Geographical Trends in Research Conferences: Closed Clubs or Open Houses?

Andrea Mannocci, Francesco Osborne, Enrico Motta

Knowledge Media Institute, The Open University, Milton Keynes, UK
name.surname@open.ac.uk

1 Introduction

Over the last decade, research started to scale up in terms of produced volume of papers, authors and contributing institutions. Nowadays, research literature is estimated to include 100-150 million publications with an annual increase rate around 1.5 million new publications [1]. Such a complex, global-scale system is worth studying in order to understand its dynamics and internal equilibria [3]. In particular, the study of authors' affiliations [4,8] has concrete impact on the interpretation of research as a complex phenomenon inserted in a delicate socioeconomic and geopolitical scenario.

In this study, we present an analysis on a dataset of conference proceedings metadata covering the 1996-2017 period, which was distilled from SciGraph[1], a free linked open data (LOD) dataset about scholarly knowledge published and curated by Springer Nature. In particular, we run a *macro analysis* on the full dataset, including conference proceedings across several scientific disciplines (e.g. computer science, life sciences, chemistry, engineering) and a *micro analysis*, which focuses on three high-tier conferences close to our area of expertise: the International Semantic Web Conference (ISWC), the Extended Semantic Web Conference (ESWC), and the International Conference on Theory and Practice of Digital Libraries (TPDL).

Firstly, we found that, over the observed period, the distributions of institutions and papers among countries follow a power law, consistently to what previously demonstrated in the literature across the 1981-2010 period [7,5,2,8]. Therefore, very few subjects keep producing most of the papers accepted by scientific conferences. Secondly, we highlight an increasing gap between the number of institutions initiating and overseeing research endeavours (i.e. first and last authors' affiliations) and the total number of institutions participating in research. Finally, we show how the annual and overall turnover rate of the top 5, 10 and 25 countries is extremely low, suggesting a very static landscape in which new players struggle to emerge.

[1] Springer Nature SciGraph, https://www.springernature.com/gp/researchers/scigraph

2 Data and Methodology

The reference dataset in our analysis is SciGraph, a LOD dataset published and curated by Springer Nature. To the best of our knowledge, SciGraph is the only free, large-scale dataset providing reconciliation of authors' affiliations by disambiguating and linking them to an external authoritative datasets in terms of institutions (in this case GRID, the Global Research Identifier Database[2]).

For our analysis we focused on conferences proceedings as conferences are the focal point of networking and knowledge exchange among practitioners. Since we intended to address both general and specific trends, we performed a *macro analysis*, on the full dataset, and a *micro analysis*, on specific conferences.

In the *macro analysis* we considered all conferences in the 1996-2016 period (2017 being discarded because of low data quality and coverage). For the *micro analysis* we focused instead on three high-tier conferences in the fields of semantic web and digital libraries: ISWC, ESWC, and TPDL; three top-tier conferences close to our field of expertise that attract quite different demographics. We extracted and manually curated (with any possible mean; i.e. familiarity with the conference domain, records comparison, and ultimately web search) the datasets about the conferences taking care to check and resolve affiliations details (i.e. gridId, organisation name, city, and country) for each contribution[3]. Table 1 summarises key features of the resulting datasets.

For each dataset, we took into consideration authors' order, and hypothesised (and validated) that the first author indicates the *initiator* of a research effort, while the last author indicates the professor or the research line manager acting as an *overseer* of the work. Of course this hypothesis does not hold in all the cases and does not reflect a common custom for several academic disciplines (e.g. Humanities & Social Sciences); however, it is a good approximation that works well for this study.

Under these assumptions, we analysed trends about papers, countries, and institutions over time, as well as their distributions across the entire observed period. Moreover, we tried to assess to what extent the research landscape is open (or closed) to changes by measuring the variability of country rankings over the years. The analysis has been run within a Python notebook, available online[4]. The full study, with extensive details and pictures, can be found in [6].

3 Results and Discussion

Results from our analysis essentially confirm that what can be observed at macroscopic level is also verified for single conference series.

[2] GRID, `https://www.grid.ac`

[3] For the sake of clarity, if paper p is authored by authors a_1 and a_2, two distinct *contributions* are present in our dataset, one for each author.

[4] `http://nbviewer.jupyter.org/github/andremann/SAVE-SD-2018/blob/master/Analysis.ipynb?flush_cache=true`

First and foremost, we confirmed the power law characteristic of the distributions of papers and institutions among countries already observed by previous studies in the literature across the 1981-2010 period [7,5,2,8]. Moreover, the number of papers produced and the number of unique contributing institutions are highly correlated, hence when we address the recent explosion of academic literature we cannot overlook the fact that the same has happened to institutions participating to research.

The second aspect emerged suggests that there is a growing gap between the number of institutions that initiate (first author) and overseer (last author) a research endeavour versus the total number of institutions involved in research. Interestingly, the gap grows even though the average number of authors per paper oscillates steadily between 2.6 and 3.3 over time. This suggests that, despite the fact that the set of institutions appearing as first/last authors' affiliation differs and expands from year to year, an ever-growing number of institutions every year takes part to research as a collaborators (nor first, nor last authors) without, however, leading the effort in any other research paper (at least for the year in consideration).

Likewise, very similar trends can be observed for countries in function of author position. Also in this cases, a less evident, yet undeniable, gap is present between the number of first/last authors' countries of affiliation and the total number of countries involved in research. Again the gap appears to widen over time.

The third aspect we observed is that the average variability (i.e. new countries breaking in) of the top-5, top-10 and top-25 countries across the observed period is rather static (about 10%), and verified that the top-5 and top-10 positions are visited by a closed club of countries. Moreover, results from the macro analysis show that (i) year by year it is fairly hard for outsider countries to break in a top-n, and (ii) that it gets harder and harder as the top-n set broadens, suggesting an ever more static scenario than we would imagine. In general, this reflects the intuition that well-formed research communities exhibit a sort of resistance towards the permeation of outsiders not always sharing knowledge and best practices consolidated over the years. A much clearer view about openness/closeness of conferences and research communities could be achieved by having access to data about rejected papers held in conference management systems such as EasyChair[5] or ConfTool[6].

In conclusion, this preliminary study shows some first important results and paves the way to further, in depth analysis on less evident conference dynamics, taking into account additional factors such as authors' role, seniority and affiliation type (i.e. academic or corporate), as well as external socioeconomic and geopolitical factors peculiar to conferences such as geographic distance, budget availability for travels, travel bans.

[5] EasyChair conference management system, `http://easychair.org`
[6] ConfTool conference & event management software, `http://www.conftool.net`

	Macro analysis	Micro analysis		
		ISWC	ESWC	TPDL
observation period	1996-2016	2003-2016	2004-2017 (excl. 2007)	2003-2017
contributions	1,664,733	3,924	4,224	3,271
unique papers (DOIs)	477,921	1,028	1,141	919
countries	163	44	54	52
institutions (gridIDs)	14,773	3,739	4,076	3,208
conference series	1,016	-	-	-

Table 1: Features of the datasets used for our analysis

References

1. Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 66(11), 2215–2222 (nov 2015)
2. Egghe, L.: Power laws in the information production process: Lotkaian informetrics. Emerald Group Publishing Limited (2005)
3. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., Barabási, A.L.: Science of science. Science 359(6379) (mar 2018), `http://barabasi.com/f/939.pdfhttp://www.ncbi.nlm.nih.gov/pubmed/29496846http://www.sciencemag.org/lookup/doi/10.1126/science.aao0185`
4. Frenken, K., Hardeman, S., Hoekman, J.: Spatial scientometrics: Towards a cumulative research program. Journal of Informetrics 3(3), 222–232 (2009)
5. King, D.A.: The scientific impact of nations (2004)
6. Mannocci, A., Osborne, F., Motta, E.: Geographical trends in research: a preliminary analysis on authors' affiliations. In: Proceedings of the SAVE-SD 2018 Workshop co-located with the 27$^{\text{th}}$ edition of the Web Conference (WWW), Lyon, France (2018)
7. May, R.M.: The scientific wealth of nations. Science 275(5301), 793–796 (1997)
8. Pan, R.K., Kaski, K., Fortunato, S.: World citation and collaboration networks: Uncovering the role of geography in science. Scientific Reports 2 (2012)