STI2022
GRANADA | SEP 7+8+9

26th International Conference on Science and Technology Indicators
"From Global Indicators to Local Applications"

#STI2022GRX

*Full paper*

**Proceeding Editors**

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

**Citation:** Mazoni, A., Arroyo-Machado, W., Traag, V., A., & Costas, R.(2022). On the quest of scholarly communities of attention: large-scale clustering of Twitter users around scientific publications. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22238). https://doi.org/10.5281/zenodo.6957616

26th International Conference on Science and Technology Indicators | **STI** 2022
"From Global Indicators to Local Applications"
7-9 September 2022 | Granada, Spain
**#STI22GRX**

# On the quest of scholarly communities of attention: large-scale clustering of Twitter users around scientific publications[1]

Alysson Mazoni[*], Wenceslao Arroyo-Machado[**] , Vincent A. Traag[***] and Rodrigo Costas[****]

[*] *afmazoni@unicamp.br*
InSySPo, Institute of Geosciences, University of Campinas, Campinas, 13083-855 (Brazil)

[**] *wences@ugr.es*
Department of Information and Communication Sciences, University of Granada, Spain

[***] *V.A.Traag@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands

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

## Introduction

The science-related activity on social media has been studied since before the advent of altmetrics, whose formal birth took place in 2010 with the publication of the Altmetric manifesto (Priem, Taraborelli, Groth, & Neylon, 2010). Following the irruption of this phenomenon, several studies flourished and tried to adapt the citation model to this social media scenario, giving rise to studies that seek a relationship between academic impact, measured by citations, and social impact, measured by social media mentions to publications. In the absence of such a relationship and altmetrics criticised capacity as a measure of social impact (Sugimoto, Work, Larivière, & Haustein, 2017), a new scenario has arisen in which scientific publications are no longer the most important object of study, and social media mentions counts are not the most important attribute to measure (Díaz-Faes, Bowman, & Costas, 2019; Wouters, Zahedi, & Costas, 2019). Instead, more interactive perspectives have been proposed in order to study the  potential use of altmetrics to capture and study interactions among academic and non-academic communities (Costas, de Rijcke, & Marres, 2020).

Twitter plays a key role in altmetrics from its very origins and has attracted the most attention in research. Such is the case that several metrics are studied using Twitter data in order to analyse the impact or spread of a publication, author or institution (Costas, Zahedi, & Wouters, 2015). Twitter is by far the data source with the largest coverage of publications from among

[1] Alysson Mazoni is financially supported by Fapesp (grant number 2021/05823-1).

all altmetric sources, with the only exception of Mendeley readership (Torres-Salinas, Robinson-García, & Arroyo-Machado, 2022; Zahedi & Costas, 2018). At the same time working with Twitter data is not free of limitations, like for example the stability of the tweets (Fang, Dudek, & Costas, 2020), which may disappear as Twitter users delete or suspend their accounts. From the perspective of social network analysis, much literature has proliferated and, in the context of this new generation of social media metrics, numerous methodological proposals have been done, for example to profile researchers (Robinson-Garcia, van Leeuwen, & Ràfols, 2018), characterise communities (Díaz-Faes, Bowman, & Costas, 2019) or studying overlapping social and semantic communities (Arroyo-Machado, Torres-Salinas, & Robinson-Garcia, 2021). As part of many of these proposals, communities and research topics have been mapped on Twitter, but always applied to specific case studies, such as Information Science and Library Science (Arroyo-Machado, Torres-Salinas, & Robinson-Garcia, 2021) or Microbiology (Robinson-Garcia, Arroyo-Machado, & Torres-Salinas, 2019). The bubble filter effect among Twitter users is well known and also reveals communities of interest for particular subjects (Grossetti, 2021). In connection to scientific research and its social visibility this is particularly interesting in that it can reveal the connections among Twitter interest groups (Crockett, 2017; Jafalli, 2014).

Once the usefulness of these methodological approaches has already been demonstrated, further efforts are required to apply them at a large scale rather than a specific case study. This is precisely the ambition of this paper, leveraging on the currently available open data on Twitter mentions around scientific publications, to offer a large-scale clustering of Twitter users and the scientific publications they have tweeted, in what can be framed as the largest portrait of scientific activity on Twitter done to date.

## Objective

The main aim of this study is to present an account of the methodological workflow, challenges and possibilities, for creating a large-scale clustering of Twitter accounts and tweeted DOIs by means of advanced clustering techniques. We will also present some preliminary results of such clustering and proposals for next steps in this endeavour.

## Methodology workflow

### *1) Data collection and preparation*

In an attempt to potentially provide a future open platform for advanced social media studies of science, we have used only open data in this study. Thus, Twitter mentions (and their users) to papers' data were collected from a snapshot of Crossref Event Data (CED) up to 2021 (Hendricks, 2020). The JSON files from CED were processed to extract all Twitter events for the different Twitter accounts and their mentioned DOIs. As a result a large dataset containing the distinct set of Twitter users and the DOIs they have tweeted has been compiled. This dataset contains more than 12 million rows including DOIs and users.

### *2) Big Data analytical tools*

Given the size of files, they were stored and processed in a data warehouse such as Google BigQuery, (Bisong, 2019). A large virtual machine was used for the task of converting the JSON files to tabular format and for the clustering. In our clustering process, the Python API for the iGraph was used in a Virtual Machine with 64 processors and 416 GB of RAM delivered inside the Google Vertex AI platform. The number of processors exceeded the need for the task, however it was necessary an architecture with a large memory in order to contain all the graph for the clustering algorithm.

*3) Clustering*

It was noted by Horvát (2012) that a network of connections of users citing content unities should preferably be considered with different nature of connections for its elements. This amounts to using a multiplex graph. The multiplex graph contains nodes that are linked in a certain fashion as a layer, but can also be linked in a different way. For example, countries can be connected by scientific papers written by authors of different nationalities in collaboration, or they can be said to be connected by the nationality of their institutions. In our particular case, there are connections solely from nodes of different types, since a user cites a paper (not another user) and a paper is cited by a user. A situation like that is a special case of a multiplex graph called bipartite graph, as described by Fer & Brodley (2004).

A difficult question in the clustering of Twitter users and papers is that some users and some papers are super-connectors. For example some Twitter users have tweeted more than 170,000 DOIs like the Twitter account @BlackPhysicists, while some papers have been tweeted by a large number of Twitter users.[2] We have considered the effect of the frequency a particular node has in making connections by using a particular weight. Since every connection is formed by a paper to a user, we define that the weight is the inverse of the product of the number of papers cited by the user ($N_p$) by the number of users linking to the paper ($N_u$):

$$w = \frac{1}{N_p N_u}$$

With this weight, a user that tweets many papers will create less important connections. The same is valid for overly cited papers.

By using the IGraph package and its Python API (Csardi & Nepusz, 2006) we produced a clustering of papers and Twitter users connected. The clustering used the Leiden Algorithm by Traag (2019) with the bipartite vertex partition as instructed in its manual with a resolution parameter of 0.1.

## Preliminary results

The clustering resulted in 879931 communities. Each cluster includes two types of elements: papers and Twitter users. In order to better investigate the data, we chose to plot the communities with more than 1000 elements overall (either users or papers), which amounts to a total of 1278 communities. Figure 1 presents the sizes for the largest 50 communities sorted by the number of users, and plotting the number of papers belonging to each community. Figure 2 presents the same visualisation but in this case sorting by the number of papers included in the cluster and plotting also the number of users of each cluster. The first aspect to highlight is the relative skewness observable in the figures, with some clusters exhibiting very large numbers of nodes (either publications or users) to then sharply decrease in the other communities, which become more stable in their number of clustered nodes. A second aspect to highlight is that in both Figures it is rather apparent a sort of polarisation regarding the presence of either users or publications, with some communities having a massive number of users and relatively small number of publications, and vice versa. These two aspects (the

[2] The paper with the DOI:10.1056/nejmc1713344 has been tweeted by more than 25,000 different Twitter users.

skewness and node polarisation) is not surprising in an environment like Twitter, in which activity (e.g. tweeting papers, retweeting them, etc.) can be easily performed, become *viral* and even automated (e.g. the presence of bots and automated accounts on altmetric Twitter data has been been discussed in the literature, see Haustein et al, 2016). In principle these two aspects could be seen as problematic in the resulting clustering of Twitter communities of attention, since there could be important imbalances in the consideration of some communities over others. However, a cursory look at some of the most extreme cases shows that the specific analysis of these large and polarised communities may bring additional relevant information about the type of dissemination that is taking place within the community. For example, the 12th largest community by number of users is composed of more than 52,699 users and just 52 papers. Many of these 52 papers are kind of the viral-type of papers, all of them with very large numbers of tweets and tweeters around them, and with relatively *funny* topics. For example the DOI:10.1038/546466a corresponds to a paper in Nature about the tsunami detection. Or the DOI:10.1038/srep17890 about climate change.[3]

Regarding communities with a large number of publications and a low number of users in some cases corresponding to users expelled from Twitter for massive tweeting. These results may point that particularly those communities with extreme polarisation between the two types of entities clustered (publications or tweeterts) may be capturing relevant information about also with extreme or special features, enabling their identification and possible study.

[3] It is important to remark that results between Crossref Event Data and Altmetric may differ.

Figure 1: Largest 50 communities in number of users and number of papers. The first communities contain more than 1.98million, 220 thousand and 218 thousand users, respectively. The cut is to clarify the other sizes.

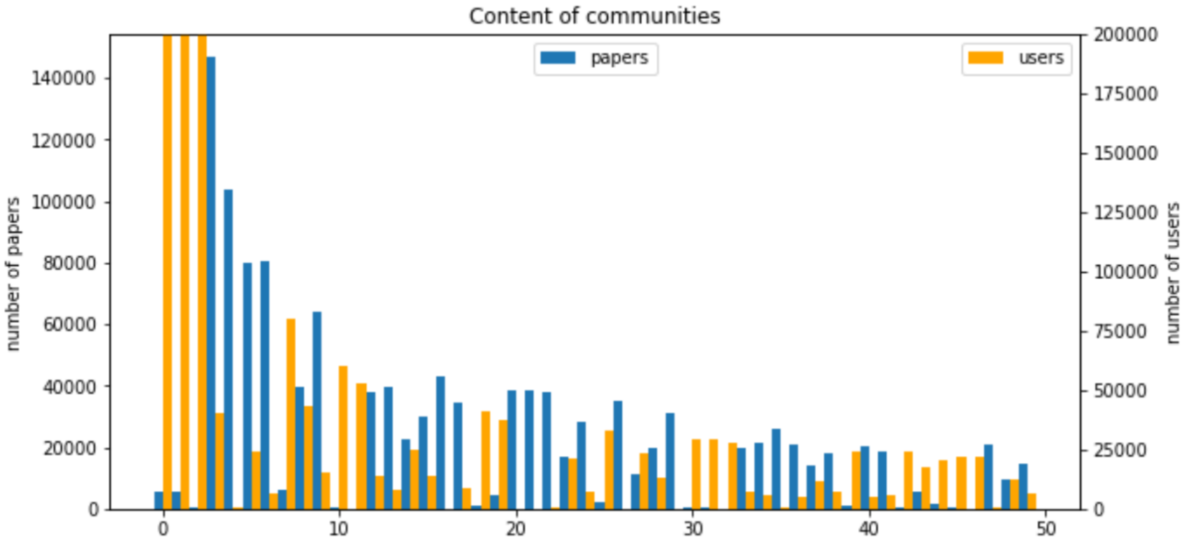
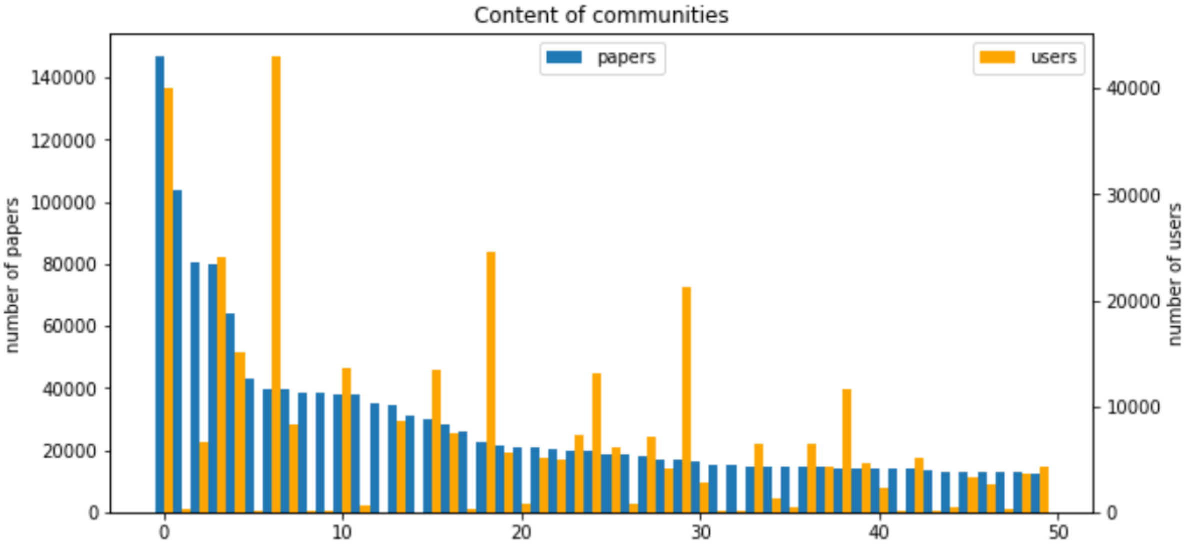


Figure 2: Largest 50 communities in number of papers and their number of users.



## Discussions

In this study we provide a first account of the methodological workflow for a large-scale clustering of the Twitter communities of attention around scientific publications as captured in the open database Crossref Event Data. To the best of our knowledge this is the largest algorithmic clustering of Twitter users and scientific publications performed to date. The availability of this type of clustering opens new analytical possibilities in the study of the Twitter dissemination of scientific publications. For example, making possible the study of the diversity of the communities in which publications have been tweeted enabling the differentiation of publications tweeted in smaller or larger communities, or the identification of those communities that tweet more superficially or automatically (Robinson-Garcia et al, 2017; Haustein et al, 2015).

From a technical point of view, the use of big data tools (Google BigQuery) was implemented given the large size of data involved in the clustering. Moreover, the use of the relative weight allowed for the determination of well connected communities without much skewness in its sizes. The sheer size and availability of open data opens the way for several kinds of analysis that demand careful use of file formats and computation resources, usually based on big data tools such as data warehouses and running multiprocessor code.

Future research will necessarily focus on two additional developments: 1) refining the clustering to include those less connected communities in a meaningful manner, making them also more balanced, and 2) implementing a labelling of the different clusters obtained. For the first, additional clustering (e.g. clustering of clusters) and reclustering of smaller clusters will be very likely the approach to go. For the second, we aim at finding potentially meaningful information by collecting metadata from papers (e.g. journals, titles, topics) and Twitter users (e.g. profile descriptions, geolocations, URLs). That information, combined with language processing techniques will potentially allow the labelling of the clusters in order to better characterise the communities and their dynamics in disseminating scientific publications on Twitter.

## References

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.

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. John Wiley & Sons, Ltd.

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, *66*(10), 2003–2019.

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.

Fang, Z., Dudek, J., & Costas, R. (2020). The Stability of Twitter Metrics: A Study on Unavailable Twitter Mentions of Scientific Publications. Journal of the Association for Information Science and Technology, *71*(12). John Wiley & Sons, Ltd. Retrieved March 4, 2020, from https://doi.org/10.1002/asi.24344

Hendricks, G., Tkaczyk, D. Lin, j, Feeney, p. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1 (1): 414–427. doi: https://doi.org/10.1162/qss_a_00022.

Bisong, E. (2019). Google BigQuery. *In: Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-4470-8_38.

Grossetti, Q., du Mouza, C., Travers, N. and Constantin, C. (2021), Reducing the filter bubble effect on Twitter by considering communities for recommendations, *International Journal of Web Information Systems*, Vol. 17 No. 6, pp. 728-752. https://doi.org/10.1108/IJWIS-06-2021-0065.

Crockett, KA and Mclean, D and Latham, A and Alnajran, N (2017) Cluster Analysis of Twitter Data: A Review of Algorithms. In: *9th International Conference on Agents and Artificial Intelligence (ICAART)*, 24 February 2017 - 26 February 2017, Portugal.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. *Altmetrics*. Retrieved from http://altmetrics.org/manifesto/

Jaffali, S. Jamoussi, A. B. Hamadou and K. Smaili, Clustering and Classification of Like-Minded People from their Tweets, 2014 *IEEE International Conference on Data Mining Workshop*, 2014, pp. 921-927, doi: 10.1109/ICDMW.2014.161.

Robinson-Garcia N, Costas R, Isett K, Melkers J, Hicks D (2017) The unbearable emptiness of tweeting—About journal articles. *PLOS ONE* 12(8): e0183551. https://doi.org/10.1371/journal.pone.0183551

Robinson-Garcia, N., Arroyo-Machado, W., & Torres-Salinas, D. (2019). Mapping social media attention in Microbiology: Identifying main topics and actors. *FEMS Microbiology Letters*, *366*(7). Retrieved January 15, 2020, from https://doi.org/10.1093/femsle/fnz075

Robinson-Garcia, N., van Leeuwen, T. N., & Ràfols, I. (2018). Using altmetrics for contextualised mapping of societal impact: From hits to networks. *Science and Public Policy*, *45*(6), 815–826.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037–2062. John Wiley & Sons, Ltd.

Torres-Salinas, D., Robinson-García, N., & Arroyo-Machado, W. (2022). Coverage and distribution of altmetric mentions in Spain: A cross-country comparison in 22 research fields. *El Profesional de la información*, e310220.

Haustein, S., Bowman, T.D., Holmberg, K., Tsou, A., Sugimoto, C.R. and Larivière, V. (2016), Tweets as Impact Indicators: Examining the Implications of Automated “bot” Accounts on Twitter. *J Assn Inf Sci Tec*, 67: 232-238. https://doi.org/10.1002/asi.23456.

E. Horvát and K. A. Zweig, One-mode Projection of Multiplex Bipartite Graphs (2012) *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012, pp. 599-606, doi: 10.1109/ASONAM.2012.101.

Fern, X. Z., & Brodley, C. E. (2004, July). Solving cluster ensemble problems by bipartite graph partitioning. *In Proceedings of the twenty-first international conference on Machine learning* (p. 36).

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5), 1-9.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1), 1-12.

Zahedi, Z., & Costas, R. (2018). General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PLOS ONE*, *13*(5), e0197326.