When Power Goes Wild Online: How Did a Voluntary Moderator's Abuse of Power Affect an Online Community?

Yukun Yang
University of North Carolina, Chapel Hill, Chapel Hill, USA
yukun@live.unc.edu

ABSTRACT

Online moderation is apropos to community curation as a way to fight against malicious behaviors bristling on User-Generated Content (UGC)-based Social Network Sites (SNS). Given the current research gap on voluntary moderation from the perspectives of power misuse, we investigate how power abuse by a moderator would affect the community dynamics in terms of participation indicators, linguistic characteristics, and network structure in a computational fashion. An event on Reddit is chosen for a case study. Using interrupted time-series analysis and social network analysis, we find moderation fueled short-term feuds and brought potential prolonged destruction to the community. People's linguistic patterns remained stable while the liberation from "tyranny" brought the community back to life and the power competition entailed negative repulsion. We also find an "Exodus" phenomenon as netizens voted with their feet and migrated to a mirror community when facing severe moderation. This preliminary research expands the connotation of moderation by addressing more forms of power abuse. We also refer to social movement and community choice theories in relevant fields and provide the insights of online moderation from interdisciplinary perspectives.

KEYWORDS

Online Moderation; Power Abuse; Reddit; Social Media; Interrupted Time-Series Analysis; Social Network Analysis

ASIS&T THESAURUS

Virtual Communities; Governance; Data Analysis

INTRODUCTION

Internet-based information and communication technologies trigger the emergence of digital spaces with the characteristics of public spheres. While democratic and decentralized virtual spaces promote the proliferation of user-generated content and stoke the participation of users, they also witness many malicious behaviors such as spam, vandals, trolls, and hate speech (Grimmelmann, 2015). As these problems are more ubiquitous across the platforms, online moderation, as a means to address these problems, becomes increasingly indispensable for community curation.

Approaches of online moderation could roughly fall into two categories: one is company-driven, which means the platform retains the privileges to regulate the forms and the moderation policy is designed by the company. This process usually involves outsourcing moderation tasks to paid workers or firms screening out the content (Roberts, 2016). This approach has drawn much attention as researchers try to understand this phenomenon from multiple perspectives (Klonick, 2017). Another type of online moderation is voluntary moderation. It is usually utilized by SNS that position themselves as content providers relying on UGC, such as Twitch, Wikipedia and Reddit (Lo, 2018). User-moderation is often characterized as bottom-up and user-run (Seering, Wang, Yoon, & Kaufman, 2019). Also, users gain the freedom to set and enforce the policy of the community (Massanari, 2017). Most research views voluntary moderation or moderators in a neutral lens. Main topics include how voluntary moderation would contribute to the community (Foerderer & Heinzl, 2018; Grimmelmann, 2015), while the negative outcome of moderation has not been exhaustively explored. Also, as most research focuses on content moderation, namely screening, modifying, removing comments, or banning users (Matias, 2016), little ink has been spilled on other types of moderation such as de-mod, the removal of one's moderator status.

Building on the literature review, we find "power" a good substitute for moderation and could serve as an umbrella term for all the forceful actions made by an online moderator. As a "basic force of social relationships" (Fiske, 1993), power accentuates the social and interactive nature of the online community. Also, it highlights the essence of moderation and thus leads the discussion of moderation into an interdisciplinary vision. Moderation boils down to power usage, the ability to "modify others' states by providing or withholding resources or administering punishments" (Keltner, Gruenfeld, & Anderson, 2003), since moderators can coercively control a user's rights in the virtual world. Under this scope, "power abuse" refers to the destructions made by moderators when they misbehave with their privileges. Based on the availability of data, we strive to answer these questions:

1. How does power abuse by user moderators affect community participation?
2. How does power abuse by user moderators affect a community's linguistic characteristics?

82nd Annual Meeting of the Association for Information Science & Technology | Melbourne, Australia | 19–23 October, 2019

DOI: 10.1002/pra2.00055

3. How does power abuse by user moderators affect a community's network structure?

In this preliminary research, we make contributions in the following ways. First, we contribute to research regarding voluntary moderation in a computational fashion using time-series analysis. Second, our research distinguishes from other voluntary moderation research for it focuses on the destructive side of moderator engagement. Finally, we enrich the connotation of moderation by adding other types of power usage neglected by previous research. On top of that, we link to relevant fields like sociology and provide new insights to study online moderation and online communities.

BACKGROUND & DATA

An infamous moderation event on Reddit is chosen for a case study. Subreddit r/xkcd was a placid comic forum until user u/soccer gained control of it in June 2011. After being the head moderator for 1.5 years, he began revising r/xkcd's sidebar by adding links to contentious subreddits like r/mensrights. Concurrently, he began repressing the voices of the opponents. In late 2013, user u/Wyboth was granted the permission to co-mod r/xkcd. However, after removing the inappropriate sidebar links, he was de-moded by u/soccer in January of 2014. This purge action attracted widespread reproof. The drama ended in August when the petition to remove u/soccer was agreed by the Reddit official. This scandal has been recognized as one of the worst moderation events that even made news headlines at the time. This historical event renders us opportunities to observe the wax and wane of the power usage in different phases. Unlike other intervention experiments (Matias & Mou, 2018), this event occurred without the researchers' intervention. Also, u/soccer is a "subreddit squatter", a person who maintains control of subreddits without active participation. It gives us new perspectives on moderator activities because of his non-typical nature. The de-mod action also set itself apart from other types of moderations as it is a form of power abuse toward colleagues, not the users.

Submissions and comments in r/xkcd are scraped using Google's Big Query (bigquery.cloud.google.com/dataset/fh-bigquery:reddit) and pushshift's API (pushshift.io/). Additionally, because of the online migrant pattern observed in the event, data of r/xkcdcomic is also obtained. In total, our dataset contains 5,259 submissions and 109,937 comments of r/xkcd, 387 submissions and 9,156 comments of r/xkcdcomic.

METHODOLOGY

Time-series analysis with intervention and outlier detection is used to answer RQ1 and RQ2. We use this method to avoid the over-simplification and auto-correlation issues which may produce problematic results. After filtering out variables with many missing values, we plot the time-series data, conduct the stationary test, and tune the ARIMA model by observing their ACF and PACF plots and comparing each model's BIC value. "tsoutliers" package in R is used to tune ARIMA model and detect novelties. Three common kinds of abnormal disturbances are examined in this research: additive outlier (AO), transitory change (TC), and level shift (LS) (Chen & Liu, 1993). An AO represents an isolated abnormal awry off the trend while a TC signifies a more influential disturbance that takes a few periods to disappear. An LS stands for an abrupt change in the mean level. After that, we also conduct model diagnostics using the Ljung-Box test. With iterative rounds of trials, we select the attributes that respond to RQs and are well-fitted for the model. For participation indicators, counts of comments, counts of submissions, the average score of the submissions, unique commenters, and unique submission posters are selected. For linguistic features, we merge all the texts from submissions and comments on a monthly basis, then calculate the number of unique words, number of sentences, average sentence length, and lexical diversity (MLTD).

Social Network Analysis is used for RQ3. We pick out 4 timings for network analysis: T1:1 month before the sidebar event, T2: the middle point between sidebar event and de-mod event, T3: the middle point between the de-mod event and dethrone event, and T4: 1 month after the dethroning.

RESULT

Participation Indicators

Figure 1 shows the effects of the outliers detected as interruptions in the time series. All three kinds of disturbance are captured in our dataset. While the inauguration of u/soccer did not translate to much influence, the forceful political indoctrination in mid 2013 stirred up controversies as we observe two AOs right after the sidebar event. Eliminationist action produces a similar result, since both submissions and comments have AOs around that time, and this effect also spread to the number of unique submission posters. However, the number of unique posters then plummet afterward, with a higher negative effect coefficient. The last momentous event, the dethroning of the moderator, proved to be the most influential one. All five metrics pertaining to participation have greatly increased, and most of them are LS and TC, the more intensive interruption.

In this section, we discern the transitory dispute induced by the power misuse, including the inappropriate use of the public sidebar and the removal of the other moderator; the latter one causes the decrease of thread posters. The emancipation from the authority intensively reinvigorates the whole community.

Lexical Features

Average sentence length and lexical diversity show no difference before and after the key event because the only AO is not detected around the time power abuse actions were executed. However, the number of unique words and the number of sentences do not keep stable over time. The sidebar event entailed a temporary increase on the

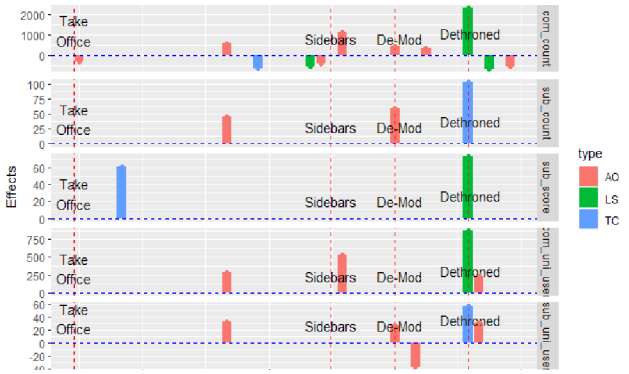

Figure 1. Outliers and their effects in community participation metrics from June 2011 to December 2014.

vocabulary of the corpus and number of sentences, which could be an indicator of activity intensity on the forum. Barring that, the dethronement of authority induced the highest impact on the user's linguistic pattern, as the abnormal effects all peaked right after August 2014. The influence is abrupt and intense and showed in the form of an AO in consecutive time sessions. In this part, we find that bad moderation only affects the vocabulary and the number of sentences. We also find the violent damaging effect of the power seizing event, since the only LS happened after that and lasted for 6 months.

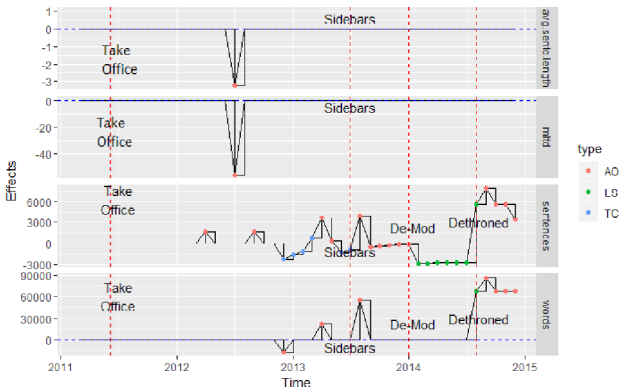

Figure 2. Outliers and their effects in lexical characteristics from June 2011 to December 2014.

Network Structure

We observe an interesting phenomenon that when people were undertaking severe moderation, they created another mirror subreddit. Based on this observation, r/xkcdcomic is also included for analysis in this part for comparison.

Figure 3 shows the social network of r/xkcd and r/xkcdcomic over time. We scale the size of the nodes based on their degrees and the width of edges based on their weights. The appearance of nodes with higher degrees indicated that the sidebar event, as a secretive action, slightly influenced the network. The brazen de-mod action aroused the controversies as the network was more stretched and more nodes were

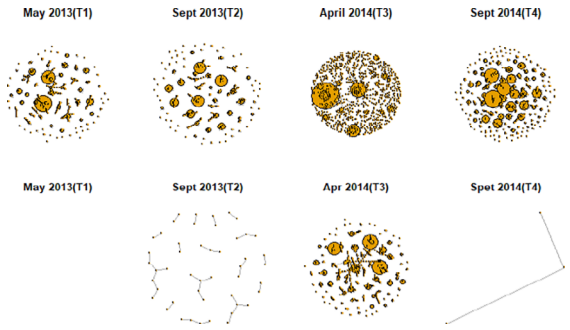

Figure 3. Network of r/xkcd(upper) and r/xkcdcomic(lower) in four chosen timings.

connected. The prevailing large nodes showed that dethroning produced many influential agents. We also witness the development of r/xkcdcomic from cradle to grave. As nonexistence in May 2013, it began to flourish after the sidebar event and then evolved into a vigorous community after the de-mod event. It finally diminished after the dethroning event retaining with negligible users.

To better understand the evolution of networks, we calculate the network statistics of these networks as Table 1 shown. Nodes and edges numbers in r/xkcd kept dwindling and touched the floor after the de-mod event, while those of r/xkcdcomic peaked at that time. The dethroning produced a record high interaction between users within the r/xkcd, while the r/xkcdcomic barely being active at the same time. Density, the indicator for the network cohesiveness, is not informative as the numbers barely nudge. Centralization is a supplement of density, explaining the level of connectedness centered on a node (Chung, Piraveenan, Levula, & Uddin, 2013). Our dataset shows that the de-mod event saw a sudden increment in the r/xkcd, while the highest value of that in r/xkcdcomic occurred on the initial phase. Assortativity measures the homophily inside the group based on degree. We witness its steadiness in both subreddits. Modularity is also calculated after community detection. Its value stays stationary and high.

	r/xkcd				r/xkcdcomic			
	T1	T2	T3	T4	T1	T2	T3	T4
Nodes	1550	1390	925	2766	/	48	2090	3
Edges	1477	1304	864	2605	/	32	2010	2
Density	0.012	0.014	0.002	0	/	0.028	0.001	0.670
Centralization	0.021	0.016	0.067	0.013	/	0.037	0.014	1
Assortativity	0.099	0.155	0.135	0.150	/	0.077	0.110	1
Modularity	0.968	0.966	0.950	0.982	/	0.904	0.973	0

Table 1. Network statistics.

DISCUSSIONS & CONCLUSIONS

Based on the data analysis, we show that different forms of online power abuse indeed exert influence on the online community. Several implications stand out. In the participation indicators, we find that heated participation was evoked by power misuse that encompasses both covert misuses of the public sidebar and the overt power grab. The transitory dispute induced by the two kinds of power misuse aligns with the social movement theory proposed by Blumber, since the first stage "social ferment" also attracts great attention of "agitators" (Blumer, 1995). However, the number of unique posters plummeted after the event, with a higher effect coefficient, indicating a loss of the participants. This decrease potentially shows the hallmark of the fourth stage "the decline" as people choose a different community. Additionally, the uprooting of the absolute ruler has the most profound influence, as outliers are identified as LS and TC, not one-time additions, but longitude impacts. The dethroning event resembles the truth that a movement, in reality, could go back and forth. The huge volume of the participation, along with the chronological effects shares some similarity with the fourth stage that falls into the slot of the "establishment with the mainstream" (Macionis, 2000).

Although previous research indicates that lexical diversity is influenced by persuasion in an online forum (Tan, Niculae, Danescu-Niculescu-Mizil, & Lee, 2016), which is similar to the compelling nature of excessive moderation, it is not salient that lexical diversity has changed abnormally around the key events. On the other hand, the numbers of words and sentences are more sensitive and flow with the changes in almost every event. In these outliers, it is shown that the furtive scheme of adding hatred-provoking content is less long-lasting than the blatantly de-mod action. Also, the only LS in the linguistic features suggests the possible damage that excessive moderation could bring out.

The digital migration "Exodus" implies that the dynamics within social media resemble the real-world community choice when people are facing a social dilemma (Gürerk, Irlenbusch, & Rockenbach, 2009). The statistics of a network show the continual loss of users under moderation - the decreasing nodes and edges in the moderation period, signaled a loss of users. The short lifespan of r/xkcdcomic shows that it may primarily function as a hub of online social movement and a haven for online "refugees". As some researchers point out, the modularity bears the relationship to the network's robustness (Paranyushkin, 2012), so the success of the dethroning in the end, may have a relationship with the high modularity values in this event.

In conclusion, we explore the aftermath of different kinds of power abuses by a voluntary moderator. We might be wary of excessive moderation since it may bring potential long-term destruction on the community dynamics, especially the blatant action of eliminationism. Our analysis also echoes with other social theories. Additionally, an online "Exodus" phenomenon is observed as netizens' voting with their feet and migrating to a mirror community when facing severe moderation. The new community serves as temporary-use and drains the netizens from the repressed one. Liberation from tyranny is the most influential event as it brings the strongest positivity into all three aspects of the community dynamics. However, because this paper is based on a single case study, its anecdotal nature reduces the generality of the conclusions to some extent. Future studies could conduct similar case analysis for comparison. Also, the "Exodus" phenomenon could be selected to depict the user differences between the leavers and stayers.

REFERENCES

Blumer, H. (1995). Social movements. In S. M. Lyman, *Social movements: Critiques, concepts, case-studies* (pp. 60–83).

Chen, C., & Liu, L.-M. (1993). Forecasting time series with outliers. *Journal of Forecasting*, *12*(1), 13–35.

Chung, K. S. K., Piraveenan, M., Levula, A. V., & Uddin, S. (2013). Assessing online community-building through assortativity, density and centralization in social networks. In *2013 46th Hawaii International conference on system sciences* (pp. 1993–2002).

Fiske, S. T. (1993). Controlling other people: The impact of power on stereotyping. *American Psychologist*, *48*(6), 621–628.

Foerderer, J., & Heinzl, A. (2018). *Do online communities benefit from appointing volunteer moderators? evidence from a regression discontinuity design*. 12.

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology*, *17*, 42.

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2009). *Voting with feet – Community choice in social dilemmas* (Vol. *46*).

Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review*, *110*(2), 265–284.

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*, 1598.

Lo, C. (2018). *When all you have is a banhammer: The social and communicative work of volunteer moderators*.

Macionis, J. J. (2000). *Sociology* (8th ed.). Bergen, NJ: Prentice Hall.

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329–346.

Matias, J. N. (2016). *The civic labor of online moderators* (Vol. *10*). Oxford, England.

Matias, J. N., & Mou, M. (2018). CivilServant: Community-led experiments in platform governance. In *Proceedings of the 2018 CHI conference on human factors in computing systems - CHI '18* (pp. 1–13).

Paranyushkin, D. (2012). *Metastability of cognition in the body-mind-environment network!* 35.

Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. *Dirty Work*, 12.

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, 1–27.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web - WWW '16* (pp. 613–624).