The bigger your circle, the more you post?

Jérémy Bensoussan*, Ekaterina Kryukova† and Jules Triomphe‡
EPFL
Email: *jeremy.bensoussan@epfl.ch, †ekaterina.kryukova@epfl.ch, ‡jules.triomphe@epfl.ch

***Abstract*—Replication is an important condition for scientific findings. The goal of the current study is to replicate an attention and productivity proposition made in the previous Twitter studies with a new dataset obtained using the Twitter API. Our findings demonstrate the replicability of the attention and productivity hypothesis on the correlation between number of alters or number of friends and number of tweets. This study assessed the state of the influence of attention on productivity, which can inform future studies. It also examined a new approach proposed by Liang and Fu [LF15] to generating a random sample of Twitter users and its associated alters, which might serve as a solution for answering social-scientific questions at the individual level without accessing the complete data archive.**

I. Introduction

The current paper is an extension of Liang and Fu [LF15] analysis on attention and productivity on Twitter. They state that "Twitter users who receive attention from more people will post more frequently than those who receive less attention". This claim is backed by several papers ([HRW09b] and [HRW09a]), however the authors proposed a new strategy to sample data that should be validated and they suggest that this proposition is insensitive to sampling strategies and thus, should be retested [LF15].

Hypothesis validation is approached conceptually. It means that previous findings are not merely reproduced but that the previous conceptual claim is replicated using independent data, and the former statement is rephrased to hypotheses and propositions that can be tested at the individual level. In this way, all analyses in the current study were based on a new dataset.

First, relevant works pertaining to attention and productivity on Twitter will be explored. Then, the data collection method will be described, following which the method used for the analysis will be detailed. Finally, results will be compared to Liang and Fu's [LF15] work.

II. Related Work

Previous studies suggest a strong association between social media users' productivity and the attention they receive. For example, Huberman et al. [HRW09a] claimed that the total number of posts increases with the number of followers and friends. Nevertheless, the total number of posts as a function of the number of followers saturates in the end. Kwak et al. [Kwa+10] found that there are saturation points both for followers and followees. To verify this, Liang and Fu [LF15] analyze the total number of tweets and daily average number of tweets compared to the number of followers, followees and friends[1] with a Twitter dataset of 2015.

Findings of Liang and Fu [LF15] confirm the claims of Huberman et al. [HRW09a] and Kwak et al. [Kwa+10] with the correlation between the number of tweets and followers, followees and friends. Liang and Fu [LF15] found that the Pearson correlation between number of followers and the average number of tweets is 0.70 ($t = 181.86$, $df = 34{,}004$, $p < 0.001$), while the correlation between number of followees and the average number of tweets is 0.55 ($t = 121.24$, $df = 34{,}004$, $p < 0.001$), and the correlation between number of mentioned users and the average number of tweets is 0.77 ($t = 224.38$, $df = 34,004$, $p < 0.001$).

III. Data Collection

The free versions of Twitter's REST APIs are used to collect a representative Twitter dataset. The method used was the same as the one reported in Liang and Fu [LF15], as from October 21, 2013 [Twi13a], Twitter started assigning unique IDs to users from 64-bit integers made up of a combination of timestamp (41 bits), worker number (server id: 10 bits) and sequence number (12 bits rolled-over) [Twi10; Twi13b]. The sparsity of resulting

[1]Friends are users who have been mentioned at least twice in an ego's timeline.

user IDs made it unreasonable to collect a sufficient number of real users within the provided time frame given the rate limits imposed by the API.

First, $3 \times 30,000$ random numbers from 0 to 3 billion were generated. A search using the **GET /2/users** endpoint showed that there were 33,511 valid users among them, including 31,230 public users, of which 26,825 users with at least one alter and 17,397 users with at least one tweet. The latter's timelines were then pulled using the **GET /1.1/statuses/user_timeline.json** endpoint. A total of 4,999,405 tweets were pulled, of which 2,426,816 tweets are unique and have user mentions. These 2,426,816 unique tweets belong to 6,532 unique egos and 173,537 unique ego friends.

IV. Dataset description

The distributions of mean number of daily tweets, followers and following count as well as tweet count for users that have at least one alter are positively skewed. The mean of daily tweets is 0.272171 with a large standard deviation of 2.22 and maximum value of 160 average tweets per day. The mean number of followers is 69 with a large standard deviation of 599 and maximum value of 44,835 followers. The mean number of followees is 95 with a large standard deviation of 467 and maximum value of 34,035 followees. **Figure 1** displays the users distribution for users that posted at least once.

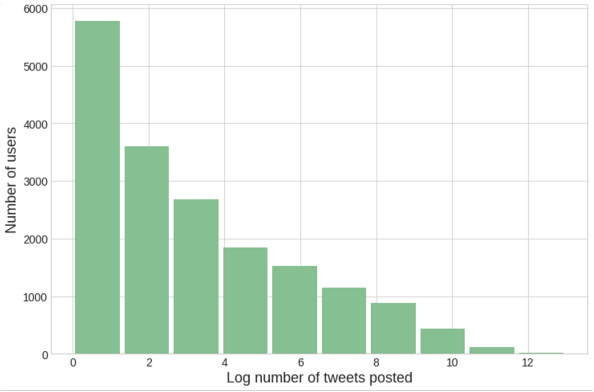

Fig. 1: Number of users per log number of tweets posted

V. Method

The purpose of the analysis is to verify if the results found by Liang and Fu [LF15] are still relevant. Therefore, the processing method is similar to the one they used. Considering the size of the data, PySpark was mainly used to enable distributed calculus.

After having pulled the data from the Twitter API, users having at least one follower or one followee were filtered to remove ghost accounts[2] which might pollute the analysis.

While Kwak and al. [Kwa+10] calculated the number of tweets per user and plot that against the number of followers these users have, here a mean is performed on the number of tweets per users with the same number of followers, followees or friends. Moreover, considering that an older user certainly posted more tweets than a younger one, this bias is mitigated by computing the average number of daily tweets per user, as Liang and Fu [LF15] did (# of total statuses/days since created).

Since all count variables are right-skewed, the correlation coefficients were calculated after log-transformations. To assess the correlations between each factor, the Pearson correlation method is used (**Equation 1**).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

VI. Results and findings

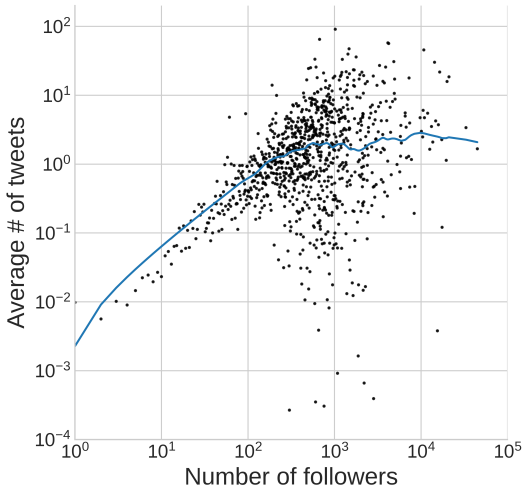

Fig. 2: The average number of tweets as a function of the number of followers

The results reported in **Figures 2**, **3** and **4** corroborate the claims made by Huberman et al. [HRW09a] and the analysis of Liang and Fu [LF15]. Indeed, the trend curves for each graph are increasing, meaning that the the total number of

[2] A ghost account is an account that does not have either a follower or a followee.

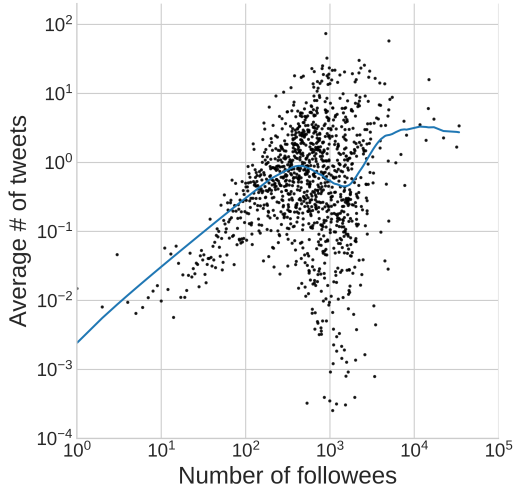

Fig. 3: The average number of tweets as a function of the number of followees

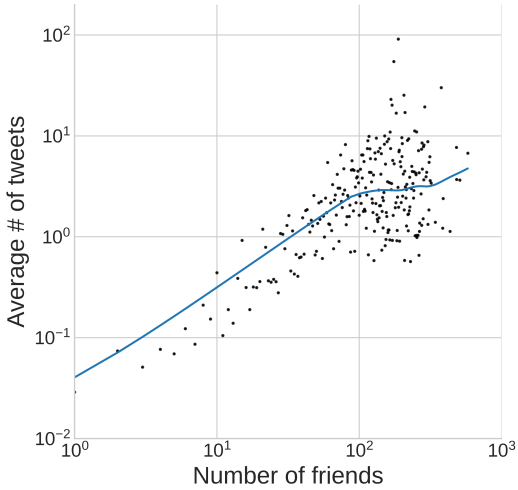

Fig. 4: The average number of tweets as a function of the number of friends.

posts indeed increases with the number of followers, followees and mentioned friends. However, the saturation of the number of posts as a function of the number of followers mentionned by Huberman et al. [HRW09a] is less visible than in their paper. The claim made by Kwak et al. [Kwa+10] extending this saturation effect to the number of followees and friends is clearly visible on **Figures 3** and **4** however.

TABLE I: Comparison of Pearson correlations between the current analysis and Liang and Fu's [LF15].

Pearson correlation between:	Liang and Fu [LF15] dataset	New dataset
Avg. #tweet / #followers	0,70	0,75
Avg. #tweets / #followees	0,55	0,64
Avg. #tweets / #friends	0,77	0,71
Total. #tweet / #followers	0.81	0,75
Total. #tweets / #followees	0.68	0,64

Pearson correlations between average number of tweets and the number of followers, followees and friends are compared in **Table I**. Results from the new dataset are higher than in Liang and Fu [LF15]. The Pearson correlations between average number of tweets and the number of friends are lower than in Liang and Fu [LF15] however. Correlations also differ from Liang and Fu [LF15] correlations between total number of tweets and followers or followees.

VII. Conclusion

This study replicated attention and productivity propositions using a truly representative updated Twitter dataset. The effects of having more followers and followees on the average number of daily tweets posted is strengthened since Liang and Fu [LF15], as the Pearson correlation between variables increased.

The saturation points found by Liang and Fu [LF15] and Huberman et al. [HRW09a] on the number of followers are less visible however. This may be because of a shift in the use of social medias. As the number of users interacting on Twitter and other platforms significantly increased [Seh20], the saturation fades. Since the exact method to create the figures in Liang and Fu's [LF15] study was not published, the current replication may also contain differences explaining this effect. Indeed, here a local regressor is used, whereas Kwak et al. [Kwa+10] use median values of bins for example.

This analysis therefore extends the findings of Liang and Fu [LF15] on attention and productivity, showing that the productivity of social medias users is more dependant of the attention they received than in 2015 when their study was conducted. The reasons for this change are however unclear and require further analyses. Further studies may consider the link between the number of users on social media, the number of interactions, and the attention productivity effect.

References

[HRW09a] Bernardo Huberman, Daniel Romero, and Fang Wu. "Social Networks that Matter: Twitter Under the Microscope". In: *First Monday* 14 (Jan. 2009). DOI: 10.2139/ssrn.1313405.

[HRW09b] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. "Crowdsourcing, attention and productivity". en. In: *Journal of Information Science* 35.6 (Dec. 2009), pp. 758–765. ISSN: 0165-5515, 1741-6485. DOI: 10.1177/0165551509346786. URL: http://journals.sagepub.com/doi/10.1177/0165551509346786 (visited on 12/17/2020).

[Kwa+10] Haewoon Kwak et al. "What is Twitter, a social network or a news media?" In: *Proceedings of the 19th international conference on World wide web*. WWW '10. New York, NY, USA: Association for Computing Machinery, Apr. 2010, pp. 591–600. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772751. URL: https://doi.org/10.1145/1772690.1772751 (visited on 12/17/2020).

[Twi10] Twitter. *Announcing Snowflake*. en_us. 2010. URL: https://blog.twitter.com/engineering/en_us/a/2010/announcing-snowflake.html (visited on 12/17/2020).

[Twi13a] Twitter. *Moving to 64-bit Twitter User IDs*. en_us. 2013. URL: https://blog.twitter.com/developer/en_us/a/2013/64-bit-twitter-user-idpocalypse.html (visited on 12/17/2020).

[Twi13b] Twitter. *twitter-archive/snowflake*. en. 2013. URL: https://github.com/twitter-archive/snowflake (visited on 12/17/2020).

[LF15] Hai Liang and King-wa Fu. "Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science". en. In: *PLOS ONE* 10.8 (Aug. 2015), e0134270. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0134270. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0134270 (visited on 12/17/2020).

[Seh20] Katie Sehl. *Top Twitter Demographics That Matter to Social Media Marketers*. en-US. May 2020. URL: https://blog.hootsuite.com/twitter-demographics/ (visited on 12/18/2020).