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

Zhichao Fang, Jonathan Dudek, & Rodrigo Costas

Centre for Science and Technology Studies (CWTS)

10-11-2019
Background

- **Data consistency** is of great concern in studies of altmetric data (Wouters, Zahedi, & Costas, 2018).

- The lack of consistency is seen as one of the most noteworthy data quality challenges that all altmetric indicators have to confront (Haustein, 2016).

- Previous studies mainly focused on the consistency of various altmetric data among different *data aggregators* (Zahedi & Costas, 2018; Ortega, 2018).

---


Data

- Over **2.6 million Twitter mentions** of the 1,154 most tweeted scientific publications (with mentions posted by at least 1,000 unique Twitter users) recorded in the historical data file by Altmetric.com until October 2017 were selected as research objects.

- Statuses of 2.6 million Twitter mentions were rechecked through **Twitter API** in April 2019, error codes and error messages for unavailable tweets are recorded.
Preliminary findings

- 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.

- Deletion of tweets by users is the main reason for unavailability, followed by suspension and protection of Twitter user accounts.

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

- Older Twitter mentions (e.g. from years 2011, 2012, or 2013) exhibit higher proportions of unavailable tweets.

- The longer the time between the tweet and the data collection, the larger the chances of finding unavailable tweets.
Preliminary findings

- All selected highly tweeted publications have a certain share of Twitter mentions unavailable, and most of them have less than 20% of tweets that have become unavailable.

- The top 10 publications have over 90% of Twitter mentions directing to them have become unavailable.
Twitter dissemination structure

• Is it possible to determine which scientific publications are at a higher risk of substantially decreased Twitter metrics due to unavailable tweets?

• Twitter dissemination structure refers to the dissemination form of research outputs on Twitter over time, which is composed of original tweets, retweets, and the retweeting links.
Twitter dissemination structure

10 Twitter mentions

Publication A

Publication B

Publication C

Publication D
Twitter dissemination structure

*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.

*Concentration* is proposed to show the extent to which a publication’s retweets are linked to its most retweeted original tweet.
 Calculation of DO and DC

• The **Degree of Originality** (DO) of publication $x$ is calculated as follows. The more original tweets a publication has, the higher its degree of originality.

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

• The **Degree of Concentration** (DC) of publication $x$ is calculated as follows. The more retweets concentrate on the most retweeted original tweet, the higher the publication’s degree of concentration.

\[
\text{Degree of Concentration}_x = \max \left( \frac{N(RTOT_i)}{TN(RT_x)} \right) \quad (i = 1, 2, ..., n)
\]
Calculation of DO and DC

DO = 3/10 = 0.3
DC = 6/7 = 0.86

DO = 3/10 = 0.3
DC = 3/7 = 0.43

DO = 6/10 = 0.6
DC = 4/4 = 1.0

DO = 6/10 = 0.6
DC = 1/4 = 0.25
TUR of publications with different DO and DC

• Most publications with high Twitter unavailability rates (proportion of unavailable Twitter mentions) are located at the upper left part.

• Throughout all four fields, the Twitter metrics of publications with high DO and low DC (right lower part) seem to be the most stable.
Although most publications with different Twitter dissemination structures keep a relatively low Twitter unavailability rate, publications with extremely unstable Twitter metrics are more likely to occur when they have less original tweets and more concentrated retweets (Group B) or less original tweets and relatively deconcentrated retweets (Group C).
Conclusions

• For over 2.6 million Twitter records of the 1,154 most tweeted publications in the historical data 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 varies among publications, most of them have Twitter unavailability rates of less than 20%.

• Twitter metrics of publications with relatively low Degree of Originality and relatively high Degree of Concentration are at greater risk of becoming highly unstable.
Outlook

• Degree of concentration of Twitter mentions can be calculated through other methods.

• Twitter dissemination structure can be described at the Twitter user level as well.

• For 42.5 million Twitter mentions in the historical data file, about 5.5 million of them (account for 13%) have become unavailable in September 2019.
Thank you for your attention!
Questions or comments?

Zhichao Fang, Jonathan Dudek, & Rodrigo Costas
Centre for Science and Technology Studies (CWTS)

10-11-2019