Testing social science propositions on Reddit
CS401 - Applied Data Analysis

Clément Lhoste, Haojun Zhu, Dora Lourenço

***Abstract*—Reddit is one of the most famous social media platforms worldwide. Our goal is to use Reddit data to test social science propositions. These propositions were derived from the paper "Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science" [2]. We tested the 20/80 rule of content creation, sociability and attention and productivity. We discovered that the 20%-80% content generation rule can be replicated on the Reddit dataset, Reddit has a sociability level of 61.61%, and attention and productivity seems not related on the platform. In the end, we compared our findings to the ones obtained on the Twitter paper mentioned above.**

I. Introduction

The structure of a social media platform can affect how users interact with it and with other users. On Twitter, users are motivated to improve their own profiles and acquire more followers. This is similar to many other social media platforms like Facebook and Instagram. However, there is a platform that stands out: Reddit. On Reddit, the interaction is different. Users are pseudoanonymized, due to the use of usernames, and the concept of followers and followees does not exist. On Reddit, no one knows who is the real life person behind a profile. Therefore, it would be interesting to test if that could lead to a different user behaviour compared to other platforms.

In this project, we will compare our findings to Twitter propositions. These propositions can be found in the paper "Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science". Out of all the 10 propositions tested in the paper, we are mostly interested in testing 3: the 20/80 rule of content generation, sociability and attention and productivity. All these propositions were rephrased to be applied to Reddit. For the first two rules, we test them at the level of the whole dataset and at the level of subreddits.

II. Dataset

A. Dataset

The dataset we used for the project is available at Reddit Kaggle. This dataset contains around 54 million Reddit comments created on May 2015. Each row on the dataset corresponds to a comment. The dataset contains around 30Gb of data and 22 columns.

B. Data collection

We extract 500 000 users randomly from this dataset. To accomplish this, we take advantage of the column *author*, which indicates which user posted that comment. With this column we are able to know all the users who have posted on May 2015, which are a total of 2 611 449 users. From this group of users we randomly select 500 000 users. Afterwards, we extract their comments. However, we are not interested in all the 22 features, but only on the features author, subreddit, parent_id and score. Therefore, when extracting the comments, we select only the features that we need. The result of our extraction is a new dataset called *comments.csv*. It contains 9 692 614 comments and it is used to test all of our propositions.

III. Propositions

A. 20%-80% content generation rule

First, we wanted to test the 20%-80% content generation. From the previous study of the case of Twitter, this rule can't be verified and the result showed that 2% of the users created 80% of tweets and the distribution of the tweets per user is different from a power law. With the Reddit dataset, we first tried to fit the distribution by a power law and we used the Clauset-Shalizi-Newman method[1] as the previous study did. The exponent is 3.35 with p = 0.0003 and we concluded that the distribution cannot be fitted by a power law. Then, we decided to find the cumulative percentage of comments created by cumulative percentage of authors. The result is given in Figure 1.

The result is different from the Twitter case. For our dataset, 20% of the author created 82% of comments and it verified the content generation rule. Comparing the structure of the two datasets, we found the possible reason for this difference: our Reddit dataset only has information of the authors who had at least one comment

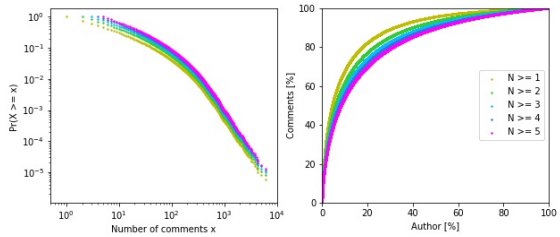

Fig. 1: Log-log plot of the complementary cumulative distribution functions of the number of comments per author(left) and cumulative percentage of comments created by cumulative percentage of authors(right)

in May 2015. It's already different to the Twitter case since there could be authors who had at least one comment but didn't post anything in the month when the data was collected. In fact, the authors with more than one comment in our dataset didn't represent the same sort of population of users as the Twitter study did. So, what we can conclude from our result was, for the active users in a given period of time, the content generation verified the 20%-80% rule. We also tested this content generation rule in some of the most popular subreddits and the result was consistent.

B. Sociability

To test the level of sociability on Reddit, we use the following formula:

$$\frac{\text{Number of comment replies}}{\text{Total number of comments}} \times 100$$

The comments can be root comments of a post or comment replies. To determine if a comment is a root comment or a comment reply, we use the feature *parent_id*. On this dataset, the id of a post starts with *t3* and the id of a comment starts with *t1*. We use this information to know which comment are root comments and replies. A root comment has a post as a parent and a comment reply has a comment as a parent.

On dataset contains 5 971 918 comment replies and 9 692 614 comments in total. The level of sociability obtained is 61.61%.

We are also interested in testing sociability on a subreddit level. To achieve this we group all the comments by subreddit, whose name is given by the feature *subreddit*. However, we don't consider all subreddits in our analysis. We don't consider subreddits which:

- only contain replies to comments. This is done, because it is not possible in a real scenario to have replies to comments without any root comments on a post.
- contain more than 100 comments. This cutoff value excludes 75% of all subreddits.

We then compute the sociability level for each subreddit.

In particular, we are interested in checking whether the ranking of a subreddit affects its level of sociability. We defined the ranking based on the total number of comments. The more comments a subreddit contains, the more popular it is and the higher the ranking. Our results are plotted in Figure 2.

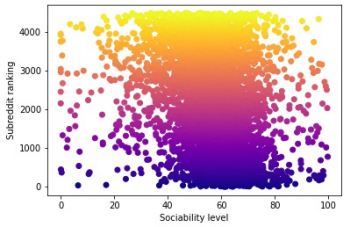

Fig. 2: Sociability level and raking of each subreddit. The closer the color to purple the higher the ranking. The closer the color to yellow the lower the ranking

In Figure 2, we can see that the range of sociability level slightly shifts to the right as the ranking increases. To analyse this data in more detail, we equally split it in 50 categories based on ranking. We then assign rankings to the categories from 1 to 49. By plotting the sociability values for each category with histograms, we could verify that they follow a Gaussian distribution. Therefore, we decided to obtain the mean and variance for each category and check whether or not there is a correlation between them and the categories' rankings. We obtained the plots shown in Figure 3.

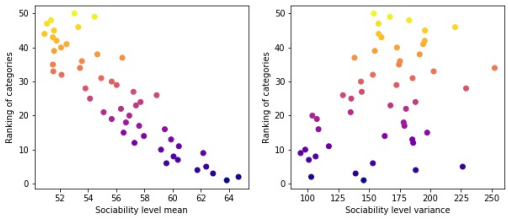

Fig. 3: Mean of the sociability level per category (on the left) and the variance of the sociability level per category (on the right)

By analysing the data in the two plots, we discovered that the sociability level mean has high correlation with

the category of subreddits. We obtained a Pearson R score of -0.9039 and a p-value of 2.4875e-19. When analysing the plot on the right on Figure 3, we obtained a Pearson R score of 0.3819 and a p-value of 0.0062. Therefore, there is no correlation between sociability level variance and category of subreddits.

C. Attention and productivity

Finally, we were interested in how attention is impacting productivity. In Twitter, the average number of tweet is correlated with : the number of followers, followees and mentioned friends [2]. In Reddit, as such metrics does not exist, we decided to measure how the score of a Reddit's comment (sum of up and down votes) would have an impact on the number of comments posted by the user. To do so, We decided to create a scatterplot of the number of comments as a function of the mean score of each users, for the whole dataset of May 2015. The result is shown in Figure 4.

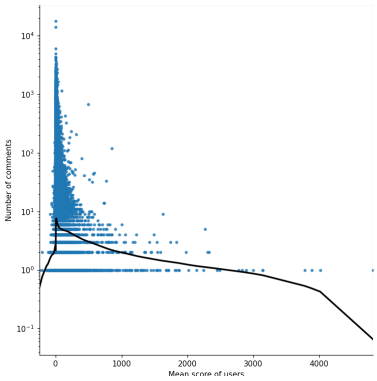

Fig. 4: Testing attention and productivity on Reddit

We can observe an important concentration of users around null or small scores. Multiple users have a negative mean score. The number of comments seems to increase for these negative mean scores, then saturates around a mean score of 20 and finally decrease with higher mean score of users. There is no evident positive or negative relation graphically. This is confirmed by the very low Pearson correlation coefficient: 0.0059, even the correlation can be considered as significant ($p < 0.05$). We then decided to apply the same analysis on subsets of our dataset: the five most popular subreddits. We obtained consistent results similar to the Figure 4 (no evidence of Simpson's paradox).

We need to understand that there is a difference in the nature of the metrics used. In Reddit, users with positive but relatively low mean score (around 10), seems to be the most active. They post a lot of comments, and probably some are unnoticed and get lower scores. Therefore, their mean score is lower by definition. In Twitter, we measure their number of followers for example, which is independent of the fact that some of the tweets can get unnoticed as well.

Beyond this difference of metrics, we can believe there is also a difference in the way the users are dealing with attention in the two platforms. Actually, the number of followers is an important metric in Twitter, as it is a score that appears on every user profile. In Reddit, the system is quite different, as the mean score is a metric we produced for this analysis. Users with the higher mean score in Reddit have only posted few comments. We then produced an analysis only on users with at least one comment with a score superior to 2000 (a quarter of the users), in order to understand if this event could lead to more comments. It is interesting to note that their mean number of comments is 3.5, lower than the mean of the whole distribution: 65. To conclude, users from Reddit that stand out thanks to their high mean score, a mark of attention, tends to be more focus on the content itself rather than their productivity.

IV. Conclusion

Our study focuses on three important propositions related to computational social science, originally tested on Twitter [2], and extended by our work to Reddit. We could replicate the 20%-80% content generation rule, that was not fitting the Twitter dataset. We could also obtain a level of sociability of 61.61%, higher than what was measured on Twitter, where the number of reply and mentions was 31.0%. Finally, we could not find any linear relation between attention and productivity, as observed on Twitter, because of the difference of metrics and the nature of the Reddit platform. But due to the limit of corresponding features in the Reddit dataset, many other propositions cannot be attempted to replicate. It could be possible to improve our work with the help of new datasets or new methods to obtain key features.

References

[1] Newman MEJ Clauset A Shalizi CR. "Power-law distributions in empirical data". In: *SIAM Review* (2009), pp. 661–703.

[2] Hai Liang and King-wa Fu. "Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science". In: *PLOS ONE* 10.8 (Aug. 2015), pp. 1–14. DOI: 10.1371/journal.pone.0134270. URL: https://doi.org/10.1371/journal.pone.0134270.