STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

An empirical investigation of the Tribes and their Territories: are research specialisms rural and urban?[1]

Giovanni Colavizza[*], Thomas Franssen[**] and Thed van Leeuwen[**]

[*] *gcolavizza@turing.ac.uk*
The Alan Turing Institute, British Library, 96 Euston Road NW1 2DB, London UK.

[**] *t.p.franssen@cwts.leidenuniv.nl; leeuwen@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Leiden University, Willem Einthoven Building, Kolffpad 1, 2333 BN Leiden NL.

Introduction

A long tradition of sociological research aims to understand the differences in the organizational and cognitive structure of scientific fields (Merton, 1974; Whitley, 1984; Fuchs, 1993). This sociological tradition was in its earlier years intimately connected with the emerging field of bibliometric methods and applications, originated in the 1960s with the work of Storer and Price (Storer, 1967; Price, 1970). However, the sociology of science and scientometrics have since the early 1980s drifted apart and attempts to reconcile them, or to reconcile the more theoretically inclined field of science and technology studies with scientometrics, have not had the desired effect (e.g. Leydesdorff, 1989). Recently, scholars have again argued for the need for interdisciplinary work bridging the sociology of science (Gläser & Laudel, 2016) or science and technology studies (Wyatt et al., 2016) with scientometrics. We take up these calls and explore ways to bridge the sociology of science with scientometrics.

One of the most influential sociological frameworks is presented in the work of Becher and Trowler (2001). They argue that epistemic structures have both a cognitive and a social dimension and that communication practices of their tribes, mirror (and thus reproduce) these structures. In their view, rural areas organize in many, small topics, thus resulting in fragmentation, urban specialisms instead organize into few, more populated topics. The main distinction between rural and urban specialism is made based on the number of topics studied within a community at a given time—low for urban specialisms, high for rural specialisms—and

[1] This work was in part supported by the Swiss National Fund with grants 205121_159961 and P1ELP2_168489.

the "people-to-problem" ratio, meaning the number of researchers involved in a research topic at any one time—high for urban specialism and low for rural specialisms. We hypothesize:

> Hypothesis 1: The number of topics being researched is high for rural specialisms and low for urban specialisms. In rural specialisms more, smaller topics are expected to be found, everything else being equal, while in urban specialism fewer, larger topics are expected to be found.
>
> Hypothesis 2: Rural specialisms have a low people-to-problem ratio and urban specialisms a high people-to-problem ratio.

The authors subsequently suggest that this difference has implications for publication practices. They argue that rural authors have a broader scope intellectually and move more freely between topics. As there is no clear agreement about the core problems, in each publication the argument has to be embedded explicitly in the previous literature of the specialism, across topical boundaries. In urban specialisms, on the other hand, references are highly specialized, as there is no need to contextualize the publication by referencing outside of the topic. We hypothesize:

> Hypothesis 3: in rural specialisms there are comparatively more core publications that are shared beyond topics, making the specialism more reliant on them overall. This is less the case in urban specialisms, where core publications are mostly restricted to a topic. By core we mean highly cited publications. Intuitively, there are more weak ties across topics in rural specialisms than urban ones, due to the need to embed arguments within the broader specialism and not just within the specific topic.

Becker and Trowler (2001) make several more points, but we limit ourselves to these hypotheses here. Furthermore, the question why a specialism is urban or rural in the first place is a crucial one that we cannot answer in the present analysis.

Operationalization

The first step into the operationalization of the rural and urban conceptualization of the structure of scientific fields is to proxy its two basic units of analysis: specialisms and topics. A specialism is a group of people (a community) focusing on related topics of research which communicates this research internally through specialized journals, conferences and seminars. A topic of research is a well-identified set of problems and related questions, recognized by the community as being of interest and part of it. For example, in the specialism of natural language processing, speech recognition is a topic. Topics can be individuated at different granularities.

> *We proxy a specialism by considering a community producing publications which are a-priori well-individuated (e.g. by publication venue).*
>
> *We then proxy a topic as a well-connected cluster in the bibliographic coupling network of the publications published by authors active in the specialism.*

Working with reference overlap bibliographic coupling networks, *we consider a well-connected cluster to be a connected component with a minimum edge weight on every internal edge. We thus use connected components to approximate topics,* where a connected component is a sub-graph where every node is connected to other nodes by at least a path. Crucially, specialisms and

topics must be defined identically for every field under consideration, to allow for comparisons. The proposed method preserves the benefit of simplicity of interpretation and does not require us to judge whether a topic should be identified as such but rather assumes that overlap in references identifies similarity between publications. In what follows we focus on networks of publications.

Given our operative definition of specialisms and topics, we propose to operationalize the hypotheses derived from the rural and urban analogy as follows:

1. *Number and size of topics (hp. 1)*: we remove edges at increasing weight thresholds. The connectivity of the network in terms of the number and size of its connected components gives us a way to measure the relative number and size of topics. According to hypothesis 1, rural specialisms will fragment into more topics given the same weight threshold than urban specialisms.

2. *People-to-problem ratio (hp. 2)*: we operationalize the second hypothesis through the number of authors (people) active per topic (connected component) at the same edge weight threshold.

3. *Core publications (hp. 3)*: we compare the concentration of citations across the whole specialism to identify core sources that are highly cited. We measure the effect of core sources on the overall network connectivity by removing them in order of received citations or at random. We expect the global reliance on core sources to be greater in rural specialisms, thus the impact of removing them first to be comparatively lower on the overall network structure and size of the giant component at the specialism level. By impact here we mean the relative importance of sources into connecting the network, therefore a rural specialism will initially disconnect less rapidly by removing core sources first. Urban specialisms should be more reliant on core sources at the level of large topics, therefore they shall be less impacted by the removal of core sources as soon as the network has fragmented into topics.

Data

We select ten specialisms, and corresponding datasets, within five disciplines (two specialisms each): history, computer science, astrophysics, literature and biology. Each dataset is extracted from Scopus and contains representative publications for every specialism covering 5 years, from 2011 to 2015 included.[2] These datasets are not comprehensive, but hopefully representative of the research published in the respective specialism. We consider the following specialisms: *A1, economic history*; *A2, history of science*; *B1, computer science, neural networks and machine learning*; *B2, computer science, natural language processing*; *C1, astrophysics, solar system*; *C2, astrophysics, cosmology and astroparticle physics*; *D1, literature, classics*; *D2, English literature*; *E1, biology, neuroscience*; *E2, molecular biology*.[3] Summary statistics for the

[2] Scopus has a coverage comparable to the Web of Science, especially for materials from 1996 onwards (Harzing & Alakangas, 2016). Furthermore, its coverage of conferences in computer science is better than the Web of Science, considering the specialisms being here investigated.

[3] The full list of journals is given in Colavizza et al. (2018).

datasets under consideration are given in **Error! Reference source not found.**. Importantly, the overall number of articles is comparable.

Specialism / Statistic	A1-ec_hist	A2-hist_sci	B1-NIPS	B2-ACL	C1-icarus	C2-JCAP	D1-classics	D2-eng_lit	E1_neuron	E2_MBE
Number of articles	816	1753	1845	1410	2088	2834	1369	924	1522	1326
Number of articles 2015	143	367	403	316	454	648	375	152	318	261
Number of articles 2014	137	357	411	286	431	658	229	160	343	277
Number of articles 2013	175	384	360	328	388	570	211	233	309	234
Number of articles 2012	178	352	370	188	402	538	308	207	286	261
Number of articles 2011	183	293	301	292	413	420	246	172	266	293

Table 1: Summary statistics for the datasets under consideration.

From the Scopus interface, all relevant research articles or conference papers are downloaded, including their references. In order to include source and non-source items in our analysis merging references to the same object is need. To do so we proceed as follows. Firstly, the authors are separated from the rest of the reference. Secondly, references without author are discarded. For two references to be merged into the same object cluster, three things need to happen: 1) the surnames of the first authors need to match; 2) the two lists of authors need to have a Jaro-Winkler score of 0.9 or above; 3) the rest of the reference text needs to have a Jaro-Winkler score above a threshold determined for each specialism/dataset. This last threshold is established empirically by finding the score yielding an accuracy of less than 0.5 in the 100 pairs of references to be merged with a score just below that threshold. Similarly, the 100 pairs immediately above the threshold must yield an accuracy above 0.5. The intuition is that the accuracy of matches above the threshold rapidly improves, as it rapidly deteriorates below the threshold, therefore yielding acceptable results.

Methods

In this section we introduce the bibliographic coupling network relying on reference overlap. Take $B = (V, E, W)$, the weighted bibliographic coupling network made of the publications of a specialism in a given year or over few contiguous years. W is the weighted, symmetric

adjacency matrix. We weight the edges by considering the cosine similarity over the references that two publications i and j have in common:

$$W_{i,j} = \frac{R_{i,j}}{\sqrt{R_i}\sqrt{R_j}} \qquad (1)$$

Where $R_{i,j}$ is the number of references in common between i and j, R_i the number of references of i. The cosine similarity is particularly appropriate as it allows to evenly compare the weight of edges among publications with different reference list lengths.

A connected component of B is a sub-graph whose nodes are all connected, i.e. there exists a path between every pair of nodes in the component. An isolated node is an individual connected component. The giant component is the largest connected component measured by the number of nodes it contains (Newman 2010, 142-3). In order to explore the connectivity property of different specialisms, we measure the proportion of connected components over the total possible, and the proportion of nodes in the giant component, at steps in which we remove all edges below a certain weight threshold t. Given our operationalization of topics, the method allows to compare the number and size of topics at different granularities, across specialisms.

In practice, given an edge weight threshold t, we are interested in two measures, calculated at increasing t:

$$c(t) = \frac{C^t}{N} \qquad (2)$$

$$g(t) = \frac{G^t}{N} \qquad (3)$$

Where N denotes the number of publications, or nodes in the specialism network, which is also equal to the maximum number of connected components in the disconnected network; C^t denotes the number of connected components after removal of edges with weight strictly below t; G^t denotes the number of nodes in the giant component after removal of edges with weight strictly below t.

A complementary view on the granularity of topics in different specialisms can be given by considering the connectivity properties of the network when removing highly cited sources. The network will fragment after the removal of a proportion of highly cited sources, but it will do so at different speeds and times. Crucially, the more the specialism globally relies on shared sources (i.e. cited across topics), the less rapidly the network will initially fragment during such process; the more the specialism topically relies on core sources (i.e. cited within topics), the less rapidly the network will fragment once topics have been reached during such process.

We compare two processes considering the directed citation network of a specialism: one where we remove increasing fractions of cited sources in reverse order by the number of citations they received (from high to low), another where we remove cited sources at random. We then construct the reference overlap bibliographic coupling network and inspect its connectivity properties at regular intervals, as done in the previous subsection.

Results

We consider the reference overlap network (built using Eq. 1). We group humanities specialisms (A and D) and science specialisms (B, C and E), following the hypothesis that humanities specialisms are more rural, and science specialisms more urban. Consider first Equation 2. In Figure 1 we plot $c(t)$ on the y axis versus t on the x axis, averaging results over the humanities and the sciences. In Figure 2 we give the same results, averaged over every specialism instead. Consider next Equation 3. In figure 3**Error! Reference source not found.**, we plot $g(t)$ on the y axis versus t on the x axis, with the same set-up as in Figure 1. Results for the size of the giant component are coherent with those for connectivity.

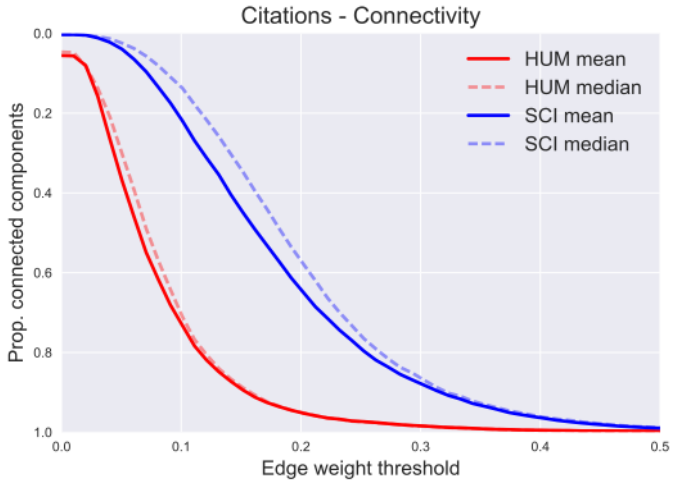

Figure 1: Mean and median connectivity of the network, grouped into the humanities (red/grey) and the sciences (blue/dark grey).

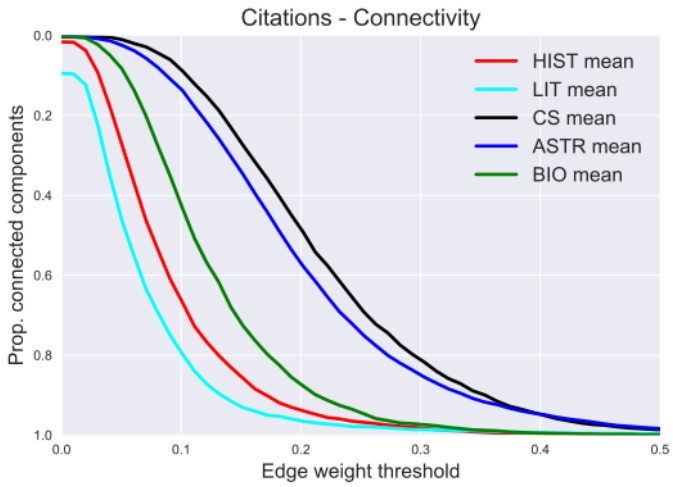

Figure 2: Mean connectivity of the network, by specialism. Legend: HIST: A, LIT: D, CS: B, ASTR: D, BIO: E.

Our results clearly highlight a lower overall connectivity for specialisms in the humanities. Individually, specialisms behave differently, with biology (E) being closer to history than astrophysics in this respect. Nevertheless, all scientific specialisms have higher connectivity than specialisms in the humanities, indicating that research topics are finer-grained in the humanities than in the sciences, as discussed in hypothesis 1.

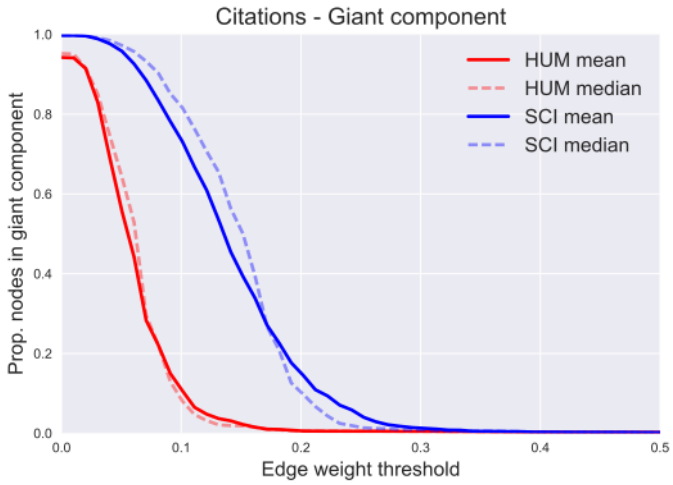

Figure 3: Mean and median size of the giant component of the network, grouped into the humanities (red/grey) and the sciences (blue/dark grey).

Given hypothesis 1 (i.e. urban specialisms maintain a higher connectivity, due to the presence of a lower number of larger topics), hypothesis 2 follows if urban specialisms also have more authors overall and more co-authorships, at the same weight threshold. Urban specialisms in our dataset indeed possess a higher number of unique authors (detected by exact matching within specialisms). In

we report the average and median size of the connected components with more than one node, and the corresponding people-to-problem ratio (number of unique authors per connected component), at different weight thresholds. The size of topics and, especially, the people to problem ratio are sensibly higher for scientific specialisms, as expected.

Specialism / Statistic	Threshold t=0.1		Threshold t=0.2		Threshold t=0.3	
	Topic size	p-t-p	Topic size	p-t-p	Topic size	p-t-p
A1 Economic History	4 (2)	5.9 (4)	2.3 (2)	3.5 (3)	2 (2)	2.9 (3)
A2 History of Science	7.7 (2)	7.9 (2)	2.8 (2)	2.9 (2)	2.4 (2)	2.2 (2)
B1 NIPS	93.4 (2)	159.5 (5)	6.6 (2)	14.9 (7)	3 (2)	6.9 (6)

B2 ACL	100.8 (2)	164.4 (6)	10.3 (2)	20.3 (6)	3.8 (2)	9.2 (6)
C1 Icarus	69.1 (2)	164.8 (10)	5.4 (3)	19 (11)	2.8 (2)	9.6 (7)
C2 JCAP	34.2 (2)	79.1 (5.5)	6.2 (2)	15.6 (6)	3.1 (2)	7 (5)
D1 Classics	4.8 (2)	4.4 (2)	2.6 (2)	2.3 (2)	2.6 (2)	2.4 (2)
D2 English Literature	3.4 (2)	3.6 (2)	2.3 (2)	3 (2)	2.2 (2)	3.2 (2)
E1 Neuron	14.8 (2)	89.3 (20)	3 (2)	19 (13)	2.5 (2)	15.6 (11)
E2 MBE	8.6 (2)	35.2 (12)	2.8 (2)	9.6 (7)	2.4 (2)	9.1 (5)

Table 2: Size of connected components mean (median) and people-to-problem ratio mean (median), calculated at different thresholds considering components with more than one node. By people we consider unique authors active in the component.

Moving to consider the reliance of specialisms on their cited sources, we show results in Figure 4 for connectivity, averaging as before over the humanities and the sciences. In this case, a process of removal in order of received citations is compared with one where edges where removed at random.

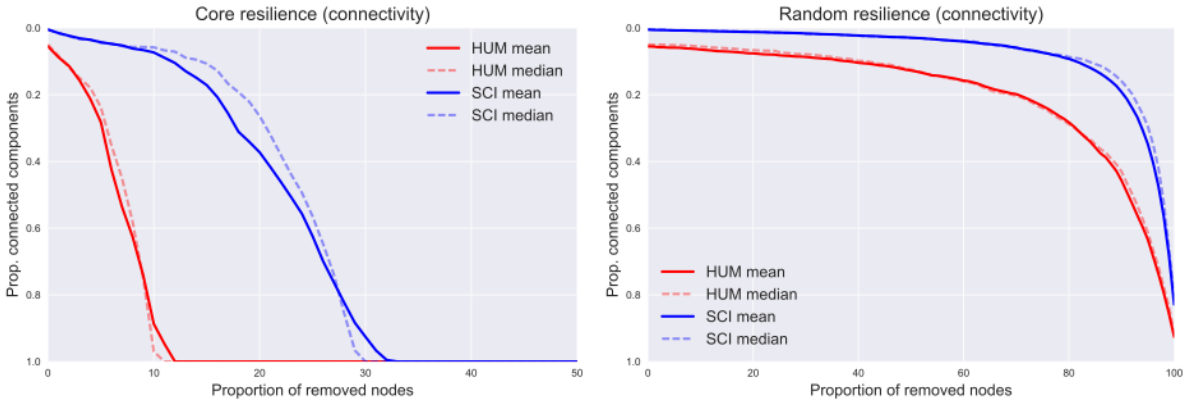

Figure 4: Connectivity of the networks to the removal of highly cited sources first (left) or at random (right), divided in the humanities (red/grey) and the sciences (blue/dark grey). The proportion of removed nodes is in %.

Two results emerge. Firstly, scientific specialisms are more reliant on core literature both at the specialism level and at the topic level, witness the higher resilience of their bibliographic coupling networks to the removal of the core literature at all stages. Secondly, the humanities present less well-connected bibliographic coupling networks in general, as shown by their lower connectivity apparent at all stages of the random removal process. To be sure, this phenomenon is much stronger in literature than history. We might conclude that the humanities possess a more fragmented intellectual base in terms of reference overlap, and in particular share fewer core sources at the specialism level, which in part contradicts what was stated in hypothesis 3. If rural

specialisms are fragmented into small topics, as shown, and at the same time possess fewer core sources, it follows that their fragmentation is in part due to more focused, either topic-specific or unique referencing behaviour.

Discussion

We have offered possible quantitative operationalization of Becher and Trowler's (2001) conceptualization of the social and cognitive structure of research specialisms as rural or urban, by comparing the reference connectivity among publications within ten humanities and science specialisms. We used publication venues (journals) to proxy specialisms and well-connected clusters in the bibliographic coupling network of publications to proxy topics. We find that science specialisms are overall better connected than in the humanities, with some disciplinary variations. We also find strong supporting evidence for a considerably higher people-to-problem ratio in urban specialisms, or the number of active authors per topic. However, topics are not that easily defined and we do not find many distinct clusters in any specialism. This leads us to argue that within the sciences, specialisms are comparatively well-connected at both the level of general, larger topics and the level of smaller, tighter ones, but that the distinction between these two levels of the cognitive structure is not as clear as Becher and Trowler suggest. Within the humanities we find a comparatively lower connectivity also at the general level of the specialism. This means that we do not find any evidence for the idea that humanities scholars tend to cite more broadly to establish an intellectual base for their contribution within the specialism as a whole.

In light of these findings we suggest to re-evaluate the use of the rural versus urban conceptualization. Rather, we find that science specialisms show an overall cohesion that suggests that scholars work in a particular paradigm in which topics are not necessarily clearly distinguished. The overall fragmentation of the humanities specialism suggests instead a less unified cognitive structure, at least to the extent to which this is articulated through reference lists and textual similarity. It remains an open question if in these humanities fields a particular paradigm is dominant without scholars having to articulate it or having to make reference the historical sources that lay at the basis of this paradigm.

Bibliography

Becher T. and Trowler P. (2001). *Academic tribes and territories: intellectual enquiry and the culture of disciplines*. Open University Press, Philadelphia.

Colavizza G., Franssen T. and van Leeuwen T. (2018). "An empirical investigation of the Tribes and their Territories: are research specialisms rural and urban?" Under review for the *Journal of Informetrics*.

Fuchs S. (1993). "A sociological theory of scientific change." *Social Forces*, 71(4): 933–953.

Gläser J. and Laudel G. (2016). "Governing science." *European Journal of Sociology*, 57(01): 117–168.

Harzing, A.-W. and Alakangas, S. (2016). "Google Scholar, Scopus and the Web of Science: A

Longitudinal and Cross-Disciplinary Comparison." *Scientometrics* 106(2): 787–804.

Leydesdorff L. (1989). "The relations between qualitative theory and scientometric methods in science and technology studies: Introduction to the topical issue." *Scientometrics*, 15(5-6): 333–347.

Merton R. K. (1974). *The sociology of science: theoretical and empirical investigations*. University of Chicago Press, Chicago.

Newman M. (2010). *Networks: an introduction*. Oxford University Press, Oxford.

Price D. De Solla. (1970). "Citation measures of hard science, soft science, technology, and nanoscience." In C. E. Nelson and D. K. Pollock (editors), *Communication among scientists and engineers*, pages 3–22. Heath Lexington Books, Lexington Mass.

Storer N. W. (1967). "The hard sciences and the soft: some sociological observations." *Bulletin of the Medical Library Association*, 55(1): 75–84.

Whitley R. (1984). *The intellectual and social organization of the sciences*. Oxford University Press, Oxford.

Wyatt S., Milojevic S., Park H. and Leydesdorff L. (2016). "The intellectual and practical contributions of scientometrics to STS." In U. Felt, R. Fouché, C. A. Miller and L. Smith-Doerr (editors), *The handbook of science and technology studies*, pages 87–112. The MIT Press, Cambridge MA.