The Path to Power: Elite Reddit Users Curate Conflict to Drive Division

Anonymous Author(s)

ABSTRACT

In online forums, the top percentile of users disproportionately generate content and capture users' attention. They wield outsized influence, shaping conversations, worldviews, and agendas within their communities and beyond. However, these elite users are understudied in the literature. How can we identify these powerful users and characterize their community influence? What sets them apart from an average user? Does their ascension to power change their behavior within their communities and across the system? And what are systemic consequences of power user behavior for online communities? In this paper, we use data from Reddit to answer these questions: First, we demonstrate the use of the HITS algorithm on a graph of social interaction to identify users with significant community influence. Next, we show that influential users markedly differ from other users, initially having a tendency to search across communities before adapting to maintain a broad, but stable presence across communities as they specialize. We also find that power users work strategically to bring contrasting ideas to their chosen communities and that ascension to and fall from power are both associated with increased non-conformity relative to their community bases of power. We consider two hypotheses to explain this behavior: first, that power users act as "cosmopolitans" exploring Reddit and bringing back new, and exciting ideas; second, that the ideas power users import are not used to expand, but rather to entrench their communities in established ways of thinking by curating conflict that foments persistent division and inoculating them against competing ideas. Based on the tendency of power users to court controversy as they rise in and maintain influence over their communities, our findings provide stronger support for the latter, raising concerns about increasingly robust polarization of online communication.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Networks** → **Social media networks**; • **Information systems** → **Data mining**.

KEYWORDS

computational social science, social media, data mining

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Anonymous Author(s). 2022. The Path to Power: Elite Reddit Users Curate Conflict to Drive Division. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The vast majority of recent research on user behavior in online communities has sought to characterize the behavior of the typical online user [4–6, 12, 13]. While this yields valuable insight regarding how users engage with communities, the kind of language they adopt, and the communities in which they participate, this dominant research agenda ignores the '90-9-1' dynamic of the internet [2, 11]: the majority of content is produced by 1% of users, with another 9% generating the minority as the remainder lurk, reading and amplifying, but rarely generating original content of their own [7, 8].

What explains this dissonance between how individuals generate content in online communities and how researchers study behavior there? Online Social Networks (OSNs) are frequently characterized as faceless crowds in the research imaginary. This is with justification: users experience enormous turnover in the generation of content and individual participation, all attributable to the scale of these platforms. Individual posts or interactions are not persistent, with their shelf-life measured in hours. This holds both for platforms where users are associated with their names (Twitter, Facebook) and those on (semi)anonymous platforms (Reddit, Discord) where users can employ pseudonyms. In both, name-recognition is low or limited, suggesting that reputation (as suggested by]) is insufficient for distinguishing power users in the online context. And popularity, [suggested by]] fares no better:

And while] establish the importance of studying the role of elite individuals in community formation and agenda-setting

rpopularity alone are

neither activity,

With such great turnover in content, reliable inferences on the of participation can prove to be tricky. The community itself provides a more reliable/sturdy/persistent unit of analysis. But when we describe these communities and the nature of interactions within them, are we attributing these results to a mythical median-user, one who may not even exist? While these results might give us an idea of the centroid of the community in some hypothetical space, we cannot reliably or reasonably attribute this to an individual or collective subgroup. As such, much of the current literature, especially on community health and communication on social platforms, ignores the outsized influence of a small collection of 'elite' users on the entire discursive system online. As online interactions grow in scale and importance to inform social, political and economic activities in the material offline world [1, 3], we argue that the path to a better understanding of information ecosystems lies through an intimate analysis of these influencers, a yet-to-be-addressed need.

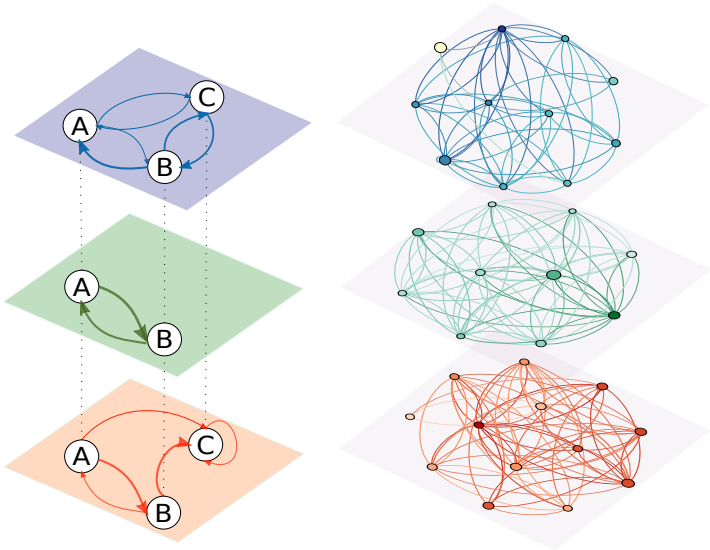


Figure 1: Representation of multi-layer interaction on Reddit. Each layer represents a subreddit.

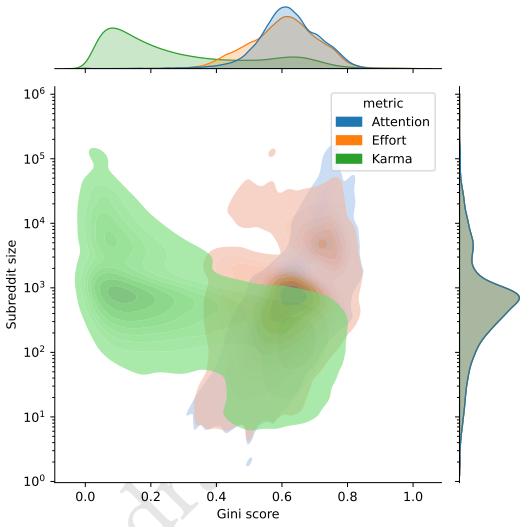


Figure 2: Distribution of Gini scores for Effort, Attention, Karma by size across all subreddits

Critical issues of content moderation and misinformation caused by scale are best addressed by looking at these influential users closely.

In this study, we propose a model of social interaction on multi-community OSNs and provide a measure of user influence within their respective communities. The questions we pose in this study are motivated by the rift between the faceless-crowds paradigm and the theorized existence & importance of community elites.

We study how these elite influencers different from the average user; if they are not, maybe these communities are better described as faceless crowds. If they are different, what makes them so. We find that they are indeed different, showcasing greater exploration and participation.

We also study the role they play within communities: content generation, mediation, ideological framing, etc. And we study how they ascend to power: if they are merely following the crowd and reflecting their choices back to them, these elites are not influential in any meaningful way, being more akin to Rogerian therapists with minimal sway. We find that their ascent to power is closely associated with them providing new, surprising language to the community. Interestingly, this is also associated with greater controversy, i.e. their content is not widely greeted with positive feedback/votes alone.

We also try to understand the consequences of the influence these elites have on the relationship between users in online crowds and communities.

1.1 Related Work

Historical studies of offline communities assumed their natural state to be one of disconnected isolation. Pathways to influence off-line involved either intensive *local* involvement in community management or extensive *global* engagement by cosmopolitans who brought valuable insight from and relevance to the rest of the world [9]. Much additional work promoted the local-cosmopolitan distinction across offline setting, from work teams and organizations to nations (CITE Hannerz 1990, Theory, Culture & Society).

The study of interactions between individuals embedded within a community or common space has a longstanding tradition in classical work on social networks of the offline kind . These studies highlight the importance of power in these slow-moving networks, and the role of elites and entrepreneurs in a host of critical functions: mediating conflict, delineating the boundaries of the base and community, framing issues and grievances, to name some. Katz & Lazarsfeld [?] identified the importance of the personal influence of individuals in interpersonal networks in the processing and acceptance of mass media, and provide a two-step flow model of communication, where opinion leaders act as intermediaries to play a critical role by parsing and amplifying news to their networks.

1.2 Contributions

By examining power user behavior and its deviation from typical behavior online, we discover the distinctive roles that power and typical users play in the creation of intransigent and polarized online communities. Online echo-chambers, as defined by C Thi Nguyen [?], actively work to reinforce distrust and inoculate its members from competing arguments by pre-emptively discrediting these sources. They recirculate assumptions and accepted understandings by those within them, which are constructed positively by typical users who speak in conformity to community norms, but negatively by power users who source contrasting material from across the online system against which community norms can be rehearsed and performed. This paper helps identify emergent leaders of such communities, while providing a host of metrics to study their structure of interaction and power.

2 MEASURING INTERACTION ON REDDIT

In this section, we describe our approach to measuring interactions and attention on Reddit. We examine user interaction on Reddit in order to characterize power users and their distinct influence on the construction of online community.

Reddit is an online social media platform with tens of thousands of sub-communities, referred to as "subreddits." Subreddits can be created by any user, and although some focus on broad sets of interests (such as "r/pics" and "r/politics") many are narrowly focused. Subreddits can bring together users based on geography ("r/nyc"), entertainment interests ("r/OnePiece"), sports, games, interest in a particular kind of joke, and many other topics. Users on Reddit can make "submissions" to a community, sharing pictures, videos, links, or text. Users can then comment on these submissions or original posts, and reply to comments made by others users. The majority of content on Reddit is contained within these comment threads.

Reddit is well-suited to the study of interaction across communities and over time because it allows us to view how a given user's behavior differs between communities where they wield influence, and those where they do not. It also enables a study of the mobility dynamics of users as they move between communities, ostensibly searching for those most aligned with their interests and ideologies. Indeed, Reddit has been used as a dataset for the study of social media user mobility, language, and ideology. [10, 14].

2.1 Data

We use Reddit data collected from `Pushshift.io` [**?**], for a 3+ year period ranging from January 2016 to March 2019. The dataset consists primarily of information on the comments posted by users, primarily: their username, the time of their comment, the community in which they made the comment, the text of the comment, and the "parent" comment or post to which they are replying.

2.2 Social Interaction Graph

Interaction on Reddit follows a tree-like structure of replies as is common across the internet. All replies to a particular comment can be thought of as leafs sprouting from that originating comment. At the level of a submission, this representation shows us who are the participants within a conversation and which users responded to each other, along with a host of additional metadata like the aggregate upvotes for each comment.

For ease of analysis, we collapse these individual subreddit-level conversation trees by month to center the interactions between users within that community. For each subreddit & month, we construct a graph where users are nodes, with an edge being drawn between any two users if they have directly replied to each other. Specifically, we draw a directed edge from user B to user A if B replies to a comment made by A. The resulting directed graph is the *subreddit interaction graph*. Since we consider the top 1000 subreddits by comment volume, the global representation of Reddit for a particular month would be a multi-layer graph, where each layer is a subreddit and the user-nodes are shared between layers (as shown in Figure 1).

2.3 Effort, Attention, and Karma

We have two variants of edges in our representation of monthly interactions by subreddit: the comment count, which is the total number of comments from B to A, and the aggregate comment score, which is the sum of up- and down-votes for all comments from B to A. To contextualize what these edge-types mean, we look at them from the standpoint of an individual user embedded in this graph: outgoing edges encode their contribution to the community via a comment, while incoming edges encode the replies they have received.

It is helpful to draw a distinction between these edge-types, as they each tell a part of the user story. A user who has many more outgoing edges than incoming ones is somebody who generates a lot of content, but does not get any engagement, which could be an indicator for content not of interest to the community. Conversely, somebody who has many more incoming edges than outgoing ones is a user whose content is timely and engaging, drawing in participation from others. But we also want to distinguish between high-quality content the community approves of and low-effort comments by high-volume trolls. The former would show up in our data as edges with a high aggregate karma score, and the latter as ones with low/negative score. As such, the raw count of outgoing edges from a user is representative of the effort they have expended in the subreddit, while the raw count of incoming edges is the attention/engagement they have received. High karma edges represent comments/interactions that are viewed positively by the community (positive feedback), while low karma edges indicate negative feedback.

For a user U_i, we represent their effort E and attention A activity over time across Reddit as two matrices:

$$E_{ijt} = [e_{i,1}, e_{i,2}, ...e_{i,j}, ...e_{i,1000}]$$
$$A_{ijt} = [a_{i,1}, a_{i,2}, ...a_{i,j}, ...a_{i,1000}]$$

where $e_{i,j,t}$ is the proportion of U_i's comments in month t that were made in subreddit S_j, and $a_{i,j,t}$ the proportion of replies received.

As predicted by the 90-9-1 principle, Figure 2 showcases the extent of inequality on Reddit across these metrics. Larger subreddits are more unequal in both the number of comments generated and replies received. Surprisingly, the trend does not hold for karma: larger subreddits are more egalitarian in their distribution of upvotes.

2.4 Identifying Influential Users

The representation described above makes certain claims about interactions between users: by responding to a user's comment, you are making your contribution to the community and directing attention to the parent comment and user. Every reply your comment receives is in turn an inflow of attention to you and your comment. Under this framing, users who receive a lot of responses capture attention, be it the good and/or bad kind. They drive the discourse by creating content that generates other content.

The metaphor of giving and receiving attention parallels the hyperlinked landscape of the internet. Under the Hubs-and-Authorities model of the internet, a page is deemed authoritative if it is linked to (has an incoming edge) by a large number of other pages. Similarly,

a page serves as a hub for directing attention if it points to (has an outgoing edge) a large number of authoritative pages.

Unlike the hyperlinked internet, edges in the subreddit interaction graph have positive and negative scores. We argue that this score is not as relevant to identifying users who create and drive 'discourse', which is a value-neutral construct. Even if your comment is heavily down-voted but ends up getting a lot of attention/engagement, you are driving discourse, even if it is discourse created in opposition to what you were saying. Celebrities are identified by the amount of publicity they receive, good or bad. Ergo, celebrities in these subreddits are determined by comment count alone, and not score.

Thus, we assign two scores to each user in a subreddit for a given month: an authority score and a hub score, which indicate their relative stature in generating attention vs directing it. We use the iterative HITS centrality measure over comment count edges to calculate these measures.

Power Users: For this paper, a power user is one whose authority centrality score within a subreddit is in the top 1% in any month in our period of analysis. Since we know the distribution of effort, attention, and feedback are highly imbalanced on Reddit, these are the users who generate a disproportionate amount of activity within that subreddit in a particular month.

3 ANALYZING THE MOBILITY OF POWER USERS

In their pursuit of power, users are split for choice in how they choose to spend their time on the platform: do they maintain their presence and status in the subreddits they are a part of, do they choose to strengthen or weaken it, or do they choose to find new communities to explore. In particular, we want to study how they shift their activity across subreddits, how they move across Reddit itself. To measure this shift in effort and attention, we propose two measures for both: focus and stability.

3.1 Measuring User Focus

In our multi-community context, users can pick from over 100,000 niche communities to to participate in. As such, we want to quantify the degree of specialization a user might engage in: do they only participate in a handful of subreddits, or do they participate more widely. For example, one user could be active in only select political subreddits. They would be markedly different from another type of user who showcases similar activity in those same political subreddits, but also actively engages in others like gaming and sports. Our measure of Focus here is an indicator of such specialization, it measures the degree of dispersion of activity and participation.

Given a user U_i whose effort and attention sequences for month t can be represented as E_t, A_t, Focus is the negative shannon entropy of these sequences. Greater focus indicates lower entropy, i.e their participation is limited to a low number of subreddits. Focus is calculated for both Effort & Attention, each indicating the degree of concentration for comments made and replies received. We have

$$\text{focus}(U_i, t) = \sum_{j=1}^{N} (p_{i,j,t} \times \log p_{i,j,t})$$

where $N = 1000$ (number of subreddits), and $p_{i,j,t}$ is the effort or attention value for user U_i in subreddit j for month t. Focus ($f_{i,t}$) ranges from -6.91 (equal participation in all 1000 subreddits) to 0 (participation in one subreddit alone).

3.2 Measuring User Stability

Participation on Reddit is both multi-context and longitudinal. We would expect that after an initial period of exploration, once users find their chosen corners of Reddit, they 'settle' within them, with their monthly activity limited to these communities and not changing by much. On the flipside, a user could also regularly change up their base of participation, choosing to engage in completely different communities than the ones they did in the previous month. Each represents a different class of users. We term this relative change as Stability, a measure of how much a user's activity footprint has changed relative to their past.

Given a user U_i whose effort and attention sequences for month t can be represented as E_t, A_t, Stability is the negative Kullback-Leibler divergence between the sequences for month t and $t-1$. Greater stability indicates low or limited change in activity. It is calculated for both Effort & Attention, each indicating the degree of change for comments made and replies received. We have

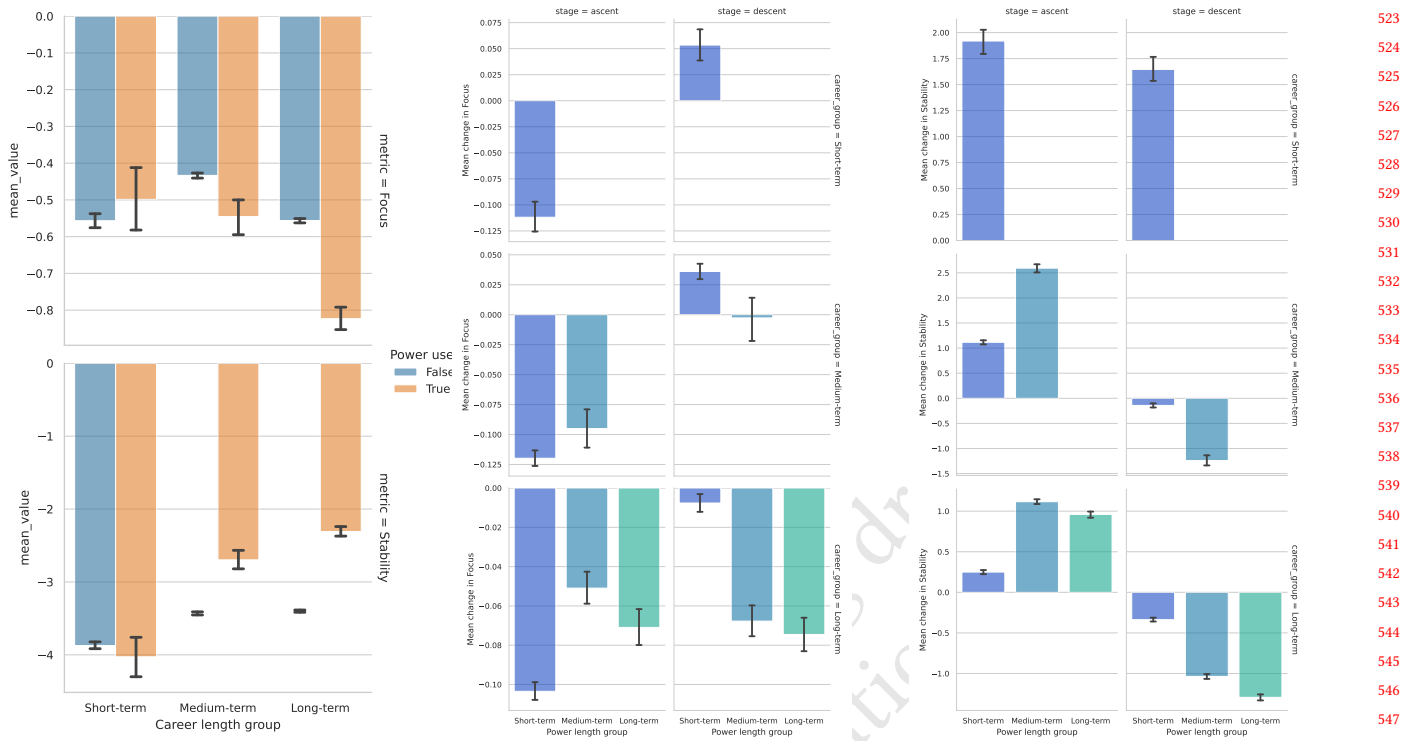
$$\text{stability}(U_i, t) = -\sum_{j=1}^{N} (p_{i,j,t} \times \log \frac{p_{i,j,t}}{p_{i,j,t-1}})$$

where $N = 1000$ (number of subreddits), and $p_{i,j,t}$ is the effort or attention value for user U_i in subreddit j for month t. Stability ($s_{i,t}$) takes a maximum value of 0 (no change from month $t-1$ to t).

3.3 Analytical Setup

To study how power users differ from their counterparts, we construct a matched sample of non power-users who have a similar profile as them across our dataset. That is, we compare them to users who show a similar volume and distribution of activity over 4 years. We generate propensity scores using a Logistic classifier where a user's power status is the label and a vector concatenation of their career duration, and the aggregate effort and attention sequences over the data. The model is trained on 5% of the data. We compute their average focus and stability across 3+ years on Reddit, while categorizing users as short-, medium-, and long-term users based on the duration of activity in our data.

Additionally, we study how change in power status affects a user's behaviour. We ignore changes in the location of power: if a user holds power in subreddit S_1 in month t_1 and holds power again in subreddit S_2 in month t_2, we consider them to be a power user from t_1 to t_2. We treat the period of power as continuous, regardless of the continuity of their power status, until they permanently fall out of power across Reddit. We justify this by arguing that power has an indelible and sustained effect on a user. If it comes from a knowledge and appreciation of what their subreddit needs, that understanding does not change unless they permanently fall out of it. We consider the period before and after as pre- and post-power stages. We examine the difference in average focus and stability when power users transition from one phase to another. In addition to the career length feature described above, we add a similar power



(a) Mean Focus & Stability for Power users and similar non-power users across career lengths. Long-term power users are less focussed and more stable.

(b) Mean change in Focus of power users as they ascend and descend from power. Long-term power users are less focussed across both phases.

(c) Mean change in Stability of power users as they ascend and descend from power. Long-term power users are more stable when they rise to power, and less when they fall.

Figure 3

length category to distinguish users by the length of their in-power period.

3.4 Results

3.4.1 Power Users are less focused. We find that power users, on average, show less focus than their non power-user counterparts. They tend to participate more equally across communities and do not show a tendency to narrow it down to a select few: they comment in and receive replies more uniformly from a wide collection of subreddits, proving that they are truly more cosmopolitan than their peers. Figure 3a shows that when we control for career length, the effect increases for long-term users, regardless of power status. Thus we can say that longevity on Reddit, irrespective of power status, is associated with a wider and more uniform base of participation. This is to be expected, since we assume these users identify and create new communities of interest as their interests broaden or change.

3.4.2 Power Users are more stable. Power users, on average, show greater stability than their non-power user counterparts (Figure 3a). While both types of users show a persistent decrease in stability, the decrease is lower for medium- and long-term power-users. They tend to change their distribution of participation much less than similar users, who tend to be more volatile and don't easily set roots. This highlights the tension between the desire to explore across communities (as described in the result above) and that of finding a groove and sticking with it. When we control for career length, we find that long-term power users are more stable, and that non-power users across career lengths have similar levels of stability. Interestingly, only short-term power users are less stable than their counterparts. These results indicate that a consequence of attaining power is having to maintain and cultivate it, which is evidenced by greater stability and hints to greater rigidity in behavior.

3.4.3 Power users widen their focus across phases. As users transition into power, their focus widens. The decrease in focus is most stark for short- & medium-term power users. Surprisingly, this trend inverts partially when they fall out of power: short- and medium-term users show an increase in focus, but long-term power users continue to show a decrease (Figure 3b). This indicates two key findings: one, that long-term power users are constantly exploring and participating in a wide range of communities, and the attainment or loss of power does not change that behaviour. Second, that short- & medium-term power users are much more sensitive to losing power, which could explain why they see an increase in focus. These effects hold true for effort and attention.

3.4.4 Power users are more stable when in power. While users may widen their participation as they rise to power, they do not disturb their distribution of effort by much. In their ascent to power, there is little monthly change in activity: they seem to maintain links and activity in a wide collection of communities, but are reluctant to rock the boat in how they behave in there. This effect inverts when they fall out of power – their stability decreases, regardless of their change in focus (Figure 3c). When we control for the duration they spend in power, we notice that the changes in both phases are more muted for long-term power users, while medium-term ones show the greatest variation. We can conclude that users are extremely sensitive to the gain or loss of power and alter their behavior accordingly. When they rise to power, they freeze their relative participation across subreddits, and they rapidly change it up when they fall from power, presumably to regain status.

4 ANALYZING THE LANGUAGE OF POWER USERS

In this section, we first describe the subset of data we rely on, and then a series of experiments designed to measure the effect of rising to power on a user's behavior. These experiments are primarily based on the use of language models to analyze trends in the commenting behavior of power users over time, and with a particular focus on the events surrounding their "ascension" and "fall" from power. Many of the experiments in this section were inspired by the work of Danescu-Niculescu-Mizil et al. [6], which analyzed trends in the behavior of users in two online communities.

4.1 Data subset

For this section, we select the top 1000 communities on Reddit, measured by comment count in our data set. Additionally, we identify all the users who were power users for at least one month, in at least one community, during this time span. We then separate users into categories based on the duration of their tenure as power users: < 3 months, 3-12 months, and > 12 months. We refer to these users later as "short-term," "medium-term," and "long-term" power users, respectively. A user may be a "long-term" power user in one community, a "short-term" power user in another, and not a power user at all in another still. Table ?? shows the number of users in each group. To simplify the reporting of these measurements, we group users by their most-powerful group – e.g., a user who has been a "long-term" power user at any time is reported as one.

4.2 Experimental Design

Community standards and social norms are always changing in online communities. Additionally, in multi-community settings such as Reddit, users can be engaging with more than one community at a time. This creates the potential for users to "code-switch," or modify their commenting behavior, as they move between communities. To better understand the social dynamics of elite users, it is important to understand their behavior relative to *all* the communities in which they participate.

To analyze the commenting behavior of users relative to their communities, we make user of *snapshot language models* to measure the difference between comments and the communities in which they appear. [6] The goal of the *snapshot language model*

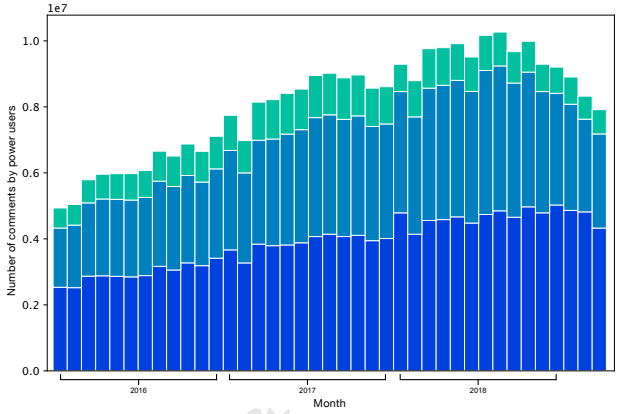


Figure 4: A stacked bar plot displaying the total number of comments produced by users who achieve power in at least one month over the 3+ year period. Short-term power users are on the bottom (in blue), mid-term power users in the middle (in teal), and long-term power users are on the top (in cyan).

is to measure the linguistic norms of a community in each month with a simple language model. Thus, the similarity of a comment to a community's language can be given by the cross-entropy of the comment, given the community's *snapshot language model*. A higher cross-entropy corresponds to more unexpected or unlikely comments for the community at the given time; a lower cross-entropy represents a comment that is highly predictable, or expected.

We implement bigram language models with backoff. We train a model for each community in each month in our four-year span, yielding nearly 40,000 different models; not all of the top 1000 communities have data in every month of the 3+ year span. Having a model for each community, in each month, allows us to evaluate the commenting behavior of users relative to both the communities in which they post the comments, and to the other communities in which they participate.

4.3 Measuring Linguistic Similarity and Difference

For each comment in our dataset, we measure the similarity of that comment to the community in which it was posted. Additionally, for identified power users, we measure the similarity of *all* their comments to the language models of the communities in which they have power. To evaluate the linguistic difference between a given user and a community's language, we extend the Cross-entropy formulation from Danescu-Niculescu-Mizil et al. [6] to average over a user's activity:

$$\text{score}(u, m, LM) = \frac{1}{N} \sum_{i} H(c, LM)$$

where $c_i, ..., c_N$ are the comments made by user u in month m, LM is the language model for community in month m, and $H(c, LM)$ is the length-normalized cross-entropy for the comment.

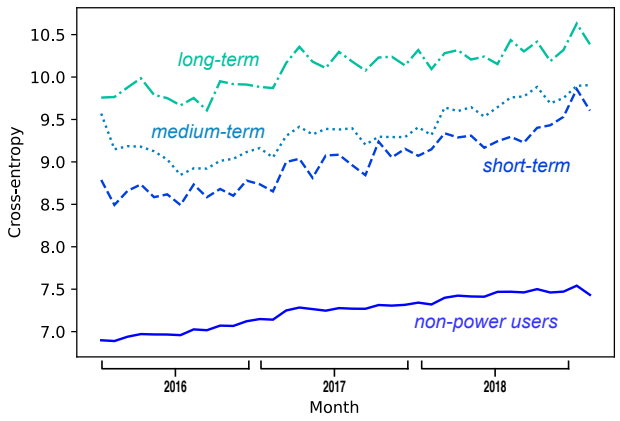


Figure 5: A comparison of language model entropy for different classes of users over time. Power users are highly distinct from non-power users, with the effect most pronounced for long-term power users.

4.3.1 Power users talk differently from normal users. To evaluate the linguistic patterns of both power users and non-power users, we began my modeling the snapshot language model cross-entropy for every comment in the subset corpus. For each month, we then group the users into four classes: (1) non-power user, (2) short-career power user, (3) medium-career power user, and (4) long-career power user. For each community in the corpus, we analyze the difference in cross-entropy trends between the different user categories over the four year span. Figure 5 shows an aggregation of this data across communities. Importantly, we find that these trends continue to hold when we control for each community's month-to-month baseline cross-entropy. 6 shows the difference in entropy distributions for different user classes; non power users are generally predictable, while power users comment in more unexpected ways.

4.3.2 Power users diverge as they rise to power. We also investigate how the ascension to power affects the commenting behavior of power users. In order to measure this effect, we identify the beginning of each user's "power career" for each community they participate in. (Most users have no "power careers" in any community.) We consider the "beginning" of a career to be the first month in which that user is a power user in the community. We then investigate the user's behavior in the k months leading up to their rise in influence, as well as the k months immediately following. Again, we compare across three classes of power users: those with short careers of less than 3 months, medium careers of 3-12 months, and long careers of > 12 months. 7 shows the cross-entropy of power users (in the communities where they wield power) in the months leading up to and following their ascension. 10 shows the cross-entropy of power users in the months leading up to their fall from power.

The measurements indicate that as a user rises to power in a community, they behave in less predictable ways within the confines of that community. Interestingly, this change does not carry

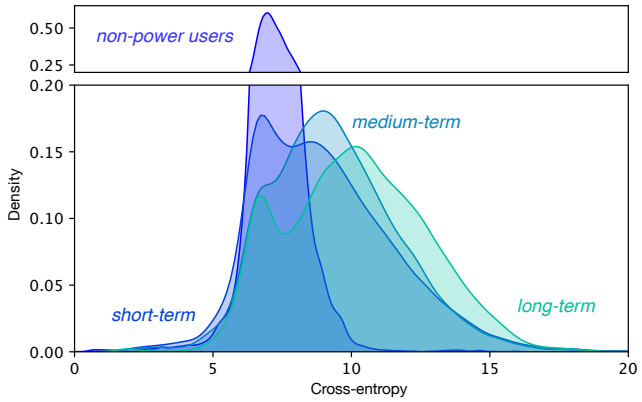


Figure 6: Comparison of the cross-entropy for different user classes. Non-power users typically have lower cross-entropy, whereas power users, with high cross-entropy, diverge from linguistic norms. Note the scale change between the upper and lower plots.

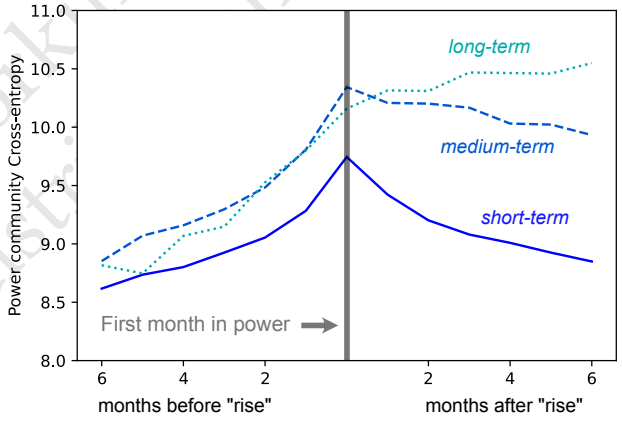


Figure 7: Power users cross-entropy in the months leading to, and following, their rise to power. Power users diverge as they rise in power, with the effect most pronounced and long-lasting for long-career power users.

over to their other communities - over the same time period, in communities where they *do not* achieve power, comments remain as predictable as before. We verify this trend by measuring the correlation between cross-entropy and time t relative to the ascension month for each user. Among long-career power users, the relationship is strongest: a Spearman's rank correlation coefficient of 0.1894 with $p < 5 \cdot 10^{50}$. The correlation remains significant, but weaker, for each descending class of power user. However, across all power user classes, there is not a strong correlation (at most, $r = .01$ for mid-term power users) between t and cross-entropy outside of the community in which they achieve power: their rise to power only has bearing on the divergence of their behavior in the community in which they ascend. The relationship goes in the opposite direction as users fall from power: for long career

power users, the Spearman's rank correlation coefficient between Cross-Entropy and t is -0.0619, with $p < 10^{30}$

4.3.3 Power Users Behave Strategically. Ascension to power also creates increased visibility for elite users. While the ability to blend in with the mass of other users may have afforded them a kind of anonymity in the past, once they become notable – and noteworthy - members of a community, they may face increased scrutiny, even when participating in other communities. To measure this effect, we investigate whether power users remain consistent with the language of their "power communities" even when commenting in other spaces. We use the same strategy as before, this time measuring the Cross-Entropy of each comment made *outside of power communities* $S_0^t, ..., S_N^t$ relative to their "power communities" $C_0^t, ..., C_N^t$ in a given month.

$$\text{score}(\text{user}, t) = \frac{1}{MN} \sum_{i}^{N} \sum_{j}^{M} \text{Cross-Entropy}(S_i, C_j, t)$$

These experiments reveal a dramatic difference in the power user's language across communities even as they ascend to power. In their *home communities*, where they wield power, thier language diverges from the home-community norms as they rise to power. However, their language in other communities does *not* diverge from their home community norms any more than before. This indicates strategic behavior on the part of power users, where they intentionally choose to diverge from a target community *while in that community*, but do not behave in the same way *outside* of the target community.

4.4 Measuring Power-user Controversiality

To better contextualize the strategic behavior of power users, we leverage the binary 'controversiality' flag in the Reddit API. According to Reddit's original source code [1], as well as the Reddit API [2], a comment is deemed controversial if it meets a threshold for number of votes and has an upvote:downvote ratio between 0.4 and 0.6. Comments flagged in this manner indicate divisive content, with a sizeable number of users up- and down-voting it. If power users are in fact creating in-group conflict, we would expect them to have a greater number of controversial comments than non-power users. If they are instead generating divergent, but broadly accepted comments, we expect that their controversiality would not change as they rise to power. Additionally, we would expect the efforts towards controversiality being geared towards the community where they hold power. We follow the analytical setup described in Section 3.3 to study these questions.

4.4.1 Power users are more controversial. When we compare power users to similar non-power users, we find that they generate more controversial comments (Figure 8). The difference is most stark for short-term power users, but is sizeable and significant for long-term power users too. Additionally, we find that their propensity to generate controversy is higher (2.5×10^{-2}) in their home community than outside of it. This reinforces the idea that power users generate divergent content that is not widely accepted in their communities.

[1]

[2]

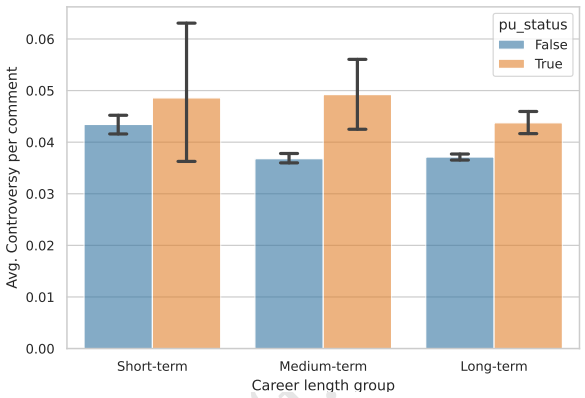


Figure 8: A comparison of average controversiality per comment (99% CI) for power and non-power users of different career lengths. Power users generate more controversy per comment.

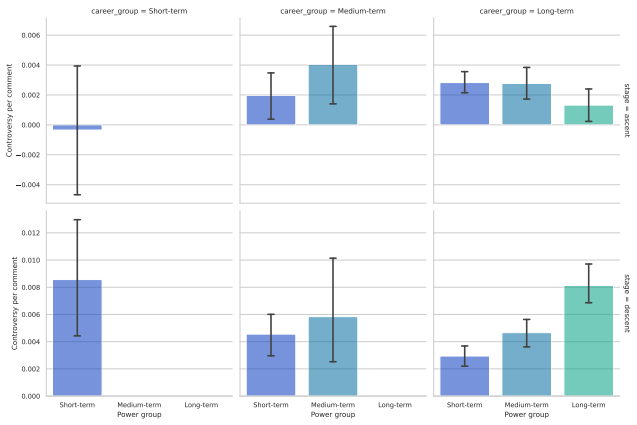


Figure 9: A comparison of mean change in controversiality per comment (99% CI) when power users ascend and descend from power. Power users generate more controversy when they rise to and fall from power.

A user showing a high level of controversiality is associated with them being destined for power.

4.4.2 Power users generate more controversy when in power. Power users generate more controversy when they both rise to and fall from power, with the latter stage showcasing a greater increase across users of different career and power lengths. The associated increase when they rise to power is much lower (Figure 9). This shows us that the path to power is not always lined with self-congratulatory approval alone. Instead, it relies on identifying wedge issues within the community and anchoring the discourse around them. By curating and highlighting these high-temperature, polarizing issues and inviting engagement over them, power users showcase their knowledge of deep debates and disagreements within the community, and are rewarded for it.

REFERENCES

[1] Tim Althoff, Pranav Jindal, and Jure Leskovec. 2017. Online actions with offline impact: How online social networks influence online and offline user behavior. In *Proceedings of the tenth ACM international conference on web search and data mining*. 537–546.

[2] Alessia Antelmi, Delfina Malandrino, and Vittorio Scarano. 2019. Characterizing the behavioral evolution of Twitter users and the truth behind the 90-9-1 rule. In *Companion Proceedings of The 2019 World Wide Web Conference*. 1035–1038.

[3] Dawn Beverley Branley and Judith Covey. 2017. Is exposure to online content depicting risky behavior related to viewers' own risky behavior offline? *Computers in Human Behavior* 75 (2017), 283–287.

[4] Yan Chen, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review* 100, 4 (2010), 1358–98.

[5] Tiago Cunha, David Jurgens, Chenhao Tan, and Daniel Romero. 2019. Are all successful communities alike? Characterizing and predicting the success of online communities. In *The World Wide Web Conference*. 318–328.

[6] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*. 307–318.

[7] Stan Garfield. 2020. The 90-9-1 rule of thumb for community participation. In *Handbook of Community Management*. De Gruyter Saur, 117–126.

[8] Mattia Gasparini, Robert Clarisó, Marco Brambilla, and Jordi Cabot. 2020. Participation Inequality and the 90-9-1 Principle in Open Source. In *Proceedings of the 16th International Symposium on Open Collaboration*. 1–7.

[9] Robert King Merton. 1968. *Social theory and social structure*. Simon and Schuster.

[10] Jeremiah Milbauer, Adarsh Mathew, and James Evans. 2021. Aligning Multidimensional Worldviews and Discovering Ideological Differences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 4832–4845.

[11] Jakob Nielsen. 2006. Participation inequality: Encouraging more users to contribute. *http://www. useit. com/alertbox/participation_inequality. html* (2006).

[12] Eunho Park, Rishika Rishika, Ramkumar Janakiraman, Mark B Houston, and Byungjoon Yoo. 2018. Social dollars in online communities: The effect of product, user, and network characteristics. *Journal of Marketing* 82, 1 (2018), 93–114.

[13] Catherine Ridings, David Gefen, and Bay Arinze. 2006. Psychological barriers: Lurker and poster motivation and behavior in online communities. *Communications of the association for Information Systems* 18, 1 (2006), 16.

[14] Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*. 1056–1066.

A1: SUPPLEMENTARY MATERIALS

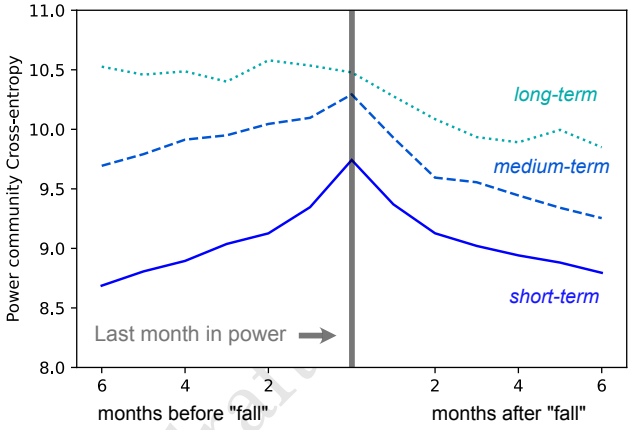


Figure 10: Power users cross-entropy in the months leading to, and following, their fall from power. Long- and medium-term power users language becomes more predictable as they fall from power.