Do popular research topics attract the most social attention? A first proposal based on OpenAlex and Wikipedia

Wenceslao Arroyo-Machado[*] and Rodrigo Costas[**]

[*]*wences@ugr.es*
0000-0001-9437-8757
Department of Information and Communication Sciences, University of Granada, Spain

[**]*rcostas@cwts.leidenuniv.nl*
0000-0002-7465-6462
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherland
DST-NRF SciSTIP, Stellenbosch University, South Africa

Abstract
Altmetric research has seen its horizons expanded to the heterogeneity of interactions produced between scientific and non-scientific entities. In this context, Wikipedia stands out as a social media of particular interest as the page views of its articles have proven to be a valuable metric of social attention. The aim of this paper is to contribute to this new research stream by analysing whether the research topics of greatest academic interest align with those that attract the most social attention. To this end, the OpenAlex concepts are explored by comparing their work counts with the page views of their respective Wikipedia articles. As a result, a correlation analysis between the two metrics reveals a lack of connection between the two realms. Likewise, root-level concepts are explored to illustrate such a difference.

1. Introduction

Altmetric studies have moved away from the mere counting of social media mentions to scholarly outputs, to now encompass the broad research on the social interactions surrounding science (Díaz-Faes et al., 2019). The diversity of interactions and relationships produced between scientific and non-scientific entities is thus seen as a new paradigm for approaching altmetrics studies (Costas et al., 2020). Multiple proposals are emerging from this perspective, seeking to explore and take advantage of the vast universe of possibilities that this new paradigm opens. Twitter, the most prolific altmetric source in terms of both activity and studies, has demonstrated the potential and usefulness of this new conceptualization through proposals that delve into the different types of engagement produced around tweets and users (Fang et al., 2022), as well as social network analysis that overlap topic interests with social relations (Arroyo-Machado et al., 2021). However, a major question to be resolved in Twitter is how representative its activity is of general social attention and how far it extends beyond the academic world as far as scientific discussion is concerned.

This problem is diluted in the case of other social media, such as Wikipedia. This free encyclopaedia has a whole universe of metric possibilities from which the different interactions of the activity produced in its contents can be captured (Arroyo-Machado, Torres-Salinas, et al., 2022b). The *page views* of Wikipedia have been proposed as a valuable metric of broader social attention, even capable of capturing global trends (Yoshida et al., 2015). There are several cases in which a correlation has been found between Wikipedia *page views* and other phenomena of a markedly social nature, such as market trends (Gómez-Martínez et al., 2022), tourism (Donovan et al., 2017), or disease monitoring (Generous et al., 2014). Positive relationships have also been found with science, for example between the academic prestige of universities and their social attention (Arroyo-Machado, Díaz-Faes, et al., 2022). However, further efforts are still needed not only to explore Wikipedia's potential as a source for

measuring and tracking social attention but also to shed light on the science-society relationships.

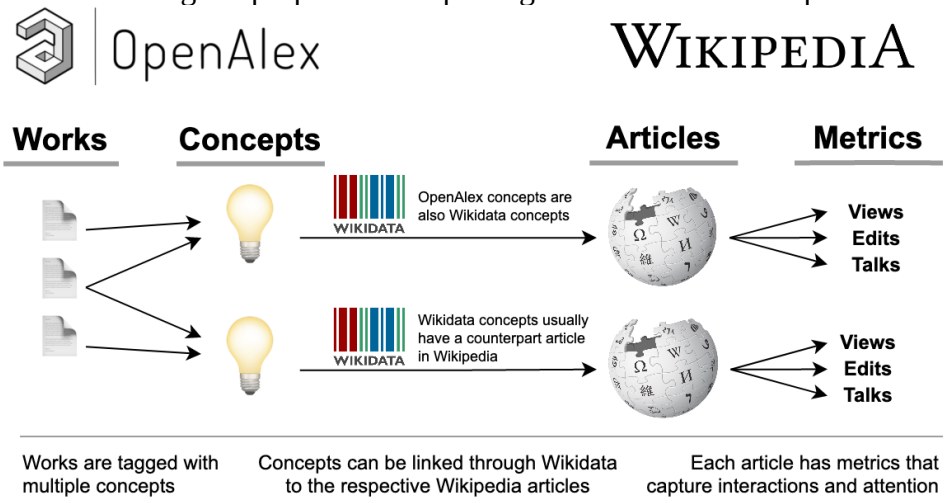
To further explore the science-society relations capturable through Wikipedia, the main objective of this article is to analyse whether the research topics of greatest academic interest are in line with those that attract most social attention on Wikipedia. For this purpose, scholarly outputs counts and Wikipedia *page views* are used as a proxy for academic interest and social attention, respectively. The following specific objectives have been established:

- To calculate the correlation between the volume of scholarly outputs and Wikipedia *page views* of research topics.
- Explore how the attention offered by the two metrics may differ by comparing the volume of academic publications and *page views* on Wikipedia for major and broad research topics.

2. Methodology

For this analysis, data from OpenAlex and Wikipedia were combined. From OpenAlex we extracted the metadata of its *concepts*, the hierarchical thematic classification inherited from Microsoft Academic Graph and which are assigned to the scholarly outputs by means of an automated classifier. On 24 August 2022, a total of 65,073 concepts were retrieved from OpenAlex through its API. From all the metadata, the level to which the concept belongs (there is a hierarchy of 6 levels), total counts of scholarly outputs assigned to the concept (*works count*) and total citations (*cited by count*) have been selected. Moreover, an important advantage of OpenAlex is that each OpenAlex concept is a Wikidata concept, a term describing a real or abstract conceptual entity, most of which are linked to a counterpart Wikipedia article (Figure 1). For example, the concept *Computer science* (Q21198[1]) is directly linked to the article *Computer science*[2]. Thus, from this relationship we have identified a total of 63,933 Wikipedia English articles (98% of the total), whose associated *page views* have been used as a proxy for the social attention of such a concept. Specifically, through the Wikipedia Knowledge Graph dataset (Arroyo-Machado, Torres-Salinas, et al., 2022a) we have obtained for each Wikipedia article its number of *page views*, which is representative of the period from April 1 to June 30, 2021. Table 1 summarises the number of concepts and Wikipedia articles by level.

Figure 1: Methodological proposal for capturing social attention of OpenAlex concepts



[1] https://www.wikidata.org/wiki/Q21198
[2] https://en.wikipedia.org/wiki/Computer_science

Table 1. Number of OpenAlex concepts and respective English Wikipedia articles per level.

	Number of concepts	%	English Wikipedia articles
Level 0	19	0,03%	19
Level 1	284	0,44%	271
Level 2	21,460	32,98%	21,090
Level 3	24,768	38,06%	24,372
Level 4	12,406	19,06%	12,196
Level 5	6136	9,43%	5985
Total	65,073	100%	63,933

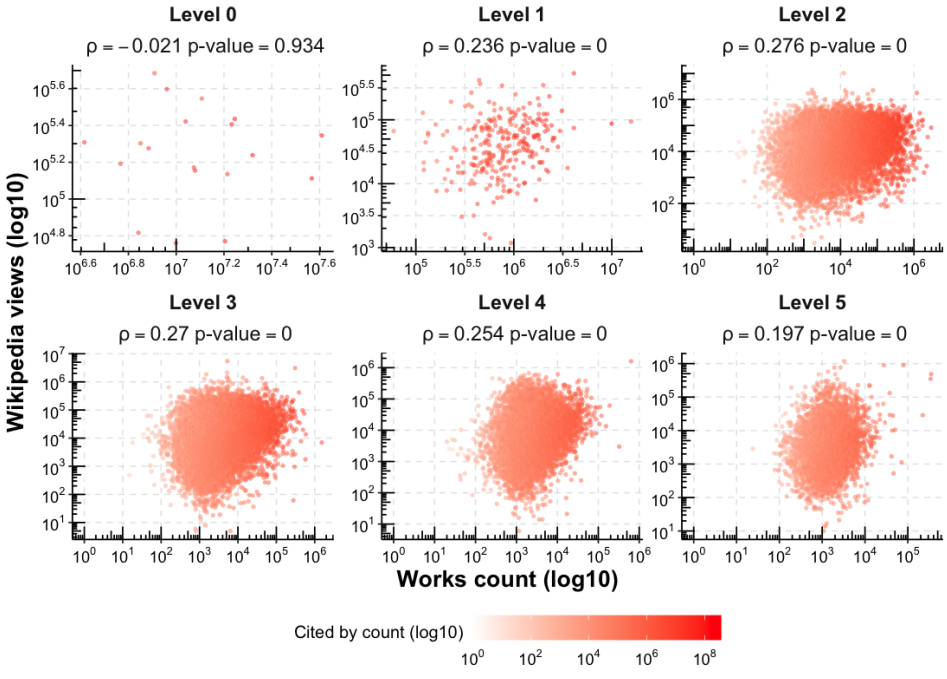
From the dataset generated, the relationship between academic interest and social attention has been studied. To do so, we explored the relationship between *works count* and Wikipedia *page views* for each concept, calculating the Spearman correlation for each level to determine the existence or not of such a relationship. After this, the differences and similarities in the attention offered by both metrics were observed using the 19 root-level concepts (level 0). Firstly, the total *works count* and *page views* of these concepts have been compared, and secondly, the concepts of levels 1 and 2 have been aggregated under the root-level concepts in order to carry out a more precise comparison between academic interest and social attention.

3. Results

3.1. Academic interest–works count–and social attention–Wikipedia page views–correlations

Overall, there is no clear relationship or trend between the volume of publications (*works count*) of the concepts and the social attention the articles of those concepts attract on Wikipedia (Wikipedia *page views*). Figure 2 shows the association between the two metrics, differentiated by the six levels of concepts. In none of them is a clear pattern discernible that shows that those concepts that are of greater scientific interest, that is, those with a higher works count, attract more social attention, that is, have greater *page views* on Wikipedia.

Figure 2: Distribution of concepts by *works count* and Wikipedia *page views* per level.

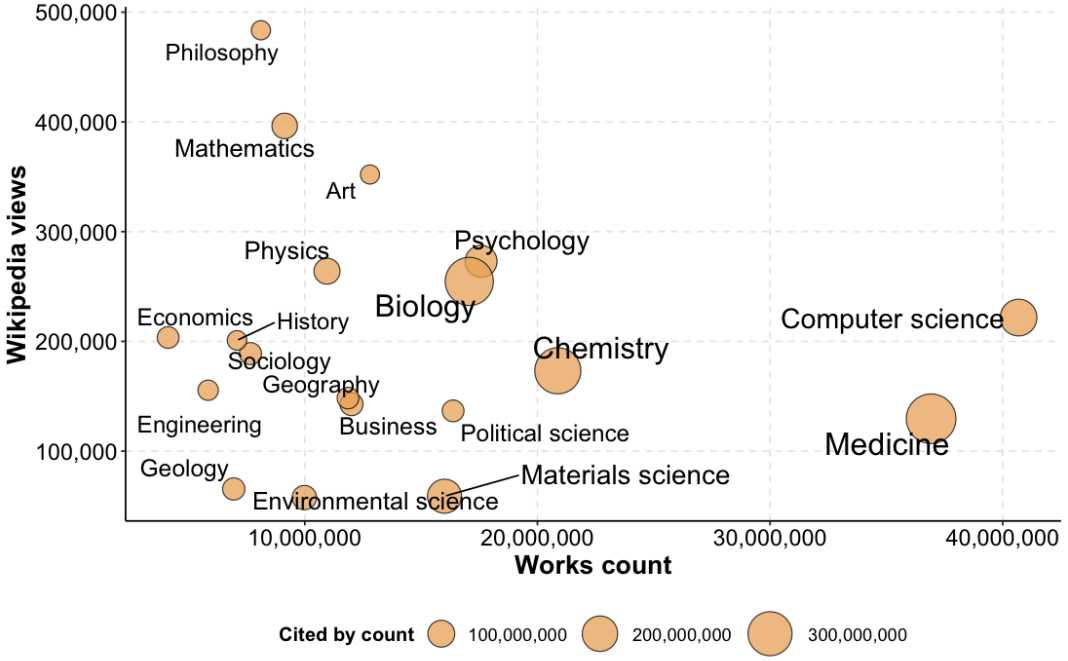


This lack of relationship is also highlighted by an analysis of Spearman correlations between *works counts* and Wikipedia *page views*, differentiating by concept level. The highest value is in the case of level 2 (ρ=0.276) and level 3 (ρ=0.27) concepts. The rest of the correlations are at values close to ρ=0.2, except for level 0 (ρ=-0.021), which is also the only one of all the correlations that is not significant. This can be explained by the fact that it only has 19 very broad and different concepts. Therefore, in view of these results, especially this slight positive correlation, we cannot conclude that there is clear evidence of association between academic interest and social attention and that therefore important discrepancies seem to exist between the two realms.

3.2. Analysis of major research topics

When further exploring the case of the major OpenAlex concepts, which is the 19 root-level concepts, the differences between scientific production and social attention become evident (Figure 3). On the one hand, we find concepts such as *Medicine* and *Computer science*, which have the highest number of works, but a reduced number of *page views* in their corresponding Wikipedia articles, especially the former. In contrast, there are concepts such as *Philosophy*, *Mathematics* or *Art*, which, although they have a comparatively small number of works, have the highest number of Wikipedia *page views*. However, this approach is limited because it directly compares each of the major scientific areas into which OpenAlex is thematically structured with a single, holistic Wikipedia article. See in this sense the case of *Medicine*, where on one side we have 36,922,842 works covering the whole research topic, while on the other side we have a single Wikipedia article[3] whose *page views* are limited as a reflection of the general domain of medicine that considers the large number of topics of a medical nature.

Figure 3: Distribution of root-level concepts by *works count* and Wikipedia *page views*.

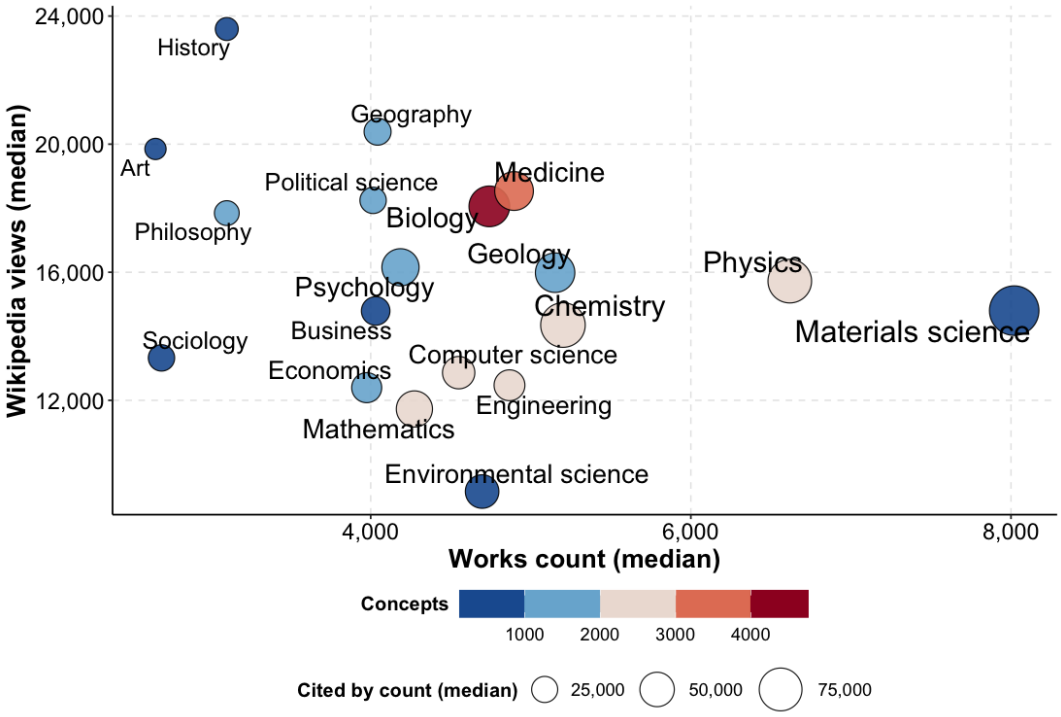


For this reason, an aggregate analysis of concepts from immediately lower levels has been carried out. This allows a more accurate representation of social attention by considering the multiplicity of concepts, as is the case in *Medicine* where for example the concepts *Nursing* and *Virology* are found in level 1 and *Cancer* and *Ibuprofen* in level 2. Figure 4 shows the concepts

[3] https://en.wikipedia.org/wiki/Medicine

of levels 1 and 2 grouped by the 19 root-level concepts. *Material science* now stands out as the concept with the highest median number of papers, while *History* stands out in *page views*. However, both two have fewer concepts, whereas *Biology* and *Medicine* have the largest number of concepts. Also noteworthy is *Environment science*, which, with a production in an intermediate position to the rest, has the lowest median number of *page views*, as well as the lowest number of concepts.

Figure 4: Distribution of aggregated level 1 and 2 concepts by *works count* and Wikipedia *page views*.



4. Discussion

In this paper we have made a first attempt at studying of the relationship between scientific topics that have interest on the part of researchers (as measured by the number of publications published) and those that receive the greatest social attention (as measured by the number of views of the Wikipedia articles of those topics). To this end, the *works count* of the OpenAlex concepts and the *page views* of the English Wikipedia articles associated with these concepts have been compared. The slight correlation found and the differences seen across both metrics suggest that academic and social interest may be governed by different forces. Particularly interesting is that topics from the Humanities and Social Sciences seem to receive an attention on Wikipedia that is not corresponded with the output production in those fields, reinforcing the idea that social sciences and humanities, as well as medical and health topics, tend to have a stronger affinity with social media dynamics (cf. Fang et al, 2022).

Overall this research is still under development and has several methodological limitations that will be addressed to provide further support and context of these findings. In terms of the data, firstly, we are working to include the time periods of scholarly outputs and match them with those of *page views*, as well as to increase the coverage of *page views*. Secondly, we also intend to incorporate into the analysis the attention provided by the rest of the Wikipedia language editions. Lastly, it is intended to include more metadata and metrics from Wikipedia (e.g. size

of the Wikipedia articles, their age, etc.). In terms of methods, we will strengthen statistical methods and cover the overall concepts more comprehensively.

Open science practices
All data used in this research are open access. OpenAlex has been used to retrieve concepts and works count while Wikipedia page views were retrieved from the Wikipedia Knowledge Graph dataset (doi:10.5281/zenodo.6346900). Scripts for data retrieval, processing and analysis are available at the GitHub repository: https://github.com/Wences91/wikipedia_concepts

Acknowledgments
Rodrigo Costas was partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP).

Author contributions
Wenceslao Arroyo-Machado: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing—original draft.
Rodrigo Costas: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing—review & editing.

Competing interests
The authors have no competing interests.

References
Arroyo-Machado, W., Díaz-Faes, A. A., & Costas, R. (2022). New insights on social media metrics: Examining the relationship between universities academic reputation and Wikipedia attention. In N. Robinson-Garcia, D. Torres-Salinas, & W. A. Arroyo-Machado (Eds.), *26th International Conference on Science, Technology and Innovation Indicators (STI 2022)* (p. sti22159). https://doi.org/10.5281/zenodo.6962442

Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022a). *Wikipedia Knowledge Graph dataset* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6346900

Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022b). Wikinformetrics: Construction and description of an open Wikipedia knowledge graph data set for informetric purposes. *Quantitative Science Studies*, 1–22. https://doi.org/10.1162/qss_a_00226

Arroyo-Machado, W., Torres-Salinas, D., & Robinson-Garcia, N. (2021). Identifying and characterizing social media communities: A socio-semantic network approach to altmetrics. *Scientometrics*, *126*(11), 9267–9289. https://doi.org/10.1007/s11192-021-04167-8

Costas, R., de Rijcke, S., & Marres, N. (2020). "Heterogeneous couplings": Operationalizing network perspectives to study science-society interactions through social media metrics. *Journal of the Association for Information Science and Technology*, *72*(5), 595–610. https://doi.org/10.1002/asi.24427

Díaz-Faes, A. A., Bowman, T. D., & Costas, R. (2019). Towards a second generation of 'social media metrics': Characterizing Twitter communities of attention around science. *PLOS ONE*, *14*(5), e0216408. https://doi.org/10.1371/journal.pone.0216408

Donovan, C., Flaherty, E. T., & Quinn Healy, E. (2017). Using big data from Wikipedia page views for official tourism statistics. *Statistical Journal of the IAOS*, *33*(4), 997–1003. https://doi.org/10.3233/SJI-160320

Fang, Z., Costas, R., & Wouters, P. (2022). User engagement with scholarly tweets of scientific papers: A large-scale and cross-disciplinary analysis. *Scientometrics*, *127*(8), 4523–4546. https://doi.org/10.1007/s11192-022-04468-6

Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., & Priedhorsky, R. (2014). Global Disease Monitoring and Forecasting with Wikipedia. *PLOS Computational Biology*, *10*(11), e1003892. https://doi.org/10.1371/journal.pcbi.1003892

Gómez-Martínez, R., Orden-Cruz, C., & Martínez-Navalón, J. G. (2022). Wikipedia pageviews as investors' attention indicator for Nasdaq. *Intelligent Systems in Accounting, Finance and Management*, *29*(1), 41–49. https://doi.org/10.1002/isaf.1508

Yoshida, M., Arase, Y., Tsunoda, T., & Yamamoto, M. (2015). Wikipedia Page View Reflects Web Search Trend. *Proceedings of the ACM Web Science Conference*. https://doi.org/10.1145/2786451.2786495