Multi-Scale User Migration on Reddit

Cai Davies[1], James Ashford[1], Luis Espinosa-Anke[1], Alun Preece[1]
Liam D. Turner[1], Roger M. Whitaker[1], Mudhakar Srivatsa[2], Diane Felmlee[3]
[1]School of Computer Science and Informatics, Cardiff University, UK
[2]Pennsylvania State University, USA
[3]IBM TJ Watson Research Center, Yorktown Heights, NY, USA

Abstract

Migration is a natural phenomenon in online social networks where users move across virtual spaces within or across platforms. Platforms such as Reddit drive migration behaviour by facilitating the formation of communities around discussion topics. However, migration also occurs in relation to the emergence of cyber-social threats. For example, 'trolling' involves users intentionally moving in and out of spaces to provoke conflict. In extreme cases, platforms move to shut down spaces due to user behaviour, forcing those users to migrate elsewhere. This paper provides a set of methods to help study migration at two scales: micro-scale involving individual users moving between spaces over relatively short time periods (between posts), and macro-scale involving groups of users moving over relatively longer time periods (changing their posting habits). We take Reddit as an example platform due to its community orientation, and COVID-19 as a means of qualitatively illustrating the results of our methods. We show how micro-scale analyses reveal sub-network structures indicative of overlapping user communities and how macro-scale analyses provide important context to understand the micro-scale patterns, while also revealing consequences of platform moderation on user migration. For COVID-19, the micro-scale analyses reveal an interplay of controversies, politics and humour, while the macro-scale analyses point to increasing politicisation of the pandemic and a rise of conspiracy theories.

Introduction

Migration in online social networks involves users moving across virtual spaces within or across online platforms. It is a natural phenomenon driven by platforms (e.g., Reddit) grouping similar content together where users may interact with other users with similar dispositions on a particular topic. This structure can create a sense of community, but also opportunities for cyber-social threats to emerge. For example, as users can openly contribute to virtual spaces, they can intentionally move in and out to add content intended to provoke conflict (e.g., 'trolling'). This can also be coordinated across groups, creating inter-community conflicts (Datta and Adar 2019) or migration can be forced by moderation measures that shut down spaces which force users to move elsewhere. The focus of this study is to examine how spatio-temporal patterns in migration data can be used to detect potentially disruptive behaviour using Reddit data as an example platform.

Reddit is a class of social media known as social content aggregators, which depend entirely on user generated content and discussion. It is now acknowledged as one of the world's most popular platform for online social interaction; as of January 2021, Reddit has 430 million users[1], which places it higher than Twitter.

Reddit is divided into virtual spaces called subreddits which are created and moderated by users, which center around user-defined topics. Users can post a submission within a subreddit with text, images, or a link to external content. Comments can be added to submissions in a tree structure. Reddit staff have oversight over subreddits and have the ability to quarantine or ban entire subreddits. A ban prevents access to a subreddit, while quarantining involves limiting access by removing and disallowing links or searches that direct to it; this is usually the case when content is deemed offensive, misleading etc. but not to an extent that requires the full banning of that subreddit.

The topic-oriented subreddit structure of Reddit provides an opportunity to explore how users navigate content. Note that alongside direct participation through content creation, another key aspect of user behaviour is their switching between participation in alternative subreddit. Knowledge of this behaviour is valuable because it provides additional context on a user's contributions, and gives insights into their motivation and similarity with other users. This switching behaviour constitutes user migration; in this paper we pro-

This research was sponsored by the U.S. Army Research Laboratory and the UK Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the UK Ministry of Defence or the UK Government. The U.S. and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

[1]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

vide a means of analysing migration patterns over different scales. Specifically, *micro migration* addresses near-term (sequential) transition between subreddits for a single user, while *macro migration* addresses the longer term migration trends aggregated across multiple users. For macro migration, we distinguish between *natural migration* and *forced migration*, where the latter involves migration from a subreddit that has been banned or quarantined.

The primary contribution of this work is to provide a set of methods for studying migration at micro- and macro- scales. We illustrate the utility of the methods by taking Reddit as a focal platform due to its community orientation, and COVID-19 as an illustrative topic area for qualitative understanding of our results. We show (i) how our micro-scale analysis methods highlight sub-network features indicative of overlapping user communities, and (ii) how our macro-scale analysis methods contextualise the micro-scale patterns while also aiding understanding of the consequences of platform moderation interventions.

Related Work

Migration in the context of social media is usually studied in the form of migration between platforms, which is of particular relevance following the Washington DC Capitol Hill riot in January 2021, which was preceded by users migrating from platforms that they perceived as 'restrictive' such as Twitter to Parler (Aliapoulios et al. 2021). This shows the potential impact social media migration has outside of virtual space.

Online user migration behaviour is heavily dependant on graph-based metrics (Cai, Decker, and Zheng 2019) (Guimarães, da Silva, and Almeida 2015). Research has shown how cross-posting on Reddit can be used to model the flow of information from one subreddit to the next as new subreddits are created as others are banned (Cai, Decker, and Zheng 2019). In doing so, a simple classification task is used to predict which subreddits are likely to become banned based upon the source and destination. Furthermore, similar behavior can be observed through ad-hoc communities on knowledge-sharing networks (Guimarães, da Silva, and Almeida 2015).

(Newell et al. 2016) studies the migration of users on Reddit to alternative platforms following a major ban event in 2015, which caused users to find less moderated platforms. The study relied on matching user accounts across multiple platforms by algorithmically matching usernames that were likely to be shared by a single user. (Kumar, Zafarani, and Liu 2011) looked to a less permanent form of migration defined by the movement of attention from one platform to another, where a user active on platform A at time t_i and at time $t_j > t_i$ is active on platform B is said to have migrated from A to B. In our terms, these two works focus on macro- and micro- timescales respectively, as defined in the Introduction.

From (Datta and Adar 2019), we find the idea of users having a 'home' in a subreddit that arises from frequent interaction with that community, and is distinct from 'drive-by' posting. This work discusses the need for this notion of 'membership-by-proxy' due to the problems with membership on Reddit. That is, while Reddit has a system to 'join' a subreddit, this information is not made public; regardless, this type of membership does not guarantee the user is engaged in the community, and may give little or no interaction.

Community influence in online discussion is studied in (Belák, Lam, and Hayes 2012), which looked to a platform very similar to Reddit (Boards.ie) and concluded that some communities are highly influential or highly dependent on others, which relied on cross-community posting.

Finally, a body of works focuses on content analysis in Reddit using natural language processing (NLP) techniques. The insights to be gained from studying Reddit posts is varied, with some of the most prominent being: studying users' responses to various real-world and platform-specific events (Musco, Musco, and Tsourakakis 2018; Shen and Rose 2019), modeling discourse and conversation dynamics using LSTMs (Zayats and Ostendorf 2018), and humour detection with transfer learning (Wang et al. 2020).

Data

As indicated in the Introduction, we focus on COVID-related subreddits. Subreddits such as r/Coronavirus and r/COVID19 were easily found due to the use of the common name for the virus, and from these a 'snowball sampling' method was used; from links and user posting habits, we could find other popular COVID-oriented subreddits. While some of these subreddits cover the topic of COVID-19 in general, other subreddits are more focused, such as around a geographic area (r/CoronavirusUK) or theme such as anti-lockdown (r/LockdownSkepticism) or anti-masks (r/NoLockdownsNoMasks). We collected data from the start of 2020 to March 2021. For geographical-focused subreddits, we focus on US, UK, Canada and Australia, which are the top 4 countries for Reddit traffic (as of December 2020).[2] The full set of focal subreddits can be seen in Table 1.

Reddit provides an API[3] to download submissions and comments given parameters such as the subreddit and date range, as well as user information. However, the API provides no access to content from subreddits that have been quarantined or banned, as well as no endpoint to search all submissions from a subreddit. To overcome these limitations, the Pushshift API[4] was used to download submissions and comments where the Reddit API had no access; Pushshift archives submissions and comments posted to Reddit and so allows access to data pre-ban.

From the submissions and comments downloaded for a subreddit, we can build a set of *top users* by engagement, that is we take the most active users of a subreddit. For windowed submissions and comments where $t = 1$ month, we can select the top n users based on submission activity and top n users based on comment activity ($n = 500$). For each of these users, the Reddit API was used to download the set

[2]https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

[3]https://www.reddit.com/dev/api/

[4]https://pushshift.io/api-parameters/

Index	Subreddit	Submissions		Comments	
		Count	Mean	Count	Mean
1	r/NoNewNormal	33,420	4.3	234,432	10.1
2	r/COVID19	32,580	2.1	100,454	5.0
3	r/Coronavirus	364,275	3.6	2,785,219	7.2
4	r/LockdownSkepticism	27,232	4.0	230,265	18.5
5	r/CovIdiots	17,044	1.9	73,111	3.7
6	r/CanadaCoronavirus	13,221	3.6	66,291	9.4
7	r/CoronavirusUK	35,974	3.1	222,808	10.1
8	r/CoronavirusCA	10,046	2.5	39,019	4.4
9	r/China_Flu	84,162	4.2	427,326	10.5
10	r/COVID	12,213	1.5	15,759	3.2
11	r/CoronavirusUS	47,268	2.1	180,701	5.2
12	r/CoronavirusDownunder	18,234	3.9	122,559	10.6
13	r/CoronavirusCirclejerk	15,437	5.8	88,924	13.3
14	r/PublicFreakout	131,896	2.5	4,794,530	5.8
15	r/conspiracy	272,584	3.8	2,424,871	12.4
16	r/politics	504,814	5.6	15,165,267	14.3
17	r/NoLockdownsNoMasks	1,194	45.9	1,310	7.2

Table 1: Table of subreddit indexes for our COVID-19 focal set where 'X' denotes 'other', i.e., all subreddits not in this set. Submissions and Comments columns formatted as count and mean across users

of comments and submissions across all subreddits made by this user that had not been deleted or removed. The sampling date for the user data was from November 2019 to March 2021, as this gives some insight into the user behaviour pre-pandemic. While this method may not accurately represent the actual population of users who interact with a particular subreddit, given the emphasis on high activity, these top users do represent a majority of the content created in the subreddit; a higher value for n can mitigate this problem. (Newell et al. 2016) describes a 'silent majority' where the majority of users are infrequent posters, and instead browse content silently, so the data for these users is likely to be far from the ground truth of subreddits they visit. Also is the lack of public membership to particular subreddit, as mentioned in Related Work with (Datta and Adar 2019), and so activity is used as a proxy for membership.

Methods

In this section, we will describe the methods used for *micro migration*, *macro migration*, and distinguish *natural* and *forced* migration. We will also discuss at a higher level the implications and observations these methods are attempting to uncover. First we define the set S, which is the set of all subreddits that have ever existed, regardless of quarantined or banned status. We then define S_c, which is a subset of S and is used as the focus for these methods; for this paper we have chosen to focus on COVID-related subreddits, and the behaviours that arise from users in these communities during the period of the COVID-19 global pandemic. It is worth noting that for other studies, this set S_c can be any non-strict subset of S to provide focus to other topics found in Reddit.

We use the term 'post' to mean a comment *or* submission, but for this study we focus on comments, given the tendency of disruptive behaviour to arise from inflammatory comments rather than submissions (Datta and Adar 2019), as well giving the study a larger sample size.

Migration

Micro migration is the sequential movement of a user between subreddits from post to post, so if a user posts in subreddit B where the previous post was in subreddit A, then the user is said to have (micro-)migrated from A to B. We can construct a graph of micro migration behaviour for a user where nodes denote subreddits and edges represent the counts of that user posting sequentially in A then B. These counts can be scaled by the total number of posts made by that user to get the probability of migration, that is given a post in subreddit A that the next post is in subreddit B.

Formally, we can first define the graph as $G = (N, E)$ where N is the set of targeted subreddits $S_c \subseteq S$ if S is the set of all subreddits, and $E = \{(x, y)|x, y \in N^2\}$; for this paper, the set S_c can be seen in Table 1. For a user u and the set of associated posts P_u, we can define a weight function: $\omega : E \mapsto \mathbb{R}^{\geq 0}$ where $\omega_u(x, y)$ is the count of posts in P_u in which y follows x divided by $|P_u|$ if $|P_u| > 0$ otherwise $\omega_u(x, y) = 0$.

From this, we can average over a set of users U to define a weight function of group micro migration behaviour

$$\omega(x, y) = \frac{\sum_{u \in U} \omega_u(x, y)}{|U|}$$

We also introduce the distinction between the use of 0 and NaN, where we can can redefine the edge weight function ω_u to be

$$\omega'_u(x, y) = \begin{cases} \omega_u(x, y) & \omega_u(x, y) > 0 \\ \text{NaN} & \text{otherwise.} \end{cases}$$

This results in a graph where the edges are the probability of a post from A to B given the user has made the migration from A to B at least once. The purpose of this is to remove the contribution of users to subreddits they have never posted to as a way of reducing the differences in popularity between subreddits.

Macro migration occurs at larger scales than micro: larger timescales and involving groups of users rather than individuals. Three motivating examples of macro migration on Reddit are (i) where a popular subreddit is banned, forcing its regular users to migrate elsewhere in terms of their posting behaviour, (ii) where a formerly-popular subreddit loses its popularity over time, as its regular users naturally migrate elsewhere, and (iii) where a new subreddit is created and users migrate in from other subreddits.

Observing aggregated patterns of migration behaviour in Reddit is a challenge because there is no public notion of membership to subreddits. Therefore we align users to their 'favorite' subreddits, based on their posting activity. To achieve this, we look to create an aggregation of activity volume for a group of users across all subreddits they interact with, which can be visualised as the daily, weekly, monthly activity of these users. This allows us to observe where the majority of interactions are occurring in Reddit-space for this set of users at a certain time, and how this evolves temporally as external (bans, real-world events) and internal (subreddit popularity, changing views) factors come in to play.

Macro migration is therefore the observation of aggregated user behaviour defined relative to (i) a set of users of interest U_i and (ii) a timescale over which to observe them. For this paper, we use the set of top users of a chosen subreddit $s \in S_c$ as the set of users of interest, but note that this set can be defined by other means, for example, a set of known troll or bot accounts. For a user $u \in U_i$ and the set of associated posts P_u, we can use a sliding window method to count the activity of the user in a time frame, for every subreddit $s' \in S$ the user interacts with, which can be daily, weekly etc.

From this, we can aggregate over all users in U_i by the summation of absolute counts, which results in the windowed activity for a group of users. Visualised, we can observe where the users of interest spend the majority of interaction at time t, and how this changes across time $t + x$.

We can also take the mean over normalised counts, where the activity of each user is normalised to the range $[0, 1]$ through dividing by the maximum value, which gives equal weighting to users regardless of posting frequency, since a very small minority of users are substantially more active, as seen across the subreddits in Table 1.

Natural and forced migration are classifications based on the causes of migration, where natural migration can be caused by many factors, as mentioned previously, with the important distinction being these causes do not *force* the user to migrate away from a subreddit. As mentioned in the Introduction, Reddit moderators can ban or quarantine subreddits if they break content policy. Following a ban, the users of that subreddit are forced to migrate their primary attention to other subreddits, given the banned subreddit is no longer accessible. It may be the case that a user becomes inactive or is suspended or deleted following a ban, and this is potentially further encouraged by the ease of creating multiple so-called 'throwaway' accounts (Leavitt 2015) in Reddit without the need of email verification. Quarantined subreddits, while not directly forcing a user away from a subreddit, may encourage this with a warning page before directing to the subreddit. In the results, we shall discuss differences in the migration behaviour of users sampled from both normal and banned or quarantined subreddits.

Results

Micro migration

Below, we can see the resulting graphs of micro migration for both a set of subreddits and the ego-centric case. These graphs are calculated as described in the Methods section by averaging the micro migration graphs across a set of users, which here would be the set of top users across a set of subreddits or the set of top users for the ego-centric subreddit. We also define the node 'other' which is the set of all subreddits not in the set S_c. In Figure 1 we show the visualisation of the graph G across the set of top users of all subreddits in S_c.

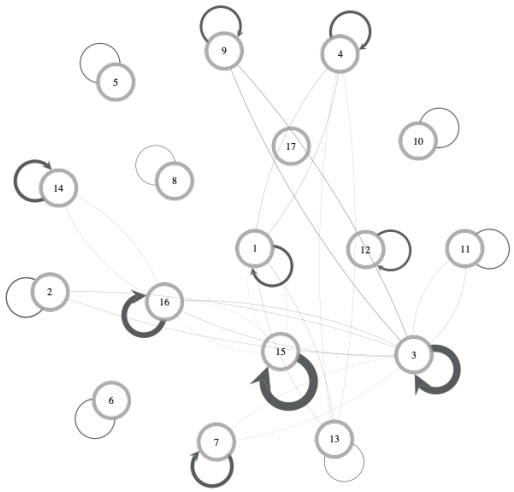

Figure 1: Micro migration for all users with 'other' ignored

Immediately noticeable are the loop edges, which indicates the highest probability after posting in x is to post again in x, and so no micro migration has occurred. We also leave out the node 'other' since the edges to and from this node are substantially larger than all other edges. Disregarding the loop which indicate no migration has occurred, we notice the largest edge is between r/Coronavirus (3) and r/China_Flu (9), which potentially indicates the highest rate of back-and-forth posting between these two subreddits.

However, these are also the two largest COVID-specfic subreddits in S_c by number of comments, that is disregarding r/PublicFreakout (14), r/conspiracy (15) and r/politics (16), and so we investigate whether this is the cause by using ω'_u. As mentioned, this is a method of mitigating the

popularity differences between subreddits by removing the contribution of users to the mean of edges they never migrate across. We found that edges with a very small number of users contributing however resulted in disproportionately large probabilities, and so we introduce a minimum number of users n for an edge to be kept. Figure 2 shows the same set of top users with the ω'_u weight function and a user limit $n = 12$.

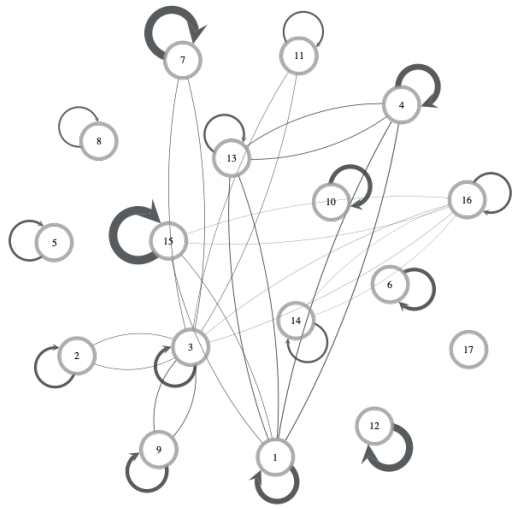

Figure 2: Micro migration for all users using ω'_u with 'other' ignored

We can see the difference the use of ω'_u makes in mitigating the difference in subreddit popularity, with many more paths of micro migration being visualised. While the r/Coronavirus (3) and r/China_Flu (9) edge remains strong, we notice the largest edges are now between r/NoNewNormal (1), r/LockdownSkepticism (4) and r/CoronavirusCirclejerk (13), which indicates a very strong back-and-forth posting for the users who have engaged at least once in these communities; while the first two are intuitively anti-lockdown subreddits, r/CoronavirusCirclejerk (13) covers memes and 'stupid, ridiculous and amusing' posts from 'panic-filled and alamarist Coronavirus-related sub', which has implications of anti-lockdown and COVID-skeptical themes. Reducing the set of users to the set of top users of a subreddit, we can investigate the ego-centric case; for this, we look to r/CoronavirusCirclejerk (13) to focus on the triad mentioned above, and is shown in Figure 3.

This ego-centric graph reinforces the back-and-forth posting shown in the edges between r/CoronavirusCirclejerk (13), r/NoNewNormal (1) and r/LockdownSkepticism (4) as seen in Figure 2. What we notice however, is the probability of the top users of r/CoronavirusCirclejerk is higher for posting and staying in r/NoNewNormal rather than r/CoronavirusCirclejerk, which indicates that this is not their main subreddit to engage with. For this graph, we include the node 'other' to show the exploratory nature of these users, that is they are not merely engaging in a very small set of communities, but rather have a higher probability of migrating to and posting in 'other'; the loop for this node was removed for easier visualisation, as it is substantially larger than all other edges. This problem may be mitigated by increasing the size of the set S_c, or indeed allowing S_c to be dynamic in the ego-centric case to better capture the micro migration patterns of these users without the need for 'other'.

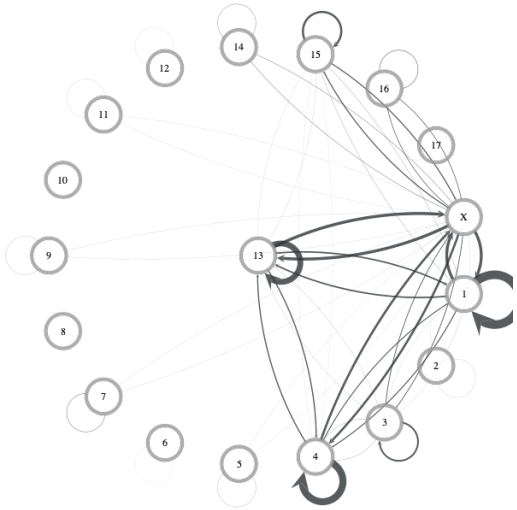

Figure 3: Ego-centric micro migration for r/CoronavirusCirclejerk with loop removed for 'other'

While the use of ω'_u is used to mitigate the popularity differences between subreddits, for the ego-centric case this results in a graph where meaningful behaviour cannot easily be seen. It also defeats the purpose of reducing the set of users to an ego-centric subreddit in visualising micro migration at the community level.

To compare the micro migration graphs of the subreddits in S_c, we can define an adjacency matrix A_s where $A_s(x, y) = \omega(x, y)$, and calculate the distance between adjacency matrices by using the Frobenius norm of $A_x - A_y$ if $x, y \in S_c$. In Figure 4, we show these distances for all $x, y \in S_c$.

Here we can see quantitatively the similarity between subreddits based on the macro migration behaviour of their respective top users. We notice a high similarity between between r/NoNewNormal (1), r/LockdownSkepticism (4) and r/CoronavirusCirclejerk (13), which follows from the observation of high back-and-forth posting between these subreddits in Figure 2, as well as a high similarity between r/Coronavirus (3) and r/CoronavirusUS (10). We also notice that r/PublicFreakout (14), r/conspiracy (15) and r/politics (16) typically show a typically higher distance from the directly COVID-related subreddits, which indicates a measurable difference in the posting behaviour of those users.

To calculate the similarity of the user bases between $x, y \in S_c$, we use the Jaccard index (Hamers and others 1989) on the sets U_x and U_y which is defined as the size of the intersection divided by the size of the union of two sets.

The high similarity in the top user sets between

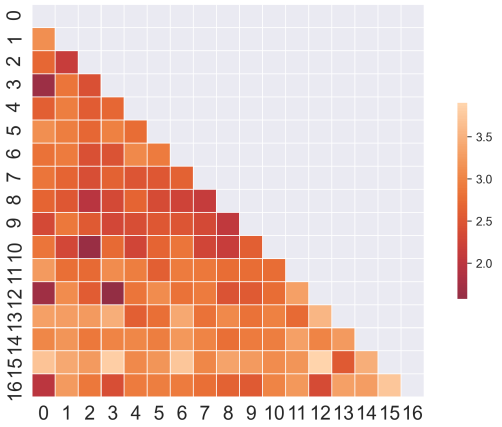

Figure 4: Frobenius distance between ego-centric micro migratio graphs

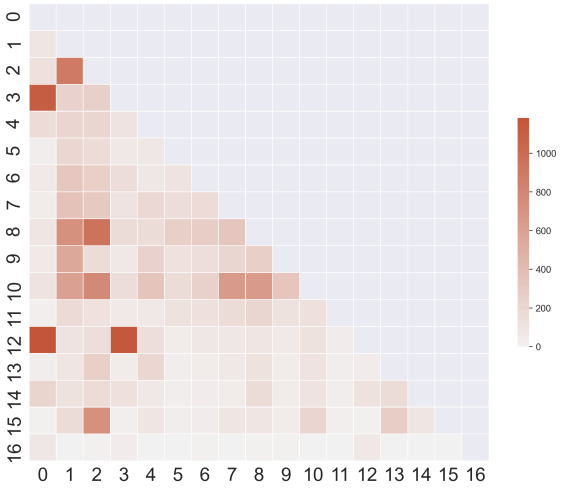

Figure 5: Jaccard index of top user sets

r/NoNewNormal (1), r/LockdownSkepticism (4) and r/CoronavirusCirclejerk (13), as well as between r/Coronavirus (3) and r/CoronavirusUS (10), may explain to some extent the similarity seen in the micro migration graphs. However, we notice a low Frobenius distance between r/NoNewNormal (1) and r/NoLockdownsNoMasks (16) while also showing a low similarity in the top users sets.

Macro migration

For macro migration, we take the set of top users U_s of a subreddit $s \in S_c$ and create the set of time series for each user $u \in U_s$ with a window size of 1 day, as described in the methodology section. We can then sum across all users in U_s for each time series to get the daily activity for the top users of s across all subreddits. Since the set of top users is built by the union of the top n commentors and top n submittors each month, the set used for macro migration can be restricted to the top users at month t, however for this paper we visualise across the entire set of top users U_s. In Figure 6 we show the result of this for r/NoNewNormal, which aligns with motivating example (iii) as mentioned in the section Methods.

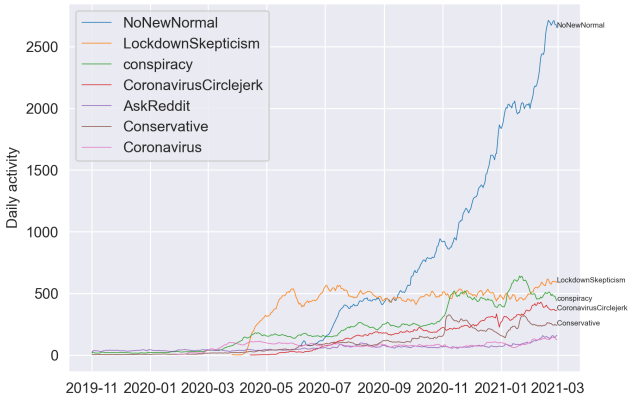

Figure 6: Macro migration of top users of r/NoNewNormal with 7-day moving average

We can see clearly the behaviour of $U_{\text{NoNewNormal}}$ showing the migration from r/LockdownSkepticism (4) into r/NoNewNormal (1), and before that the migration from r/Coronavirus (3) and r/conspiracy (15) into r/LockdownSkepticism. As lockdowns became global and seemingly never-ending, anti-lockdown movements grew in popularity, particularly in the US with the far-right (Vieten 2020), and so this increasing migration we see from general COVID subreddits into communities surrounding topics of lockdown skepticism and the anti-'new normal' views aligns with this real-world trend.

Figure 7 shows the method of averaging over the set of user time series rather than using summation, where the activity of each user is normalised to the range $[0, 1]$, and so giving equal weight to users rather than weight on activity. Interestingly, we see similar behaviour but with slight differences, most notably the subreddit r/Conservative in

November seeing normalised activity almost as high as r/NoNewNormal, as well as the exclusion of r/Coronavirus in the top 7 subreddits for commenting activity.

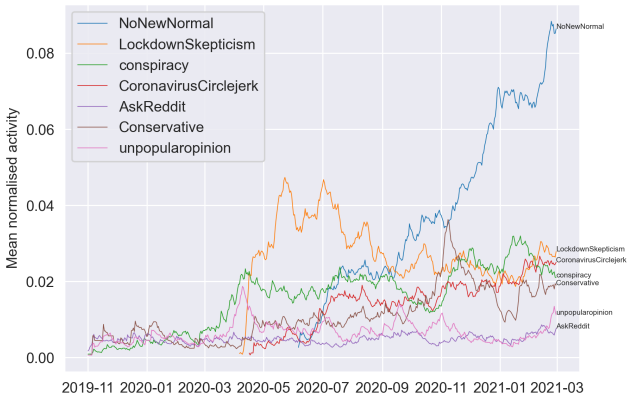

Figure 7: Mean per-user normalised macro migration of top users of r/NoNewNormal with 7-day moving average

The presence of r/Conservative in the set of 'favourite' subreddits for these users may somewhat indicate the political bias and potentially the ideologies of these subreddits that extend past the topics of anti-lockdown. Also we notice the activity on r/conspiracy and r/unpopularopinion prior to r/LockdownSkepticism and r/NoNewNormal, as well as the increasing activity on r/conspiracy over this time frame, which again may give some indication on the type of user and the content generated. This analysis is simply a basic qualitative view of these results, and as discussed in Future Work would benefit by extending the analysis of the content through NLP methods.

Natural vs Forced migration

For this section, we focus on the subreddit r/Wuhan_Flu for the macro migration behaviour of a quarantined subreddit, given that the focus has been on COVID-related subreddits. We also investigate the trend of user activity following a ban when compared to normal or quarantined subreddits; for this we use r/The_Donald, r/DebateAltRight and r/okbuddyanarchy as example cases, since these subreddits faced bans during 2020. For quarantined subreddits, we include r/TheRedPill, r/CoronavirusConspiracy, r/MGTOW and r/FULLCOMMUNISM. This analysis aligns with motivating examples (i) and (ii) from the section Methods.

Figure 8 shows the macro migration times series for Wuhan_Flu. We can see early on, around February, a dip in daily activity which potentially indicates the date that the subreddit was quarantined; although this information is not publicly available to be sure. Post-quarantine, the daily activity continues to rise to become the main subreddit of these top users. We can also see the migration of these users from r/China_Flu and r/Coronavirus, and before that r/conspiracy; albeit that r/Wuhan_Flu came shortly after r/China_Flu and r/Coronavirus and activity of all 3 grew to a peak in early March of 2020.

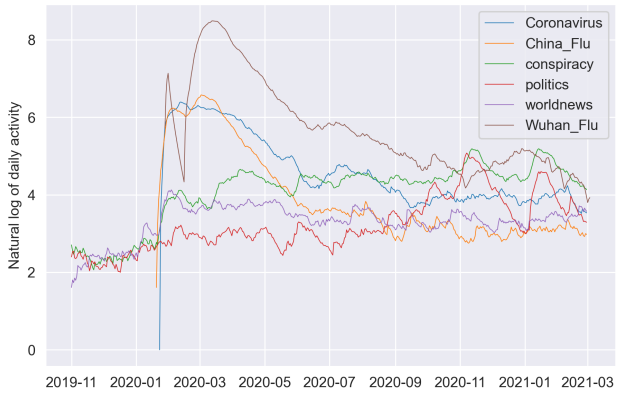

Figure 8: Forced macro migration of top users of r/Wuhan_Flu

What is noticeable is a drop in overall activity that is much more significant than most other subreddits from the set S_c, with the exception of r/China_Flu. Given that subreddits are quarantined and banned due to the content, and the content is user generated, it makes sense that users will face suspension pre or post this action. Here, we will investigate the activity of top users for banned or quarantined subreddits in comparison to the normal subreddits in S_c.

We define the active range of a user to be between the first and last post made by the user on any subreddit, at the time of the data being downloaded. From this, we can count the number of users in the set U_s for a particular subreddit s at a particular time t that are considered to be 'active'. In Figure 9, we plot these counts with a window size of 1 day across all the subreddits in S_c as well as the banned and quarantined subreddits mentioned. Each time series is scaled by its maximum value to normalize to the range $[0, 1]$.

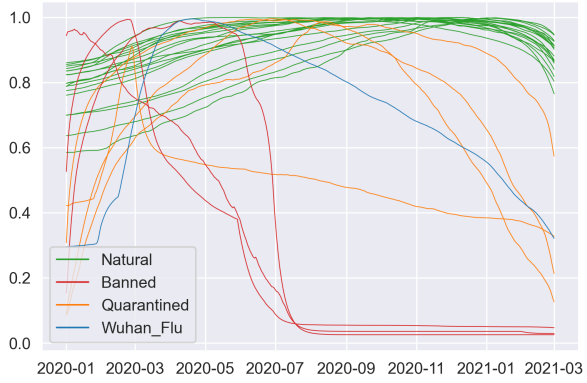

Figure 9: Normalised count of active top users

Here we see a clear difference between the 'activeness' of banned, quarantined and normal subreddits. From this small sample of banned subreddits, we can see that the top users tend to become inactive very quickly following a ban, which is likely aided by idea of 'throwaway' accounts (Leavitt 2015). As mentioned in the section Methods, users are very

likely to face suspensions pre or post a subreddit ban given the content on that subreddit is user generated, and so preemptive 'throwaway' accounts are a way of mitigating this.

Interestingly, quarantined subreddits seem to fall somewhere between normal and banned subreddits here, with r/MGTOW behaving closest to a normal subreddit. However, the majority of these subreddits seem to be converging towards 0 in the same way we see with the banned subreddits, which may indicate the 'throwaway' account trend is similar.

We do however see the number of active top users beginning to drop around the end of 2020 for the set of normal subreddits, including established ones such as r/politics and r/conspiracy, so the time at which the data is downloaded may have a small impact on the number of active users, however the difference is significant enough that this would have minimal impact.

Discussion

From the above illustrative examples, we see that micro migration analysis is able to highlight sub-network features revealing inter-community relationships. By setting S_c to a collection of communities-of-interest, we can readily see where active users migrate over short timescales. In our COVID study, this highlighted the strong relationships (i) between the general r/Coronavirus and specfic r/China_flu subreddits, the latter having in previous studies been shown to be more prone to racist posting behaviour (Zhang et al. 2020), and (ii) a tight interplay between the anti-lockdown r/LockdownSkepticism subreddit, the more generally COVID-skeptic r/NoNewNormal subreddit, and the humour-oriented r/CoronavirusCirclejerk subreddit. Our macro-scale analysis provided important context for the micro-scale migration patterns. Over a longer time period, we saw that the rise of user engagement in the r/NoNewNormal subreddit as the pandemic progressed was mirrored by increased engagement of those users with r/conspiracy and r/Conservative among other subreddits. Moreover, we observed that Reddit moderation actions—'quarantines' and bans on subreddits such as r/Wuhan_Flu—tended to result in users shutting down their accounts or subsequently reducing their posting activity on those accounts, likely indicating their switching to newly-created accounts.

The preceding examples also show how the micro and macro analysis methods can be used to 'zoom in' and 'zoom out' as an analyst examines a focal set of community interactions. The progressive 'zooming in' shown in Figures 1'to 3 led to a focus on the subreddits mostly closely inter-connected with r/CoronavirusCirclejerk, including r/NoNewNormal. We then took r/NoNewNormal as a focus for 'zooming out' to the macro migration analysis in Figure 6. That macro-scale view brought additional subreddit outside the original focal set—such as r/conspiracy—and the analyst can then modify the set S_c as desired to 'zoom in' once again for a micro-scale migration view.

Conclusion and Future Work

This paper has introduced two methods for analysing migration in Reddit on multiple scales, and shown how they can be used in conjunction to provide focus for observation on interesting behaviours. For COVID-19, the micro-migration revealed an interplay of controversies, politics and humour, while the macro-scale migration highlighted increasing politicisation of the pandemic and a rise of conspiracy theories. We note that these methods are potentially useful in future work for areas such as modelling and prediction of user behaviour online, tracking the activity of bot and bad faith actor accounts, and enhancing analysis of polarisation and the conflict that arises from it. This work can be extended with the introduction of NLP for content analysis, through techniques such as word embeddings, e.g., (Mikolov et al. 2013; Bojanowski et al. 2017) or language models, e.g., (Radford et al. 2019; Liu et al. 2019), which would open up questions such as, "What topics are contributing to a migration trend?" or, "What are the most frequently used terms across source and target subreddits in migratory terms?"

References

Aliapoulios, M.; Bevensee, E.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and Zannettou, S. 2021. An early look at the Parler online social network. *arXiv preprint arXiv:2101.03820*.

Belák, V.; Lam, S.; and Hayes, C. 2012. Cross-community influence in discussion fora. *Proceedings of the International AAAI Conference on Web and Social Media* 6(1).

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics* 5(1):135–146.

Cai, B.; Decker, S.; and Zheng, C. 2019. The migrants of Reddit: An analysis of user migration effects of subreddit bans. *preprint*.

Datta, S., and Adar, E. 2019. Extracting inter-community conflicts in Reddit. *Proceedings of the International AAAI Conference on Web and Social Media* 13(01):146–157.

Guimarães, A.; da Silva, A. P. C.; and Almeida, J. M. 2015. Temporal analysis of inter-community user flows in online knowledge-sharing networks. In *SIGIR 2015 Workshop on Temp, Soc and Spatially-aware Info Access*.

Hamers, L., et al. 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management* 25(3):315–18.

Kumar, S.; Zafarani, R.; and Liu, H. 2011. Understanding user migration patterns in social media. *Proceedings of the AAAI Conference on Artificial Intelligence* 25(1).

Leavitt, A. 2015. "This is a throwaway account": Temporary technical identities and perceptions of anonymity in a massive online community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, 317–327. New York, NY, USA: Association for Computing Machinery.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Musco, C.; Musco, C.; and Tsourakakis, C. E. 2018. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 World Wide Web Conference*, 369–378.

Newell, E.; Jurgens, D.; Saleem, H.; Vala, H.; Sassine, J.; Armstrong, C.; and Ruths, D. 2016. User migration in online social networks: A case study on Reddit during a period of community unrest. *Proceedings of the International AAAI Conference on Web and Social Media* 10(1).

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Shen, Q., and Rose, C. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online*, 58–69.

Vieten, U. M. 2020. The "new normal" and "pandemic populism": The COVID-19 crisis and anti-hygienic mobilisation of the far-right. *Social Sciences* 9(9).

Wang, M.; Yang, H.; Qin, Y.; Sun, S.; and Deng, Y. 2020. Unified humor detection based on sentence-pair augmentation and transfer learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 53–59.

Zayats, V., and Ostendorf, M. 2018. Conversation modeling on Reddit using a graph-structured LSTM. *Transactions of the Association for Computational Linguistics* 6:121–132.

Zhang, J. S.; Keegan, B. C.; Lv, Q.; and Tan, C. 2020. A tale of two communities: Characterizing Reddit response to COVID-19 through /r/China_Flu and /r/Coronavirus. *arXiv preprint arXiv:2006.04816*.