The stability of Twitter metrics: A study on unavailable Twitter mentions of scientific publications

Zhichao Fang¹, Jonathan Dudek¹, Rodrigo Costas¹,²
z.fang@cwts.leidenuniv.nl; j.dudek@cwts.leidenuniv.nl; rcostas@cwts.leidenuniv.nl

¹ Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands
² DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Stellenbosch, South Africa

Abstract

This paper investigates the stability of Twitter counts over time through the analysis of the availability status of over 2.6 million Twitter mentions received by the 1,154 most tweeted scientific publications recorded by Altmetric.com. Results show that for these highly tweeted publications’ Twitter mentions recorded by Altmetric.com up to October 2017, about 14.3% have become unavailable by April 2019. Deletion of tweets by users is the main reason for unavailability, followed by suspension and protection of Twitter user accounts. This study proposes two measures for describing the Twitter dissemination structures of publications: Degree of Originality (i.e. the degree of original tweets of a paper) and Degree of Concentration (i.e. the degree of concentration of retweets in a few original tweets). Twitter metrics of publications with relatively low Degree of Originality and relatively high Degree of Concentration are observed to be at greater risk of becoming unstable due to the potential disappearance of their Twitter mentions. In light of these results, we highlight the importance of paying attention to the potential risks of instability of Twitter counts, and the significance of identifying the different Twitter dissemination structures to study the Twitter metrics of scientific publications.

Keywords

Twitter metrics, Altmetrics, data stability, Twitter unavailability rate, Twitter dissemination structure

Introduction

Data consistency, as one of the challenges in altmetrics (Haustein, 2016), is of great concern in studies of altmetric data. Considering the strong dependency of altmetric data on commercial data providers, previous studies mainly focused on the consistency of various altmetric data among different data aggregators (Ortega, 2018; Meschede & Siebenlist, 2018; Zahedi & Costas, 2018), and the inconsistencies of metrics across data providers were observed. Moreover, as explained by Chamberlain (2013), altmetric data can be collected at different times, which potentially can also end up in obtaining different values of social media metrics, even when collected from the same source and for the same set of publications. This is one of the explanations for the differences in the data collected by different aggregators (Zahedi & Costas, 2018).

In this paper we introduce a different form of altmetric data inconsistency, related to the ever-changing nature of social media data, in which data records and social media events can easily be deleted by their creators, or users may abandon the social media platforms removing all their records from the platform. This form of inconsistency is therefore related to the stability of altmetric data, and more specifically of Twitter metrics. To the best of our knowledge, research on this type of inconsistency of Twitter metric
data, as well as on their underlying causes, is still lacking in the social media metrics literature. In this paper we intend to fill this gap through a large-scale study of Twitter counts of publications collected at different times, focusing also on conceptualizing the potential reasons and risks that the observed instability may pose for the consistent calculation of Twitter metrics.

The main objectives of this study are: (1) to investigate the stability of Twitter metrics by identifying Twitter mentions that have become unavailable over time, and (2) to explore the potential influence that these unavailable tweets may have on the overall Twitter metrics of publications. We addressed the following specific research questions:

Q1. What is the number and share of Twitter mentions of highly tweeted scientific publications in Altmetric.com that have become unavailable over time?

Q2. What are the most common reasons for tweets becoming unavailable?

Q3. To what extent do unavailable Twitter mentions influence the temporal stability of Twitter metrics of scientific publications?

Q4. Is it possible to determine which scientific publications are at a higher risk of substantially decreasing their Twitter metrics due to unavailable tweets?

Data and methods

Availability of Twitter mentions of the most tweeted scientific publications

The Twitter mention data of scientific publications used in this study were extracted from the historical data files provided by Altmetric.com in 2017. Since 2016, Altmetric.com has made annual snapshots of its database available for researchers to study. These snapshots serve as an important point-in-time reference to study tweets that were later unavailable, because in the snapshot data there is still the evidence that a paper was tweeted even in the case when the tweet has been removed from Twitter.

Until October 2017, there are 1,154 scientific publications with Twitter mentions posted by at least 1,000 unique Twitter users, all the tweet IDs (unique identifier of tweets) of tweets to these most tweeted scientific publications were collected from the data files provided by Altmetric.com (version: October 2017).

On the basis of the tweet IDs previously identified by Altmetric.com, in April 2019 we rechecked all the tweets through the Twitter API in order to examine which tweets have their statuses changed. For all tweets that are still available, detailed meta data can be acquired, and for those that are no longer available, error codes and error messages would be responded by the API. Both unavailable tweet IDs and their error codes were recorded for further analysis. For the 2,643,531 Twitter mentions recorded by Altmetric.com until October 2017, a total of 378,766 (14.3%) were unavailable by April 2019.

Indicators for describing Twitter dissemination structure

In order to provide some understanding of the influence that unavailable tweets can have for the calculation of Twitter metrics, we study the Twitter dissemination structures of scientific publications. Twitter dissemination structure refers to the dissemination form of research outputs on Twitter, which is composed of original tweets, retweets, and the retweeting links. Original tweets are defined as Twitter mentions of scientific publications originally posted by Twitter user; while retweets refer to the re-dissemination of original tweets by other Twitter users. Twitter dissemination structure tells the story about how many original tweets a paper has accrued, how many retweets each original tweet has received, and how do these original tweets and retweets connect.
A common Twitter metric for scientific publication is the total count of tweets it has accumulated. However, the dissemination process of a scientific publication on Twitter is too intricate to be explained with a simple number. Studying the Twitter dissemination structure of scientific publications on Twitter can be seen as a more advanced approach to characterize the Twitter metrics of scientific publications. Originality and Concentration are proposed as two dimensions for describing Twitter dissemination structures, which are based on the varieties that can be observed with scientific publications’ original tweets, retweets, and their connections. Figure 1 illustrates four hypothetical examples of original tweet and retweet combinations in order to explain the two main dimensions for describing Twitter dissemination structures of publications. Blue nodes and yellow nodes represent original tweets and their related retweets, respectively. The four publications in the example (publication A, B, C, and D) do all have the same total number of Twitter mentions (TN = 10). From the perspective of total tweet counts they show the same impact on Twitter, but they perform differently through the lens of Originality and Concentration.

Figure 1. Two dimensions for describing Twitter dissemination structures of publications.

Originality is proposed to represent how many Twitter mentions of a specific scientific publication are posted originally by Twitter users rather than retweeting previous tweets. The more original tweets a publication has, the higher its degree of originality is. Degree of Originality (DO) of publication x is calculated as follows:

$$\text{Degree of Originality}_x = \frac{N(OT_x)}{TN(OT_x + RT_x)}$$

$N(OT_x)$ denotes the number of original tweets that publication x has received, while $TN(OT_x + RT_x)$ refers to the total number of Twitter mentions (including all original tweets and retweets) that publication x has accumulated. DO reflects the percentage of original tweets a publication has. In Figure 1, publication A (DO = 0.6) and publication D (DO = 0.6) fall into the category that has accumulated more original tweets; while publication B (DO = 0.3) and publication C (DO = 0.3) belong to the category that has less original tweets.
Concentration is proposed to show the extent to which a publication’s retweets can be linked to a certain amount of original tweets. The more retweets concentrate on a certain original tweet, the higher its degree of concentration. Degree of Concentration (DC) of publication x is given by:

\[ \text{Degree of Concentration}_x = \max \left( \frac{N(RT_{OT_i})}{TN(RT_x)} \right) (i = 1, 2, ..., n) \]

\(N(RT_{OT_i})\) denotes the number of retweets that original tweet \(i\) \((i = 1, 2, ..., n)\) for publication \(x\) has received, \(TN(RT_x)\) denotes the total number of retweets that publication \(x\) has accumulated. DC reflects the maximum percentage of retweets linking to an original tweet. The higher the maximum percentage is, the more retweets a few original tweets have received, while a low maximum percentage reflects a more disperse distribution of retweets. For each publication in Figure 1, the proportions of retweets that every original tweet received are calculated and the maximum one is the DC of that publication. Therefore, DC of publication A and publication B are 1.0 and 0.86, respectively, most retweets of these two publications concentrate on a certain original tweet; while for publication C (DC = 0.43) and publication D (DC = 0.25), their retweets are distributed dispersedly.

In order to study the dissemination structures of sample highly tweeted publications, their original tweets and retweets were distinguished at first. For Twitter mentions that are still available, their collected metadata were used to distinguish if they are original tweets or retweets. Besides, retweets and their corresponding original tweets were also connected, whenever the available metadata would allow for unambiguously determine such connection. For Twitter mentions that are not available on Twitter any more, their status of original tweet and retweet, and their original tweet-retweet connections were established based on the data recorded by Altmetric.com, whenever this was possible.

Results

Reasons for the unavailability of Twitter mentions

Table 1 presents the number of unavailable Twitter mentions arranged by the specific error codes directly provided by the Twitter API. There are four main error codes that signal the unavailability of Twitter mentions. The major reason for the unavailability is that the tweet has been deleted, around 54.7% of unavailable Twitter mention records fall into this category. The second major reason is that the Twitter user accounts have been suspended because of violation against Twitter rules\(^1\), leading to the unavailability of all their tweets. This accounts for 25.9% of all errors returned, and is followed by the protection of tweets implemented by users\(^2\). Once a Twitter user has chosen this setting, unauthorized users cannot get access to their tweets (anymore), although the tweets themselves still exist. During our data collection, this error was found for in 16.7% of all unavailable Twitter mentions. Lastly, 2.7% of unavailable tweet IDs could not be found because the tweet IDs were directing to a page that does not longer exist. It should be noted that in those cases where the tweet IDs are no longer existent (with error codes of 144 and 34), the related Twitter mentions about scientific publications are unrecoverable. Concerning unavailable tweet IDs due to user suspension or tweet protection (with error codes of 63 and 179), it is still possible that they become available to the public again once the suspended user accounts are unlocked or users cancel the protection of their tweets. Nevertheless, whether such reversion will

---

1 See more information about suspended Twitter accounts at: https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts
2 See more information about public and protected tweets at: https://help.twitter.com/en/safety-and-security/public-and-protected-tweets
take place is uncertain, thus the unavailability of these tweet IDs still has a negative effect on the stability of the Twitter metrics.

Table 1. Numbers of unavailable Twitter mentions and reasons for their unavailability.

| Error code | Twitter Error message                                      | Description                                                                 | N    | P   |
|------------|------------------------------------------------------------|----------------------------------------------------------------------------|------|-----|
| 144        | No status found with that ID.                              | The requested Tweet ID is not found (if it existed, it was probably deleted). | 207,147 | 54.7% |
| 63         | User has been suspended.                                   | The user account has been suspended and information cannot be retrieved.   | 98,194 | 25.9% |
| 179        | Sorry, you are not authorized to see this status.         | Thrown when a Tweet cannot be viewed by the authenticating user, usually due to the Tweet’s author having protected their Tweets. | 63,393 | 16.7% |
| 34         | Sorry, that page does not exist.                           | The specified resource was not found.                                     | 10,032 | 2.7% |
| Total      |                                                            |                                                                            | 378,766 | 100.0% |

**Influence of unavailable Twitter mentions on the stability of Twitter metrics**

Figure 2 shows the total number of Twitter mentions (blue line) and still available Twitter mentions (orange line) for 1,154 Altmetric IDs. The Twitter unavailability rate, namely the percentage of unavailable Twitter mentions of each scientific publication, is presented as a yellow dashed line. For clearer visualization, the 1,154 publications are divided into three parts in the order of their total number of Twitter mentions and shown in Figure 2(a), 2(b), and 2(c), respectively. All highly tweeted publications have a certain amount of unavailable tweets, and the amounts vary greatly across publications. Most publications have less than 20% of their Twitter mentions unavailable, their Twitter metrics are relatively stable regarding minor losses. Peaks of yellow dashed line represent those publications with a large share of unavailable Twitter mentions. The top 10 Altmetric IDs with the highest unavailability rates are highlighted with red diamonds, each of them has over 90% of Twitter mentions unavailable. Due to the high unavailability rates, it can be argued that the Twitter metrics of these publications are unstable.

![Figure 2. Twitter unavailability rates of the 1,154 most tweeted scientific publications.](image-url)
Twitter unavailability rates of publications with different Twitter dissemination structures

In order to further investigate the potential influence of different Twitter dissemination structures on the stability of Twitter metrics, we calculated the Degree of Originality (DO) and Degree of Concentration (DC) for 1,154 sample scientific publications and the distribution is shown in Figure 3. Each dot represents a publication, and its color is determined by the Twitter unavailability rate shown in the color bar on the right. The dashed vertical and horizontal lines indicate the median DO (0.284) and median DC (0.203) of all publications, respectively. Moreover, the top 10 publications with the highest unavailability rate of Twitter mentions highlighted in Figure 2 are marked by stars. Most publications with high Twitter unavailability rates are located at the upper left part, especially for the eight starred publications with the highest unavailability rates. Their Twitter dissemination structures have very low DO and quite high DC, which means that once an original tweet with lots of retweets linking to it has been removed, most of that publication’s Twitter mentions become unavailable. This results in the collapse of its Twitter metrics. Some publications at the left lower part, namely those with both low DO and low DC, also show quite high unavailability rate of Twitter mentions. This kind of publications only has few original tweets but most of them received some retweets. Here, the distribution of retweets is more balanced, meaning that the risk of losing most of the retweets received once the original tweet becomes unavailable is not as high as for the publications at the upper left part. However, if the few original tweets come from a specific Twitter user, and that user account is suspended, or that user decides to protect tweets, the stability of Twitter metrics of this kind of publications would be seriously affected as well. This is the case with the two starred publications at the left lower part. There are less publications with high Twitter unavailability rates in the right part. Publications in this part accumulated more original tweets, so they have less retweets that rely on the existence of original tweets. Throughout all four fields, the Twitter metrics of publications with high DO and low DC (right lower part) seem to be the most stable, since their dissemination structures consist of more independent original tweets and more decentralized retweets, which lowers the risk of losing a mass of Twitter records caused by the unavailability of several highly retweeted original tweets.

Figure 3. Distribution of 1,154 scientific publications with different DO and DC.
Conclusions and outlook

This study examines the stability of Twitter metrics of scientific publications by rechecking the statuses of their Twitter mentions. For over 2.6 million Twitter records of the 1,154 most tweeted publications recorded by Altmetric.com until October 2017, about 14.3% of them have become unavailable by April 2019. The main reason for the high unavailability rate is deletion of tweets, followed by suspension and protection of Twitter user accounts. The stability of Twitter metrics vary among publications, most of them have Twitter unavailability rates of less than 20%, but there are some publications showing extremely high unavailability rates. The potential influence of Twitter dissemination structures on the stability of Twitter metrics is investigated. Degree of Originality and Degree of Concentration are proposed to describe Twitter dissemination structures based on original tweets, retweets, and original tweet-retweet connections. Twitter metrics of publications with relatively low Degree of Originality and relatively high Degree of Concentration are at greater risk of being highly unstable. The results of our study do not only discuss the stability and persistency of Twitter metrics of scientific publications and its potential risks, but also shed light on the importance of distinguishing dissemination structures in the consideration of Twitter-based indicators.

It should be noted that though the value of Altmetric.com database snapshots is clear when studying changes in altmetrics over time, due to the new Twitter restriction, Altmetric.com will no longer be able to share tweets related to particular papers over time for tweets that have been removed from Twitter, and researchers are now required to delete unavailable tweets from their locally hosted snapshot files. In the future study, we will pay more attention on the Twitter dissemination process of papers, including optimize the indicators for describing Twitter dissemination structure and explore its possible application in the measurement of scientific outputs’ Twitter performance and impact.

Acknowledgements

Zhichao Fang is supported by the China Scholarship Council. The authors thank Altmetric.com for providing the historical Twitter data of scientific publications, and thank to Stacy Konkiel from Altmetric.com for her explanation of snapshot files and their solutions to historical Twitter records.

References

Chamberlain, S. (2013). Consuming article-level metrics: Observations and lessons. Information Standards Quarterly, 25(2), 4-13.

Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. Scientometrics, 108(1), 413-423.

Meschede, C., & Siebenlist, T. (2018). Cross-metric compatibility and inconsistencies of altmetrics. Scientometrics, 115(1), 283-297.

Ortega, J. L. (2018). Reliability and accuracy of altmetric providers: a comparison among Altmetric.com, PlumX and Crossref Event Data. Scientometrics, 116(3), 2123-2138.

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.

3 Extracted from personal communication with Stacy Konkiel from Altmetric.com.