NETWORK SCIENCE OF ONLINE INTERACTIONS

Chapter 4 exercises + Reddit primer II

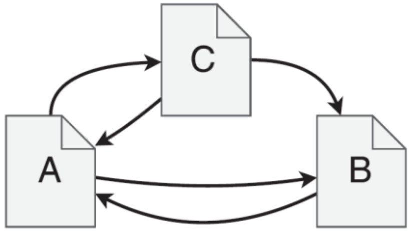
Joao Neto

17/May/2023

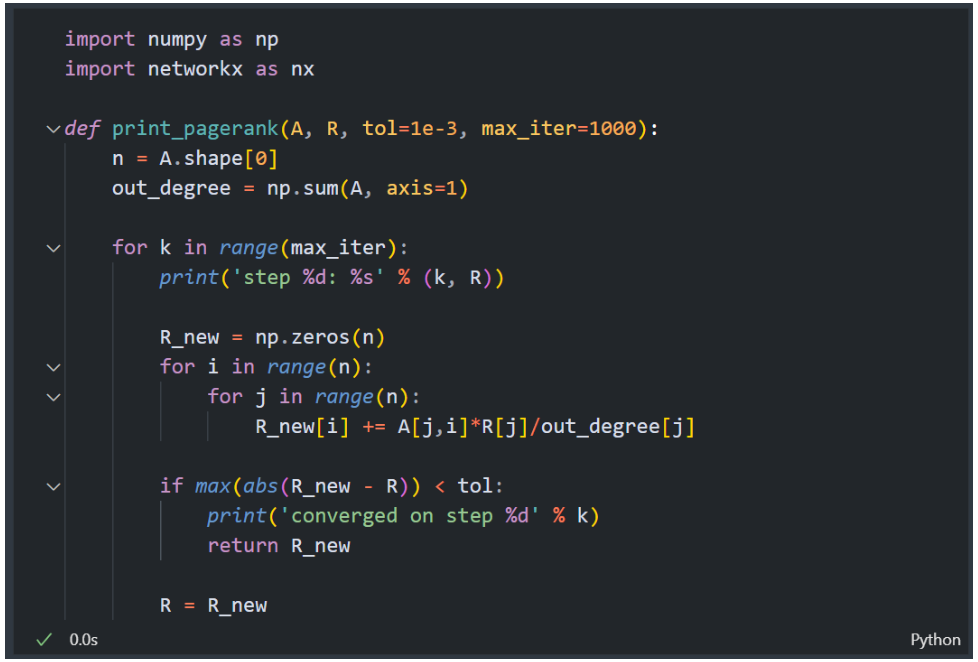
CHAPTER 4 EXERCISES

- Exercises: 4.4, 4.9

4.4 Consider the small network in Figure 4.19. Initialize the PageRank of each page with the value $R_0 = 1/3$. Apply Eq. (4.1) without teleportation ($\alpha = 0$) to calculate the values of PageRank in the next iteration ($t = 1$). Continue to update the values until convergence — assume the values have converged when there is no change in the third decimal digit of each node's PageRank. [*Hint*: Be sure to use the values from the previous iteration when computing the new values; for example, use the initial values ($t = 0$) when calculating the $t = 1$ values.] After how many iterations do the values converge? What are the final values of PageRank?

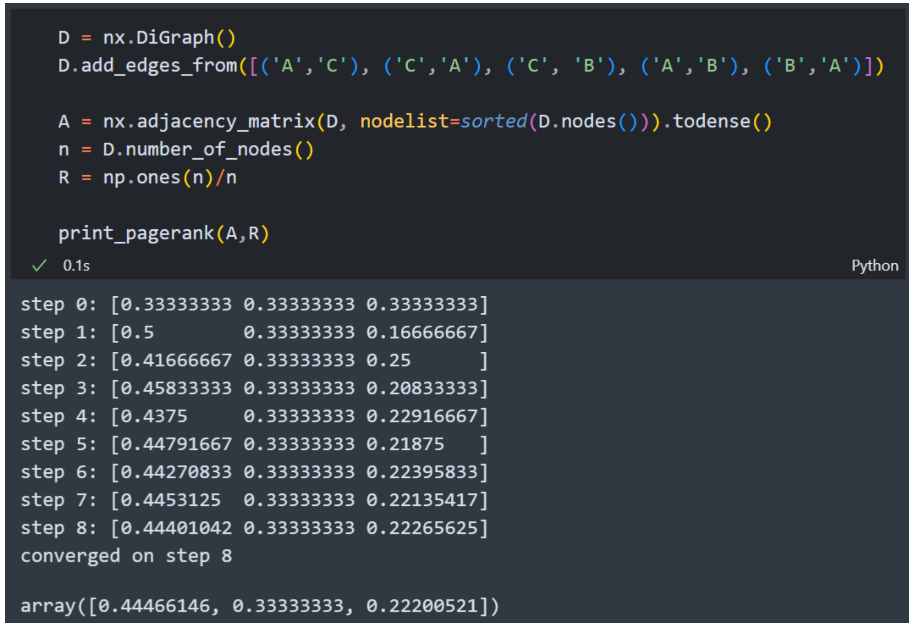


$$R_t(i) = \boxed{\frac{\alpha}{N}} + \boxed{(1-\alpha) \sum_{j\in pred(i)} \frac{R_{t-1}(j)}{k_{out}(j)}}$$



```python
import numpy as np
import networkx as nx

def print_pagerank(A, R, tol=1e-3, max_iter=1000):
    n = A.shape[0]
    out_degree = np.sum(A, axis=1)

    for k in range(max_iter):
        print('step %d: %s' % (k, R))

        R_new = np.zeros(n)
        for i in range(n):
            for j in range(n):
                R_new[i] += A[j,i]*R[j]/out_degree[j]

        if max(abs(R_new - R)) < tol:
            print('converged on step %d' % k)
            return R_new

        R = R_new
```

✓ 0.0s Python

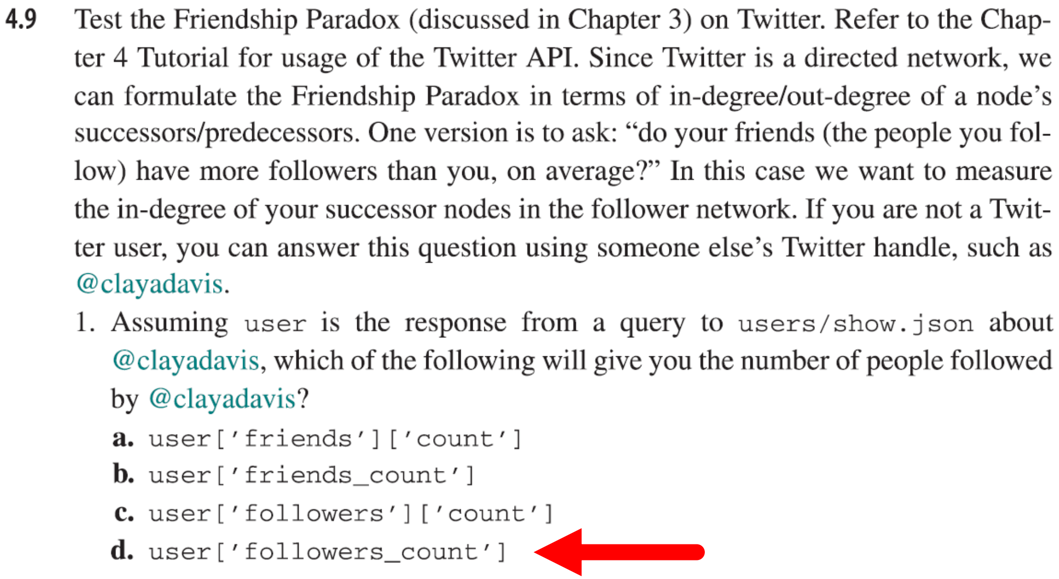


```python
D = nx.DiGraph()
D.add_edges_from([('A','C'), ('C','A'), ('C', 'B'), ('A','B'), ('B','A')])

A = nx.adjacency_matrix(D, nodelist=sorted(D.nodes())).todense()
n = D.number_of_nodes()
R = np.ones(n)/n

print_pagerank(A,R)
```

✓ 0.1s Python

```
step 0: [0.33333333 0.33333333 0.33333333]
step 1: [0.5        0.33333333 0.16666667]
step 2: [0.41666667 0.33333333 0.25      ]
step 3: [0.45833333 0.33333333 0.20833333]
step 4: [0.4375     0.33333333 0.22916667]
step 5: [0.44791667 0.33333333 0.21875   ]
step 6: [0.44270833 0.33333333 0.22395833]
step 7: [0.4453125  0.33333333 0.22135417]
step 8: [0.44401042 0.33333333 0.22265625]
converged on step 8

array([0.44466146, 0.33333333, 0.22200521])
```

CHAPTER 4 EXERCISES



4.9 Test the Friendship Paradox (discussed in Chapter 3) on Twitter. Refer to the Chapter 4 Tutorial for usage of the Twitter API. Since Twitter is a directed network, we can formulate the Friendship Paradox in terms of in-degree/out-degree of a node's successors/predecessors. One version is to ask: "do your friends (the people you follow) have more followers than you, on average?" In this case we want to measure the in-degree of your successor nodes in the follower network. If you are not a Twitter user, you can answer this question using someone else's Twitter handle, such as @clayadavis.

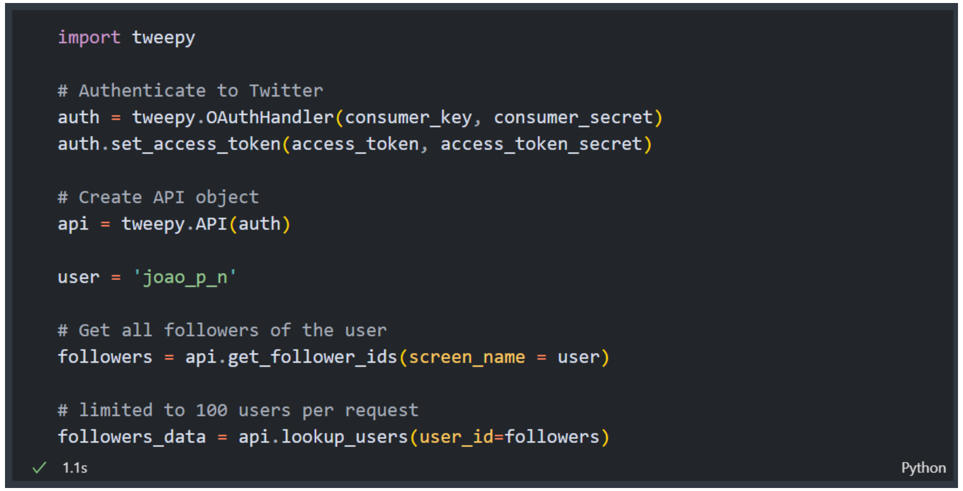
1. Assuming `user` is the response from a query to `users/show.json` about @clayadavis, which of the following will give you the number of people followed by @clayadavis?

   a. `user['friends']['count']`

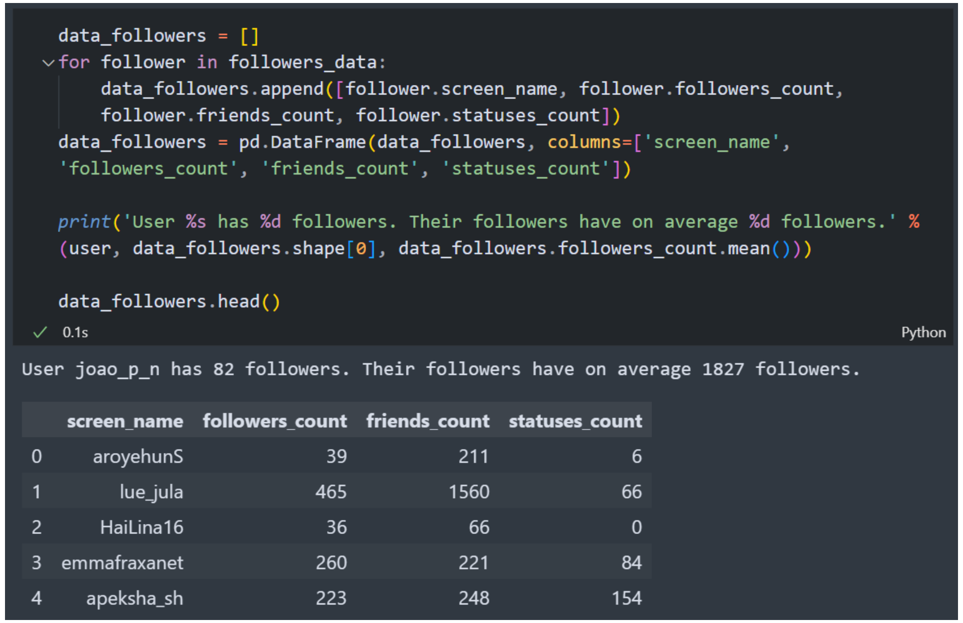
   b. `user['friends_count']`

   c. `user['followers']['count']`

   d. `user['followers_count']`



```
import tweepy

# Authenticate to Twitter
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

# Create API object
api = tweepy.API(auth)

user = 'joao_p_n'

# Get all followers of the user
followers = api.get_follower_ids(screen_name = user)

# limited to 100 users per request
followers_data = api.lookup_users(user_id=followers)
```

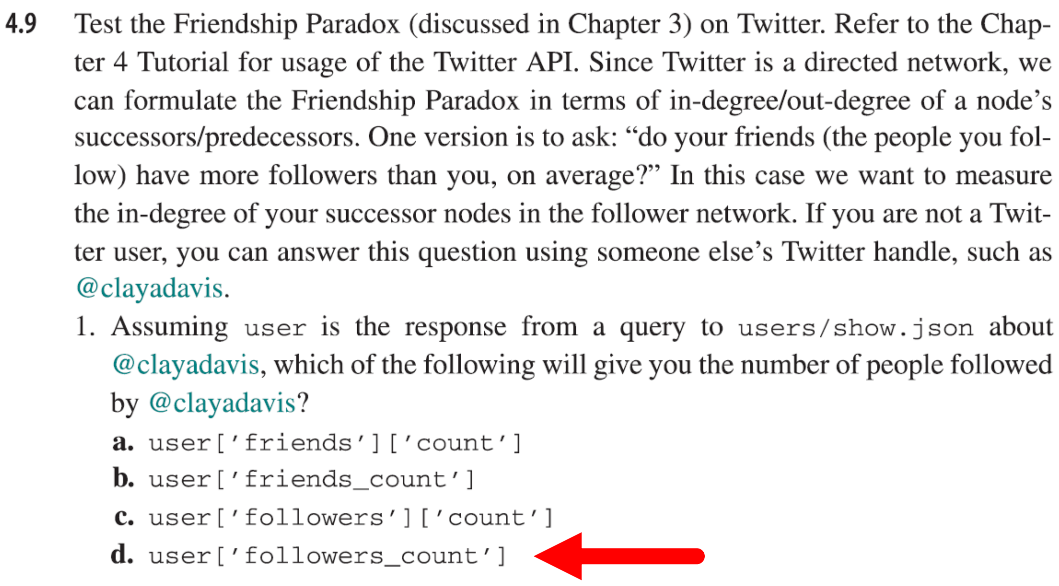


```
data_followers = []
for follower in followers_data:
    data_followers.append([follower.screen_name, follower.followers_count,
    follower.friends_count, follower.statuses_count])
data_followers = pd.DataFrame(data_followers, columns=['screen_name',
'followers_count', 'friends_count', 'statuses_count'])

print('User %s has %d followers. Their followers have on average %d followers.' %
(user, data_followers.shape[0], data_followers.followers_count.mean()))

data_followers.head()
```

```
User joao_p_n has 82 followers. Their followers have on average 1827 followers.
```

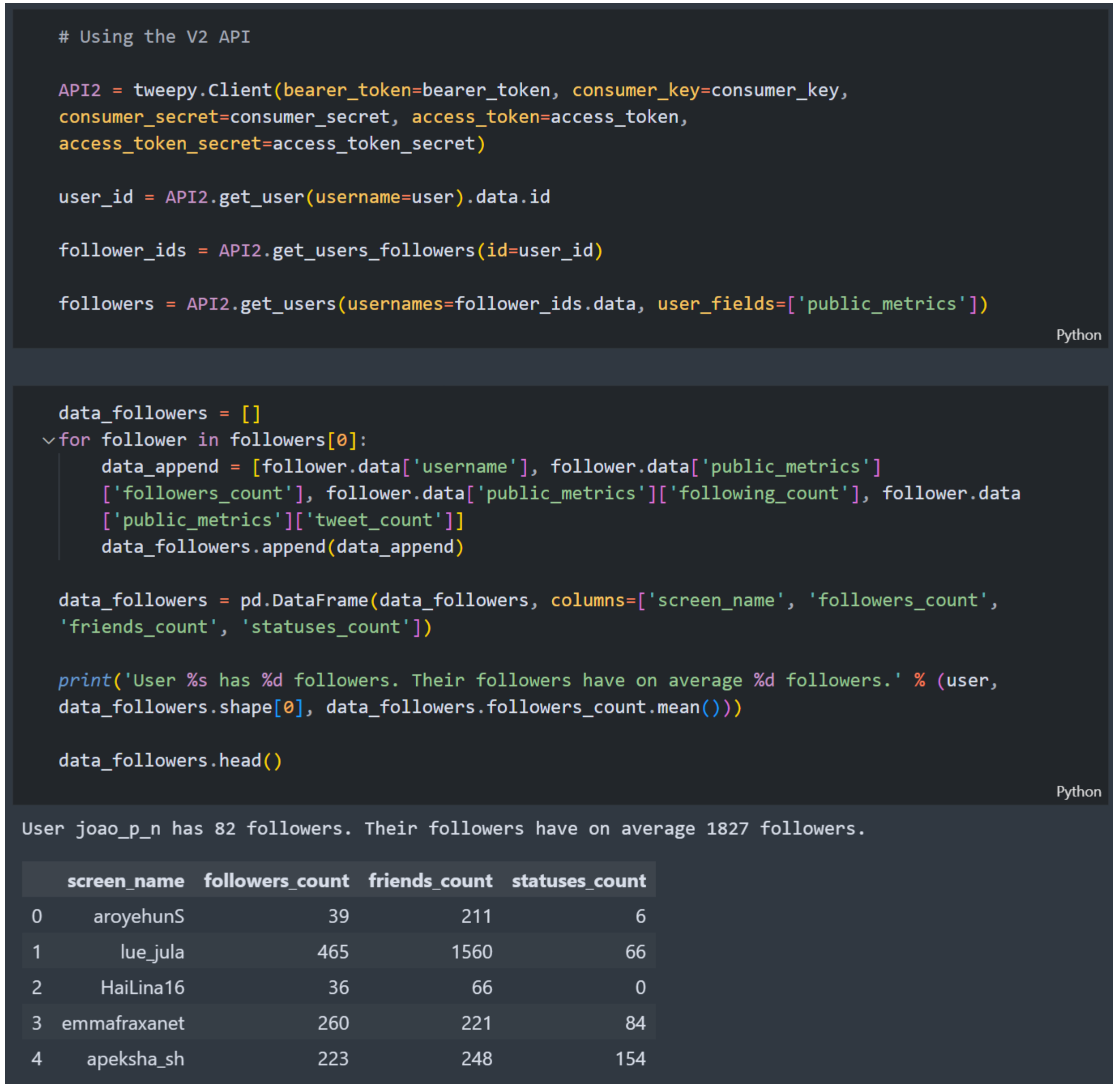
	screen_name	followers_count	friends_count	statuses_count
0	aroyehunS	39	211	6
1	lue_jula	465	1560	66
2	HaiLina16	36	66	0
3	emmafraxanet	260	221	84
4	apeksha_sh	223	248	154

CHAPTER 4 EXERCISES



4.9 Test the Friendship Paradox (discussed in Chapter 3) on Twitter. Refer to the Chapter 4 Tutorial for usage of the Twitter API. Since Twitter is a directed network, we can formulate the Friendship Paradox in terms of in-degree/out-degree of a node's successors/predecessors. One version is to ask: "do your friends (the people you follow) have more followers than you, on average?" In this case we want to measure the in-degree of your successor nodes in the follower network. If you are not a Twitter user, you can answer this question using someone else's Twitter handle, such as @clayadavis.

1. Assuming `user` is the response from a query to `users/show.json` about @clayadavis, which of the following will give you the number of people followed by @clayadavis?
   a. `user['friends']['count']`
   b. `user['friends_count']`
   c. `user['followers']['count']`
   d. `user['followers_count']`



```
# Using the V2 API

API2 = tweepy.Client(bearer_token=bearer_token, consumer_key=consumer_key,
consumer_secret=consumer_secret, access_token=access_token,
access_token_secret=access_token_secret)

user_id = API2.get_user(username=user).data.id

follower_ids = API2.get_users_followers(id=user_id)

followers = API2.get_users(usernames=follower_ids.data, user_fields=['public_metrics'])
```

```
data_followers = []
for follower in followers[0]:
    data_append = [follower.data['username'], follower.data['public_metrics']
    ['followers_count'], follower.data['public_metrics']['following_count'], follower.data
    ['public_metrics']['tweet_count']]
    data_followers.append(data_append)

data_followers = pd.DataFrame(data_followers, columns=['screen_name', 'followers_count',
'friends_count', 'statuses_count'])

print('User %s has %d followers. Their followers have on average %d followers.' % (user,
data_followers.shape[0], data_followers.followers_count.mean()))

data_followers.head()
```

```
User joao_p_n has 82 followers. Their followers have on average 1827 followers.
```

	screen_name	followers_count	friends_count	statuses_count
0	aroyehunS	39	211	6
1	lue_jula	465	1560	66
2	HaiLina16	36	66	0
3	emmafraxanet	260	221	84
4	apeksha_sh	223	248	154

CHAPTER 4 EXERCISES

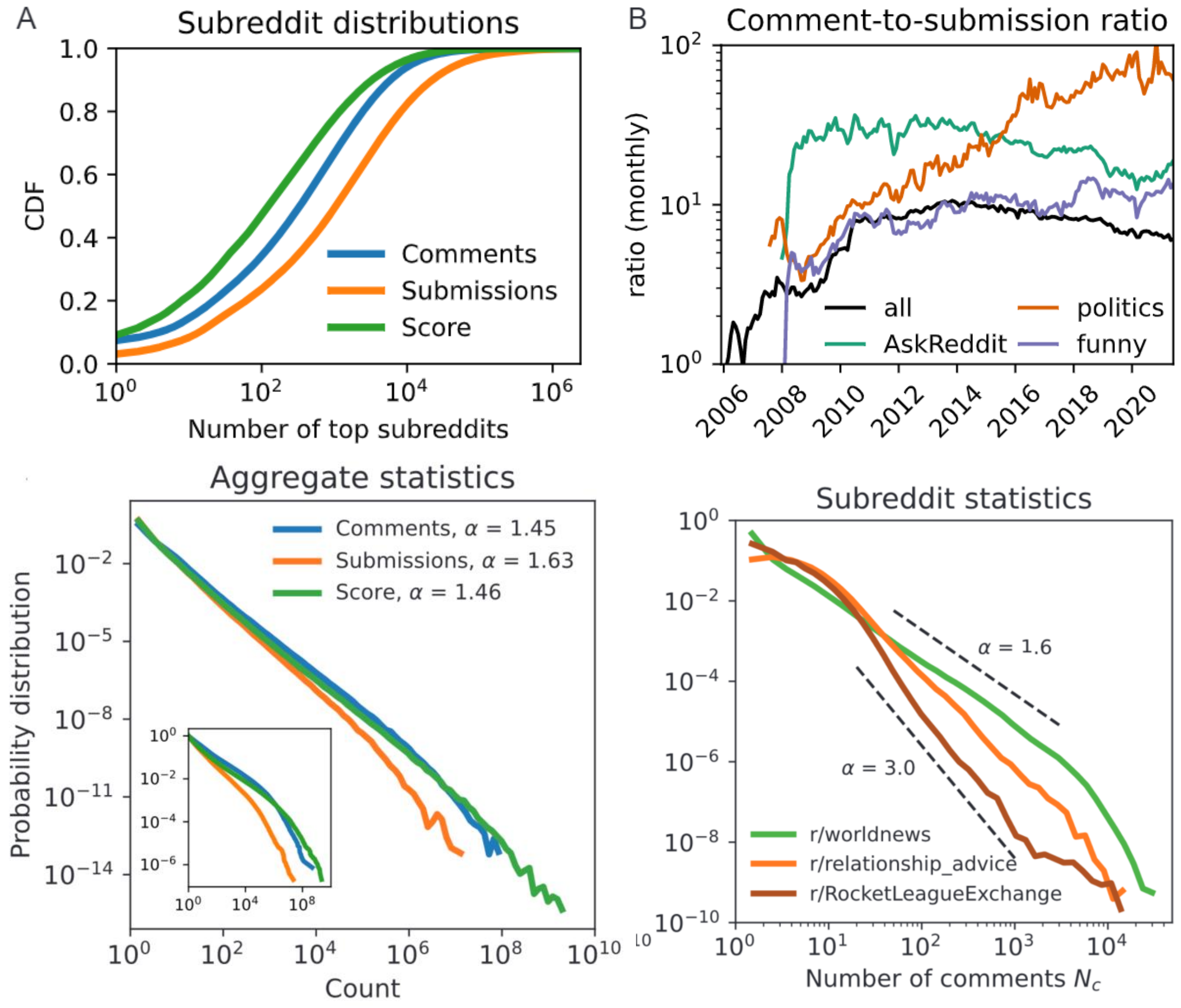
- Questions?

- Reminder
  - Project registration: **11/Jun/23**
  - Project guideline:
    - https://github.com/joaopn/teaching_networks_2023/blob/main/project_guidelines.md

REDDIT PRIMER II

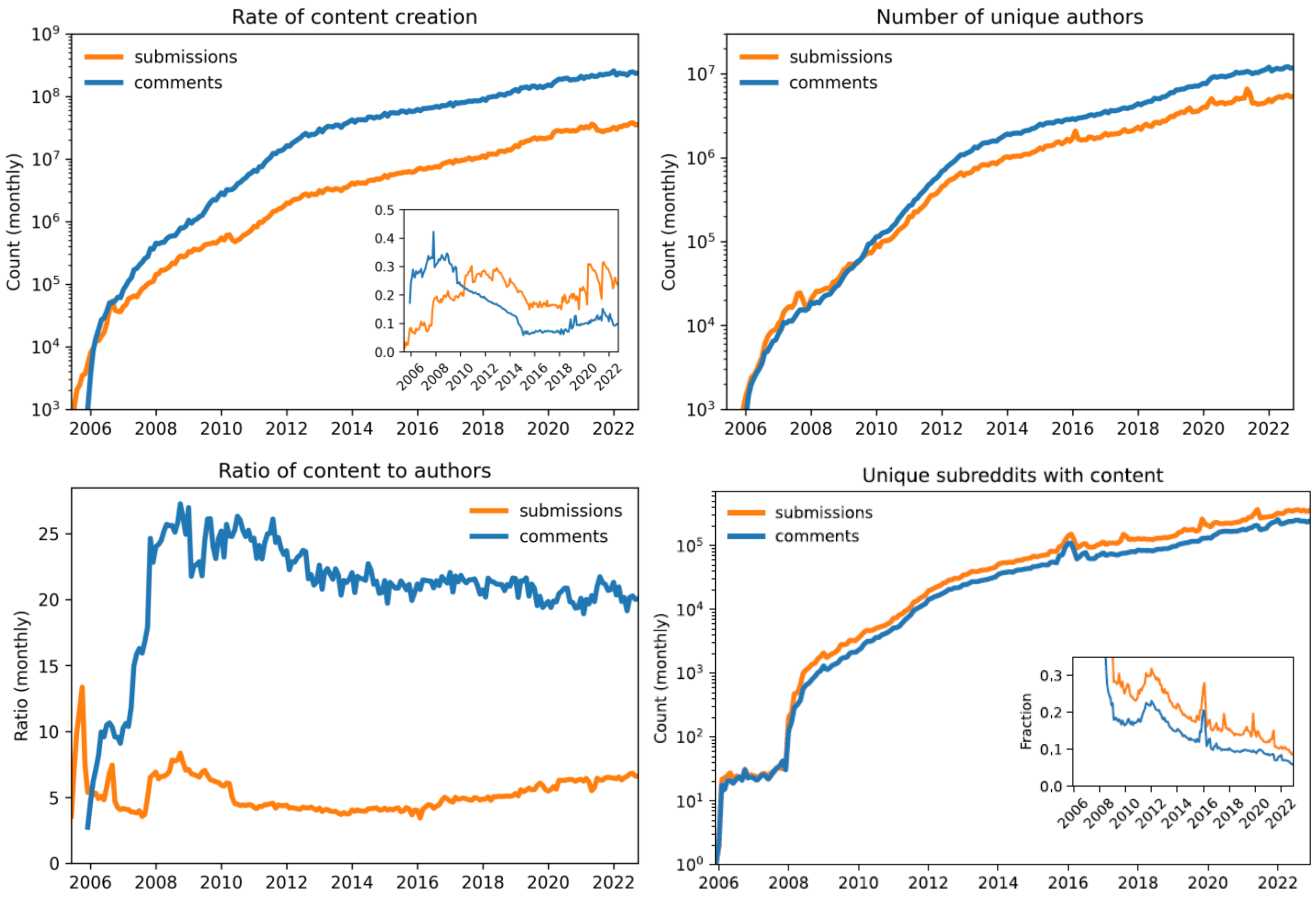
- Previously
  - Content is highly concentrated
  - Subreddits evolve
  - Almost all distributions are heavy-tailed
  - High subreddit heterogeneity

- Today
  - More statistical analyses from different POVs
    - Subreddit
    - Author
    - Submission/comment
  - Summary of relevant literature



RATE PROPERTIES

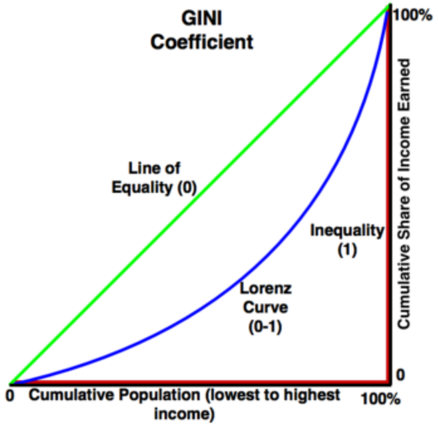
- How is Reddit content distributed over time?
  - Rate of content increases exponentially
  - About 10-30% of content is deleted
  - Number of authors increases exponentially
  - Authors create comments about 5X more than submissions
  - Number of subreddits increases over time, but only a small fraction is active

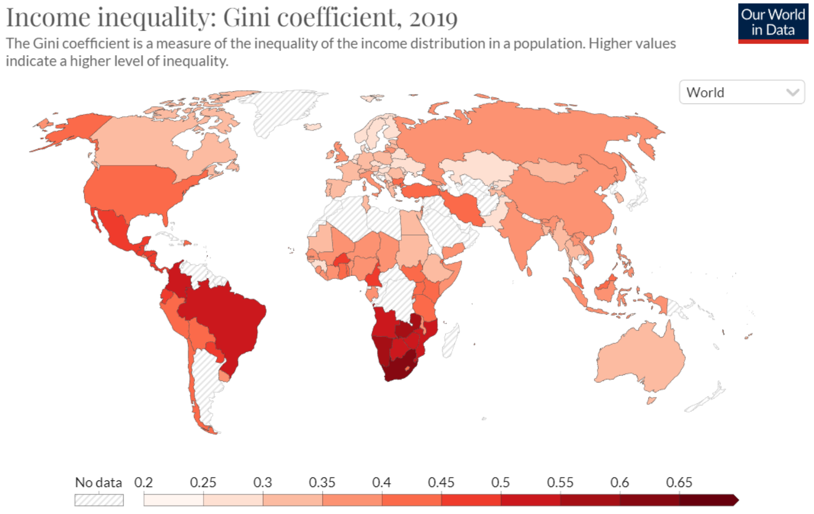


CONTENT CONCENTRATION

- Content on Reddit is concentrated
- Measure of concentration: Gini coefficient
  - $G = 0$: perfect equality
  - $G = 1$: complete inequality
  - Income inequality*: between 0.2-0.6

- How is it for social media?
  - Generally very high concentration
  - For authors: very high
  - For submissions: variable, but high
    - Score is more concentrated than comments

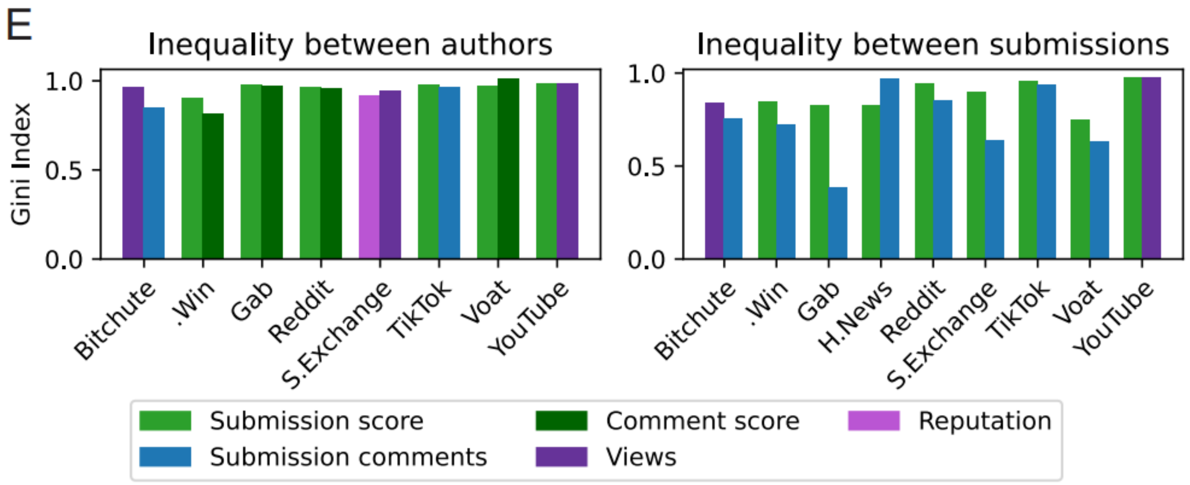

GINI Coefficient
Line of Equality (0)
Inequality (1)
Lorenz Curve (0-1)
Cumulative Share of Income Earned
Cumulative Population (lowest to highest income)
0
100%
0
100%


Income inequality: Gini coefficient, 2019

The Gini coefficient is a measure of the inequality of the income distribution in a population. Higher values indicate a higher level of inequality.

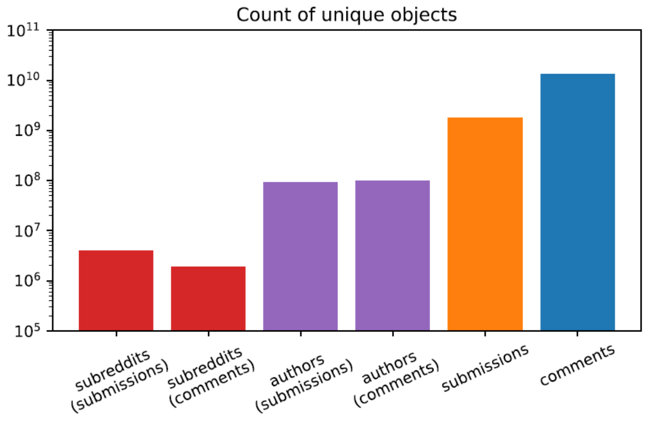
Source: World Bank Poverty and Inequality Platform OurWorldInData.org/income-inequality/ • CC BY

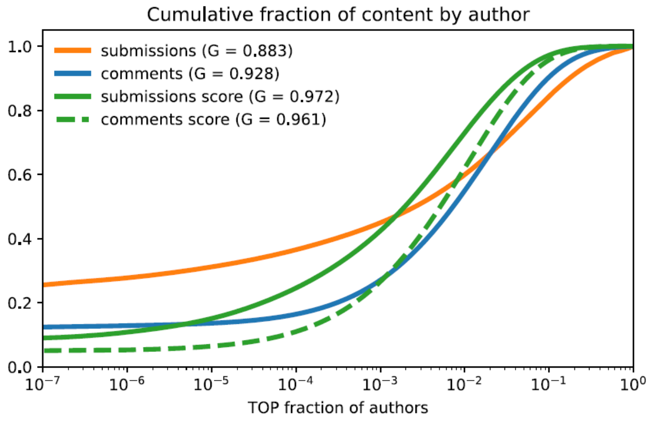
Note: Depending on the country and year, the data relates to either disposable income or consumption per capita.

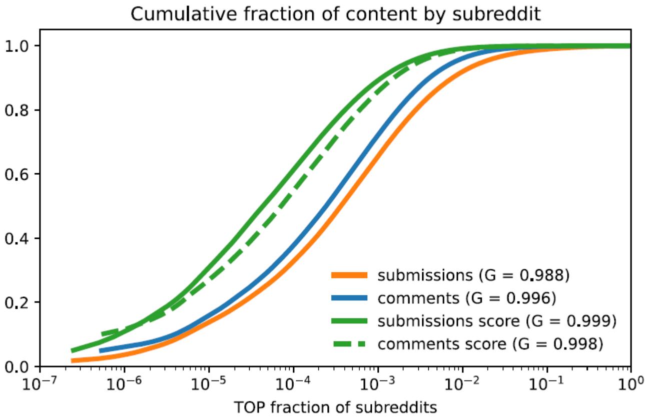


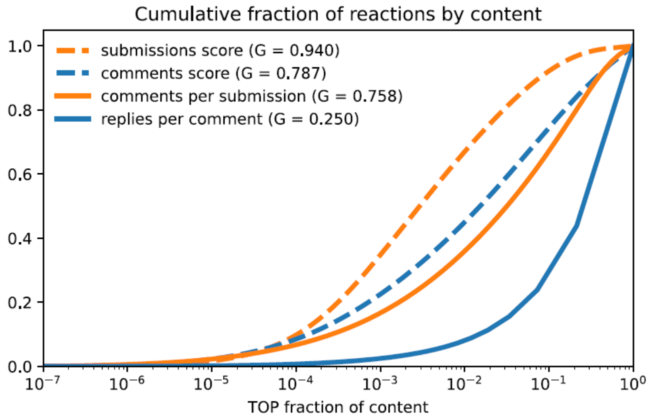
CONTENT CONCENTRATION ON REDDIT

- Many empty subreddits
- By subreddit
  - *Extremely* high concentration
  - Score is more concentrated than content
- By author
  - Still very high concentration
  - Submissions curve points towards botting
- By content
  - Variable concentration
  - Submission score is very concentrated $(G = 0.94)$
  - Comment replies is very distributed $(G = 0.25)$


Count of unique objects

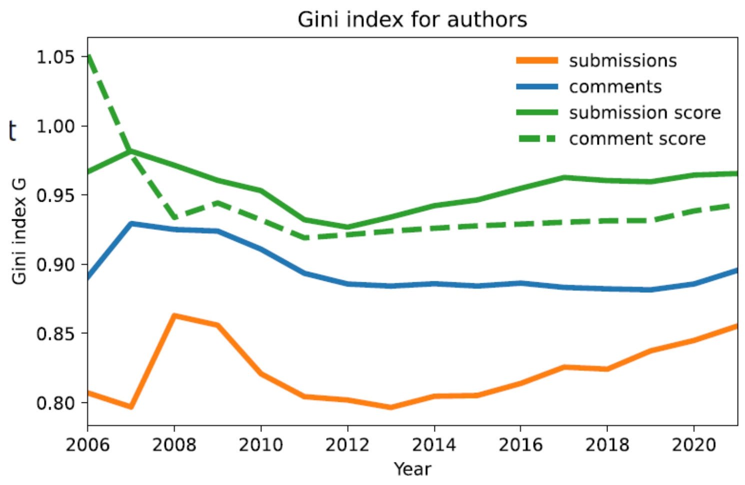

Cumulative fraction of content by author

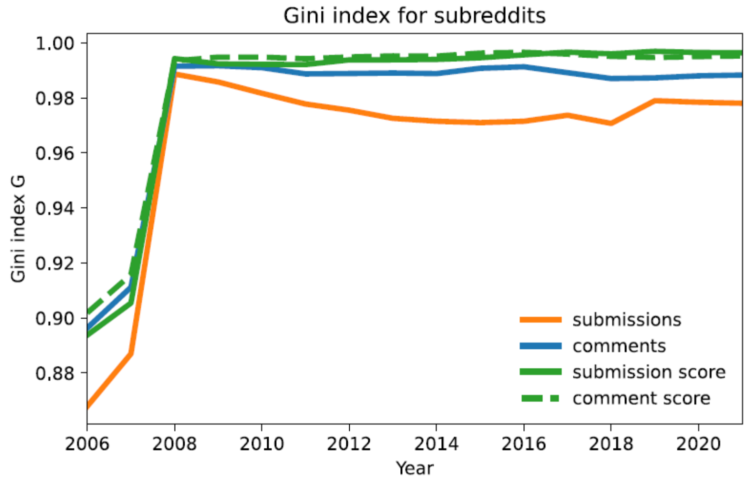

Cumulative fraction of content by subreddit

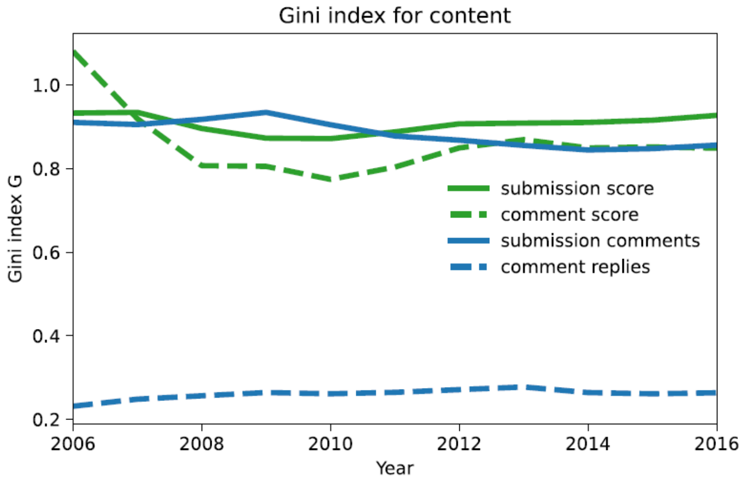

Cumulative fraction of reactions by content

CONTENT CONCENTRATION ON REDDIT

- Has it changed over time?
- Remarkably stable
- Suggests few changes in the core algorithm
- Note: $G > 1$ possible if variable can be negative

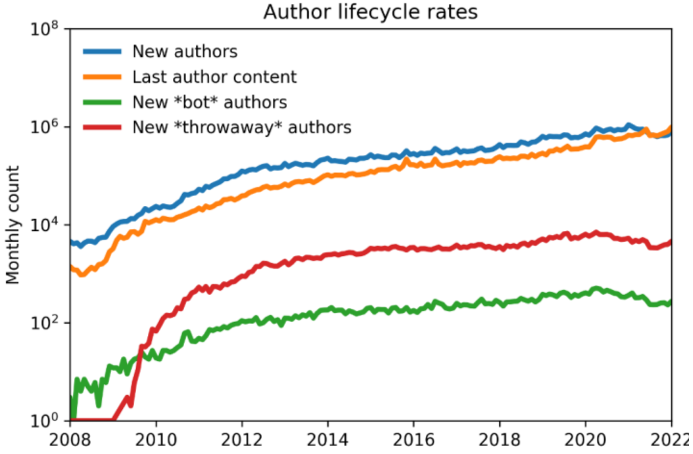


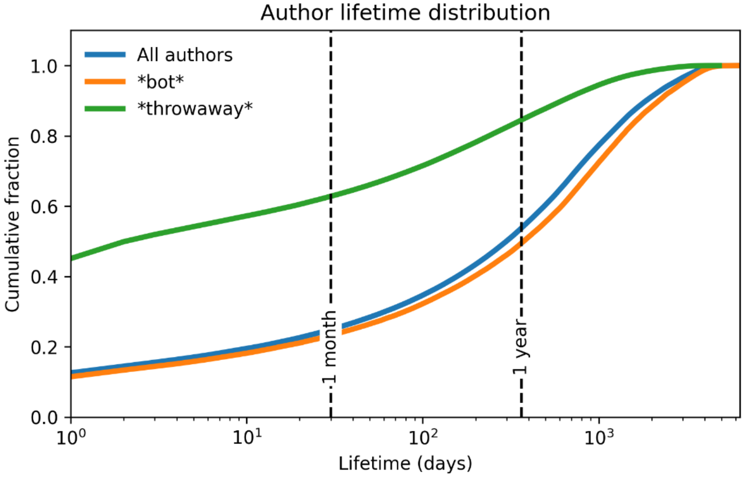


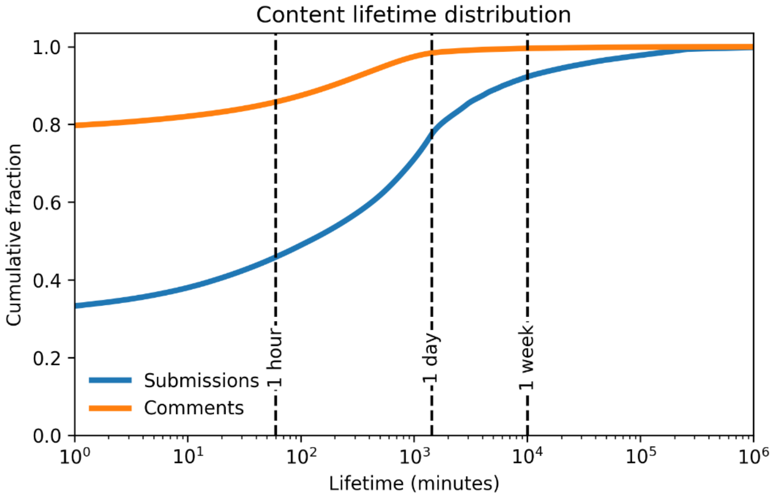


OBJECT LIFECYCLE

- Lifecycle: author/content/subreddits can leave/get inactive
- Author lifecycle rates
  - Increasing rates, but net new authors is relatively stable at ~20K per month
  - Bot and throwaway accounts are a negligible fraction
- Author cumulative distribution
  - Only includes authors inactive for > 1 year
  - 60% of authors where active for > 1 year
  - Throwaway accounts are different
- Content lifetime distribution
  - Comments are effectively dead after a day
  - Submissions have a long tail, but 90% are dead in a week

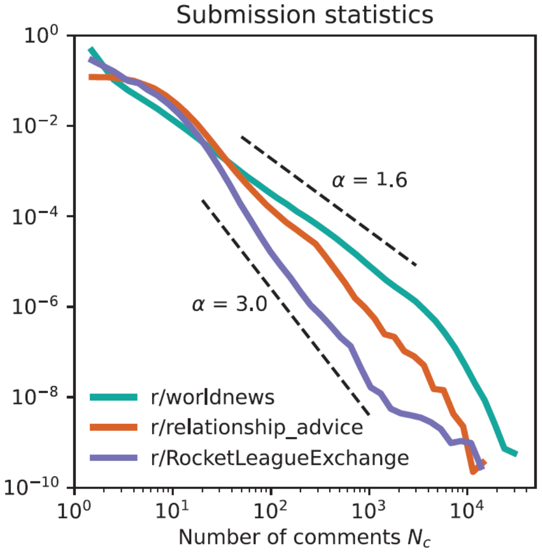

Author lifecycle rates

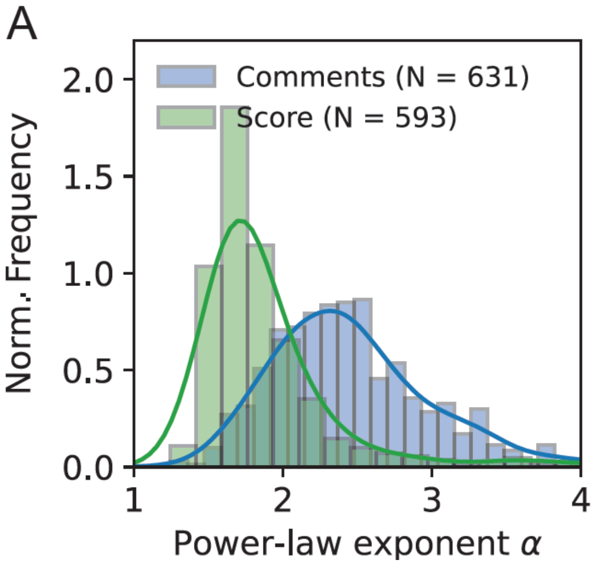

Author lifetime distribution

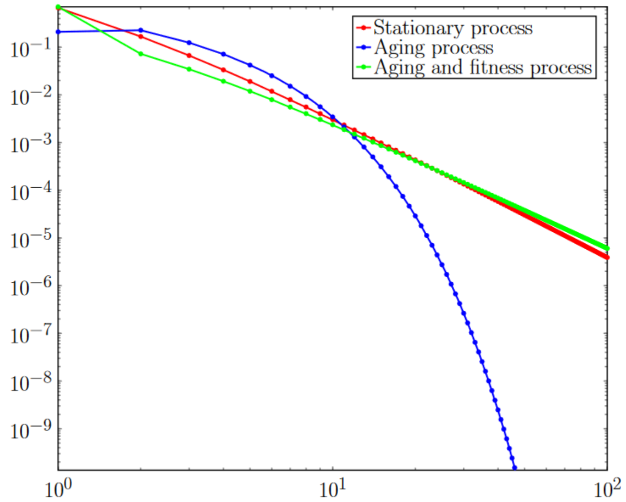

Content lifetime distribution

SUBREDDIT ANALYSIS

- Subreddits are different: how much?
  - Different shapes, different power-law exponents
- TOP1000 subreddits
  - ~60% are good power-laws
  - Exponent α distribution:
    - Comments: $\bar{\alpha} \approx 2.2$
    - Score: $\bar{\alpha} \approx 1.6$
- Importance:
  - Different exponents are associated with different models
  - Can hint towards mechanisms such as (submission) fitness and aging

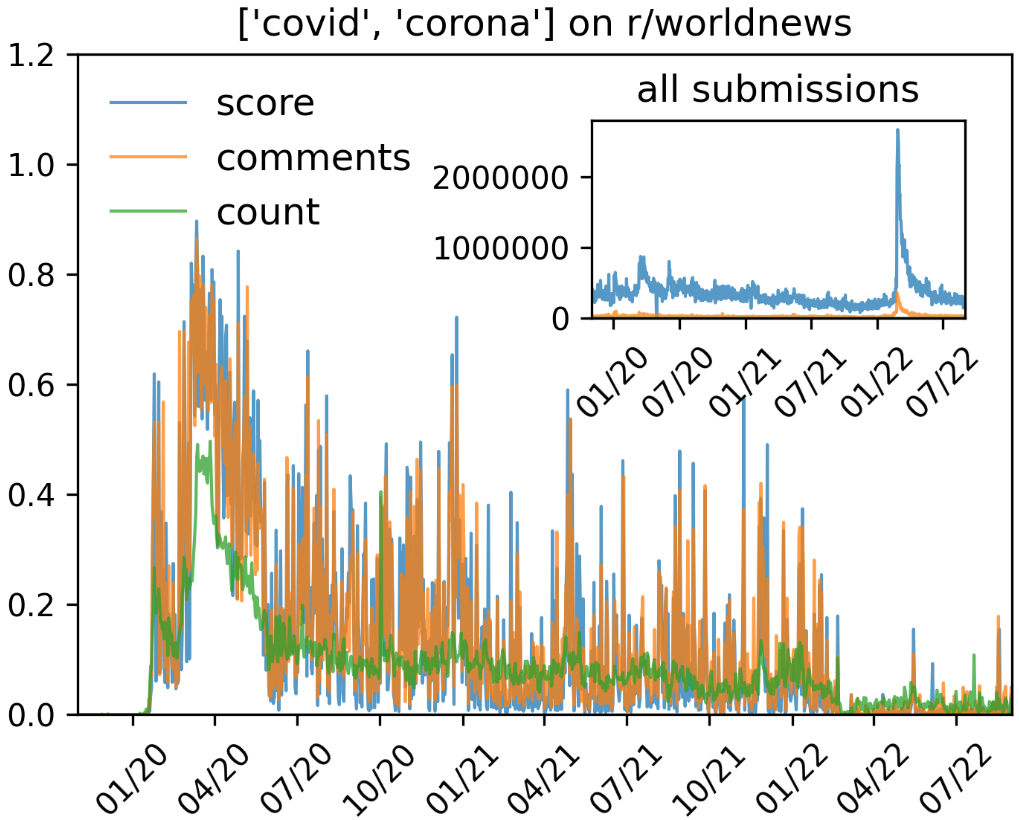






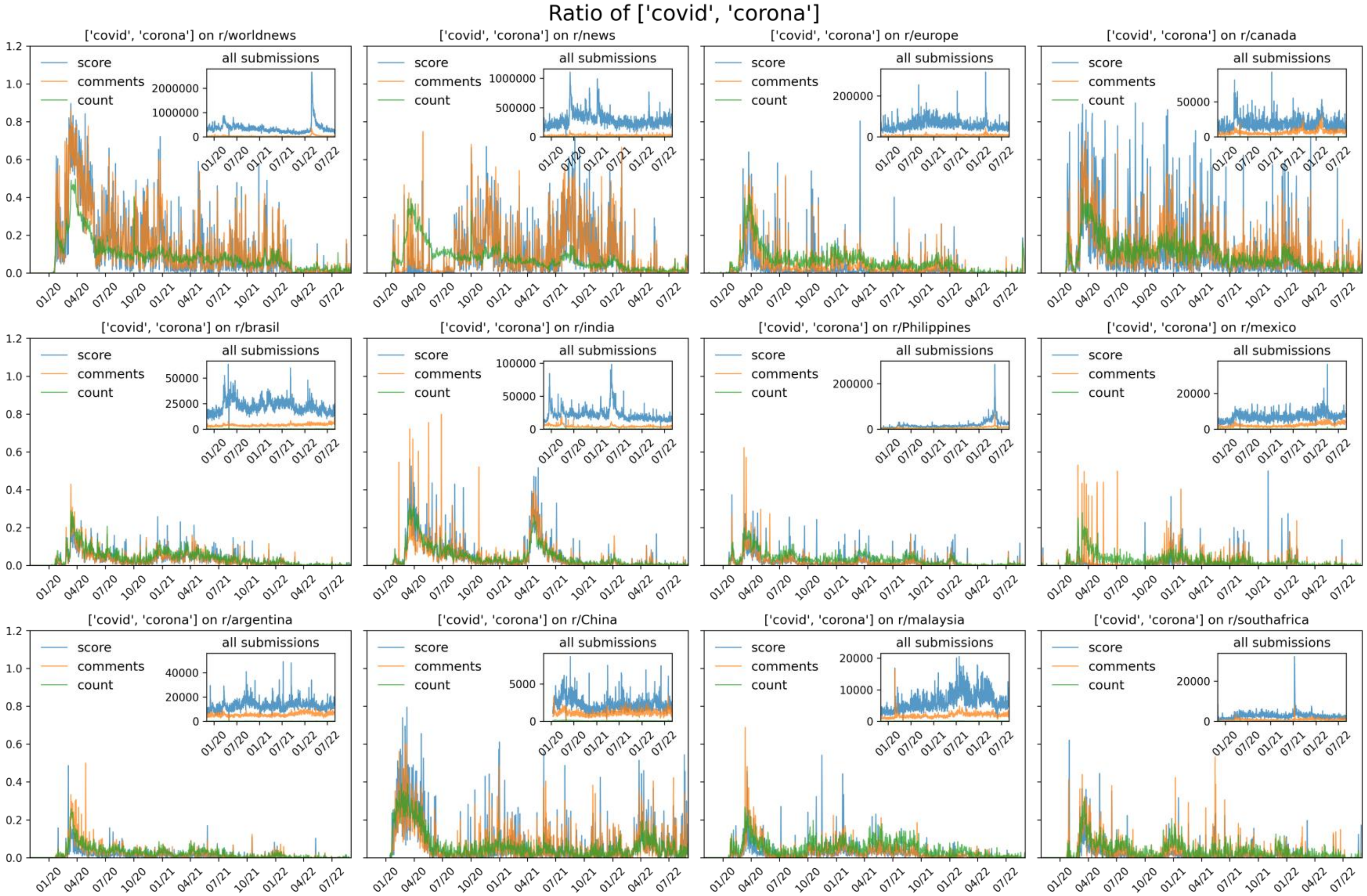
RESPONSE TO EXTERNAL EVENTS

- COVID-19 submissions
  - Very spiky: theme dominates the subreddit on some days



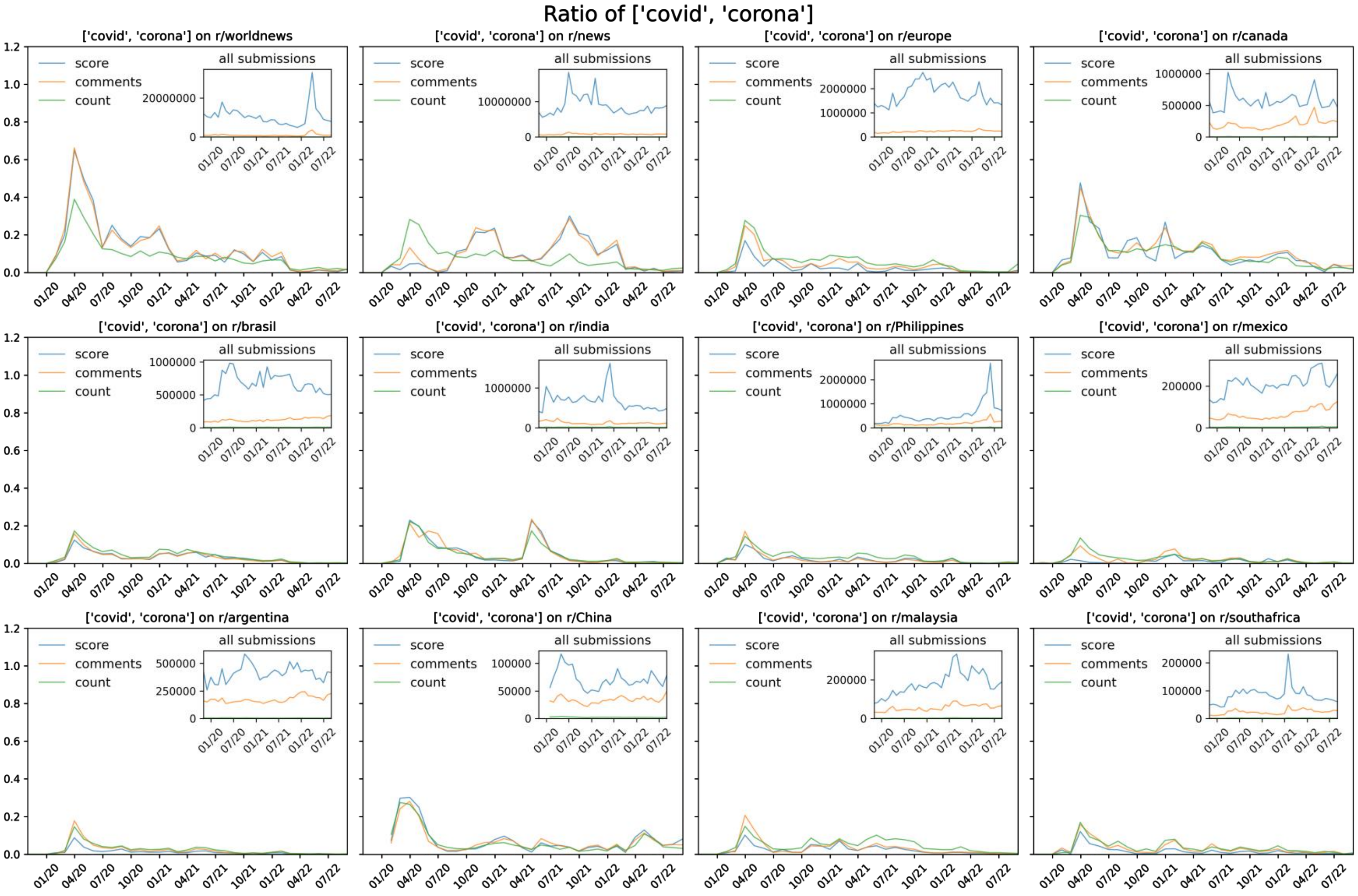
RESPONSE TO EXTERNAL EVENTS

- COVID-19 submissions
  - Very spiky: theme dominates the subreddit on some days
- International subreddits
  - Spikes do not align
  - More related to local level events
  - Relative importance varies by country
  - Proxy for media attention, can be used as changepoints for COVID-19 modeling



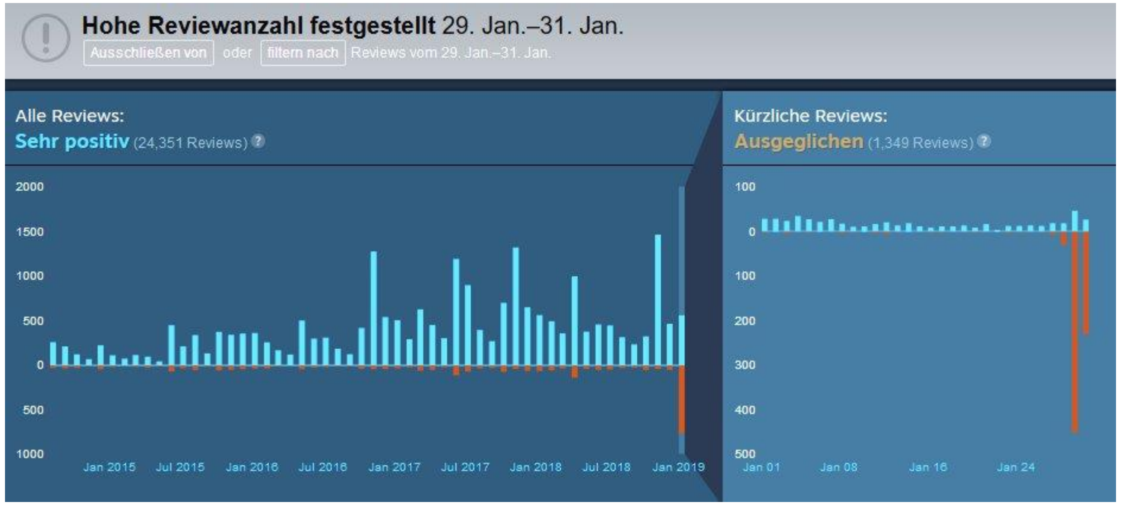
RESPONSE TO EXTERNAL EVENTS

- COVID-19 submissions
  - Very spiky: theme dominates the subreddit on some days
- International subreddits
  - Spikes do not align
  - More related to local level events
  - Relative importance varies by country
  - Proxy for media attention, can be used as changepoints for COVID-19 modeling



RESPONSE TO EXTERNAL EVENTS

- Steam review bombing
  - Coordinated response to external events
- Reddit counterpart?
  - Also very spiky
  - Doesn't only align with the bombing
  - Response varies by subreddit (PC vs console gaming)

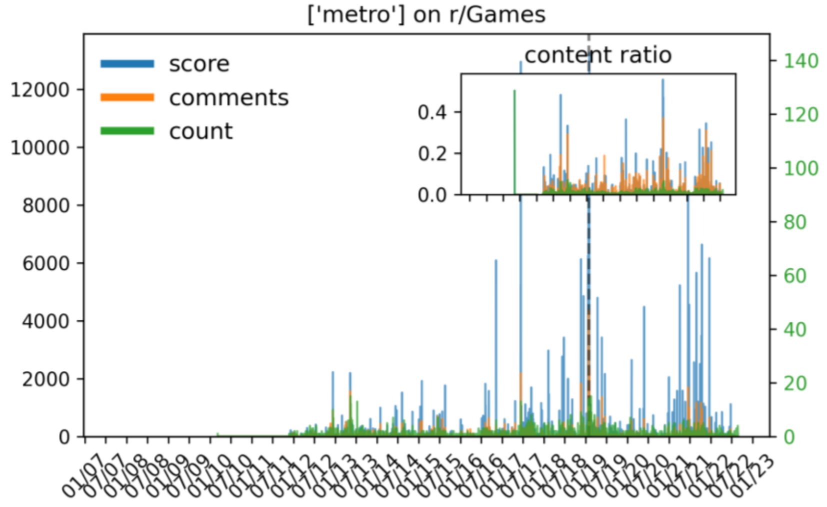


Hohe Reviewanzahl festgestellt 29. Jan.–31. Jan.

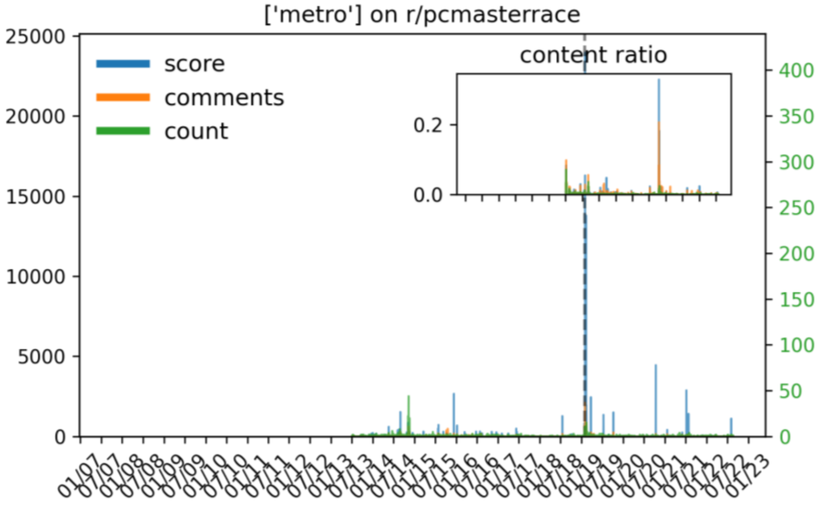
Ausschließen von oder filtern nach Reviews vom 29. Jan.–31. Jan.

Alle Reviews:
Sehr positiv (24,351 Reviews)

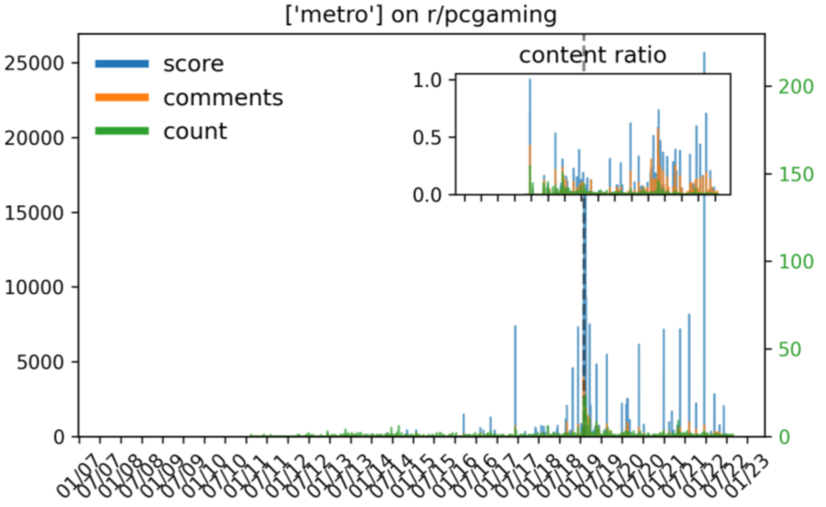
Kürzliche Reviews:
Ausgeglichen (1,349 Reviews)



['metro'] on r/Games



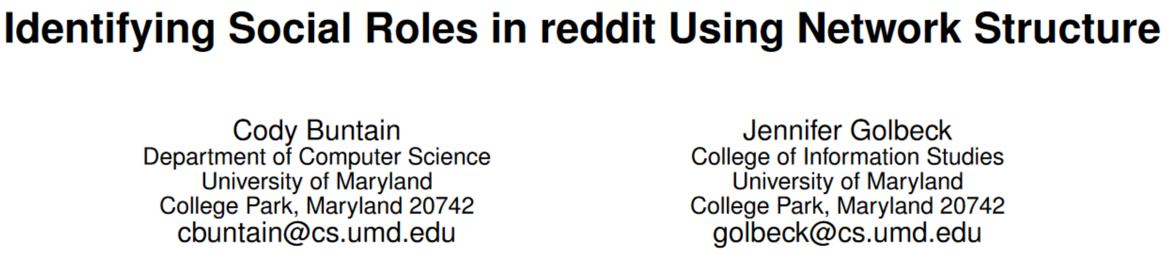
['metro'] on r/pcmasterrace



['metro'] on r/pcgaming

RELEVANT LITERATURE

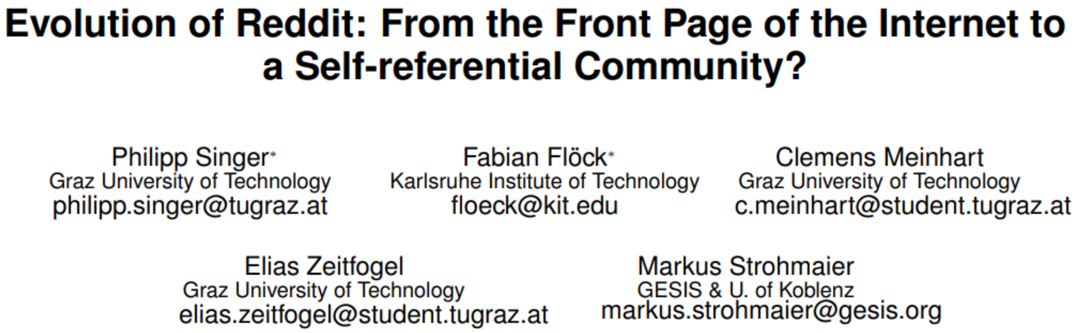
- Relevant literature
  - Buntain & Golbeck (2014)
    - Claims the existence of social roles such as "answer people" within Reddit
  - Singer et al (2014)
    - Claims that Reddit is increasingly self-referential
  - Weld et al (2022)
    - Claims that "bad content" is concentrated on an extremely small fraction of communities



Identifying Social Roles in reddit Using Network Structure

Cody Buntain
Department of Computer Science
University of Maryland
College Park, Maryland 20742
cbuntain@cs.umd.edu

Jennifer Golbeck
College of Information Studies
University of Maryland
College Park, Maryland 20742
golbeck@cs.umd.edu



Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?

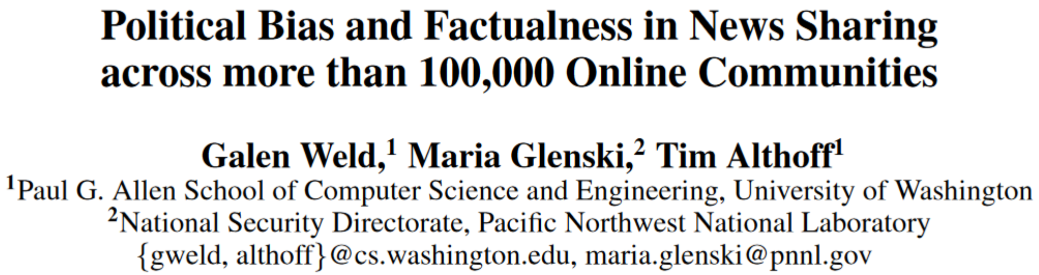
Philipp Singer*
Graz University of Technology
philipp.singer@tugraz.at

Fabian Flöck*
Karlsruhe Institute of Technology
floeck@kit.edu

Clemens Meinhart
Graz University of Technology
c.meinhart@student.tugraz.at

Elias Zeitfogel
Graz University of Technology
elias.zeitfogel@student.tugraz.at

Markus Strohmaier
GESIS & U. of Koblenz
markus.strohmaier@gesis.org

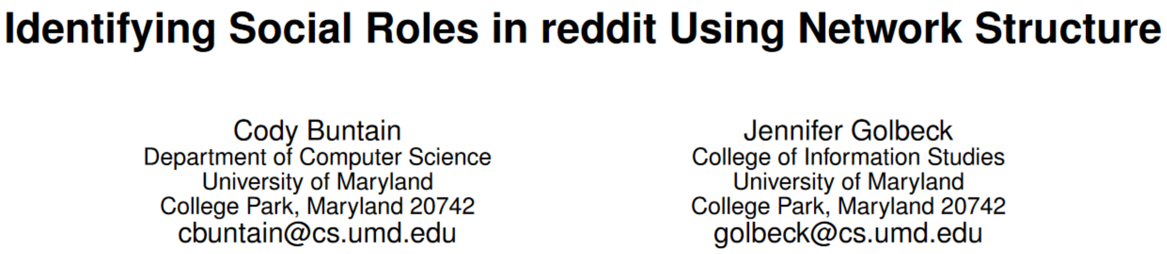


Political Bias and Factualness in News Sharing across more than 100,000 Online Communities

Galen Weld,[1] Maria Glenski,[2] Tim Althoff[1]
[1]Paul G. Allen School of Computer Science and Engineering, University of Washington
[2]National Security Directorate, Pacific Northwest National Laboratory
{gweld, althoff}@cs.washington.edu, maria.glenski@pnnl.gov

SOCIAL ROLES ON REDDIT

- Idea: Reddit has social roles
  - “answer people”: posts but doesn’t engage in discussions
  - “discussion people”: engages in discussions

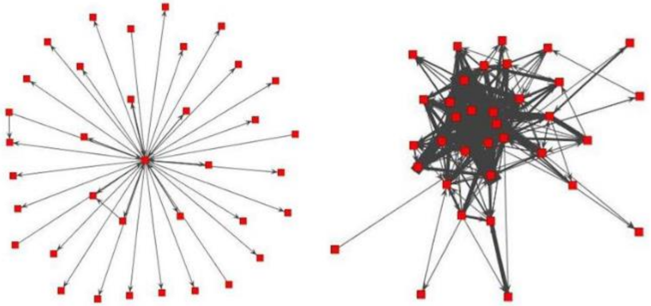


Identifying Social Roles in reddit Using Network Structure

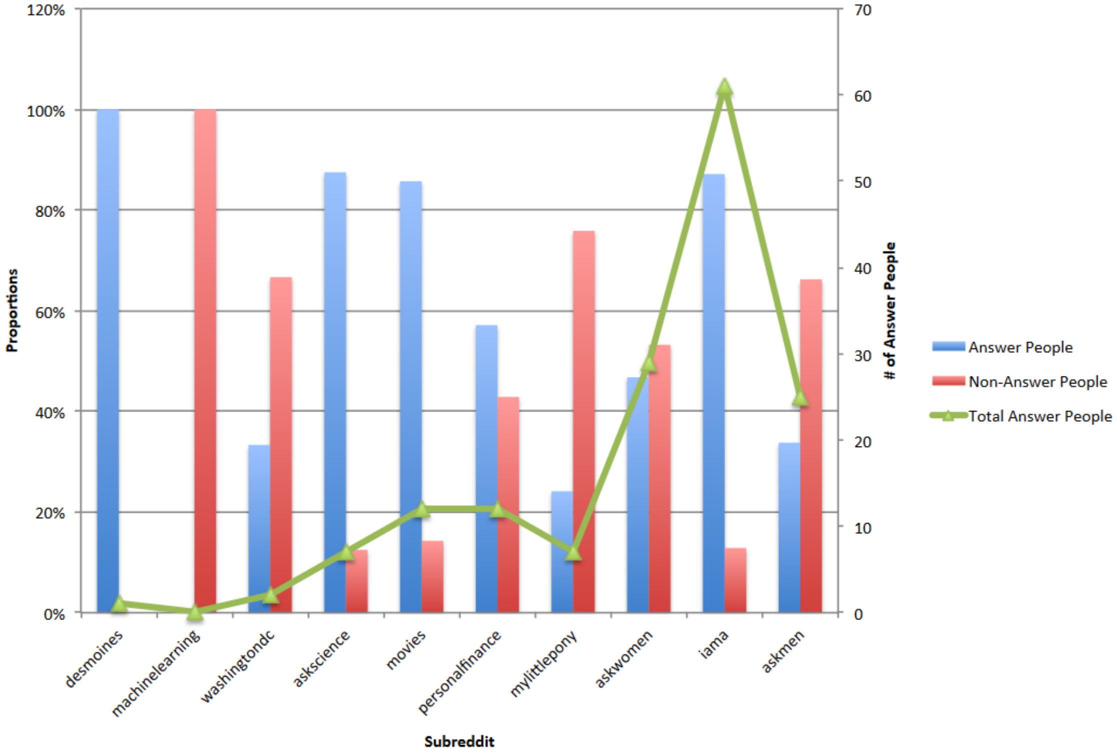
Cody Buntain
Department of Computer Science
University of Maryland
College Park, Maryland 20742
cbuntain@cs.umd.edu

Jennifer Golbeck
College of Information Studies
University of Maryland
College Park, Maryland 20742
golbeck@cs.umd.edu

- Results:
  - Fraction of answer people varies considerably between subreddits

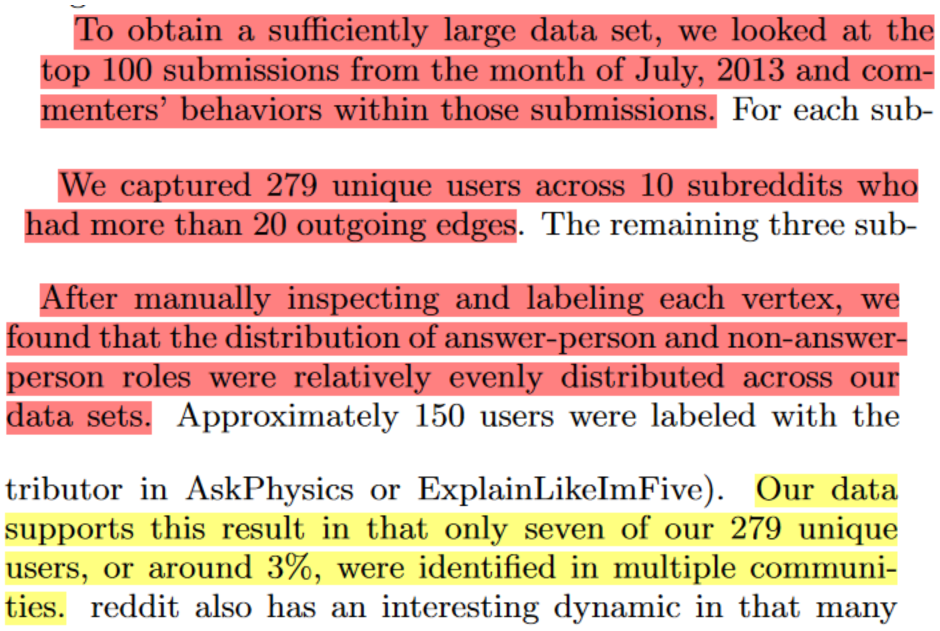


(a) Answer Person (b) Discussion Person



SOCIAL ROLES ON REDDIT

- Problems with the paper
  - Extremely sparse and biased data
  - Node classification from ML classifier with manual labelling
- Great opportunity to do it with good quality, large-scale data

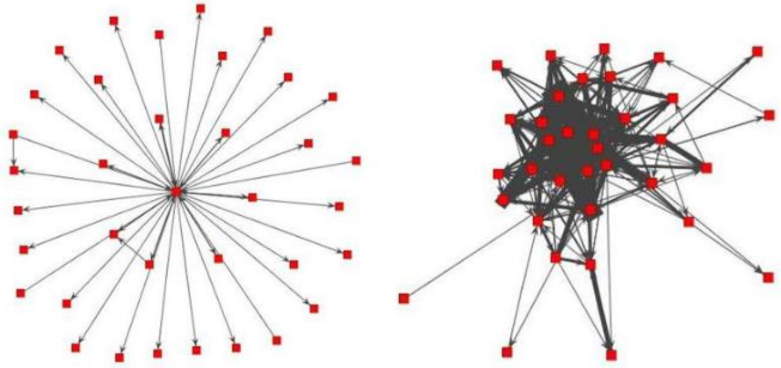


To obtain a sufficiently large data set, we looked at the top 100 submissions from the month of July, 2013 and commenters' behaviors within those submissions. For each sub-

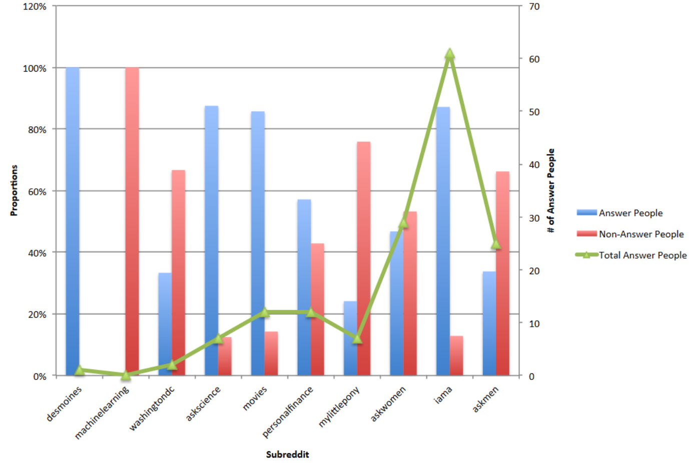
We captured 279 unique users across 10 subreddits who had more than 20 outgoing edges. The remaining three sub-

After manually inspecting and labeling each vertex, we found that the distribution of answer-person and non-answer-person roles were relatively evenly distributed across our data sets. Approximately 150 users were labeled with the

tributor in AskPhysics or ExplainLikeImFive). Our data supports this result in that only seven of our 279 unique users, or around 3%, were identified in multiple communities. reddit also has an interesting dynamic in that many

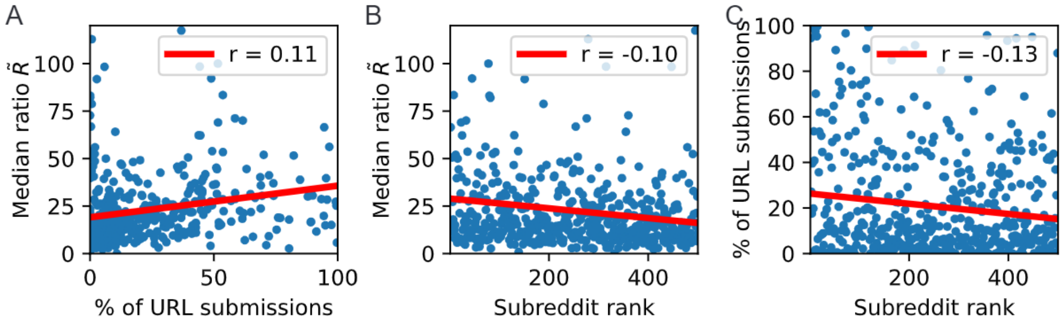


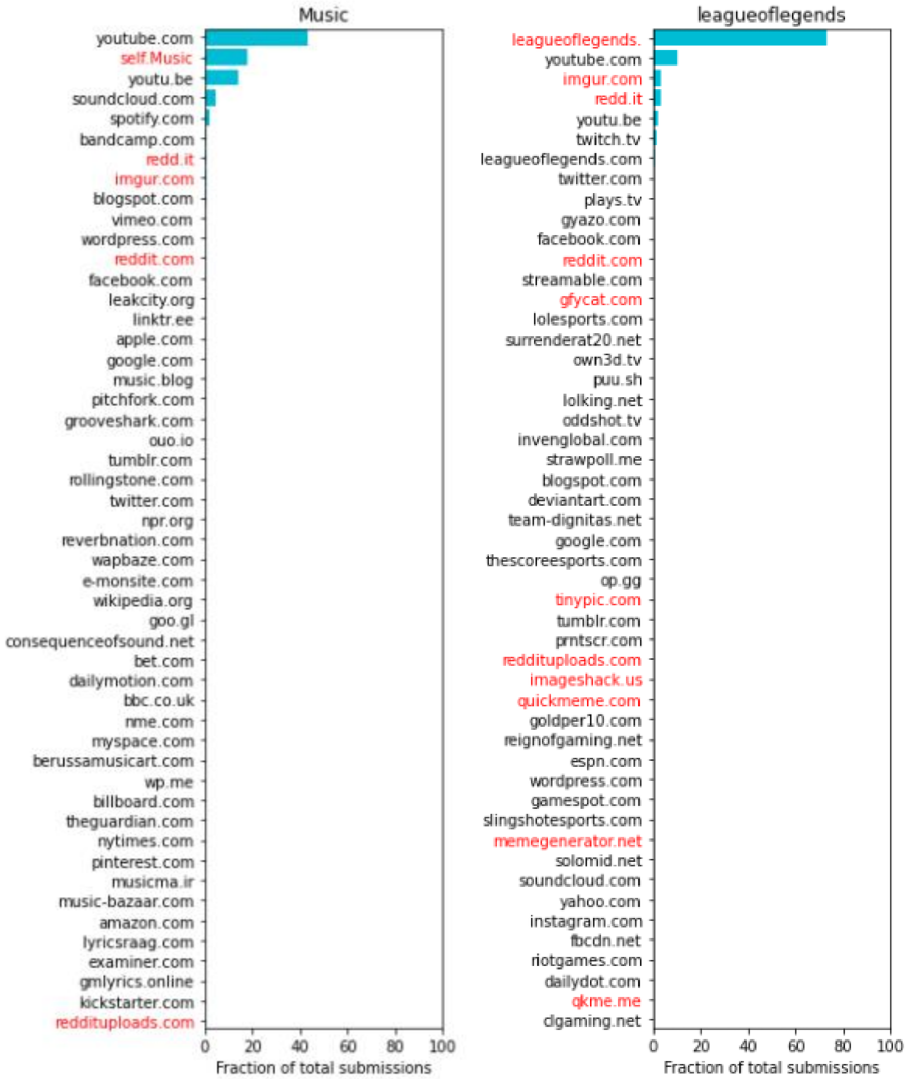
(a) Answer Person (b) Discussion Person



INTERNAL VS EXTERNAL DISCUSSIONS

- Literature claim:
  - Reddit is now mostly self-referential (Singer et al, 2014)
  - … as of 2014
- Re-analysis:
  - ~70% of submissions don’t link to external URLs
  - Highly variable between subreddits
  - Subreddit: % of URL submissions, submission-to-comment ratio R and subreddit rank are weakly correlated

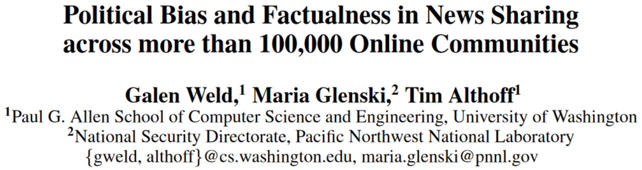




POLITICAL BIAS ON REDDIT

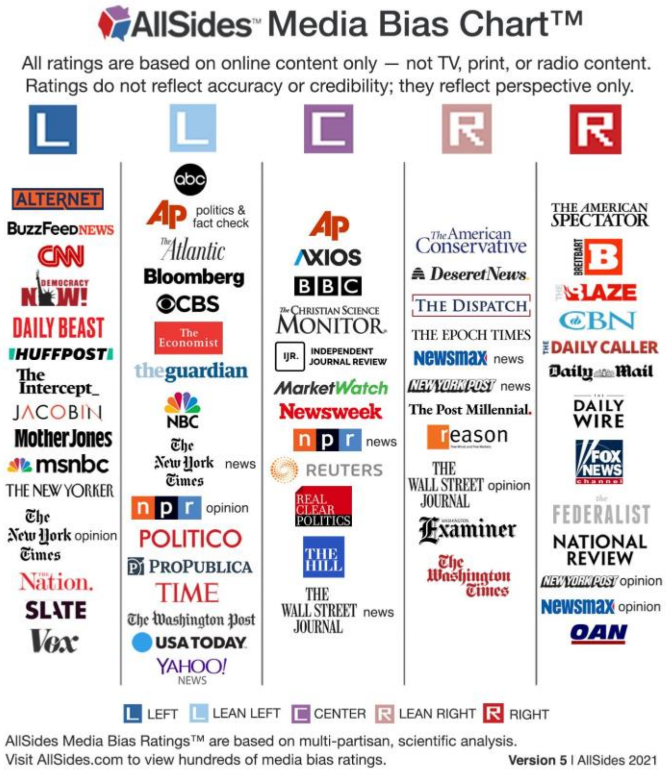
- The idea:
  - Check reddit submission links against Media Bias/Fact Check (MBFC)
  - Analyze differences between the TOP 100k subreddits

- The methods
  - MBFC database
  - Two scales:
    - Political bias (extreme left – extreme right)
    - Factualness (very low – very high)



Political Bias and Factualness in News Sharing across more than 100,000 Online Communities

Galen Weld,[1] Maria Glenski,[2] Tim Althoff[1]
[1]Paul G. Allen School of Computer Science and Engineering, University of Washington
[2]National Security Directorate, Pacific Northwest National Laboratory
{gweld, althoff}@cs.washington.edu, maria.glenski@pnnl.gov



AllSides Media Bias Chart™

All ratings are based on online content only — not TV, print, or radio content. Ratings do not reflect accuracy or credibility; they reflect perspective only.

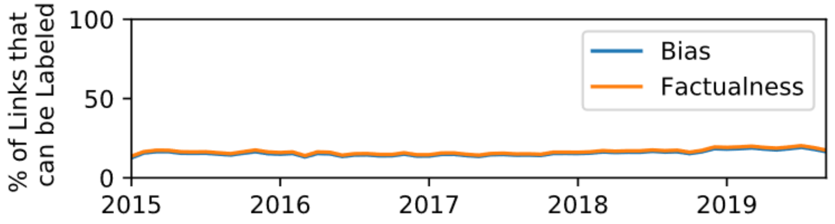
L LEFT L LEAN LEFT C CENTER R LEAN RIGHT R RIGHT

AllSides Media Bias Ratings™ are based on multi-partisan, scientific analysis.
Visit AllSides.com to view hundreds of media bias ratings.

Version 5 | AllSides 2021

POLITICAL BIAS ON REDDIT

- The caveat: only ~20% of links can be labelled
- Results
  - 74% of communities are center-left
  - Most content comes from high factual sources
  - Extreme content receives less exposure than neutral content
  - Extreme content is shared in few, contained groups
    - 99% on 0.5% of communities



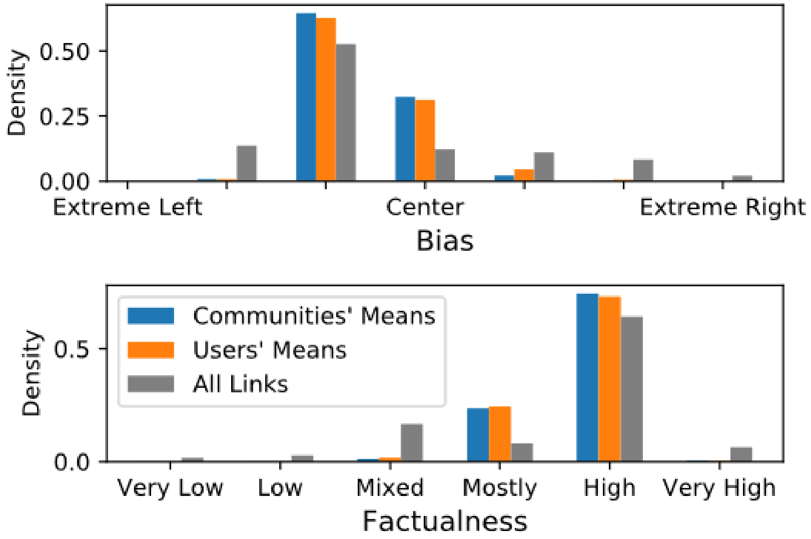


Figure 2: Distributions of mean bias and factualness are quite similar for both the user and community units of analysis. Grey bars show the normalized total counts of links of each type across all of reddit.