Unlocking the microblogging potential for science and medicine

Aditya Sarkar

Department of Computer Science, University of California Los Angeles, 580 Portola Plaza, Los Angeles, CA 90095, USA

asnov2k@gmail.com

ORCID: https://orcid.org/0000-0002-4496-8289

Twitter: SF_AdityaS

Augustin Giros†
Ecole Centrale, Paris, France
augustin.giros@student.ecp.fr

Louis Mockly†
Ecole Centrale, Paris, France
louis.mockly@student.ecp.fr

Jaden Moore#

Orange Coast College, 2701 Fairview Rd, Costa Mesa, CA 92626, USA

moorejaden5389@gmail.com

ORCID: https://orcid.org/0000-0001-7795-402X

Andrew Moore#

Viterbi School of Engineering, University of Southern California, 1985 Zonal Avenue, Room 713. Los Angeles, CA 90089-9121, USA

mooreand@usc.edu

ORCID: https://orcid.org/0000-0003-1457-4593

Anish Nagareddy[#]
Viterbi School of Engineering, University of Southern California, 1985 Zonal Avenue, Room 713. Los Angeles, CA 90089-9121, USA
nagaredd@usc.edu
ORCID: https://orcid.org/0000-0002-5835-4165

Yesha M. Patel
Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033, USA
ypatel@usc.edu
ORCID: https://orcid.org/0000-0002-4983-1222

Karishma Chhugani

Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, 1985 Zonal Avenue, Room 713. Los Angeles, CA 90089-9121, USA

chhugani@usc.edu

ORCID: https://orcid.org/0000-0001-8764-9352

Varuni Sarwal
Department of Computer Science, University of California Los Angeles, 580 Portola Plaza, Los Angeles, CA 90095, USA
vsarwal@cs.ucla.edu
ORCID: https://orcid.org/0000-0003-3302-1409

Nicholas Darci-Maher
Computational and Systems Biology Program, UCLA, Los Angeles, CA, USA.

Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA.
niko.darcimaher@gmail.com
ORCID: https://orcid.org/0000-0002-8124-030X
Twitter: shwaxter

Yutong Chang
Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern
California, 1540 Alcazar Street, Los Angeles, CA 90033, USA
yutongh@usc.edu
ORCID: https://orcid.org/0000-0002-4701-7868

Lana X. Garmire
Department of Computational Medicine and Bioinformatics, University of Michigan
1600 Huron Parkway, Ann Arbor 48105
lgarmire@umich.edu
ORCID: https://orcid.org/0000-0003-1672-6917
Twitter: garmireGroup

Riyue Bao
UPMC Hillman Cancer Center
University of Pittsburgh Department of Medicine
Pittsburgh, PA 15232
baor@upmc.edu
ORCID: https://orcid.org/0000-0002-6105-1704
Twitter: RiyueSunnyBao

Rayan Chikhi*

Department of Computational Biology, Institut Pasteur & CNRS, Paris, France

rchikhi@gmail.com

ORCID: https://orcid.org/0000-0003-1099-8735

Twitter: RayanChikhi

Serghei Mangul*

Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033, USA

serghei.mangul@gmail.com

ORCID: https://orcid.org/0000-0003-4770-3443

Twitter: smangul1

* These authors jointly supervised this work.
†These authors contributed equally
# These authors contributed equally

Abstract

Twitter is one of the most popular microblogging and social networking services, where users can post, retweet, comment, and engage in collaborative discussions. However, improper usage of Twitter can be detrimental to science and even have a negative impact on mental health. Thus, analyzing tweets and Twitter data of various researchers will help us to deduce appropriate ways of using Twitter to advance in our research careers. Existing literature has analyzed the activity of scientists on Twitter, such as studying the relationship between Twitter mentions and article citations, determining the benefits of Twitter in the development and distribution of scientific knowledge, relevant metrics for prediction of highly cited articles, and type of content that researchers tweet etc. For example, Eysenbach et al performed an analysis of tweets and citations and how one can predict citations using tweets. Most of the existing literature analyzed a limited number of researchers, compromising the generalizability of derived results. In our study, we have taken a comprehensive and systematic approach to analyze 167,000 scientists who published research papers on PubMed using data-driven methods. We observed various parameters like number of followers, number of friends, citation count and K-index, in the light of gender, ancestry and profession of the researchers.

Introduction

Twitter is one of the most popular microblogging and social networking platforms allowing users to post, retweet, comment, and engage in discussions[1]. With its advancing popularity among the general public, it has found its way to the research community and is effectively used in sharing ideas, preprints, research articles, and establishing novel collaborations. Twitter allows researchers to both propagate the work of others by retweets and also provide constructive feedback by commenting on the original post and follow the work of others, thereby creating a synergistic and open environment. Scientists use Twitter to establish professional networks, build scientific collaborations, share their experiences, seek advice from the community, and explore job options. In combination with the preprint service widely used by the biomedical community, Twitter promises to mitigate barriers in terms of accessing information, leading to a major impact especially during the early stages of one's scientific career[2]. Thus, analyzing the Twitter data is

very important and it not only ameliorates our understanding of its usage among the research community, but also allows us to infer best practices of using Twitter.

Existing analyses

Existing analyses[3–8] of the activity of scientists on Twitter has many important insights including relationships between Twitter mentions and article citations, the benefits of Twitter in the development and distribution of scientific knowledge, and the content that researchers tweet[3–8]. Some previous works have emphasized the benefits of social media-based scientometrics compared to standard metrics such as citation count[9]. As an example, Eysenbach *et al*[10] performed an analysis of tweets and citations, and proposed scientometrics based on Twitter data to better measure the impact social media has on information dissemination within the research community[10] in a timely manner. However, most existing studies are limited by sample size and have only analyzed a limited number of researchers, potentially compromising generalizability of reported results. For instance, the publication that satirically introduced the name K-index[11] only analyzed data across 40 scientists. Despite the tongue in cheek approach that led to the development of the K-index, it quantifies the visibility of scientists on Twitter relative to their citation count. The K-index of a given researcher is a ratio of the actual number of followers (F_A), over the theoretical number of followers that a researcher should have given their citation count (C), empirically determined to be $F_C=43.3C^{0.32}$. In the original K-index publication[11], it is asserted that having a K-index above 5 is more indicative of being a Twitter celebrity than a scientific one. However a low K-index could also indicate that a scientist is being undervalued; here we empirically suggest a lower K-index threshold of 0.5. Additionally, a recent study[1] performed a survey to understand the ways scientists interact on Twitter. They concluded that the platform keeps the volunteers updated on new preprints, publications, innovations, grants etc. It also helps them in self-promotion, makes them feel connected to the scientific community and interact with like-minded people. With all the benefits that come with the platform, they agree that it also acts as a time sink if not used properly. However, the survey was only able to reach a limited number of researchers (around 100) who agreed to participate in the survey, thus unable to fully capture more general trends of using the social platform by a biomedical community. Studies conducted so far have been unable to quantify the important characteristics of scientific

Twitter including the following factors: fraction of scientists using Twitter, the length of time scientists are using Twitter, their current career status, impact of gender and ancestry on Twitter engagement and how they are followed, and comparison of time scientists are spending on Twitter in regards to the career stage. We aim to overcome a key limitation of previous studies, small sample size, by analyzing substantially larger data accrued for over six thousand scientists; our ultimate goal being to identify relationships between various Twitter parameters, academic parameters and their influence on each other.

Results

In our study, we collected the information of 4.25 million biomedical researchers extracted from 11.61 million research manuscripts published in PubMed. We selected 167,000 scientists who authored at least five publications. Next, we used web scraping techniques to identify the corresponding Twitter accounts of each of 167,000 researchers based on first and last name. Using our method, we obtained 873,000 candidate Twitter handles matching the names of scientists. Next, we gathered the authors' Twitter data, including profile description using the web scraping tool Tweepy[12]. As an additional step, we used the Twitter descriptions to conservatively select scientists by finding specific keywords related to academic titles, for example: "professor", "postdoc", "Ph.D.", etc to ensure that they belong to the scientific background. Next, we removed Twitter handle names that did not match a PubMed author. We resolved duplicate accounts by considering the first search engine hit for the author name as the most likely one to be correct, because it is the most nearest result to the query entered. Applying all the described procedures allowed us to identify 6124 trustable (or similar term) accounts. To benchmark the proposed approach, we randomly extracted 200 scientists from the dataset of 6,123, and then manually identified the corresponding Twitter IDs, thereby obtaining an accuracy of 0.99, indicating 99 out of every 100 scientists are correctly mapped to their Twitter IDs. "Filter" denotes the filtering process performed above.

Additionally, using the metapub[13] and pmidcite[14,15] Python modules, we retrieved the number of citations for each researcher. We then quantified the academic position of scientists using their Twitter descriptions. If a particular keyword was present in their description, then we assigned it

to that category (e.g. keyword 'prof' assigned to the professors). We made four categories based on academic stages which are Professor, Postdoctoral Scholar, Doctoral and Other. The "Other" category includes the remaining scientists who do not belong to the predefined categories, but are working in research based positions as well as those who are misclassified. Our data is composed of 2,786 professors, 254 postdoctoral, 873 doctoral, and 2,210 other types of scientists. Following our recent study[16], we have also estimated the ancestry of the scientists using the ethnicolr python package using the last names of scientists. Based on the 2010 census, we were able to estimate e 4,611 scientists to be European, 1,103 Asian and Pacific Islanders (API), 27 African and 382 Hispanic. We predicted the gender of the scientists from their first names using a Naive Bayes classifier trained on the NLTK corpus names dataset. Out of 6,123 scientists, 4,000 (65%) were predicted to be male and 2,123 (35%) as Female. Despite using these measures for our predictions, we recognize the accuracy of machine learning methods used to infer gender and ancestry is yet to be validated and thus were used with caution. For 6,123 researchers, we have collected the following features: number of followers, number following, citation count, status count, number of days the scientist is active on Twitter, academic profession, ancestry, and gender.

To estimate the fraction of scientists that are present on Twitter, we calculated its lower i.e. the minimum number of scientists on Twitter and the upper bound i.e. the maximum number of scientists on Twitter. Lower bound will give us a limit below which the number of scientists on Twitter cannot fall. Similarly, the upper bound will give us a limit above which the number of scientists cannot increase. A lower bound is given by the number of scientists that are correctly predicted; as estimated by the accuracy of our filters (99%), 6,061 (99% of 6,123) scientists represent 3.63% of the initial pool of 167,000 scientists with five or more publications on PubMed. Our filters conservatively remove accounts with non-scientific keywords in description. If we generously consider that all of these Twitter accounts belong to scientists, we get an upper threshold of 37,743 or 22.5% of the total 167,000. Therefore, we estimate that 3.6-22.5% of biomedical scientists present on Twitter.

We further investigated patterns in Twitter parameters among various groups of scientists based on their academic stage, gender and ancestry. In our data, scientists of European ancestry have

the most number of professors among all of the groups (79.3%), followed by Asian Pacific Islanders (13.7%). Hispanic and African ancestry groups have very few postdoctoral scholars i.e. 6.2% and 0.4% respectively. There are few postdoctoral scholars (6.22%) in Hispanic and African (0.4%) ancestry groups compared to other academic fields, for example, Europeans have 58.66% and API have 34.6% postdoctoral scholars. As expected, Professors have the highest median value of the K-index (1.43) and number of followers (3,460) compared to other academic stage categories lending to their popularity & social presence among the scientific community. On a different note, doctoral students have the highest mean status count per day (mean=0.9, std dev=7.3), followed by Professors (mean=0.7, std dev=2.4), and postdoctoral scholars with the lowest values (mean=0.3, std dev=0.9). We noticed that only a few scientists tend to follow many people, for example, there are only 18 scientists who have less than 150 followers but are following more than 900 individuals indicating scientists tend to be selective in whom they follow.

We also implemented an ordinary least squares regression model to investigate the relationship and impact parameters such as the number of followers, citation count, status count per day, academic stage, ancestry, and gender have on the number of followers. Our target was to analyze the variability of different parameters with respect to followers. It is expected and important to note that in our data, there is a high Pearson's correlation (r=94.3%, p-value < 0.05) between the K-index and number of followers, however the K-index may perform slightly better as a measure and will be our outcome of interest. We used the R-squared metric to quantify the amount of variability explained by the parameters in relation to the number of followers, results of which are discussed below. Min-max Normalization (where the minimum value gets transformed to 0, maximum value gets transformed to 1 and the remaining values lie between 0 and 1) was performed on the data to avoid any bias from the values of the features. In case of categorical variables such as career stages, race and gender, we have used label encoding to assign them a numerical value. Similarly, we find that the number of following, average status per day and gender are significantly associated with the K-index (with $p < 0.05$). So, we examined the variability of the parameters for each gender. Among both males and females, the most significantly associated predictors are following ($p < 0.05$), and average status per day ($p < 0.05$). And for males, ancestry was another significant contributing factor (p=0.4). It is likely a

limitation of a small sample size among stratified observations that we didn't observe a significant association with ancestry among females.

Evaluating the duration of presence of scientists on Twitter and their K-index can determine if their work has influenced their status on the platform. Thus we started with analyzing how long scientists were present on Twitter, where we further assume that the scientists have continuously used Twitter after creating an account. On average, African and Hispanic male scientists have been present on Twitter for the longest duration (11.1 and 9.6 years respectively), whereas male scientists of Asian Pacific Islander and European ancestry were active on Twitter for the least amount of time (9.3 and 9.2 years respectively). Despite being present on Twitter for a long time, scientists from underrepresented ancestry groups still do not have as strong of a social presence (measured by K-index) as European scientists, particularly European Professors who have the highest K-index as previously noted. Overall, male scientists have joined and were present on Twitter (2,773 days), using it earlier than female scientists (2,513 days) with a p-value of 1.1×10^{-7} (Student's T-test). Average K-index of males is 23.4 (median = 1.3, std = 831.9) and females is 3.1 (median = 0.9, std = 12.3). Additionally, we noticed that all scientists had a median K-index value of 1.1 and roughly 32% of scientists had a K-index less than 0.5, indicating these scientists have fewer followers relative to the larger number of citation counts, alluding to their lack of social presence in relation to their publication history. This can happen if a scientist is new to the Twitter environment and has few followers, or if they are posting less frequently on Twitter or their work is not being largely recognized on social media. Approximately 17% of scientists, a majority of whom are males (75%), more than half are professors and are predominantly European had a K-index value greater than 5 and could be considered among those who possess high-profile scientific blogs or Twitter feeds, but have not published many peer-reviewed papers[11]. On average, these scientists engage in 2.6 tweets or retweets per day, which is double the frequency of the average scientist. Remaining 49% have their K-index ranging from 0.5 to 5.

Conclusion

Our analysis provides appealing arguments that Science Twitter provides an effective platform for researchers to disseminate new ideas, establish new collaborations and discuss scientific ideas[7,17]. Ability to leverage Twitter an other social media platform by scientific community is essential skills of scientist of 21 century and will not only help accelerate the spread of broad scientific ideas among diverse communities, but also gain communal feedback, providing an effective means to substantially improve scientific work which can be used in addition to or instead of traditional peer review methods[9,10,18]. Engaging in field-specific discussions on Twitter allows researchers to develop and expand their professional network, launch new research projects and collaborations, communicate with researchers, and receive help from the community at various stages of their projects. It aids researchers in presenting their work and supports international scientists. For example, the Twitter account (@Sci_for_Ukraine) was created to support scholars and students affected by Russia's war in Ukraine; it gained 7900 followers in 40 days. Together with the website and other social media accounts, it is curated by an international community of volunteering scholars. Our analysis indicated that gender and ancestral backgrounds had some impact on the Twitter parameters. Future work is needed to fully understand and interpret the differences we have noted across gender and ancestry groups. We hope our work provides an important and unique insight into the tremendous use of Twitter to effectively communicate scientific ideas to benefit the broad biomedical community. We also hope the results presented here will encourage researchers to join Twitter and become fully engaged on social media for sharing their ideas and contribution across the global scientific community. We believe social media can be effectively used as a harbor of scientific knowledge to widely disseminate novel ideas and innovations among the global research community beyond geographical and administrative boundaries.

ACKNOWLEDGEMENTS:

We thank Jeremy Rotman for his assistance in data analysis. Our paper is dedicated to all freedom-loving people around the world. We would like to thank Sanita Reinsone (@sanitare) for helpful discussions.

DECLARATIONS:

Competing Interests

The authors declare that they have no competing interests.

Data and Material Availability

All supporting material for this analysis can be found at:
https://github.com/Mangul-Lab-USC/Twitter-editorial

Contributions:

SM and RC conceived of the presented idea. NDM prepared the datasets. AS conducted the major analysis of the datasets and wrote the scripts for initial data preprocessing and analysis. YC assisted in curating the data. LM and AG contributed to data analysis. RC, AS, KC, VS and SM contributed to the writing of the manuscript. All authors discussed the text and commented on the manuscript. All authors read and approved the final manuscript.

REFERENCES:

1. Tweet, tweet. *Nat. Methods* **17**, 553–553 (2020).
2. Cheplygina, V., Hermans, F., Albers, C., Bielczyk, N. & Smeets, I. Ten simple rules for getting started on Twitter as a scientist. *PLoS Comput. Biol.* **16**, (2020).
3. Shuai, X., Pepe, A. & Bollen, J. How the Scientific Community Reacts to Newly Submitted

Preprints: Article Downloads, Twitter Mentions, and Citations. *PLOS ONE* **7**, e47523 (2012).

4. *Twitter and society*. (Peter Lang, 2013).
5. Ebner, M. & Reinhard, W. Social networking in scientific conferences - Twitter as tool for strengthen a scientific community. *Sci. TEL Workshop* (2009).
6. Letierce, J., Passant, A., Decker, S. & Breslin, J. G. Understanding how Twitter is used to spread scientific messages. 8.
7. Darling1, E. S., Shiffman, D., Côté, I. M. & Drew, J. A. The role of twitter in the life cycle of a scientific publication. *ArXiv13050435 Phys. Q-Bio* (2013).
8. Average Twitter user has 126 followers, and only 20% of users go via website. *the Guardian* http://www.theguardian.com/technology/blog/2009/jun/29/twitter-users-average-api-traffic (2009).
9. Todd, P. A. & Ladle, R. J. Hidden dangers of a 'citation culture'. *Ethics Sci. Environ. Polit.* **8**, 13–16 (2008).
10. Eysenbach, G. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *J. Med. Internet Res.* **13**, e123 (2011).
11. Hall, N. The Kardashian index: a measure of discrepant social media profile for scientists. *Genome Biol.* **15**, 424 (2014).
12. API Reference — tweepy 3.10.0 documentation. https://docs.tweepy.org/en/latest/api.html.
13. Most, N. *metapub: Pubmed / NCBI / eutils interaction library, handling the metadata of pubmed papers.*
14. Klopfenstein, D. V. & Dampier, W. Commentary to Gusenbauer and Haddaway 2020:

Evaluating retrieval qualities of Google Scholar and PubMed. *Res. Synth. Methods* **12**, 126–135 (2021).

15. Hutchins, B. I. *et al.* The NIH Open Citation Collection: A public access, broad coverage resource. *PLoS Biol.* **17**, e3000385 (2019).
16. Wu, J. *et al.* Virtual meetings promise to eliminate geographical and administrative barriers and increase accessibility, diversity and inclusivity. *Nat. Biotechnol.* **40**, 133–137 (2022).
17. Le, T. T., Himmelstein, D. S., Hippen, A. A., Gazzara, M. R. & Greene, C. S. Analysis of scientific society honors reveals disparities. *Cell Syst.* **12**, 900-906.e5 (2021).
18. 20 Essential Twitter Rules You've Probably Never Heard. *The Muse* https://www.themuse.com/advice/20-essential-twitter-rules-youve-probably-never-heard.