at the (rightmost) box plot in figure 4*b*. We see that there are both positive and negative variations, consistent with our qualitative observations in the example of *Physical Sciences* in figure 4*a*. However, the average variation $\langle\nu\rangle \approx -0.00025$ over all pairs of disciplines (i,j) is not distinguishable from zero (the null hypothesis of $\langle\nu\rangle = 0$ has a p-value = 0.07 in the t-test for the mean of one sample and a p-value = 0.21 in the non-parametric Wilcoxon test), i.e. the typical dissimilarity remains unchanged. This result suggests that, while there are systematic trends for individual pairs of disciplines, on average there is no significant increase or decrease in the interdisciplinarity for the science as a whole in the last three decades as measured by the language.

On a more fine-grained level, however, we observe systematic trends that suggest that individual disciplines tend to become more (less) central. For this, we focus on the discipline pairs (i,j) which experienced the most extreme variation in the last decade (one standard deviation away from $\langle\nu\rangle$). These pairs have typically $|\nu| \gtrsim 0.003$, meaning that their (normalized) dissimilarity changes roughly 3% in a decade. The three disciplines that are most frequently seen in the left tail ($\nu < 0$) are: *1-02 Computer and Information Sciences*, *2-08 Environmental Biotechnology* and 3-01 *Basic Medicine*. The language of these disciplines became significantly more similar to the language of other disciplines in the last three decades, suggesting that these disciplines became more central. By contrast, the three disciplines that experienced most strongly the opposite effect (most frequently seen in the right tail, $\nu > 0$) are (in decreasing order): 5-01 *Psychology*, 2-05 *Materials Engineering* and 2-02 *Electrical Engineering, Electronic Engineering, Information Engineering*.

In the interpretation of the results reported in this section it is crucial to take into account that the measure D_{lang} we use depends only on the frequency of the words in each of the fields and in each year.

[3] We work at the level of disciplines because most specialities fail to have enough publications in a single year.

In particular, this means that the results can be interpreted as an absolute dissimilarity independent of the content or volume of other fields. Another advantage of our measure D_{lang} is that it allows us to quantify the contribution of individual words [28]. This general feature of our method is illustrated in table 2 and allows for a deeper interpretation of the meaning of D_{lang} (e.g. the contribution of topical words and stylistic differences).

4. Discussion

We investigated the similarity between scientific fields from different perspectives: an expert classification, a citation analysis and a newly proposed measure of linguistic similarity. We found that these different dimensions are related yet different, yielding thus new insights on the relationship between disciplines, their hierarchical organization and their temporal evolution.

Our first main finding is that the language and citation relationships between disciplines are similar and substantially different from the expert classification. This is consistent with the motivation exposed in our introduction which associated the expert classification to the (largely idealized) essentialist view of scientific disciplines, while the citation (social) and language (cognitive) were closer to dimensions that play a more important role in the relationship between fields. Interestingly, our results indicate that the language-relation of fields is more distinct from the expert classification than the citation-relation is, especially in the natural sciences.

Our second main finding is that in the last 30 years the language of different scientific fields remain, on average, at the same distance from all other fields. While individual disciplines show clear trends of increasing (or decreasing) centrality, this suggests that, overall, diverging tendencies in science (e.g. specialization) are in balance with converging tendencies (e.g. multidisciplinarism). This is a remarkable quantitative finding because of the substantial changes observed in this period.

The latter result demonstrates that our textual measure is of practical relevance for the study of interdisciplinarity. In recent years, interdisciplinary research achieved a central position [10] due to its broader relation to the concept of diversity [37] and its effect on the impact [38,39] and performance of teams [40] as well as its implications for policymaking, e.g. in terms of funding [41]. Is it just a fashion or science is really getting more and more interdisciplinary? A usual way to assess interdisciplinarity is based on citation networks using heuristic approaches [9,32,42] or methods from complex networks [43–46]. In line with the arguments exposed in the introduction, interdisciplinarity can be viewed through different dimensions and the cognitive dimension would be best measured using textual data. However, there are only very few works [47–49] relating textual measures with interdisciplinarity, despite the increasing availability of the text of scientific articles. In this view, the significance of our approach is that it provides a measure of interdisciplinarity based on how much the usage of words in different disciplines overlap.

Finally, we hope our results and methodology will stimulate a multiple-dimensional approach in other problems related to the study of sciences, profiting from the modern availability of large (textual) databases of scientific publications that allow us to go beyond traditional bibliometric analysis [1,9]. These include, but are not limited to, the formulation of more meaningful bibliometric indicators [50], the identification and prediction of influential papers and disciplines [51–53], or the inclusion of textual information in recommending related scientific papers [54].

5. Material and methods

5.1. Data and grouping of corpora

We use the Web of Science database [30] and explore the following information available for individual articles: citations, title, abstract and the classification in one scientific speciality (per OECD classification [31]). We use all papers published between 1991 and 2014 because, the number of articles with text in the abstract is substantial only after 1991 and because, at the time we started our analysis, 2014 was the last complete year available to us. The text of an article was built concatenating its title and abstract. The corpus representing a speciality in a given year is obtained from the concatenation of the text of all articles for that specialty in that year. The corpus for one discipline (or domain) concatenates all articles in all specialities belonging to that discipline (or domain).

Our analysis is based on 19 589 166 articles for each of the textual and classification information were available (92% of all articles indexed in Web of Science during 1991–2014). In our analysis we considered only citations from and to the papers in our list because only for these papers we had a reliable classification of specialties. These citations corresponded to roughly half of the ≈625 M citations associated with these papers. See [55] for the divergences we obtained from this dataset.

5.2. Data processing

For each article in our database, we performed the following steps to process the textual information:

(i) the abstracts written in a language different from English were excluded;
(ii) the copyright information contained in the abstract was removed;
(iii) title and abstract were concatenated;
(iv) the text was converted to lowercase;
(v) contractions were replaced by their non-contracted form, exact list and details are available in [55];
(vi) the text was tokenized, and the nouns and verbs were lemmatized using the Natural Language Toolkit [56] for English;
(vii) non-alphanumeric symbols (except hyphen) inside tokens were replaced by white space, therefore generating two or more distinct tokens;
(viii) tokens composed exclusively by numbers or single letter were removed; and
(ix) tokens belonging to a preset stop-word list were discarded; exact list and details available in [55].

5.3. Minimum corpus size

We computed D_{lang} using only the 20 000 most frequent word types, disregarding the scientific fields for which there were not enough data to achieve this cut-off. This is necessary in order to ensure that the estimations of D_{lang} are reliable. The crucial problem is the slow convergence of entropy estimations (and thus D_{lang}) which leads to strong uncertainties in entropy estimations even for large corpora [27]. By choosing a fixed number of word types, we reduce the effect of the remaining bias (in the estimation of D_{lang}) on our comparative analysis of textual dissimilarity between pairs of fields. This happens because the residual bias acts as an offset in all cases (when a fixed cut-off is chosen) instead of affecting differently each case (as obtained if the maximum amount of data is used in each case). The bias decays with the number of word types used because the more frequent types are responsible for almost all the dissimilarity, specially for $\alpha = 2$ [28]. Using 10 000 types as a cut-off, we estimated the textual dissimilarity relative standard deviation, computed over multiple samples of the same scientific field, to be $\hat{\sigma}(D_{\text{lang}})/D_{\text{lang}} \approx 1\%$. Our choice for a larger cut-off of 20 000 types is a conservative choice to ensure a smaller uncertainty of the estimations, i.e. $\hat{\sigma}(D_{\text{lang}})/D_{\text{lang}} < 1\%$.

5.4. Generalized Jensen–Shannon divergence

Given two texts (indexed by p and q), we define the probability distributions over all word types $w = 1, \ldots, V$ as $\mathbf{p} = (p_w)$ and $\mathbf{q} = (q_w)$. An information-theoretic measure to quantify their similarity is the generalized Jensen–Shannon divergence:

$$D_\alpha(\mathbf{p}, \mathbf{q}) = H_\alpha\left(\frac{\mathbf{p}+\mathbf{q}}{2}\right) - \frac{1}{2}H_\alpha(\mathbf{p}) - \frac{1}{2}H_\alpha(\mathbf{q}), \tag{5.1}$$

based on the generalized entropy of order α ($\in \mathbb{R}$), where

$$H_\alpha(\mathbf{p}) = \frac{1}{1-\alpha}\left(\sum_w p_w^\alpha - 1\right). \tag{5.2}$$

Here, we consider a normalized similarity [27]

$$\tilde{D}_\alpha(\mathbf{p}, \mathbf{q}) = \frac{D_\alpha(\mathbf{p}, \mathbf{q})}{D_\alpha^{\max}(\mathbf{p}, \mathbf{q})} \tag{5.3}$$

such that $\tilde{D}_\alpha \in [0, 1]$, where $D_\alpha^{\max}(\mathbf{p}, \mathbf{q}) = ((2^{1-\alpha} - 1)/2)(H_\alpha(\mathbf{p}) + H_\alpha(\mathbf{q}) + 2/(1-\alpha))$ is the maximum possible D_α between $\mathbf{p}$ and $\mathbf{q}$, assuming that the set of symbols in each distribution (i.e. the support of $\mathbf{p}$ and $\mathbf{q}$) are disjoint. In other words, if the two corpora have no common word, $\tilde{D}_\alpha = 1$.

Note that, for $\alpha = 1$, equation (5.2) yields the Shannon entropy [57], i.e. $H_{\alpha=1}(\mathbf{p}) = -\sum_w p_w \log p_w$, and $D_{\alpha=1}$ is the well-known Jensen–Shannon divergence [58]. Gerlach *et al.* [27] shows that $\alpha = 2$ provides the most robust statistical measure of similarity of texts. This motivates our choice of $D_{\text{lang}} = \tilde{D}_{\alpha=2}$.

Data accessibility. Deposited at Zenodo https://dx.doi.org/10.5281/zenodo.816303 [55].
Authors' contributions. L.D. processed the data and prepared the figures. E.G.A. retrieved the data. L.D., M.G. and E.G.A. analysed the data. L.D., M. G., J.S. and E.G.A. designed the research, interpreted the results, wrote the manuscript and gave their final approval for publication.
Competing interests. The authors declare no competing interests.
Funding. L.D. received financial support from CNPq/Brazil through the programme 'Science without Borders'.
Acknowledgements. We thank M. Palzenberger and the Max Planck Digital Library for providing access to the data, M. de Domenico for insightful discussions, and S. Haan and the Centre for Translational Data Science (University of Sydney) for helping with figures 1 and 3.

References

1. Evans JA, Foster JG. 2011 Metaknowledge. *Science* **331**, 721–725. (doi:10.1126/science.1201765)
2. Börner K, Chen C, Boyack KW. 2003 Visualizing knowledge domains. *Annu. Rev. Info. Sci. Technol.* **37**, 179–255. (doi:10.1002/aris.1440370106)
3. Shiffrin RM, Borner K. 2004 Mapping knowledge domains. *Proc. Natl Acad. Sci. USA* **101**, 5183–5185. (doi:10.1073/pnas.0307852100)
4. Boyack KW, Klavans R, Börner K. 2005 Mapping the backbone of science. *Scientometrics* **64**, 351–374. (doi:10.1007/s11192-005-0255-6)
5. Rosvall M, Bergstrom CT. 2008 Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA* **105**, 1118–1123. (doi:10.1073/pnas.0706851105)
6. Gläser J, Glänzel W, Scharnhorst A. 2017 Same data — different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics* **111**, 979–979. (doi:10.1007/s11192-017-2295-0)
7. Wang D, Song C, Barabási A-L. 2013 Quantifying long-term scientific impact. *Science* **342**, 127–132. (doi:10.1126/science.1237825)
8. Moreira JAG, Zeng XHT, Amaral LAN. 2015 The distribution of the asymptotic number of citations to sets of publications by a researcher or from an academic department are consistent with a discrete lognormal model. *PLoS ONE* **10**, e0143108. (doi:10.1371/journal.pone.0143108)
9. Larivière V, Gingras Y. 2014 *Beyond bibliometrics*. Cambridge, MA: MIT Press.
10. Noorden RV. 2015 Interdisciplinary research by the numbers. *Nature* **525**, 306–307. (doi:10.1038/525306a)
11. Popper KR. 1952 The nature of philosophical problems and their roots in science. *Brit. J. Philos. Sci.* **3**, 124–156. (doi:10.1093/bjps/III.10.124)
12. Balsiger PW. 2005 Transdisziplinarität. In *Systematisch-vergleichende Untersuchung disziplinenübergreifender Wissenschaftspraxis*. Paderborn, Deutschland: Wilhelm Fink Verlag.
13. Guntau M, Laitko H. 1991 On the origin and nature of scientific disciplines. In *World views and scientific discipline formation* (eds RW Woodward, RS Cohen), pp. 17–28. Netherlands: Springer.
14. Peel L, Larremoore DB, Clauset A. 2017 The ground truth about metadata and community detection in networks. *Sci. Adv.* **3**, e1602548. (doi:10.1126/sciadv.1602548)
15. Boyack KW *et al.* 2011 Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. *PLoS ONE* **6**, e18029. (doi:10.1371/journal.pone.0018029)
16. Garfield E, Sher IH, Torpie RJ. 1964 *The use of citation data in writing the history of science*. Philadelphia, PA: Institute for Scientific Information.
17. de Solla Price DJ. 1965 Networks of scientific papers. *Science* **149**, 510–515. (doi:10.1126/science.149.3683.510)
18. Chavalarias D, Cointet J-P. 2013 Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PLoS ONE* **8**, e54847. (doi:10.1371/journal.pone.0054847)
19. Kuhn T, Perc M, Helbing D. 2014 Inheritance patterns in citation networks reveal scientific memes. *Phys. Rev. X* **4**, 041036. (doi:10.1103/PhysRevX.4.041036)
20. McKeown K *et al.* 2016 Predicting the impact of scientific concepts using full-text features. *J. Assoc. Info. Sci. Technol.* **67**, 2684–2696. (doi:10.1002/asi.23612)
21. Lippincott T, Séaghdha DÓ, Korhonen A. 2011 Exploring subdomain variation in biomedical language. *BMC Bioinf.* **12**, 212. (doi:10.1186/1471-2105-12-212)
22. Evans E, Gomez C, McFarland D. 2016 Measuring paradigmaticness of disciplines using text. *Sociol. Sci.* **3**, 757–778. (doi:10.15195/v3.a32)
23. Braam RR, Moed HF, van Raan AFJ. 1991 Mapping of science by combined co-citation and word analysis. I. Structural aspects. *J. Am. Soc. Info. Sci.* **42**, 233–251. (doi:10.1002/(SICI)1097-4571(199105)42:4<233::AID-ASI1>3.0.CO;2-I)
24. Vilhena D, Foster J, Rosvall M, West J, Evans J, Bergstrom C. 2014 Finding cultural holes: how structure and culture diverge in networks of scholarly communication. *Sociol. Sci.* **1**, 221–238. (doi:10.15195/v1.a15)
25. Silva FN, Amancio DR, Bardosova M, d. F. Costa L, Oliveira ON. 2016 Using network science and text analytics to produce surveys in a scientific topic. *J. Informetrics* **10**, 487–502. (doi:10.1016/j.joi.2016.03.008)
26. Sienkiewicz J, Altmann EG. 2016 Impact of lexical and sentiment factors on the popularity of scientific papers. *R. Soc. Open Sci.* **3**, 160140. (doi:10.1098/rsos.160140)
27. Gerlach M, Font-Clos F, Altmann EG. 2016 Similarity of symbol frequency distributions with heavy tails. *Phys. Rev. X* **6**, 021009. (doi:10.1103/PhysRevX.6.021009)
28. Altmann EG, Dias L, Gerlach M. 2017 Generalized entropies and the similarity of texts. *J. Stat. Mech.: Theory Exp.* **2017**, 014002. (doi:10.1088/1742-5468/aa53f5)
29. Webb A. 2002 *Statistical pattern recognition*. New York, NY: Wiley.
30. Web of Science. See http://apps.webofknowledge.com.
31. Web of Science OECD Catgory Scheme. See http://ipscience-help.thomsonreuters.com/incitesLive/globalComparisonsGroup/globalComparisons/subjAreaSchemesGroup/oecd.html.
32. Porter AL, Rafols I. 2009 Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics* **81**, 719–745. (doi:10.1007/s11192-008-2197-2)
33. Leydesdorff L. 2008 On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. *J. Am. Soc. Info. Sci. Technol.* **59**, 77–85. (doi:10.1002/asi.20732)
34. Grosse I, Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver J, Stanley HE. 2002 Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys. Rev. E* **65**, 041905. (doi:10.1103/PhysRevE.65.041905)
35. Landauer TK, Laham D, Derr M. 2004 From paragraph to graph: latent semantic analysis for information visualization. *Proc. Natl Acad. Sci. USA* **101**, 5214–5219. (doi:10.1073/pnas.0400341101)
36. Sokal R, Michener C. 1958 A statistical method for evaluating systematic relationships. *Univ. of Kansas Sci. Bull.* **38**, 1409–1438.
37. Stirling A. 2007 A general framework for analysing diversity in science, technology and society. *J. R. Soc. Interface* **4**, 707–719. (doi:10.1098/rsif.2007.0213)
38. Uzzi B, Mukherjee S, Stringer M, Jones B. 2013 Atypical combinations and scientific impact. *Sci. (New York, N.Y.)* **342**, 468–472. (doi:10.1126/science.1240474)
39. Wang J, Thijs B, Glänzel W. 2015 Interdisciplinarity and impact: distinct effects of variety, balance, and disparity. *PLoS ONE* **10**, e0127298. (doi:10.1371/journal.pone.0127298)
40. Lungeanu A, Huang Y, Contractor NS. 2014 Understanding the assembly of interdisciplinary teams and its impact on performance. *J. Informetrics* **8**, 59–70. (doi:10.1016/j.joi.2013.10.006)
41. Committee on Facilitating Interdisciplinary Research; Committee on Science, Engineering, P. P. I. of Medicine; Policy, G. A. N. A. of Sciences; National Academy of Engineering, *Facilitating Interdisciplinary Research* (National Academies Press, 2004).
42. Wagner CS, David Roessner J, Bobb K, ThompsonKlein J, Boyack KW, Keyton J, Rafols I, Börner K. 2011 Approaches to understanding and measuring interdisciplinary scientific research

(IDR): a review of the literature. *J. Informetrics* **5**, 14–26. (doi:10.1016/j.joi.2010.06.004)

43. Pan RK, Sinha S, Kaski K, Saramäki J. 2012 The evolution of interdisciplinarity in physics research. *Sci. Rep.* **2**, 551. (doi:10.1038/srep00551)
44. Sayama H, Akaishi J. 2012 Characterizing interdisciplinarity of researchers and research topics using web search engines. *PLoS ONE* **7**, e38747. (doi:10.1371/journal.pone.0038747)
45. Sinatra R, Deville P, Szell M, Wang D, Barabási A-L. 2015 A century of physics. *Nat. Phys.* **11**, 791–796. (doi:10.1038/nphys3494)
46. Omodei E, Domenico MD, Arenas A. 2016 Evaluating the impact of interdisciplinary research: a multilayer network approach. *Netw. Sci.* **1**, 1–12. (doi:10.1017/nws.2016.15)
47. Bache K, Newman D, Smyth P. 2013 Text-based measures of document diversity. In *Proc. of the 19th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining - KDD '13, Chicago, IL, USA, 11–14 August 2013*, pp. 23–31. New York, NY: ACM Press.
48. Nichols LG. 2014 A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics* **100**, 741–754. (doi:10.1007/s11192-014-1319-2)
49. Evans ED. 2016 Measuring interdisciplinarity using text. *Socius: Sociol. Res. Dyn. World* **2**, 1–18. (doi:10.1177/2378023116654147)
50. Mann GS, Mimno D, McCallum A. 2006 Bibliometric impact measures leveraging topic analysis. In *Proc. of the 6th ACM/IEEE-CS joint Conf. on Digital libraries - JCDL '06*. New York, NY: ACM Press.
51. Gerrish S, Blei DM. 2010 A language-based approach to measuring scholarly impact. In *Proc. of the 27th Int. Conf. on Machine Learning (ICML-10), 21–24 June, 2010, Haifa, Israel*, pp. 375–382. USA: Omnipress.
52. Foulds J, Smyth P. 2013 Modeling scientific impact with topical influence regression. In *Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing*, *18–21 October 2013, Seattle, USA*, pp. 113–123. Stroudsburg, PA: Association for Computational Linguistics (ACL).
53. Whalen R, Huang Y, Sawant A, Uzzi B, Contractor N. 2015 Natural language processing, article content & bibliometrics: predicting high impact science. In *Quantifying and analysing scholarly communication on the Web (ASCW'15)*, *June 2015, Oxford, UK*.
54. Achakulvisut T, Acuna DE, Ruangrong T, Kording K. 2016 Science concierge: a fast content-based recommendation system for scientific publications. *PLoS ONE* **11**, e0158423. (doi:10.1371/journal.pone.0158423)
55. Dias L, Gerlach M, Scharloth J, Altmann EG. 2017 Data from: Using text analysis to quantify the similarity and evolution of scientific disciplines. Zenodo Data Repository (https://dx.doi.org/10.5281/zenodo.816303).
56. Bird S, Klein E, Loper E. 2009 *Natural language processing with python*. Sebastopol, Canada: O'Reilly Media Inc. See http://www.nltk.org/.
57. Cover TM, Thomas JA. 2006 *Elements of information theory*. New York, NY: Wiley-Interscience.
58. Lin J. 1991 Divergence measures based on the Shannon entropy. *IEEE Trans. Info. Theory* **37**, 145–151. (doi:10.1109/18.61115)