Reddit Hate Speech and Russian Trolls

Wei Chi
CS416
University of Alabama at Birmingham
wchi@uab.edu

Cole O'Brian
CS416
University of Alabama at Birmingham
cobrian@uab.edu

Mohammad Abdelghani
CS416
University of Alabama at Birmingham
mohammad@uab.edu

ABSTRACT

This paper explores, firstly, whether Reddit has a 'hate' problem using the HateBase dictionary, and secondly, Russian Troll behaviour on Reddit. For the first question, we examine Reddit comments and submissions from February 2016 to February 2018. Using our defined metrics (hate level and hate score) to examine four groups: subreddits by submissions, subreddits by comments, users by submissions, and users by comments. We find that each group has low hate levels and hate scores, and Reddit is overwhelmingly disapproving of hate. Though there are hateful subreddits and users, they form only a small part of Reddit. For the second question, we examined, in particular, language use, subreddit activity, hourly posting patterns, and daily post counts. We found that Russian Trolls have acted in a coordinated effort towards US political and cryptocurrency objectives.

1 INTRODUCTION

An analysis on the hateful content on Reddit was done in an effort to determine how hateful Reddit is. Word counts were obtained based on different criteria, and were compared to a list of reference hate words in an attempt to gain better perspective on Reddit's possibly hateful nature. Counts were done based on submissions as well as comments, and organized by subreddit as well as by user. It was hypothesized that by grouping in this manner, specific hateful users and subreddits would become edge cases that could be further examined on an individual basis.

In addition to hate speech, Russian troll activity was analyzed based on user activity times as well as relative hatefulness of post content. The Russian troll activity was found by targeting known Russian troll accounts, and was similarly separated into posts by submission and post by comment. The actual content of the posts were then analyzed with regards to most used words, and the relative "hatefulness" of Russian troll posts was compared to normal user post data. Relative hatefulness of posts was determined simply by examining the distribution of hate words in each post.

1.1 Reddit and Hatespeech

Reddit, the self-proclaimed 'Front Page of The Internet' is a heavily trafficked content aggregation site. With over 234 million registered users and 1.14 billion monthly visitors since 2015, both registered and not.[2] The reach of reddit is incredibly wide. The site is structured around posts, known as "submissions," that registered users create. From there, other users can upvote or downvote the submission. In addition, there is also a comment section for each submission in which users can comment on the original post, or reply to another comment. These comments are also subject to the voting system. Reddit is largely unlike other social media platform. In most social media platforms, friendships or followers define the community relations; whereas on reddit, friendships are possible, but are far from the focal point of the platform.

Large and small communities on Reddit are organized into "subreddits." Subreddits allow users to congregate into one forum so that they may discuss, or share information on the topic of choice. Subreddits can be created by any registered user, and therefore revolve around virtually any topic one can imagine. Topics range from mundane topics like cute animals to very technical topics such as quantum physics theories. In addition to subreddits, users also have the option of owning a personal page that is similar in layout to a subreddit. These profile pages can contain submission and comment data, however a distinction should be clearly made between these pages and actual subreddits.

The freedoms afforded to Reddit's users makes for a discussion rich forum of virtually any topic one could choose to discuss. This freedom however, is not without it's negative impacts. One example of a negative impact is "The Fappening." The Fappening was a popular subreddit in which nude photos of various celebrities and famous athletes were obtained by malicious parties and were posted for any and all to see. Although the subreddit was shut down by site admins in late 2014, many criticize Reddit for allowing it to exist for as long as it did [9]. In recent months, Reddit has again found itself on the receiving end of harsh criticism, with many claiming that the site is a safe haven for racists and hateful individuals [10]. Subreddits like 'The_Donald', a support community for President Trump, and socialism, a community for advocates of socialism, are often cited as being large contributors to the alleged hateful atmosphere on Reddit.

1.2 Russian Trolls on Reddit

User accounts on various platforms have been linked to organizations operating on behalf of the Russian government. These "Russian trolls" participate in international political blogs and forums in an attempt to manipulate social media and spread computational propaganda. It was hypothesized that these Russian troll accounts may exhibit behavior uncharacteristic of normal users. These uncharacteristic behaviours could include frequency/distribution of posts, content of posts, and activity times.

State sponsored Russian propagandists must have specific targets. By analyzing their post submission content, targets may become clearer. Conclusions could then be draw about the types of disinformation they spread. Russia has seemed to have US politics as part of it's agenda. The post content can be analyzed by assigning various words to groups and determining the relative frequency of words that fit within a group to other words. Russian troll accounts also have relatively high amounts of racially related words, as well as many words associated with cryptocurrencies.

2 METHODS

2.1 Hate Speech Level

Hate speech data was obtained by processing Reddit comments and submissions from February 2016 to February 2018 using Hadoop. Output data was processed using Java and Python, and visualized using Matplotlib. The level of hate speech on Reddit was determined using the total number of 'hate words' (defined below) grouped by:

(1) Subreddits by Submissions
(2) Subreddits by Comments
(3) Users by Submissions
(4) Users by Comments

Profile pages (e.g., 'u_spez') were excluded from analysis for groups 3 and 4, as they were only introduced on 20 March 2017 [12] and are arguably different in nature to traditional subreddits[14].

Words and phrases considered hateful from the HateBase dictionary were manually filtered to remove ambiguous and context-dependent hate speech, as mutually agreed upon by the authors. The remaining words and phrases ('hate words') were converted to lowercase, stemmed using the PorterStemmer from the Lucene library, and placed in a HashSet.

Reddit comments or submissions from February 2016 to February 2018 (inclusive) were partitioned for and according to each of the four groups, and tokenized using Lucene's Classic Tokenizer, converted to lowercase, and stemmed using the PorterStemmer. For Reddit submissions, the title and body text were both considered part of the submission. Tokens were classified as hateful if it appeared in the set of hate words.

Results were normalized to enable within-groups comparison by taking the total number of hate words from a member of that group and dividing by the total number of tokens from that same member ('hate level'). Normalization was necessary since authors and subreddits respectively posted and contained different amounts of submissions and comments of different lengths.

Further, to consider the popularity of hate speech on Reddit, the hate level for each member of the four groups was multiplied against the corresponding total karma score of that member to obtain the hate score for each member.

2.2 Russian Troll Behaviour

Using a set of Reddit confirmed Russian Troll accounts and all Reddit comments and submissions from January 2016 to February 2018 (unless otherwise specified), the analyses below were conducted, grouped by Russian Troll and Non-Russian Troll (unless otherwise specified), and comments and submissions ('posts'):

(1) Proportion of users that only commented, only made submissions, or did both;
(2) Aggregate posting frequency by users;
(3) Aggregate daily post frequency normalized by the total count of comments or submissions;
(4) Aggregate hourly post frequency normalized by the total count of comments or submissions;
(5) Hourly post frequency grouped by month from March 2016 to June 2016, from July 2016 to October 2016, and from December 2017 to February 2018, normalized by the total count of comments or submissions;
(6) Aggregate hate level from January 2016 to February 2018;
(7) Hate level of individual users from February 2016 to February 2018;
(8) A preliminary analysis was first conducted of the frequency each stemmed word (stemmed using PorterStemmer) was used from March 2016 to October 2016 and from November 2017 to February 2018, normalized by the monthly total count of words for the respective month. Based on the preliminary analysis and the Russian interference in the 2016 US Federal Election [15], certain stemmed words grouped into three categories were considered for their usage over time by Russian Trolls from March 2016 to October 2016 and from November 2017 to February 2018:
   (a) *Race*: 'cop", 'polic', 'policeman', 'black';
   (b) *US politics*: 'donald', 'trump', 'hillari', 'clinton', 'berni', 'sander', 'obama'; and
   (c) *Cryptocurrency*: 'token', 'coin', 'crypto', 'cryptocurr', 'bitcoin', 'blockchain'.

   The frequency of the word group for a given month was the sum of the frequency that each word in that group was used in that month, normalized by the total word count of comments or submissions from January 2016 to February 2018.
(9) Subreddits that both groups commented and submitted in, and the subreddits that only Russian trolls commented and submitted in;
(10) Aggregate most commented in and submitted to subreddits; and
(11) Based on analysis 8 and 10, and the Russian interference in the 2016 US Federal Election [15], certain subreddits grouped into four categories were considered for the frequency that Russian Trolls posted in these subreddits over time for the same time periods as in analysis 9:
   (a) *Race*: 'racism', 'blackpower', 'blackfellas', 'AfricanAmerican', 'copwatch', 'Bad_Cop_No_Donut', 'Police_v_Video', 'police';
   (b) *International Affairs*: 'education', 'humanrights', 'worldnews', 'news', 'economy';
   (c) *US politics*: 'politics', 'The_Donald', 'PoliticalDiscussion', 'POLITIC', 'uspolitics', 'HillaryForPrison', 'Republican', 'AmericanPolitics', 'PoliticalHumor'; and
   (d) *Cryptocurrencies*: 'BitcoinAll', 'ethtrader', 'CryptoCurrencies', 'BlockChain', 'btc', 'Bitcoin', 'CryptoCurrency'.

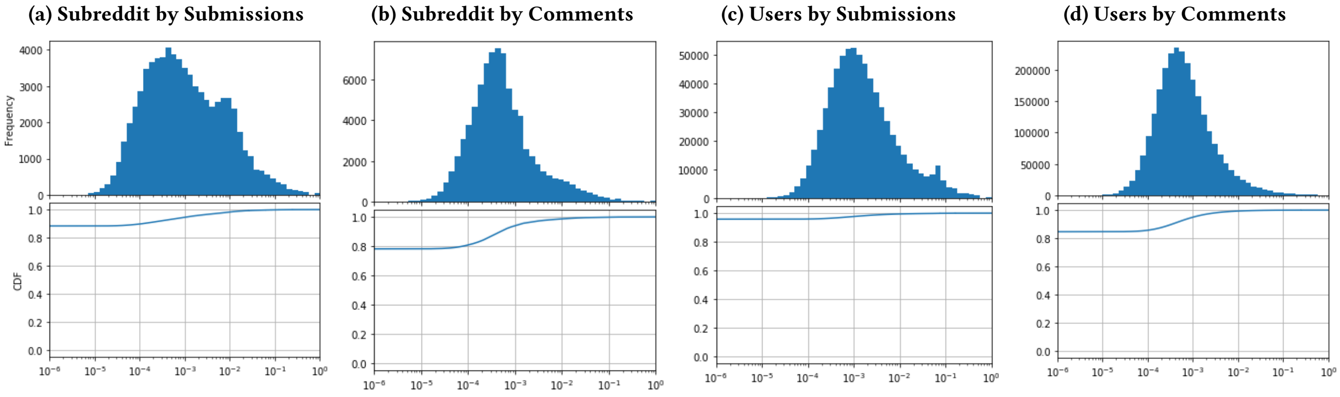
   The frequency of the subreddit group for a given month was the sum of the frequency that a post appeared in that subreddit for a given month, normalized by the total count of comments or submissions from January 2016 to February 2018.

The time periods chosen in analyses 5, 8, 9, 11 were based on periods of high posting activity from the results of analysis 3.

3 RESULTS

3.1 Hate Level

Based on Reddit comments and submissions from February 2016 to February 2018, 78.12% and 84.61% of subreddits and users by comments, 88.30% and 95.76% of subreddits and users by submissions,



(a) Subreddit by Submissions **(b) Subreddit by Comments** **(c) Users by Submissions** **(d) Users by Comments**

Figure 1: Hate Level

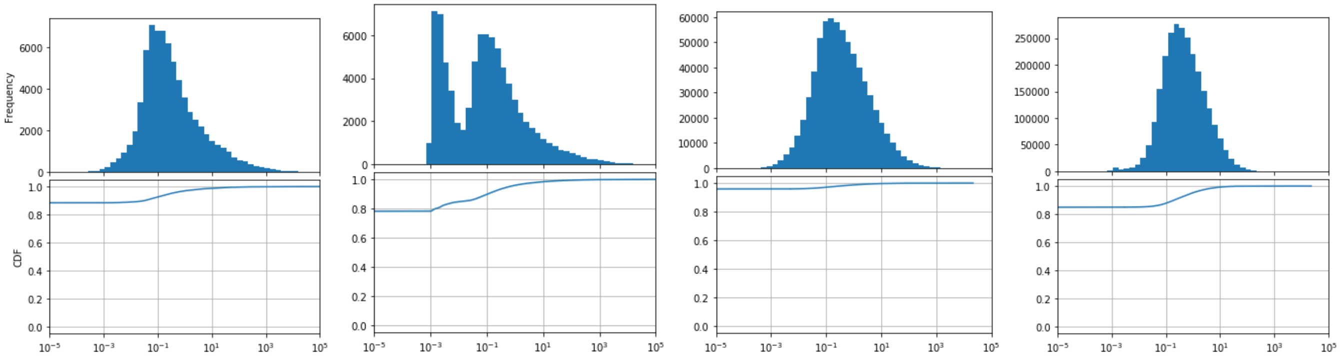


Figure 2: Hate Score

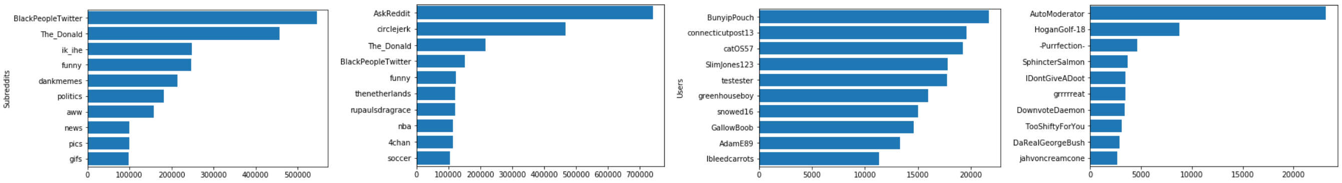
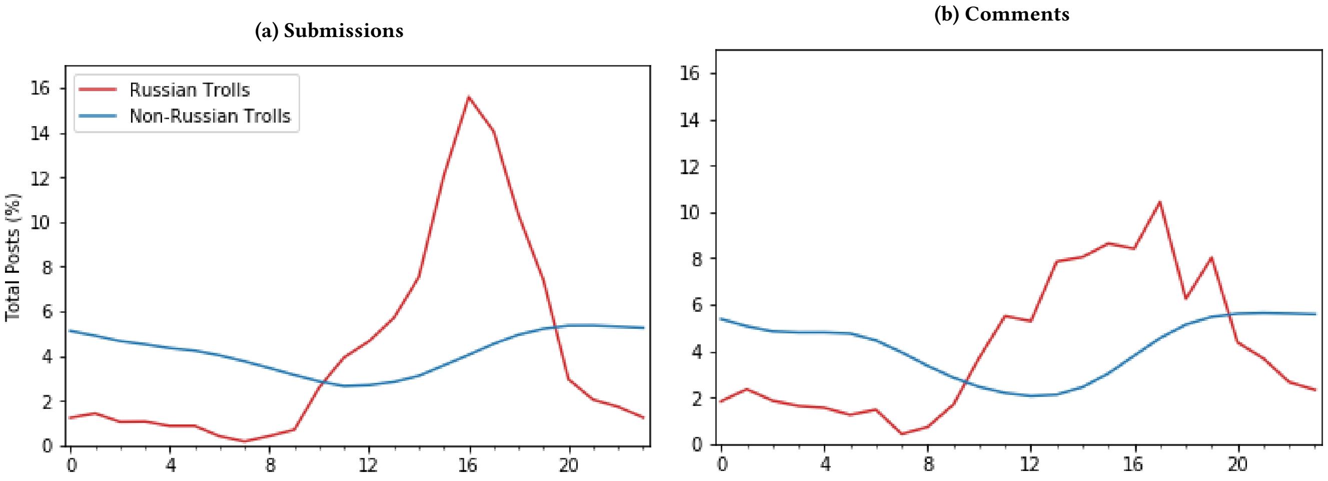


Figure 3: Top 10 Hate Score

respectively, have a hate level of 0.0. As can be seen in the CDFs in Figure 1, in all groups, the remaining amount of subreddits and users largely had a hate level between 10^{-4} and 10^{-2}. By excluding subreddits and users with hate levels of 0.0, approximately normal distributions can be seen in the histograms. These distributions are centred at 8.14×10^{-4}, 4.32×10^{-4}, 1.25×10^{-3}, and 5.5×10^{-4} for subreddits by submission, subreddits by comments, users by submissions and users by comments, respectively.

By mapping the hate level to the corresponding hate score, the subreddit distributions become more leptokurtic and positively skewed. The subreddit by comment distribution becomes bimodal. The user distributions became slightly more leptokurtic, but otherwise retain their original shape. For all groups, the range of non-zero values increased by a factor of approximately 10.

Measured by hate score, the leading subreddit by submissions (‘BlackPeopleTwitter’) has more than double the hate score of the third ranked subreddit (‘ik_ihe’). Similarly, the leading subreddit by comments (‘AskReddit’) has more than triple that of the third ranked subreddit (‘The_Donald’). In both cases, the second ranked subreddit (‘The_Donald’ and ‘circlejerk’) have similar hate scores to the first ranked subreddit. For users by comments, the leading user (‘AutoModerator’) has more than double that of the second


(a) Submissions

(b) Comments

Figure 4: Aggregate Post Hours (24h, Moscow GMT+3) from January 2016 to February 2018

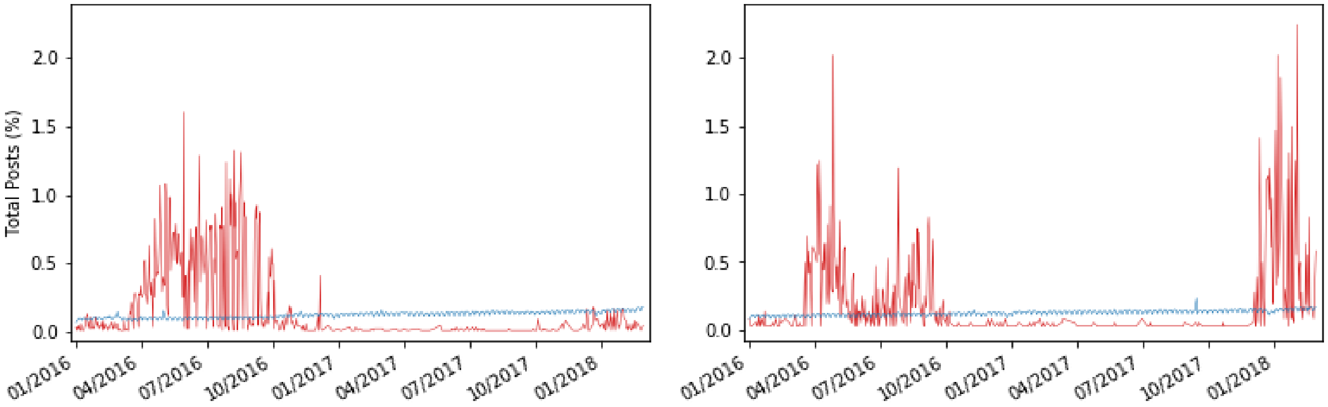
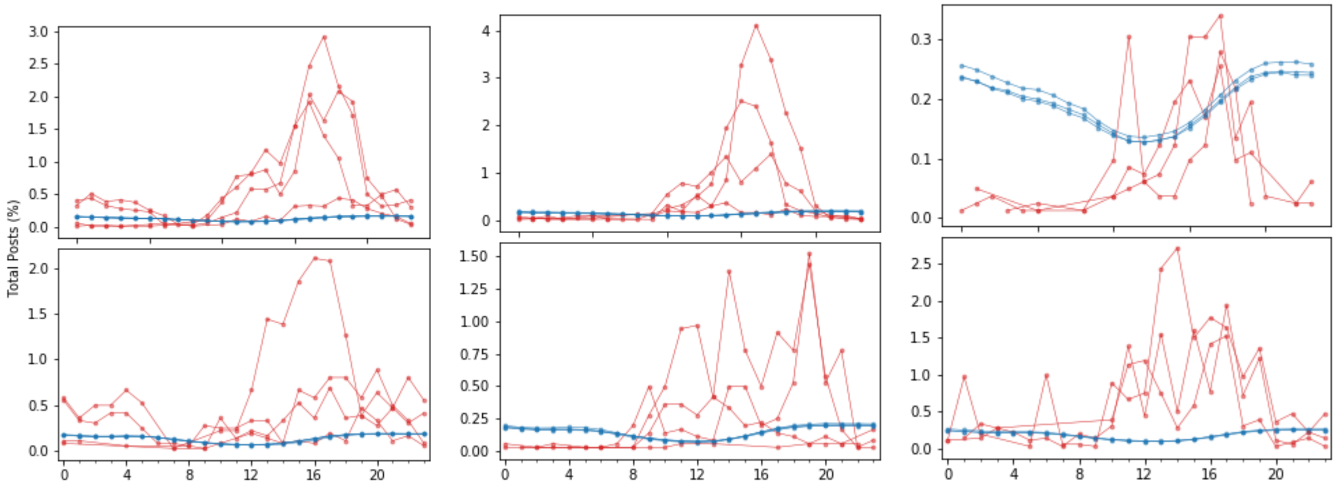


Figure 5: Aggregate Daily Posting Counts



(a) Mar 2016 - Jun 2016

(b) Jul 2016 - Oct 2016

(c) Dec 2017 - Feb 2018

Figure 6: Monthly Submission (top) and Comment (bottom) Hours (24h, Moscow GMT+3)

ranked user (‘HoganGolf-18’), and almost four times that of the third ranked user (‘-Purrfection’). However, for users by submissions, the leading user (‘BunyipPouch’) was only almost double that of the tenth ranked user (‘Ibleedcarrots’).

3.2 Russian Troll Behaviour

3.2.1 Analysis 1: Posting Preferences. The type of post (comment vs submission) is dependent on the type of user (Russian Troll vs Non-Russian Troll), $\chi^2(1) = 29.764$, $p < .01$. A total of 127 Russian Trolls posted on Reddit between January 2016 and February 2018, with 74 (58.27%) making submissions only, 4 (3.15%) making comments only and 49 (38.58%) making submissions and comments. By contrast, 22,836,649 Non-Russian Trolls posted during that period, with 5,636,441 (24.68%) making submissions only, 6,281,721 (27.51%) making comments only, and 10,918,487 making submissions and comments (47.81%).

3.2.2 Analysis 2: User Post Counts. Non-Russian Troll posting frequency by user appears to follow an exponential decay (see Figure 7), with a longer tail for comments. On the other hand, the Russian Trolls are comprised of a relatively large proportion of users that have made from 10^{-1} to 10^{-4} and 10^{-3} submissions and comments, respectively.

The two-sample Kolmogorov-Smirnov test was statistically significant for submissions ($D = .14$, $p < .01$) and comments ($D = .20$, $p < .05$). The user post counts were drawn from two different distributions.

3.2.3 Analysis 3: Aggregate Daily Post Counts. Most of the Russian Troll posting activity occurred between March 2016 and October 2016, and November 2017 to February 2018. Commenting activity in the latter period exceeded the former. Conversely, submitting activity in the latter period was significantly less than the former. Commenting activity in the latter period also appears to be greater than in any other period from January 2016 to February 2018.

Non-Russian Troll posting activity increased steadily over time for comments and submissions, with no clear differences between them.

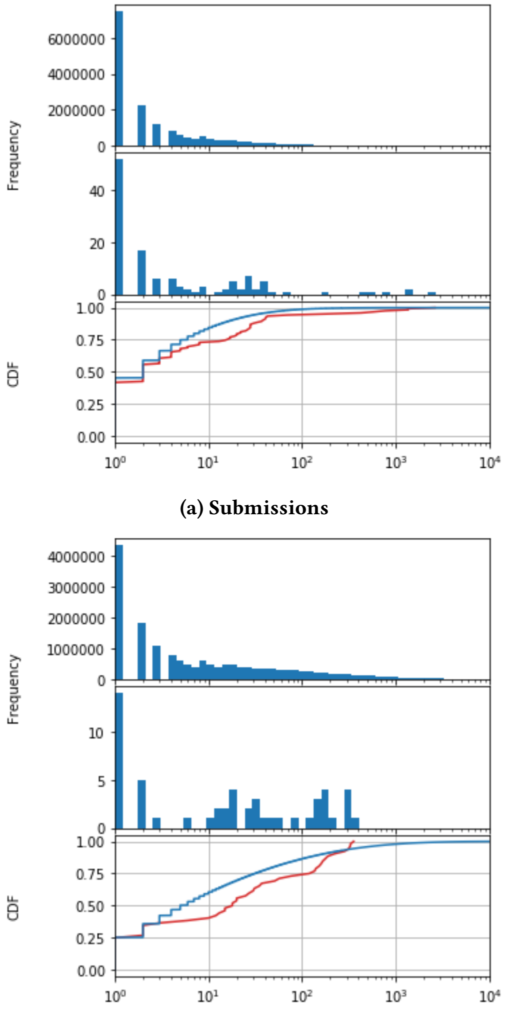
The two-sample Kolmogorov-Smirnov test (‘K-S test’) was statistically significant for submissions and ($D = .64$, $p < .01$) and comments ($D = .55$, $p < .01$). The aggregate daily post counts were drawn from two different distributions.

3.2.4 Analysis 4: Aggregate Hourly Post Counts. Russian Trolls primarily posted between 8AM and 8PM (GMT+3) from January 2016 to February 2018, with a peak of submission traffic (16%) from 4:00PM to 4:59PM and comment traffic (~10.5%) from 5:00PM to 5:59PM. Submission activity is minimal at all other periods, while low volume (2%) hourly comment activity continues.

Non-Russian Troll posting activity followed a shifted sine wave, with a peak (~5.5-6.0%) from 8:00PM to 8:59PM and a trough (~2.0-2.5%) from 12:00PM to 12:59PM.

The two-sample K-S test was statistically significant for submissions ($D = .58$, $p < .01$) and comments ($D = .38$, $p < .01$). The aggregate hourly post counts were drawn from two different distributions.

3.2.5 Analysis 5: Aggregate Hourly Post Frequency Grouped by Certain Months. For March to June 2016, Russian Troll posting patterns closely resembled that of the aggregate hourly posting patterns in Analysis 4. Most activity was concentrated between 8AM and 8PM, with a steady stream of activity continuing over night until 5AM. From July to October 2016, posting activity was mostly limited to 8AM to 8PM for submissions and 10PM for comments. For December 2017 to February 2018, a similar pattern as March to June 2016 was observed, though submission rates during these months were considerably less. In all cases, minimal activity was observed between 7:00AM and 7:59AM.



(a) Submissions

(b) Comments

Figure 7: Russian Troll (middle) and Non-Russian Troll (top) Post Counts

Some unusual ‘jagged’ activity can be observed for one month in particular from December 2017 to February 2018 from 10AM to 6PM.

3.2.6 Analysis 6: Aggregate Hate Level. For comments, Russian Trolls had a higher aggregate hate level of 0.1632 (4dp), compared to 0.0381 (4dp) for Non-Russian Trolls. Similarly, Russian Trolls reported a higher hate level of 0.0242 (4dp), compared to 0.0263

(4dp). However, as these values were calculated directly by counting all hate words and immediately dividing by the total word count, statistical significance cannot be determined without further computation.

3.2.7 Analysis 7: Hate Level by Individual User. The two-sample K-S test conducted on the hate level distribution of Russian and Non-Russian Trolls was not statistically significant for submissions ($D = .02$, $p > .05$) and comments ($D = .10$, $p > .05$). The hate level of Russian vs Non-Russian trolls were not drawn from different distributions.

3.2.8 Analysis 8: Word Group Frequency over Time. From March 2016 to October 2016, with the exception of May for submissions and April for comments, Russian Trolls primarily posted with an amount of racial words that matched or exceeded that of US politics related words. In May and April, this reversed. No posts contained cryptocurrency related words during this time.

From November 2017 to February 2018, word group dominance entirely reversed. Cryptocurrency related words were the only words used with the exception of December for comments, where about 0.03% of US politics words appeared.

3.2.9 Analysis 9: General Subreddit Activity. Russian Trolls submitted and commented in 715 and 456 subreddits (including user profiles), respectively. There were seven subreddits or user profiles that Non-Russian Trolls did not make submissions in from January 2017 to February 2018: ‘BarnieSandlers’, ‘u_smoliangirl’, ‘MovieDownlodOnline’, ‘CaptainOnBoard’, ‘animalsbeingtrolled’, ‘u_failkate’, and ‘democracysim’. There were also six subreddits or user profiles which only Russian Trolls commented in: ‘dpos’, ‘bitcoinregrets’, ‘u_chamarahs’, ‘u_SovaCrypto’, ‘Cryptfunder’,[1] and ‘u_toroidalfield’.

3.2.10 Analysis 10: Aggregate Most Active Subreddits. Most Non-Russian Troll posts were to ‘AskReddit’. For submissions, that subreddit lead the next four ranked subreddits by almost double, and the five after that by a factor of four. For comments, the lead was even more dramatic, leading the second rank by almost triple, and the remainder by a factor of four to five. Interestingly, the second most submitted to and third most commented to subreddit is ‘The_Donald’.

The most submitted to subreddits for the Russian Trolls were all different to the top subreddits that Non-Russian Troll post to. ‘uncen’ and ‘Bad_Cop_No_Donut’ led by approximately double to triple the next eight subreddits. For comments, ‘CryptoCurrency’ and ‘AskReddit’ led by approximately double to triple for the next four ranks and approximately a factor of six for the remaining four ranks. Of these subreddits, only half appear in the subreddits that Non-Russian Trolls post to: ‘AskReddit’, ‘politics’, ‘The_Donald’, ‘pics’ and ‘news’.

3.2.11 Analysis 11: Subreddit Group Activity over Time. As apparent from Figure 10, from March to October 2016, for submissions, a clear hierarchy of subreddit posting frequency emerged, in descending order: race, US politics, international affairs, cryptocurrencies. For comments, the same hierarchy was maintained, with the exception of April and from August to October. In the former, US politics eclipsed race. In the latter, posting frequency to these three subreddit groups converged. Interestingly, posts to US politics subreddits dipped in June, and for comments, almost reached 0%. Finally, similar to the frequency of word group usage in analysis 8, no posts were made to cryptocurrency subreddits during this time.

Similar to analysis 8, for November 2017 to February 2018, the hierarchy dramatically switched. Almost all posts were to cryptocurrency related subreddits. There were almost no posts to any other subreddit.

4 DISCUSSION

4.1 Hate Level

If the most liberal interpretation of a ‘hate problem’ is taken, Reddit as a whole can arguably be considered to have a hate problem, given that about 10% and 20% of subreddits by submissions and comments have a non-zero hate level, respectively. However, these hateful subreddits largely only contain posts from a small subset of hateful users, since only about 5% and 15% of users have non-zero hate levels, respectively. Regardless, of the subreddits and users with non-zero hate levels, the hate levels are extremely low, with most being less than 1%,

Assuming that karma scores can be used as a metric of exposure and approval of a post, then the hate score can be used as a metric of exposure and approval of hate on Reddit by the Reddit community. This reveals that subreddits with a low level of hateful comments are strongly approved (see Figure 2), but users and subreddits with high levels of hateful posts are not.

The hate scores of particular subreddits and users must be given special consideration. For example, the BlackPeopleTwitter subreddit is likely not as hateful as it appears. The various forms of the word ‘nigger’ are acceptably used within the black community without hateful connotation. Consequently, the hate score of subreddits largely comprised of and dedicated towards the black community is artificially inflated.

Further, the user, Automoderator, is a bot used by subreddit moderators as a moderation tool, especially in non-hateful communities [11]. The high hate score is likely the result of quoting and removing hateful content. Even if Reddit has hateful users, quarantine policies are enforced.

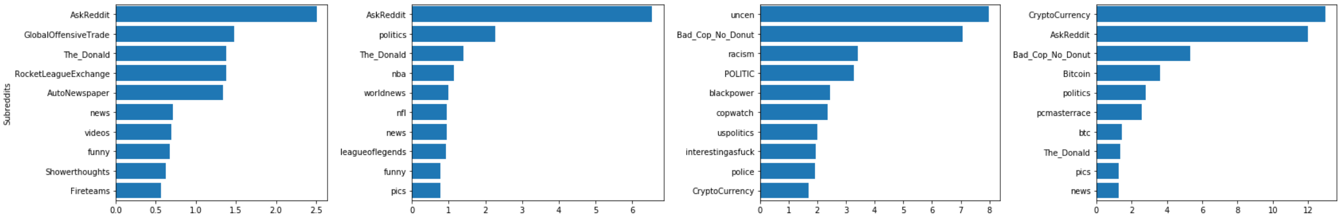
Other users and subreddits with apparently high hate scores can be countervailed by considering that hate speech is highly context dependent. If hate speech is used in a joke or a quote, it may no longer be considered hate speech. If used in a comedic or well-timed fashion, such a post would garner widespread approval, karma and a high hate score, without being genuinely hateful.

Certainly, there exist subreddits (e.g., The_Donald) and users (e.g., HoganGolf-18, BunyipPouch) with high hate scores, for which hate may genuinely be a problem, but the part is not reflective of the whole. If Reddit as a whole does not have a hate problem while taking the most liberal interpretation of a ‘hate problem’, Reddit does not have a hate problem.

4.2 Russian Trolls

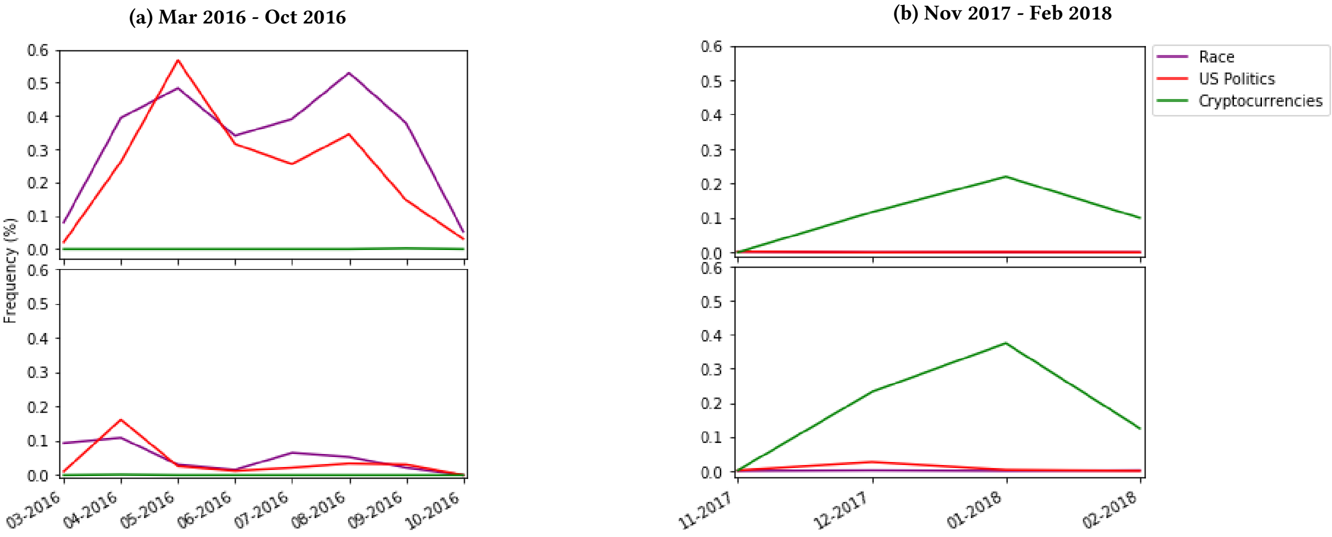
4.2.1 Aggregate Daily Post Counts. The bulk of Russian Troll submissions were made from March to October 2016 - a critical period in Donald Trump’s election. There was, however, a sudden peak

[1] See Appendix for related suspicious users, subreddits and URLs.



(a) Non-Russian Troll Submissions **(b) Non-Russian Troll Comments** **(c) Russian Troll Submissions** **(d) Russian Troll Comments**

Figure 8: Top 10 Most Posted to Subreddits (%)


(a) Mar 2016 - Oct 2016

(b) Nov 2017 - Feb 2018

Figure 9: Frequency of Word Groups in Submissions (top) and Comments (bottom)

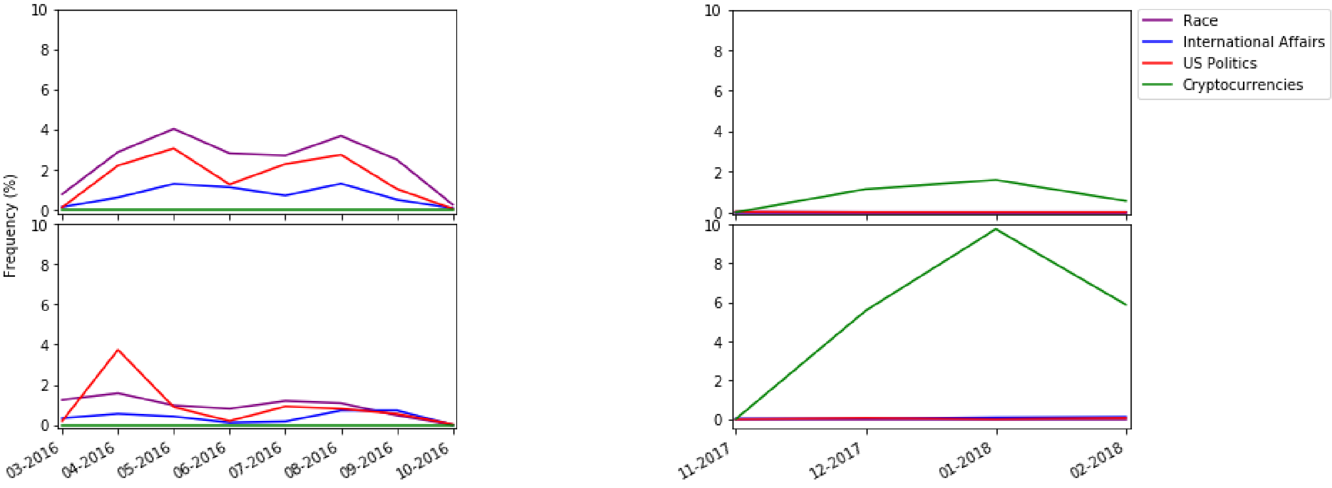


Figure 10: Frequency of Submissions (top) and Comments (bottom) in Subreddit Groups

on 6 December 2016. Given the Russian Troll's interest in African-American race issues and American police, this may have been related to the mistrial of Michael Slager (white police officer) for the shooting death of Walter Scott (African-American) [6].

Most commenting activity also occurred between March to October 2016. In both cases, there was minimal activity until November 2018. Even then, submission activity remained subdued compared to the initial period, and particularly in comparison to commenting activity. It appears that the posting patterns and perhaps objectives of the Russian Trolls had changed. Certainly, low scoring submissions do not reach as wide an audience as comments in a high scoring thread. Further, posting high quality comments with strong propaganda value is arguably lower effort and thus more efficient than posting high quality submissions.

4.2.2 Posting Hours. Superficially, it appears that Russian Trolls work long hours from 8AM to 8PM or even 10PM, and, on some occasions, even through the night until 5AM (see Figure 6). However, given the significant peak in submission activity from 4:00PM to 4:59PM, it is more likely that Russian Trolls are based out of multiple locations. Russia spans eleven time zones each with significantly different population sizes [8]. Time zones with populations above 2 million span from GMT+3 to GMT+10.

Their posting patterns consistently varied with the period in which they were posting. For July to October 2016, there was no observed activity from 12AM to 5AM, and minimal activity until 8AM. During these months, submissions ceased at 8PM. However, for March to June 2016 and December 2017 to February 2018, posting activity continued throughout the night until 5AM. This suggests coordinated activity either from multiple independent groups active during different periods or a single coordinated effort,

It should be noted that the 2016 US Primaries were held from February 1 to June 7 2016 [5]. For 2016, the Washington DC time zone was GMT-4 (DST), starting from March 13 and ending on November 6 [1]. Given this 7 hour difference, their nightly activity from 12AM to 5AM corresponds to 5PM to 10PM (GMT-4) and 2PM to 7PM (GMT-7). This is the time that Non-Russian Troll users (in particular, Americans, but not most Europeans, being 11PM to 4AM (GMT-2) [7]) are most active (see Figure 4) and are prime time US television hours.

During hours that Non-Russian Troll Users are most active (i.e., 3AM to 1PM (GMT-4), 12AM to 10AM (GMT-7)) Americans are either sleeping or starting their work day. That is, if Russian Troll content were injected onto Reddit, Americans would be waking up to it.

4.2.3 Post Counts and Subreddit Activity. Russian Trolls, as individual users, post more comments and especially more submissions than Non-Russian Trolls. This activity is especially peculiar considering that some users make continued consistent submissions to their own moderated subreddits - available to all, but viewed and posted to by virtually none. A striking example of this is u_shomyo, the moderator of the 'uncen' subreddit [4]. Approximately 8% of all Russian Troll submissions were to uncen (see Figure 8).

Russian Trolls largely fixated on race (or police, though these two facets of society are extremely intertwined in the US nowadays), US politics, international affairs (strongly related with US politics), and cryptocurrency (perhaps, money laundering). This is further discussed in the next section.

However, they also posted on more mainstream subreddits such as AskReddit, pcmasterrace, and interestingasfuck. Perhaps that is because they too cannot escape the allure of Reddit, or more likely, their accounts do not lose credibility if a random redditor investigates their post history. It is possible that different accounts have different objectives or different target audiences, reflecting different subreddit activity patterns.

4.2.4 Word and Subreddit Groups. Although Steve Huffman (u_spez), CEO and Co-Founder of Reddit, claims that the Reddit investigation into the list of Russian Trolls 'did not find any election-related advertisements' [13], their language use and subreddit activity suggests otherwise (see Figure 10).

Although classified separately in this analysis, international affairs (i.e., education, human rights, economics, world news, US news) and racial issues (especially African-American racial issues and police brutality towards African-Americans) could be comfortably categorized as part of US politics, especially given Reddit's primarily American user base [3]. Indeed, the Russian Trolls made significant reference to US politics and the aforementioned politically charged affairs from March to September 2016, but essentially simultaneously ceased all political discourse by October 2016.

Following an extended hiatus, they started posting regularly again in November 2018, but the focus shifted almost entirely to cryptocurrencies. Although they did not make as many submissions during this period, any submissions they did make were cryptocurrency related. There is further consideration of this issue in the Appendix.

5 LIMITATIONS

5.1 Hate Level

The method used to determine the level of hate on Reddit has several limitations. First, since comments and submissions were tokenized and tested for membership in a set of hate words and phrases, hate phrases were not considered and would have increased the measured level of hate speech on Reddit. However, to do so, would require N-grams of the comment and submission text, which would have increased processing time. Nonetheless, these phrases were not completely ignored; some hateful phrases were comprised of hateful and non-hateful words (e.g., "cave nigger").

Second, presumably, there were a significant number of false positives. Hate speech is context dependent. Quotes and jokes using hate speech is arguably not hate speech. This likely would have more than counterbalanced the error introduced by the first limitation.

Third, the manual filtration process used to remove ambiguous or context-dependent hate words and phrases was dependent on the authors notions of American and Australian hate and culture. This may have resulted in the inappropriate inclusion or exclusion of certain words and phrases.

Lastly, with no established objective definition of a 'hate problem', whether Reddit has a hate problem is largely a subjective analysis.

5.2 Russian Trolls

All analyses assume that the list of confirmed Reddit Russian Trolls is accurate. Although their post content and subreddit activity is atypical, and their posting hours are largely consistent with Moscow working hours, it is unclear whether they are truly Russian Trolls,

6 FUTURE WORK

It is possible that the Russian Troll accounts can be divided into at least two groups each with different objectives: US politics and cryptocurrencies. It is possible that the first group largely stopped posting after the 2016 US elections and the second group started posting around November 2017 until they were banned in February 2018. This would explain the reduction in submissions and apparent shift in focus to cryptocurrencies. Future work should identify whether this is the case.

7 CONCLUSION

Only a relatively small proportion of Reddit users and subreddits may be considered hateful. As a whole, Reddit cannot be considered as having a 'hate' problem, but perhaps, a Russian Troll problem. Given the uniformity of their language use, subreddit activity, hourly posting patterns, and daily post counts, it seems that their efforts, as one group or several, were coordinated towards American political and cryptocurrency objectives.

REFERENCES

[1] Clock Changes in Washington DC, USA in 2016. https://www.timeanddate.com/time/change/usa/washington-dc?year=2016.

[2] How Many People Use Reddit? - User and Visitor Statistics. https://techboomers.com/t/how-many-people-use-reddit.

[3] Reddit.com Desktop Traffic Share 2018 | Statistic. https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/.

[4] Shomyo (u/shomyo). https://www.reddit.com/user/shomyo.

[5] 2016 Primary Dates. http://www.ncsl.org/research/elections-and-campaigns/2016-state-primary-dates.aspx, Feb 2016.

[6] 2016 in the United States. https://en.wikipedia.org/wiki/2016_in_the_United_States, Oct 2018.

[7] Time in Europe. https://en.wikipedia.org/wiki/Time_in_Europe, Oct 2018.

[8] Time in Russia. https://en.wikipedia.org/wiki/Time_in_Russia, Oct 2018.

[9] Engel, P. Reddit Just Banned The Subreddit Where People Were Posting The Celebrity Nude Images. https://www.businessinsider.com/the-fappening-has-been-banned-from-reddit-2014-9, Sep 2014.

[10] Ghosh, S. Reddit's permissive attitude to racism is poisoning the internet. https://www.businessinsider.com/reddit-ceo-said-racism-is-okay-2018-4, Apr 2018.

[11] HideHideHidden. AutoModerator Wiki. https://www.reddit.com/wiki/automoderator/full-documentation.

[12] HideHideHidden. Reddit News. https://www.reddit.com/r/modnews/comments/60i60u/tomorrow_well_be_launching_a_new_posttoprofile/.

[13] HideHideHidden. Reddit's 2017 Transparency Report and Suspect Account Findings. https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/.

[14] Statt, N. Reddit's New Profile Pages Could Fundamentally Transform the Site. https://www.theverge.com/2017/3/21/15009388/reddit-profile-pages-social-network-facebook-twitter, Mar 2017.

[15] Yourish, K., and Griggs, T. 8 U.S. Intelligence Groups Blame Russia for Meddling, but Trump Keeps Clouding the Picture. https://www.nytimes.com/interactive/2018/07/16/us/elections/russian-interference-statements-comments.html, Jul 2018.

8 APPENDIX

One of the subreddits that only the Russian Trolls posted to from January 2016 to February 2018 was Cryptfunder: https://www.reddit.com/r/Cryptfunder/

This is the thread that one of the confirmed Russian Trolls posted in: https://www.reddit.com/r/Cryptfunder/comments/803ejh/putin_just_endorsed_blockchain_and/

They appear to have a social media presence:

(1) https://www.instagram.com/p/BkLNbC_Ah5d/
(2) https://twitter.com/cryptfunder/followers?lang=en

They apparently recruited this guy: https://medium.com/@cryptfunder/ron-ribitzky-joins-cryptfunder-as-the-technology-healthcare-segment-strategy-a

And this guy: https://www.reddit.com/r/Cryptfunder/comments/8kqmuk/nathan_christian_joins_cryptfunder_as_business/

He apparently has a twitter: https://twitter.com/ronribitzkymd?lang=en

Or two: https://twitter.com/SPHAnalyticsRon

And a website: https://rdribitzky.com/contact/

And is connected with this guy: https://blockchain-connectors.com/wp-content/uploads/2018/09/rd_ribitzky_intro.pdf

That guy has this page. He's Russian?: https://innovationstudio.ru/profile/AlexKosik/

Maybe he's real. Maybe the Russian Trolls are just using purporting to recruit people and have blockchain significant employees. Check the list at the bottom. Maybe they're just using people's online profiles: https://icobench.com/ico/cryptfunder

This is the twitter of the CEO. Note when his posts begin and end, and its content: https://twitter.com/KevinSarisky?lang=en

This guy is a moderator of cryptfunder. He likes posting memes to his own subreddit: https://www.reddit.com/user/italid

This guy also likes to post memes to his own subreddits. I think he posted on cryptfunder: https://www.reddit.com/user/justice15x?sort=new

Krypital is a moderator of Cryptfunder. He also moderates another cryptocurrency related subreddit. It seems that people have been scammed: https://www.reddit.com/r/arcblock/comments/813eg6/arcblock_rerfund/

Baanx is another cryptocurrency subreddit: https://www.reddit.com/r/BaanxICOLondon/. I can't remember the connection between these subreddits and Baanx, but I believe he promoted it in one of the other cryptocurrency subreddits, or someone else posted in Baanx.

It seems that Baanx is also scammy:

(1) https://www.reddit.com/r/BaanxICOLondon/comments/9mooo6/why_was_i_banned_from_baanx_telegram_just_for/
(2) https://www.reddit.com/r/CryptoMarkets/comments/9m4ikv/baanx_admin_just_banned_me_for_requesting_a/
(3) https://www.reddit.com/r/binance/comments/9ll8ov/is_this_breaking_binance_listing_rules/

If you follow the trail starting from cryptfunder, all you find are memes and cryptorubbish (aka memes)