A LURKING BIAS: REPRESENTATIVENESS OF USERS ACROSS SOCIAL MEDIA AND ITS IMPLICATIONS FOR SAMPLING BIAS IN COGNITIVE SCIENCE

Valtteri Vuorio & Zachary Horne (Department of Psychology, University of Edinburgh)

SOCIAL MEDIA AND COGNITIVE SCIENCE

Old news: University students are W.E.I.R.D.
Solution: Gather data from social media and MTurk?

Data from sites like Reddit, Twitter, and MTurk are widely used in behavioural and social sciences. Yet, there are great demographic differences in each with regards to age, sex, political orientation, and more. Hence, platforms's users cannot be directly compared to one another.

Furthermore, differences in supposedly more diverse samples can be made worse by poor sampling procedures and how users sites generate content, which follows the 1% rule.

90-9-1 PRINCIPLE AKA THE 1% RULE

Most user content is generated by the top 1% of users; 9% contribute rarely; 90% mainly observe.

- 97% of tweets in the 2018 U.S. midterm elections were generated by the most active 10% of U.S. Twitter users
- Wikipedia had 814 million unique users in January 2023, but only 130,218 contributors (0.00016%)
- 80% of HITs are done by 20% of MTurk workers, representing 0.6 to 1.7% of the registered Turkers.

NON-PAIRING OF EXPERIMENTS

Researchers rarely pair online studies with laboratory experiments to validate their work's generalisability.

DATASETS

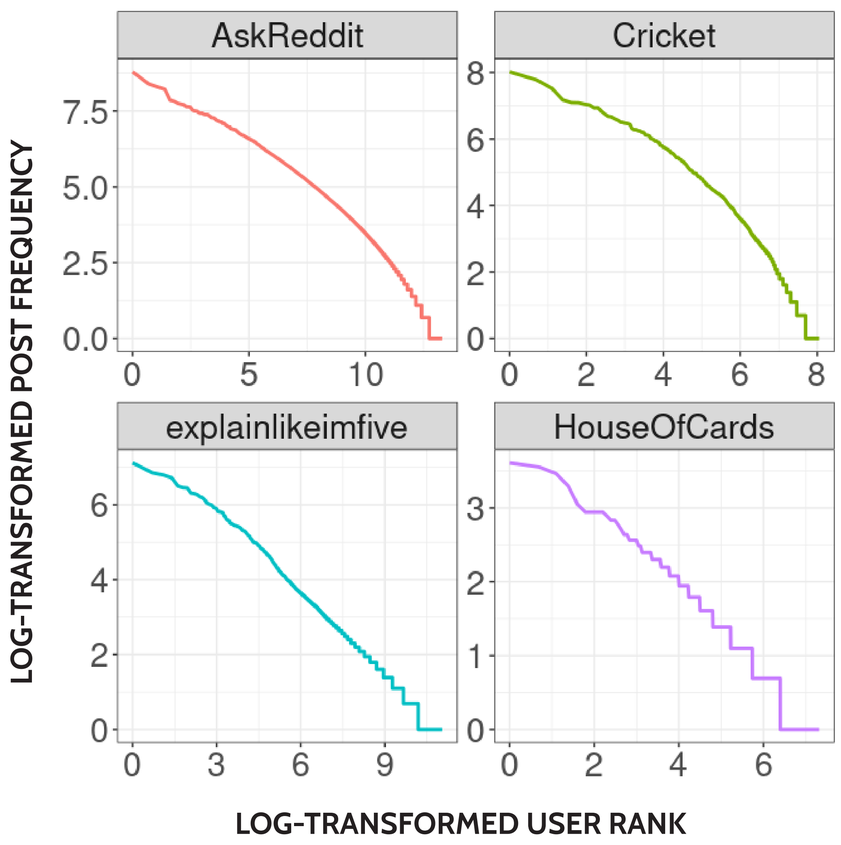
Reddit (10/2014): 42,315,878 comments from 2,209,348 users
Reddit (1/2015): 49,186,418 comments from 2,500,848 users
Twitter (10/2014): 316,669 tweets from X users

PERSONALITY DIFFERENCES BETWEEN HIGH- AND LOW-FREQUENCY USERS

Openess to Experience and Extraversion predict frequency of social media use and engagement with other users, meaning the content studied is produced by non-representative sample of the population.

Study 1: Testing the 1% rule's accuracy

We sought to confirm the validity of the 1% rule and ranked users based on their activity level. Log-transforming the values reveals user participation follows Zipf's law. Regression model including subreddit accounted for 96% of variance, with top 1% of users generating 25% of all comments. Twitter model accounted for 94% of variance, and top 1% created 37% of all comments.



Reddit had 120 million monthly active users in 2015, suggesting only 2% of users participated in active discussion.

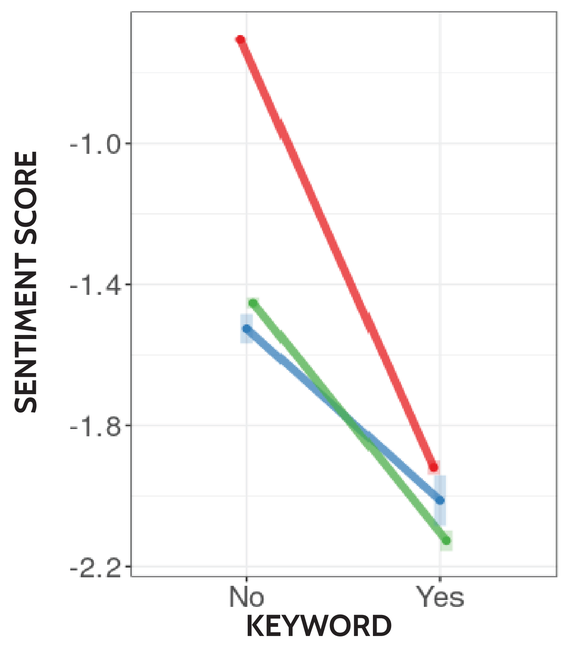
Study 2: Participation inequality's effect on discourse

We predicted high-frequency users to have a disproportionate effect on discussion (Charlie Hebdo). We filtered comments by keywords and calculated sentiment scores using BING dictionary. Most active users differed significantly in their sentiment compared to others. By only focusing on global differences in sentiment scores we could fail to have an accurate measure of attitudes in general.

QUARTILE	POSITIVE (%)	NEGATIVE (%)	COMMENTS (%)
Q1	253,538 (91)	624,522 (92)	138,260 (91)
Q2	17,304 (6.2)	40,108 (5.9)	9,346 (6.2)
Q3	4,613 (1.7)	10,661 (1.6)	2,660 (1.8)
Q4	2,322 (.84)	5,594 (.82)	1,179 (.78)

Study 3: Understanding the source of the sentiment

We investigated group differences between frequency groups and subreddits. Most active quartile produced 87% of all keyword-related discussion. There is a strong positive correlation (r = .75) between baseline sentiment and how users react to keyword presence, exarbating differences between high- and low frequency users. However, there existed no differences in sentiment ratings between subreddits relating to the keywords.



How to overcome participation inequality?

You can't.

REFERENCES

Nielsen, J. (2006). Participation inequality: Encouraging more users to contribute. Jakob Nielsen's alertbox.
Pew Research Center. (2019). National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Tweets
Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. Clinical and Translational Science, 7(4). doi: 10.1111/cts.12178