RESEARCH ARTICLE

Large-scale identification and characterization of scholars on Twitter

Rodrigo Costas[1,2], Philippe Mongeon[1,3], Márcia R. Ferreira[1,4,5], Jeroen van Honk[1], and Thomas Franssen[1]

[1]Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (the Netherlands)
[2]Centre for Research on Evaluation, Science and Technology (CREST), Stellenbosch University, Stellenbosch (South Africa)
[3]Centre for Studies in Research and Research Policy (CFA), Aarhus University, Aarhus (Denmark)
[4]Complexity Science Hub Vienna, Vienna (Austria)
[5]Institute of Information Systems Engineering, Vienna University of Technology, Vienna (Austria)

Keywords: altmetrics, bibliometrics, individual scholars, social media, Twitter

ABSTRACT

This paper presents a new method for identifying scholars who have a Twitter account from bibliometric data from Web of Science (WoS) and Twitter data from Altmetric.com. The method reliably identifies matches between Twitter accounts and scholarly authors. It consists of a matching of elements such as author names, usernames, handles, and URLs, followed by a rule-based scoring system that weights the common occurrence of these elements related to the activities of Twitter users and scholars. The method proceeds by matching the Twitter accounts against a database of millions of disambiguated bibliographic profiles from WoS. This paper describes the implementation and validation of the matching method, and performs verification through precision-recall analysis. We also explore the geographical, disciplinary, and demographic variations in the distribution of scholars matched to a Twitter account. This approach represents a step forward in the development of more advanced forms of social media studies of science by opening up an important door for studying the interactions between science and social media in general, and for studying the activities of scholars on Twitter in particular.

1. INTRODUCTION

Social media have become important for scholarly communication and dissemination. They provide researchers the opportunity to make their work widely accessible, share information with peers, and monitor the visibility of their work (Veletsianos, 2012). Popular social media tools include Twitter and Facebook. For academics, specific tools that include a social networking component are ResearchGate, Mendeley, and Academia.edu (Sugimoto, Work, et al., 2017). Some estimates indicate, for instance, that approximately 21.5% of papers from 2012 indexed in Web of Science (WoS) with a DOI have received at least one mention on Twitter (Haustein, Costas, & Larivière, 2015). Tracking and investigating social media mentions of scholarly articles as well as their relationship have become known as "alternative metrics" or altmetrics. Social media metrics have also been proposed as potentially complementary to traditional bibliometric indicators, such as citations (e.g., Priem & Costello, 2010). However, the correlations between social media indicators and bibliometric indicators have consistently been found to be low (Costas, Zahedi, & Wouters, 2015b;

an open access journal

Citation: Costas, R., Mongeon, P., Ferreira, M. R., van Honk, J., & Franssen, T. (2020). Large-scale identification and characterization of scholars on Twitter. *Quantitative Science Studies*, *1*(2), 771–791. https://doi.org/10.1162/qss_a_00047

DOI: https://doi.org/10.1162/qss_a_00047

Received: 02 October 2019
Accepted: 14 January 2020

Corresponding Author: Rodrigo Costas
rcostas@cwts.leidenuniv.nl

Handling Editor: Staša Milojević

The MIT Press

Haustein, Peters, et al., 2014; Thelwall, Haustein, et al., 2013), suggesting that social media indicators measure an additional dimension of scholarly workflows closer to public communication, socialization, networking, and engagement with wider audiences, rather than scientific impact in the narrow sense.

Despite attracting much attention, few studies have provided a comprehensive portrait of scholars active on social media because there is no database that links scholarly authors to their corresponding Twitter accounts, which is essential to our understanding of the use of social media in scholarly communication. As a result, the demographics, scientific fields, and geographical locations of the institutions of scholarly authors on Twitter remain poorly understood (Ke, Ahn, & Sugimoto, 2016).

Recently, scholars have started to reconceptualize the analysis of social media activity of research authors as part of an ambitious research agenda to study in more depth the relationship between social media and scholarly entities, in what has been termed the "social media studies of science" (Costas, 2017; Wouters, Zahedi, & Costas, 2019). This new perspective aims at understanding the added value of social media interactions to the scholarly workflow, in particular as part of communication and dissemination practices as well as network formation, rather than focusing on mere indicator development. Of particular interest is Twitter, a popular microblogging platform that provides a means for users to communicate through short 280 character messages known as *tweets*. On Twitter, users are able to "follow" each other on the platform, and thus receive notifications of their tweets, search tweets by keywords or hashtags, or link to other media or tweets (Sugimoto et al., 2017). Altmetric.com records the frequency with which a DOI (and other scholarly outputs identifiers such as PubMed-IDs) of a scientific article are mentioned on Twitter. The platform also collects Twitter metadata such as user account information (e.g., Twitter handle, username, user description, or geographical location) whenever the users have tweeted, retweeted, or mentioned a scientific article. Many of these accounts can be linked to academic user accounts such as those of scholars, academic institutions, and journals. The Altmetric.com database therefore offers a unique opportunity to identify scholars with a Twitter account at a large scale.

Earlier studies matching author-level bibliometric information to Twitter user-level information were carried out using labor-intensive approaches, such as self-identification through surveys (Collins, Shiffman, & Rock, 2016; Rowlands, Nicholas, et al., 2011; Van Noorden, 2014), or through manual verification (e.g., Haustein, Bowman, et al., 2014; Holmberg & Thelwall, 2014; Hwong, Oliver, et al., 2016; Lulic & Kovic, 2013; Veletsianos, 2012). Although these studies have provided important insights into the use of Twitter by scholars in different contexts (e.g., conferences, educational settings, sharing preprints and publications), limited response rates (in the case of survey research) and time-consuming manual approaches have resulted in data sets of matched authors and Twitter users that represent only a very small fraction of the overall universe of scholars on Twitter. A few notable exceptions are Ke et al. (2016) and Hadgu and Jäschke (2014), who used Twitter lists and conference hashtags, respectively, to identify scholars and classify similar users connected to an initial set of seed Twitter users. Despite these approaches being automated and successful in identifying large numbers of scholars (45,000 and 38,000 Twitter accounts respectively) on Twitter, they fully rely on self-reported evidence of a user identifying as a scholar, and favor more established scholars who are also more likely to be included in Twitter lists.

Accordingly, the purpose of this study is to identify Twitter accounts belonging to scholars among millions of disambiguated authors recorded in the WoS database in a fully automated way. In doing so, this paper fills a significant gap left by previous studies that depend primarily

on manual techniques and surveys, or are limited to one scientific field. To the best of the authors' knowledge, no previous studies have matched Twitter accounts with scholarly bibliometric profiles using publications in the WoS on a large scale, as is done here. Moreover, this paper also provides a more comprehensive portrait of scholars on Twitter by first creating a large-scale data set of matched Twitter accounts with WoS authors corresponding to the same individual, and then by characterizing these scholars on the basis of their field, academic age, country, and gender. The unique connection between bibliographic data and Twitter data opens up the possibility of studying not only the Twitter activities of scholars, but also their scholarly activities (as captured by bibliometric analyses) (Costas, 2017; Wouters et al., 2019).

This paper is organized as follows. Section 2 presents an overview of the literature related to matching procedures between authors of papers indexed in the WoS database and Twitter accounts. Section 3 describes the matching approach, its implementation, and validation. The results are presented in Section 4, accompanied by an empirical analysis of research authors on Twitter by field, gender, country, and academic age. In Section 5 the paper draws conclusions about our approach and discusses the method's limitations as well as plans for future research.

1.1. The Communication Context of Scholars on Twitter

Twitter use by scholars has been studied in various contexts, such as the sharing of scholarly outputs on Twitter and how scholars use Twitter to develop and maintain their professional networks. As governments and funding agencies are increasingly taking an interest in a broader view of impact (Dinsmore, Allen, & Dolby, 2014), the use of Twitter and the recording of this in altmetric indicators are sometimes perceived as being of value in this context (Das & Mishra, 2014; Priem, Piwowar, & Hemminger, 2012; Priem, Taraborelli, et al., 2010). The general presence of mentions of scholarly outputs is found to vary across different scientific disciplines (Costas, Zahedi, & Wouters, 2015a; Haustein et al., 2014; Holmberg & Thelwall, 2014), indicating the existence of different thematic interests of research topics among Twitter users or differences in the use of Twitter among different scholarly communities.

The use of Twitter by scholars shows some distinct patterns. For instance, scholars tend to share more links and retweet more than the average Twitter user (Holmberg & Thelwall, 2014). Recent studies have also shown that Twitter users who present themselves with academic and scholarly terms also tend to have a stronger focus and engagement with scientific topics on Twitter (Díaz-Faes, Bowman, & Costas, 2019). Across personal and professional tweets, the use of technological social media "affordances" on Twitter has been shown to vary based on department, gender, academic age, age, and Twitter activity (Bowman, 2015). In a study of Twitter accounts, it was reported that users who tweet academic articles describe themselves by emphasizing their occupational expertise (Vainio & Holmberg, 2017). While academic tweeters provide their full name and professional identity in their account descriptions (Bowman, 2015; Chretien, Azar, & Kind, 2011; Hadgu & Jäschke, 2014), a large share of their activity is personal as opposed to professional (Bowman, 2015; Haustein et al., 2014; Van Noorden, 2014).

When using Twitter for professional purposes, scholars discuss research-related topics and communicate with others in the field (Van Noorden, 2014). Scholarly tweets tend to contain links to both recent journal articles (Eysenbach, 2011; Holmberg & Thelwall, 2014; Priem & Costello, 2010) and blogs (Letierce, Passant, et al., 2010; Priem & Costello, 2010). The content of these tweets tends to be limited to the title, or part of the title of the scientific article being

tweeted (Friedrich, Bowman, et al., 2015; Thelwall et al., 2013) and the level of engagement of Twitter users with the content of publications, in terms of discussing particular details, is generally low (Robinson-Garcia, Costas, et al., 2017). The use of Twitter, however, does have effects on the dissemination of scientific papers. According to Ortega (2016), articles authored by Twitter users are more tweeted than those of non-Twitter users. In addition, the number of followers on Twitter is found to indirectly influence the citation impact (Ortega, 2016).

1.2. Methods for Identifying Scientists on Twitter

Earlier research has studied the use of Twitter among different scientific disciplines (Holmberg & Thelwall, 2014). The most commonly used method to identify scientists on Twitter is the manual identification of scientists' Twitter accounts. Veletsianos (2012) used snowball sampling with an initial set of four Twitter accounts with 2,000 followers or more. He then examined these followers to identify other scholars with at least 2,000 followers, resulting in a total sample of 46 Twitter accounts. Similarly, Lulic and Kovic (2013) identified 672 emergency physicians on Twitter using a keyword, manually validating the results, and examining the followers to identify other physicians. Holmberg and Thelwall (2014) first used the WoS database to identify the top 10 most productive scholars for 10 disciplines, searched for these individuals on Twitter, and complemented this data set with a keyword search and a snowball sampling method as per Veletsianos (2012) and Lulic and Kovic (2013). Their data set comprised 477 Twitter accounts. Hwong et al. (2016) manually identified 60 actively maintained Twitter accounts about space science. Past studies have also used surveys to study the Twitter uptake and activity of scientists (e.g., Rowlands et al., 2011; Van Noorden, 2014; Collins et al., 2016). These methods for identifying scientists on Twitter have some important limitations. The first is that the sample is limited by their reliance on manual selection of Twitter accounts or on self-reported information, and the second is their relatively small scale.

Twitter lists (curated groups of Twitter accounts created by Twitter users and to which other users can subscribe) have also been used to identify the Twitter accounts belonging to specific groups of users (Sharma, Ghosh, et al., 2012). Similarly, Ke et al. (2016) used Twitter lists to collect a set of 45,867 Twitter accounts belonging to scientists. The authors collected Twitter accounts with a scientific occupation (e.g., psychologist, economist, PhD, researcher) in the Twitter biographies, which were part of lists that also contained a scientific occupation in their account description. Another popular method is the use of conference hashtags or Twitter accounts. Hadgu and Jäschke (2014) have been especially successful in this regard; they used the Twitter accounts and hashtags of 98 computer science conferences to identify 38,368 Twitter accounts. To identify scholars in Education, Veletsianos and Kimmons (2016) retrieved tweets with the #aera14 hashtag. They identified 1,629 users and, after manual verification, retained the 232 graduate students and 237 professors for their study. Ross, Terras, et al. (2011) used the hashtags of three conferences in digital humanities to identify 326 Twitter accounts. Compared with a manual approach, these methods can identify larger sets of Twitter accounts belonging to scholars. However, such methods can only tell us whether a Twitter account belongs to a scholar or not. The analyses they enable are restricted to the Twitter activities of the identified scholars, and they do not provide any linkage to any other features of the scholars, which is the purpose of the present paper.

By way of summary, given the smaller scope of past methods, based on relatively small sets of scholars and Twitter accounts, they can be seen as poorly suited for large-scale analyses of the scientists' Twitter activity. A second important limitation is that these methods fail to substantially connect the Twitter information identified with scientometric and demographic

information about scholars (e.g., publications, citations, countries, affiliations, gender, disciplines), thus limiting these studies to the analysis of Twitter activities only.

Accordingly, the main objective of this paper is to introduce a new method to identify individual scholars on Twitter using data from the WoS database and Twitter data obtained from Altmetric.com. The method has three main distinctive features:

1. The matching is data driven and automatic, and is thus less labor-intensive than other methods and better suited for large-scale studies.
2. It uses various metadata elements available in the WoS database and Altmetric.com records, facilitating the identification of a larger number of scholars on Twitter than previous studies.
3. It connects two different realms of activity in which a scholar might be active: scholarly publishing and social media activities.

2. DATA SET AND METHODS

2.1. Data Sources

To match Twitter accounts with scholarly authors we use two data sources: the WoS-database and the Altmetric.com database. We use the author-name disambiguation algorithm developed by Caron and Van Eck (2014) and applied to the WoS database, resulting in a set of 25,352,720 disambiguated authors with at least one publication after 2004. We extracted all 4,117,887 distinct Twitter accounts that have tweeted at least one DOI up to October 2017 from the Altmetric.com database.

2.2. Identifying Possible Names of Twitter Users

Altmetric.com records three metadata fields with Twitter data related to the names of the users. The first is the full name field, which is an optional free text field with a maximum length of 50 characters that has no restrictions on the characters used. It may not always contain the actual name of the user and, if it does, the name can be entered in any format. We also used the Twitter handle field, which is limited to 15 alphanumeric characters or underscores. It may be less likely to contain the actual name of the user, but the character set restriction can help in cases where the users' names are usually in a language that does not use the Roman alphabet (e.g., Arabic, Chinese, and Japanese). Finally, when the URL field[1] contains a Facebook.com or Academia.edu URL, we extracted the part of the URL that potentially contains the user's name.

For each Twitter account, we created a list all of possible first names, last names, and initial(s). After replacing all nonalphabetical characters with a space, we divided each string into distinct components. For instance, "Robert J Smith" has three components: "Robert", "J", and "Smith". As the name can be concatenated (mostly in the handle or the URLs) we also parse the strings using different uppercase and lowercase patterns. For instance, the string "RobertJSmith" is divided into "Robert", "J" and "Smith". "RJSmith" is parsed in "R", "J", and "Smith".

[1] This is a field in Twitter accounts that allow users to indicate a website (e.g., personal website, professional, blogs, Facebook profiles, Research Gate profiles, ORCID profiles). Users can also add additional URLs in their Twitter bios, but those URLs are not extracted or parsed.

2.3. Matching Twitter Accounts with WoS Authors

We matched our list of names with the WoS authors' data set using the last name and the first initial, obtaining 1.06 billion potential matches. This means that the initial pairs of Twitter accounts–authors must have at least a match in the last name and first initial.

Following Caron and Van Eck (2014), we use a rule-based scoring approach in which scores are calculated (only for those pairs of Twitter accounts–authors that were matched in the last name–first initial combination) using the information available in the Twitter account and in the WoS records. There are 14 rules, presented in Table 1, which can be divided into five groups:

1. *Name matching rules* (rules 1 to 4). These rules are based on the matches found between the authors' names and the names extracted from the Twitter accounts. The scores are weighted based on the frequency of the different parts of the names in WoS[2]. The main rationale behind the scores is to weight uncommon names more and common ones less, but also without allowing for a high score based on just one of the name matching steps.
2. *Institutional and geographical rules* (rules 5 to 8). These rules are based on the matching of different elements provided both by the authors in their papers (e.g., affiliations, countries, emails) and the Twitter accounts (URLs, Twitter name and Twitter handle, and geographical or institutional information found in the Twitter accounts). The scores are weighted based on the frequency of the different elements in WoS, using the same method as for the name matching rules (rules 1 to 4). Among the rationales for the choice of the scores is also to score the less common elements more highly, but again without allowing for very high scores on just one of the institutional and geographical rules.
3. *Activity-related rules* (rules 9 to 12). These are based on the publications, fields, and journals of the authors and the publications tweeted by the Twitter account. The rationale is that the more a Twitter account has tweeted the papers of the matched author, or papers from the research fields or journals in which the matched author has published, the higher the chance that the author and the Twitter account are the same person. Additionally, a pair is likely to be valid if a tweet contains both the handle of a Twitter account and a link to a paper of the author matched with this Twitter account[3]. These rules have the highest scores because they are expected to be more accurate than names, institutions, and locations.
4. *Name commonness rule* (rule 13). If an author is only matched to one Twitter account, the matching is weighted more positively than when the author is matched to multiple Twitter accounts.
5. *Best match rules* (rules 14). We keep only the matches where the Twitter account was the best match for the WoS author (highest score based on rule 0–13) and vice versa.

[2] To calculate the weights in our methodology we have used an approach similar to the so-called Characteristic Scores and Scales (CSS) (Glänzel & Schubert, 1988), which consists in partitioning skewed distributions by using subsequent averages. Thus, the first average partitions the distribution into two parts, and the second average is calculated for the cases above the first average. As a result, elements can be weighted based on whether they belong to the group below average, to the group between the first and the second averages, and the third group above the second average. In Table 1 we specify the specific approaches and averages used for each rule that considers some weighting.

[3] See an example here: https://Twitter.com/wmijnhardt/status/781245999545212930.

Table 1. Summary of the criteria and scores for the different elements matched

Rules	Matching event	Criteria	Score
1	Last name and initial (i.e., author name, e.g., "Costas, R")	Very common full name (i.e., full names that belong to the group of the most common full names in the WoS disambiguated author database as determined by the average of the distribution)	1
		Common full name (i.e., full names that belong to the group of the second least common full names in the WoS disambiguated author database as determined by being between the two averages – higher and lower – of the distribution)	2
		Uncommon full name (i.e., full names that belong to the group of the least common full names in the WoS disambiguated author database as determined by being above the second average of the distribution)	3
2	First name	Very common first name (i.e., first names that belong to the group of the most common first names in the WoS disambiguated author database as determined by the average of the distribution)	1
		Common first name (i.e., first names that belong to the group of the second least common first names in the WoS disambiguated author database as determined by being between the two averages – higher and lower – of the distribution)	2
		Uncommon first name (i.e., first names that belong to the group of the least common first names in the WoS disambiguated author database as determined by being above the second average of the distribution)	3
3	First single name (in compound names, the first element of the name)	Common first single name (i.e., first single names that belong to the group of the most common first single names in the WoS disambiguated author database as determined by being below the average of the distribution)	1
		Uncommon first name (i.e., first single names that belong to the group of the least common first single names in the WoS disambiguated author database as determined by being above the average of the distribution)	2
4	*First single name penalization*	*The author has a first name in the papers but it does not appear in the Twitter name(s) at all*[4]	*–2*
5	Email URL (in the Twitter account and as obtained from the email server URL of the author)	Very common author URL (i.e., URLs that belong to the group of the most common URLs in the Twitter database as determined by being below the average of the distribution)	1
		Common URL (i.e., URL that belongs to the group of the second least common URLs in the Twitter database as determined by being between the two averages – higher and lower – of the distribution)	2
		Uncommon URL (i.e., URLs that belong to the group of the least common URLs in the Twitter database as determined by being above the second average of the distribution)	3

[4] We penalize when authors use their first name in their papers but use a different one (or none at all) in the Twitter name.

Table 1. (*continued*)

Rules	Matching event	Criteria	Score
6	Organization name (i.e., institutional affiliations[5] of the disambiguated authors)	Very common organization name (i.e., organization names that belong to the group of the most common organization names in the WoS disambiguated author database as determined by the average of the distribution)	1
		Common organization name (i.e., organization names that belong to the group of the second least common organization names in the WoS disambiguated author database as determined by being between the two averages – higher and lower – of the distribution)	2
		Uncommon organization name (i.e., organization names that belong to the group of the least common organization names in the WoS disambiguated author database as determined by being above the second average of the distribution)	3
7	City (i.e., cities of the institutional affiliations of the disambiguated authors)	Very common city (i.e., cities that belong to the group of the most common cities in the WoS disambiguated author database as determined by the average of the distribution)	1
		Common city (i.e., cities that belong to the group of the second least common cities in the WoS disambiguated author database as determined by being between the two averages – higher and lower – of the distribution)	2
		Uncommon city (i.e., cities that belong to the group of the least common cities in the WoS disambiguated author database as determined by being above the second average of the distribution)	3
8	Country (i.e., countries of the institutional affiliations of the disambiguated authors)	Common country (i.e., countries that belong to the group of the most common countries in the WoS disambiguated author database as determined by being below the average of the distribution)	1
		Uncommon countries (i.e., countries that belong to the group of the least common countries in the WoS disambiguated author database as determined by being above the average of the distribution)	2
9[6]	Tweeter has tweeted publications from the author (i.e., self-tweeting)	Number of self-tweeted publications: 1–2	3
		Number of self-tweeted publications: 3–5	5
		Number of self-tweeted publications: >5	7
10	Twitter user has tweeted publications from the same micro topic(s)[7] of author's activity (excluding self-tweeting)	Number of overlapping topics tweeted: 1–3	1
		Number of overlapping topics tweeted: 4–6	3
		Number of overlapping topics tweeted: >6	5

[5] For each disambiguated author we considered the most common affiliation in which the author has produced most of their scientific output.

[6] Weights for rules 9, 10, and 13 are not based on the CSS method but on rule of thumb choices.

[7] Micro topics are defined as the fields obtained in the publication-level classification developed by Waltman and Van Eck (2012).

Table 1. (*continued*)

Rules	Matching event	Criteria	Score
11	Paired by co tweeted	The tweeter has been mentioned in at least the same tweet with the paper of the author simultaneously	5
12	Tweeter has tweeted publications from the same journal(s) of author's activity (excluding self-tweeting)	Number of overlapping journals tweeted: 1–5	1
		Number of overlapping journals tweeted: >5	2
13	Commonness of the Twitter account–researcher combination	Combination of 1–2 scholars/Twitter	2
		Combination of 3–6 scholars/Twitter	1
14	The Twitter account was the best match for the WoS author (highest score based on rules 0–13) and vice versa		true/false

For example, if WoS author A is matched with Twitter account B with a score of 5. This is a best match only if A has no other match with a score greater than 5 and if B also has no other match with a score greater than 5.

The matching procedure may produce ties (i.e., matches between a scholar and multiple Twitter accounts and vice versa). Thus, ties arise when a scholar is assigned to more than one Twitter account with the same score, or when the same Twitter account is assigned to more than one (disambiguated) scholar with the same score. Ties do not necessarily mean that some of the tied pairs are invalid, because the disambiguation algorithm can sometimes split single individuals into multiple authors; moreover, an individual scholar can genuinely have multiple Twitter accounts. In the next two figures, we compare two data sets, one where we keep the ties and one where we keep only pairs where the WoS author was a best match to a single Twitter account and vice versa.

Figure 1 presents the number of distinct scholars matched with a Twitter account for different score thresholds for the two data sets. We notice that when including only matches with a final score of 5 or above, we find little difference between the data sets in terms of the number of matches, but that below this threshold removing ties reduces the size of the data set significantly. In fact, at a threshold score of 4, the cumulative number of scholars more than doubles when we include ties, suggesting that a threshold below 5 may introduce a flood of false positives.

2.3.1. Validation

We performed a precision–recall analysis using a "gold standard" of author–Twitter account matches based on ORCID data from 2017 (Haak, Brown, et al., 2016; Paglione, Peters, et al., 2015). The golden set was created by following five steps:

1. Select all ORCID profiles that contain a Twitter handle from the public file of 2017.
2. Limit to Twitter handles that are found in the Altmetric.com data (i.e., Twitter users in ORCID who have tweeted at least one paper).
3. Match the ORCID profiles with authors in the WoS database.
4. Manually verify the golden set to ensure that the Twitter handle included in the ORCID profile is the actual scholar's own Twitter account (i.e., removing those cases

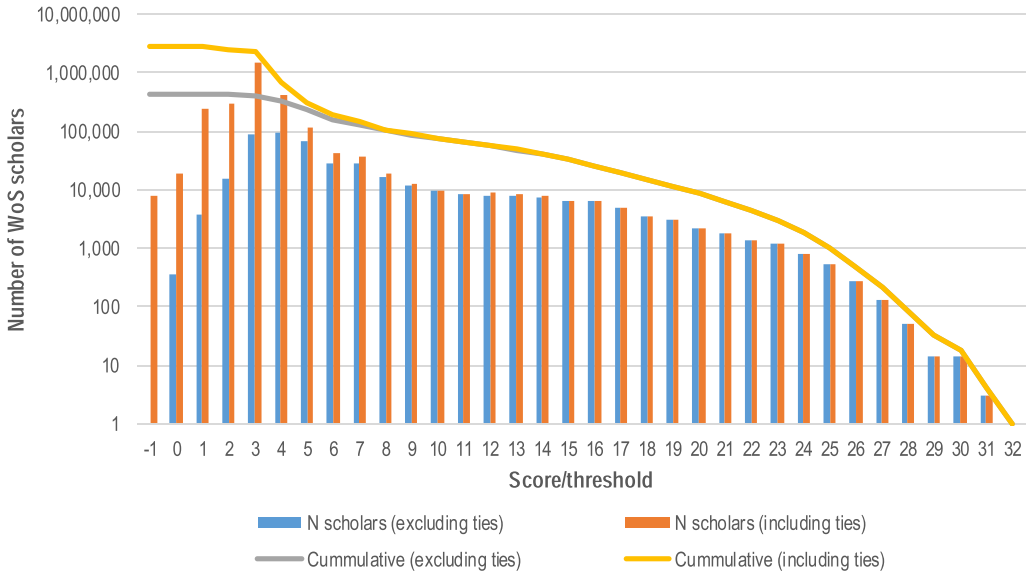

Figure 1. Number of distinct Twitter account-WoS author matches by score threshold and inclusion criteria.

in which scholars in ORCID report their group, department, or collective Twitter accounts).

5. Remove the Twitter accounts that do not include the scholar's name (either in the Twitter handle or in the "name" field).

As a result, we obtain a set of 600 validated author–Twitter account pairs (550 distinct scholars) that we use to calculate the precision and recall of our method for different scores and our two data sets.

Figure 2 presents the number of scholars and the proportion of true positives (precision) for each score, as well as the recall for the cumulative set of scholars by score threshold of the data set. As the results presented in Figure 1 suggested, the precision of the matching drops significantly for scores below 5 when ties are included, suggesting a score of 5 as a reasonable threshold. At this score removing ties does not significantly affect precision but does reduce recall significantly.

3. CHARACTERIZATION OF MATCHED SCHOLARS

3.1. Who Are the Scholars Sharing Papers on Twitter?

In this section, we present a descriptive analysis of 296,504 distinct scholars with at least one publication since 2005 for whom we matched a Twitter account with a score greater than 4 and including ties. This choice was aimed at maximizing recall without compromising precision too much[8].

[8] Other choices could also be possible, depending on the purpose of the study. Thus, studies that would require a much higher level of precision in the selection of matched scholars would be possible, this being achieved by selecting higher score values (and as a result reducing the set of matched scholars), and/or by focusing only on those matches without ties. Appendix 1 provides a breakdown of the number of matched scholars that would be available depending on the choice decided.

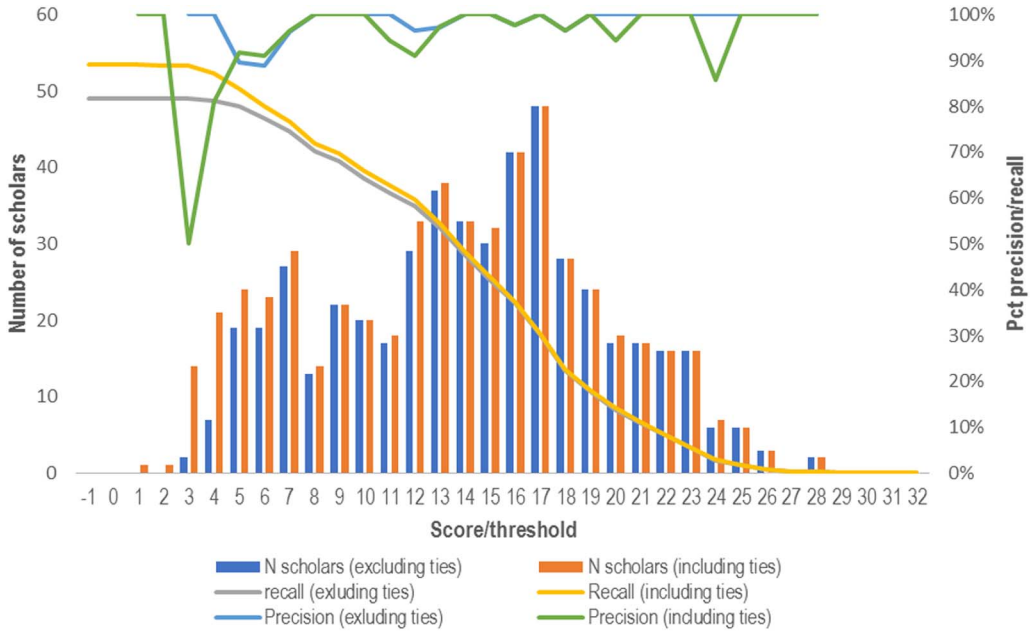

Figure 2. Number of scholars and precision and recall for the cumulative set by score threshold.

We compare the distribution of scholars by country, discipline, academic age, number of publication, and gender to those distributions for the whole set of 25,352,720 disambiguated authors in the WoS database with a publication since 2005. To check the robustness of our analysis, we compared the results presented below with those obtained when excluding ties (not shown). Although the proportions were slightly lower due to the reduced size of the data set, we did not find any discrepancies between the two sets of results.

3.2. Country

Figure 3 compares the distribution by country of scholars on Twitter and of scholars in the WoS database (see Table A2.1 in Appendix 2 for the proportion of scholars on Twitter by country). The country of scholars is determined by their most common country derived from the institutional affiliation(s) of each scholar as indicated in their publications. This is done not only for the scholars matched to a Twitter account but also for those that are not matched, thus allowing for homogeneous country-based comparisons among matched and not matched researchers.

The distribution is highly skewed, with more than 40% of the scholars active on Twitter affiliated to an institution in the United States (26.6%) or the United Kingdom (15.1%). The figure displays the proportion of scholars affiliated to each country in the WoS database. This shows which countries are over represented (e.g., United States, United Kingdom, Canada, Australia, Spain, and the Netherlands) or under represented (e.g., China, Japan, and South Korea) in the Twitter data set. The underrepresentation of China, Japan, and South Korea can to some extent be explained by their use of different alphabets, which reduces our ability to match them with WoS author names. In the case of China, this is exacerbated by the restrictions on Twitter in the country and the existence of local platforms comparable to Twitter, such as Weibo (Zahedi & Costas, 2017).

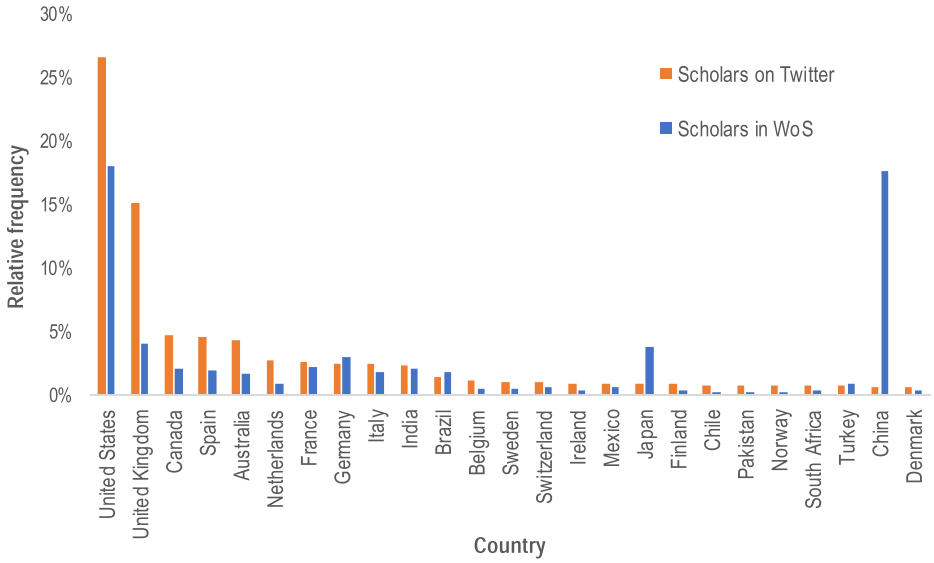

Figure 3. Relative frequency distribution of scholars in the WoS database, with and without a Twitter account, by country.

3.3. Discipline

Figure 4 presents the relative frequency distributions for scholars on Twitter and the authors in the WoS database by field (see Table A2.2 in Appendix 2 for the proportion of scholars on Twitter by field). Individuals are assigned to one of the main fields used in the 2018 version of the Leiden Ranking based on their number of publications in each field, as in Larivière and Costas (2016). However, a scholar with an equal number of publications in multiple fields is assigned to each of these fields. Scholars without any publications classified in the Leiden

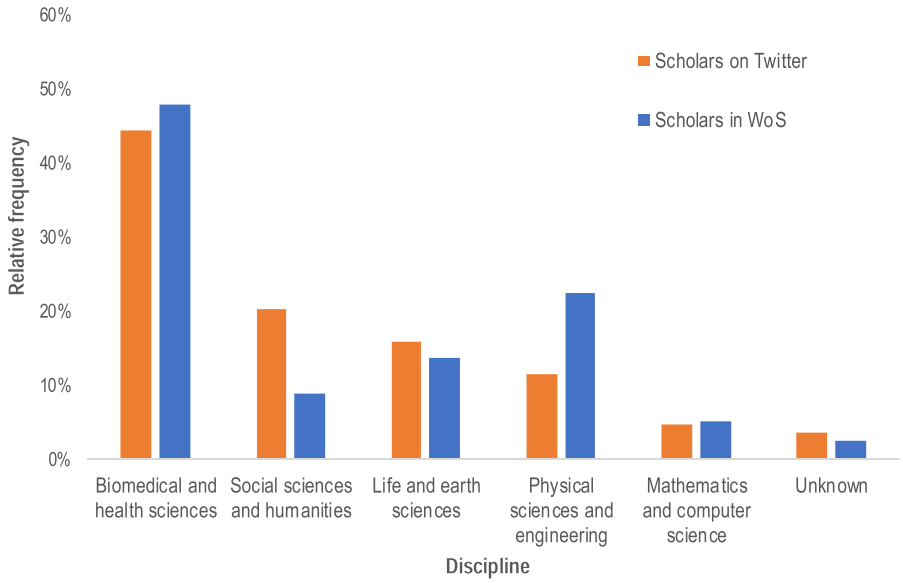

Figure 4. Relative frequency distribution of scholars in WoS with a Twitter account and overall by field.

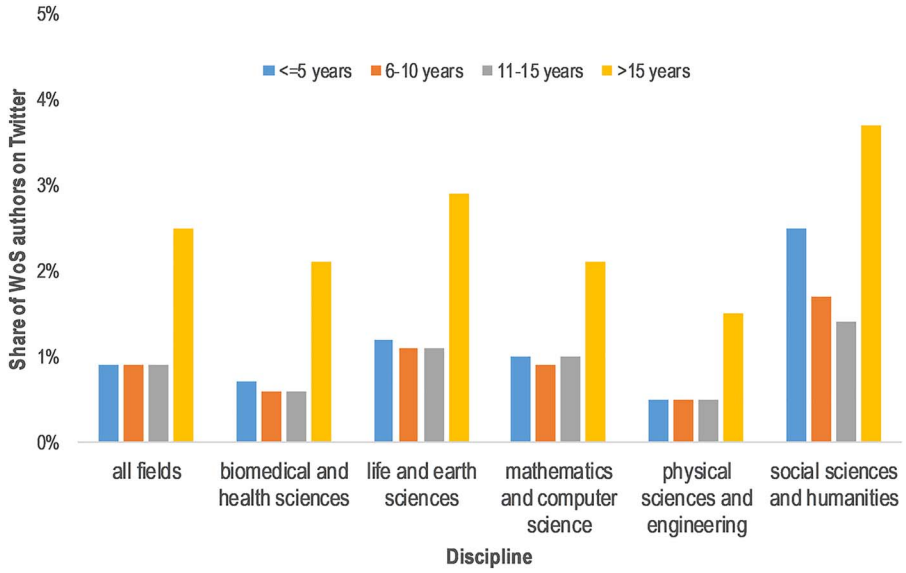

Figure 5. Share of WoS authors with a Twitter account by discipline and academic age.

Ranking form the "Unknown" category. Results show that scholars from "Life and earth sciences" and "Social sciences and humanities," as well are those scholars that could not be assigned to a discipline, are overrepresented among those who share articles on Twitter. Table A2.2 in Appendix 2 also confirms that a higher share of scholars from the "Social sciences and humanities" and "Life and earth sciences" use Twitter, while "Physical sciences and engineering" is the field with the lowest Twitter uptake.

3.4. Academic Age

Figure 5 presents the share of scholars found on Twitter by discipline and academic age, using the year of first publication as a proxy for academic age (Costas, Nane, & Larivière, 2015) and subtracting the year of first publication from 2018, thus obtaining the number of years of activity of the scholars matched on Twitter. In all fields, the older group (>15 years) has the largest share of scholars found on Twitter. There is little difference between the other academic age groups, except in "Social science and humanities," where we observe a greater Twitter uptake among academically younger scholars (<5 years). These results are, however, to be interpreted with caution, as older scholars also tend to have more publications in the WoS. This may influence the matching, because scholars with more output are more likely to be linked to their Twitter account due to the scoring rules (particularly rules 9, 10, and 12) that rely on the number of publications of scholars, as well as the bigger chances of having additional metadata elements, such as email and first names.

3.5. Gender

Figure 6 presents the share of WoS authors found on Twitter by gender[9] and discipline. In all fields, we find that men are slightly more likely than women to be on Twitter. Overall, we see that about 1.5% and 1.3% of the overall numbers of disambiguated male researchers are on Twitter, respectively.

[9] Gender has been defined by combining the data about first names from Larivière, Ni, et al. (2013) with data obtained from https://genderize.io/.

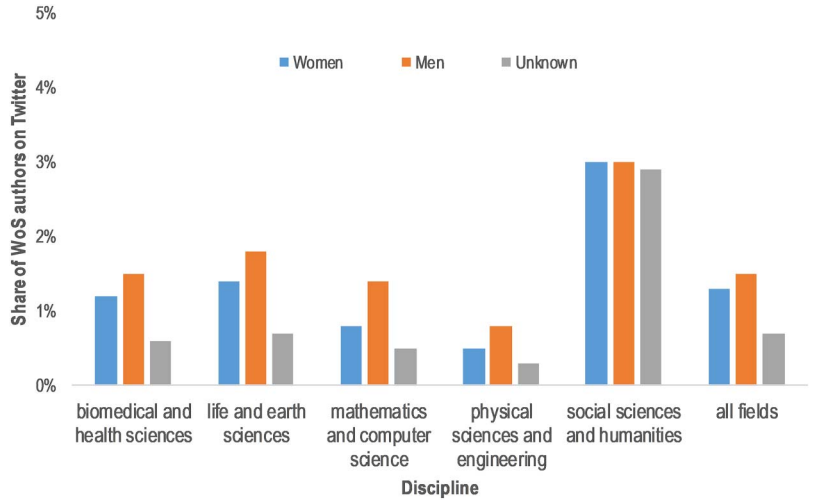

Figure 6. Share of WoS authors with a Twitter account by discipline and gender.

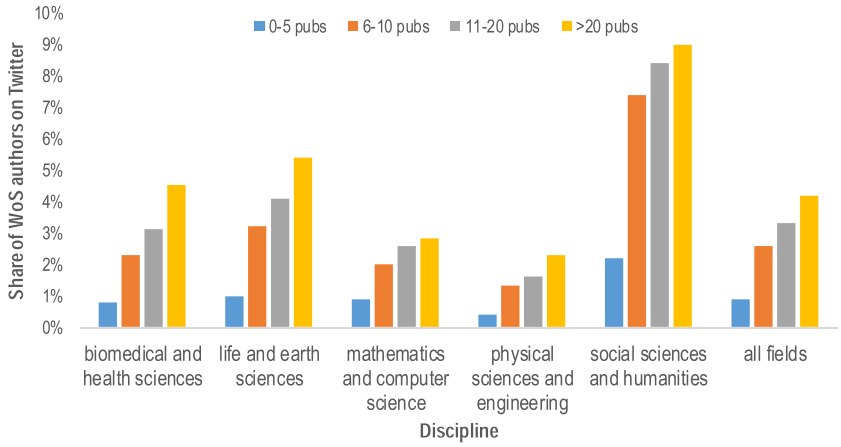

Figure 7. Share of WoS authors with a Twitter account by discipline and number of publications.

3.6. Number of Publications

Figure 7 presents the share of scholars found on Twitter account by field and number of papers published. We find that the share of scholars on Twitter increases with the number of publications. As for Figure 5, these results may be influenced by the scoring rules, which rely on the number of publications of scholars.

4. DISCUSSION AND CONCLUSION

We have presented in this paper an advanced method to match individual authors present in a bibliographic database (in this case the WoS database) and Twitter accounts (using Altmetric.com data in this case). The advantage of our methodology over previous ones is that it is systematic; it can be used with large data sets, and does not only rely on biographical descriptions of the Twitter users or their presence in lists (cf. Ke et al., 2016). As a result, we are able to

identify more scholars on Twitter than previous methods. The fact that we validated our matches using a gold standard, in this case based on ORCID data, also constitutes an advantage of this study. We rely partly on self-identification (the author has to be on Twitter using, to some extent, his or her real name and has to have tweeted a DOI of or link to a paper recorded in Altmetric.com) but also on various other factors to determine whether an account can be ascribed to a particular scientific author. Thus, we move beyond the mere identification of the Twitter user as a scholar in the Twitter biographical section to a linking of those users with their bibliometric information.

The main limitation of our matching approach is its reliance on WoS (see also Ke et al., 2016) and Altmetric.com, which means it can only be used to identify scientists who publish in journals included in the WoS-database and who have tweeted at least a paper with an identifier tracked by Altmetric.com (e.g., DOI or PubMed-ID). This means that we may fail to identify a larger amount of humanities and social science scholars active on Twitter but not publishing in journals included in the WoS database. Similarly, scholars who are on Twitter but not tweeting scientific output properly tracked by Altmetric.com would also be excluded. Also, activity on other social media platforms, such as Weibo, Facebook, and blogs, is not considered in this analysis. Overall, most of the technical challenges identified by Haustein (2016) also play a role in our matching procedure, including the availability of APIs and publication identifiers, the dependency on the decisions of social media data providers (e.g., data available on Twitter and its feasibility of extraction) as well as the dependency on the decisions of altmetric data providers (e.g., Altmetric.com). However, for the matches identified in this paper, the results of our precision and recall analysis support the validity of our methodology.

While the results obtained with our matching algorithm, in terms of the number of matches found and of precision and recall, are very promising, we believe that there is still room for improvement. Future developments necessarily include attempts to match names written in non-Roman alphabets, experimenting with different and alternative weights attributed to each element of the matching rule scores, the consideration of other scoring elements (e.g., network-based properties such as follower/followee and collaborator networks, as well as semantic and cognitive properties, such as hashtags and keywords, or citation proximity between published and tweeted publications), and the use of other gold standards (i.e., validated lists of authors and their Twitter accounts) to further assess the precision and recall of the method.

The results have shown that the numbers of scholars on Twitter vary by levels of productivity; thus scholars with higher levels of production also have a stronger probability of being identified on Twitter. This is likely to be a consequence of the reliance of the method on the number of publications of scholars. It may also be possible that scholars with low levels of output are no longer in academia, or have interrupted periods in their academic careers. In this sense, they might not discuss research on their Twitter account so often because somehow they are more detached from academic life. However, individuals with larger numbers of publications might also be individuals who have tenure and a stronger focus on scholarly research, and therefore they may use Twitter for more academic-related purposes. Further research would be needed in order to delve into the question of the dependency of the method on the number of publications of researchers, which may have repercussions for the consideration of their seniority and career stage, as well as their gender.

We found a strong presence of scholars from the social sciences and the humanities, and a lower Twitter uptake in physical sciences and engineering as well as in mathematics and

computer science. These results align with those reported by Ke et al. (2016), who also reported a higher presence of social scientists and historians on Twitter and lower levels of scholars from the life and natural sciences, as well as fewer mathematicians.

Matching individual scholars with their Twitter accounts allows us to connect two different data sets (Altmetric.com and WoS) in new ways and thus to perform large-scale empirical studies that were not possible before. Further research may seek to enhance the data set by including other databases, such as Scopus, Google Scholar, Microsoft Academic Graph, and Dimensions. This would increase the population of publications and scholars. Including other altmetric data sources (e.g., PlumX Analytics) as well as working with social media platforms directly (e.g., with Twitter, as done in Ke et al., 2016) would also increase the possibilities to identify scholars on different social media platforms and increase the effectiveness of these matching methods. The limitations of the current method could be mitigated by complementing it with the list approach (as in Ke et al., 2016) and an analysis of the biographies of the followers and followings of the identified scholars, opening the path to a more complete perspective of the engagement of scholars on social media platforms.

Once we identify the Twitter account of an author, we can link this with additional bibliometric data (e.g., affiliations, scientific domain, citation impact, collaboration patterns) related to the scientific author. It is of course also possible to extract data from the Twitter handle, thus being able to incorporate information on the online activity of the scholar (e.g., followers, followees, [re]tweeting activity, hashtags). This opens a unique possibility for exhaustive studies on the activities that scholars are performing in social media as well as in their publications.

The combination of bibliometric and altmetric information also opens a clear path to study the relationship between bibliometric performance and Twitter and social media activity. Furthermore, investigating the activities and interactions of scholars on Twitter will help us to better understand and contextualize interactions that scholars are maintaining with other societal stakeholders, as suggested by Robinson-Garcia, van Leeuwen, and Rafols (2018) and paving the way towards more advanced forms of studying the interactions between social media entities and scientific entities in what can be seen as the "social media studies of science" (Costas, 2017; Wouters et al., 2019), at the same time making the possibility of studying "science–society" interactions through social media activities more feasible.

ACKNOWLEDGMENTS

The authors acknowledge the help of Josh Brown, Adèniké Deane-Pratt, and Tom Deranville from ORCID in obtaining the gold standard database, the comments and feedback received from Bijan Ranjbar-Sahraei, Cassidy Sugimoto, and Vincent Larivière on early discussions of this paper, and the support of Jonathan Dudek in correcting the ORCID golden set. The authors thank the two anonymous reviewers of the paper for their constructive comments.

AUTHOR CONTRIBUTIONS

Rodrigo Costas: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing. Philippe Mongeon: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing. Márcia R. Ferreira: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. Jeroen van Honk: Formal analysis, Data curation, Methodology, Validation, Writing – original draft,

Writing – review & editing. Thomas Franssen: Conceptualization, Investigation, Methodology, Supervision, Resources, Writing – original draft, Writing – review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This work has been partially supported by the Eurostars-2 funded project SIA Graph. Rodrigo Costas was partially supported by funding from the DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP) (South Africa).

DATA AVAILABILITY

The data cannot be made available publicly due to the licensing contract terms of the original data.

REFERENCES

Bowman, T. D. (2015). Differences in personal and professional tweets of scholars. *Aslib Journal of Information Management*, *67*(3), 356–371.

Caron, E., & Van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *19th International Conference on Science and Technology Indicators. "Context Counts: Pathways to Master Big Data and Little Data."* Leiden: CWTS-Leiden University.

Chretien, K., Azar, J., & Kind, T. (2011). Physicians on Twitter. *Journal of the American Medical Association*, *305*(6), 566–568.

Collins, K., Shiffman, D., & Rock, J. (2016). How are scientists using social media in the workplace? *PLOS ONE*, *11*, 1–10.

Costas, R. (2017). Towards the social media studies of science: Social media metrics, present and future. *Anales de Investigación*, *13*(1), 1–5.

Costas, R., Nane, T., & Larivière, V. (2015). Is the year of first publication a good proxy of scholars' academic age? In A. A. Salah, Y. Tonta, et al. (Eds.) *Proceedings of the 15th International Conference on Scientometrics and Informetrics* (pp. 988–998). Istanbul: Bogaziçi University Printhouse.

Costas, R., Zahedi, Z., & Wouters, P. (2015a). 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.

Costas, R., Zahedi, Z., & Wouters, P. (2015b). The thematic orientation of publications mentioned on social media: Large-scale disciplinary comparison of social media metrics with citations. *Aslib Journal of Information Management*, *67*, 260–288.

Das, A. K., & Mishra, S. (2014). Genesis of altmetrics or article-level metrics for measuring efficacy of scholarly communications: Current perspectives. *arXiv preprint*, arXiv: 1408.0090.

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.

Dinsmore, A., Allen, L., & Dolby, K. (2014). Alternative perspectives on impact: The potential of ALMs and altmetrics to inform funders about research impact. *PLOS Biology*, *12*(11), e1002003.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, *13*(4), e123.

Friedrich, N., Bowman, T. D., Stock, W. G., & Haustein, S. (2015). Adapting sentiment analysis for tweets linking to scientific papers. *arXiv preprint*, arXiv: 1507.01967.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science*, *14*(2), 123–127.

Haak, L., Brown, J., Buys, M., Cardoso, A. P. Demain, P., … Wright, D. (2016). *ORCID Public Data File 2016*. figshare. https://doi.org/10.6084/m9.figshare.4134027.v1

Hadgu, A. T., & Jäschke, R. (2014). Identifying and analyzing scholars on Twitter. *CEUR Workshop Proceedings*, *1226*, 164–165.

Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Scientometrics*, *108*, 413–423. https://doi.org/10.1007/s11192-016-1910-9

Haustein, S., Bowman, T. D., Holmberg, K., Peters, I., & Larivière, V. (2014). Astrophysicists on Twitter: An in-depth analysis of tweeting and scientific publication behavior. *Aslib Journal of Information Management*, *66*(3), 279–296. https://doi.org/10.1108/AJIM-09-2013-0081

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PLOS ONE*, *10*(3), e0120495.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, *65*(4), 656–669.

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. *Scientometrics*, *101*, 1027–1042.

Hwong, Y.-L., Oliver, C., Van Kranendonk, M., Sammut, C., & Seroussi, Y. (2016). What makes you tick? The psychology of social media engagement in space science communication. *Computers in Human Behavior*, *68*, 480–492.

Ke, Q., Ahn, Y.-Y., & Sugimoto, C. R. (2016). A systematic identification and analysis of scientists on Twitter. *PLOS ONE*, *12*(4), e0175368. Retrieved from http://arxiv.org/abs/1608.06229

Larivière, V., & Costas, R. (2016). How many is too many? On the relationship between research productivity and impact. *PLOS ONE*, *11*, e0162709.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature News*, *504*(7479), 211.

Letierce, J., Passant, A., Breslin, J., & Decker, S. (2010). Understanding how Twitter is used to spread scientific messages. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*. Raleigh, North Carolina.

Lulic, I., & Kovic, I. (2013). Analysis of emergency physicians' Twitter accounts. *Emergency Medicine Journal*, *30*, 371–376.

Ortega, J. L. (2016). To be or not to be on Twitter, and its relationship with the tweeting and citation of research papers. *Scientometrics*, *109*(2), 1353–1364.

Paglione, L., Peters, R., Wilmers, C., Simpson, W., Montenegro, A., … Haak, L. (2015). *ORCID Public Data File 2015*. figshare. https://doi.org/10.6084/m9.figshare.1582705.v1

Priem, J., & Costello, K. L. (2010). How and why scholars cite on Twitter. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4.

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv preprint*. arXiv: 1203.4745.

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

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.

Robinson-Garcia, N., van Leeuwen, T., & Rafols, I. (2018). Using altmetrics for contextualized mapping of societal impact: from hits to networks. *Science and Public Policy*, *45*(6), 815–826. https://doi.org/10.1093/scipol/scy024

Ross, C., Terras, C., Warwick, M., & Welsh, A. (2011). Enabled backchannel: Conference Twitter use by digital humanists. *Journal of Documentation*, *67*, 214–237.

Rowlands, I., Nicholas, D., Russell, B., Canty, N., & Watkinson, A. (2011). Social media use in the research workflow. *Learned Publishing*, *24*(3), 183–195. https://doi.org/10.1087/20110306

Sharma, N. K., Ghosh, S., Benevenuto, F., Ganguly, N., & Gummadi, K. (2012). Inferring who-is-who in the Twitter social network. *ACM SIGCOMM Computer Communication Review*, *42*, 533.

Sugimoto, C., Work, S., Larivière, V., & Haustein. (2017). Scholarly use of social media and altmetrics: Review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037–2062. https://doi.org/10.1002/asi.23833

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLOS ONE*, *8*(5), e64841. https://doi.org/10.1371/journal.pone.0064841

Vainio, J., & Holmberg, K. (2017). Highly tweeted science articles: Who tweets them? An analysis of Twitter user profile descriptions. *Scientometrics*, *112*(1), 345–366.

Van Noorden, R. (2014). Online collaboration: Scientists and the social network. *Nature*, *512*, 126–129.

Veletsianos, G. (2012). Higher education scholars' participation and practices on Twitter. *Journal of Computer Assisted Learning*, *28*, 336–349.

Veletsianos, G., & Kimmons, R. (2016). Scholars in an increasingly open and digital world: How do education professors and students use Twitter? *Internet and Higher Education*, *30*, 1–10.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.

Wouters, P., Zahedi, Z., & Costas, R. (2019). Social media metrics for new research evaluation. In: W. Glänzle, H. Moed, & M. Thelwall (Eds.) *Springer Handbook of Science and Technology Indicators*. Dordrecht: Springer.

Zahedi, Z., & Costas, R. (2017). How visible are the research of different countries on WoS and Twitter? An analysis of global vs. local reach of WoS publications on Twitter. In: *16th International Conference on Scientometrics & Informetrics (ISS)*, Wuhan, China. Retrieved from: https://figshare.com/articles/How_visible_are_the_research_of_different_countries_on_WoS_-and_Twitter_an_analysis_of_global_vs_local_reach_of_WoS_publications_on_Twitter/5481283/files/9479545.pdf

APPENDIX 1: CUMULATIVE NUMBER OF SCHOLARS BY FINAL SCORE WITH AND WITHOUT TIES

Score	Cumulative number of scholars (excluding ties)	Cumulative number of scholars (including ties)
−1	424,264	2,789,711
0	424,263	2,782,120
1	423,905	2,763,330
2	420,220	2,525,832
3	404,801	2,226,179
4	319,351	700,817
5	225,936	296,504
6	157,741	184,364
7	130,205	142,259
8	101,618	106,620
9	85,276	88,409
10	73,447	75,714
11	64,035	65,872
12	55,865	57,471
13	47,845	48,679
14	40,062	40,499
15	32,517	32,790
16	26,060	26,232
17	19,864	19,957
18	15,030	15,090
19	11,445	11,484
20	8,443	8,472
21	6,237	6,256
22	4,425	4,441
23	3,038	3,047
24	1,840	1,843
25	1,022	1,023
26	493	493

APPENDIX 1. (*continued*)

Score	Cumulative number of scholars (excluding ties)	Cumulative number of scholars (including ties)
27	13	213
28	82	82
29	32	32
30	18	18
31	4	4
32	1	1

APPENDIX 2: SHARES OF SCHOLARS ON TWITTER BY COUNTRY AND DISCIPLINE

Table A2.1. Share of scholars on Twitter per country

Country	WoS scholars on Twitter	WoS scholars	% scholars on Twitter
United States	78,915	4,563,994	1.7%
United Kingdom	44,761	1,021,768	4.4%
Canada	13,886	511,327	2.7%
Spain	13,730	483,578	2.8%
Australia	12,744	406,635	3.1%
Netherlands	8,053	208,638	3.9%
France	7,617	548,725	1.4%
Germany	7,246	770,668	0.9%
Italy	7,164	446,403	1.6%
India	6,940	533,621	1.3%
Brazil	4,224	445,767	0.9%
Belgium	3,197	116,431	2.7%
Sweden	2,880	135,504	2.1%
Switzerland	2,824	160,008	1.8%
Ireland	2,585	72,309	3.6%
Mexico	2,526	142,181	1.8%
Japan	2,781	971,080	0.3%

Table A2.1. (*continued*)

Country	WoS scholars on Twitter	WoS scholars	% scholars on Twitter
Finland	2,617	63,482	4.1%
Chile	2,027	52,950	3.8%
Pakistan	2,023	55,196	3.7%
Norway	2,090	57,873	3.6%
South Africa	1,974	64,390	3.1%
Turkey	2,113	208,865	1.0%
China	1,918	4466,730	0.0%
Denmark	1,849	87,523	2.1%
Other countries	29,598	3,818,557	1.7%

Table A2.2 Share of scholars on Twitter per discipline

Discipline	WoS scholars on Twitter	WoS scholars	% scholars on Twitter
Biomedical and health sciences	142,877	13,345,714	1.1%
Life and earth sciences	51,109	3,790,900	1.3%
Mathematics and computer science	15,071	1,399,935	1.1%
Physical sciences and engineering	36,446	6,227,769	0.6%
Social sciences and humanities	65,048	2,464,955	2.6%
Unknown	11,212	670,803	1.7%