Essays on Trustworthy Online Platforms

Thesis by
Zhuofang Li

In Partial Fulfillment of the Requirements for the
Degree of
PhD in Social Science

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2024
Defended May 6th, 2024

Zhuofang Li
ORCID: 0009-0004-0970-9064

ACKNOWLEDGEMENTS

[Intentionally left blank.]

ABSTRACT

This thesis investigates strategies to enhance the trustworthiness of data-driven online platforms, focusing on combating misinformation, managing disruptive behavior, and fostering positive interactions. I explore multiple approaches, including the analysis of Twitter's misinformation mitigation efforts, the evaluation of moderation practices within a popular online game, and the development of innovative methods for analyzing text data on online platforms.

The first chapter is a study of Twitter. In this study, we examine the effect of actions of misinformation mitigation. We use three datasets that contain a wide range of misinformation stories during the 2020 election, and we use synthetic controls to examine the causal effect of Twitter's restrictions on Trump's tweets in the 2020 presidential election on the spread of misinformation. We find a nuanced set of results. While it is not always the case that Twitter's actions reduced the subsequent flow of misinformation about the election, we find that in a number of instances content moderation reduced the flow of social media misinformation. We estimate that Twitter's actions, on the universe of tweets we study in our paper, reduced the flow of misinformation on Twitter by approximately 15%.

In the second chapter, I shift my attention to online gaming platforms, examining how moderation strategies affect player behavior. Online competitive action games have flourished as a space for entertainment and social connections, yet they face challenges from a small percentage of players engaging in disruptive behaviors. Our research delves into the under-studied question of how moderation for disruptive behavior affects online player behavior. We use real-world game play and moderation data from a popular title—2022's Call of Duty®: Modern Warfare®II. We employ a quasi-experimental design and causal inference techniques to examine the trade-offs between reducing disruptive behavior and player experience when applying moderation. We study both the impact of delayed moderation as well as the severity of applied punishment using our real-world data. We examine these effects on a set of four disruptive behaviors including cheating, offensive user name, chat, and voice. Our findings uncover a nuanced relationship between moderation and player experience, revealing trade-offs between immediate and delayed moderation and the varying severity of punishment. Our examination of real-world gaming interactions sets a precedent in understanding the effectiveness of moderation and its impact on player behavior. Our insights offer valuable contributions toward rethinking

moderation practices and enhancing the online gaming experience for the wider community.

In the third chapter, I propose a novel topic modeling framework for analyzing social science text data, which incorporates temporal and cross-sectional structures to uncover latent topics while accounting for high-dimensional metadata such as temporal dynamics, sentiment, spatial location, and partisanship. The framework provides not only the word components of topics but also a summary of each topic in every extra meta-layer. To accomplish this, high-order tensor decomposition is utilized to account for the complex interplay between the text data and the associated metadata. Through a series of experiments using synthetic and real-world data, the proposed framework demonstrates its ability to capture high-dimensional latent information and outperforms the current state-of-the-art method in terms of topic coherence and interpretability. These findings have important implications for the analysis of social science text data, as the incorporation of rich metadata into topic modeling analyses enables more nuanced and insightful investigations.

PUBLISHED CONTENT AND CONTRIBUTIONS

Li, Zhuofang. et al. (2023). "The Effect of Misinformation Intervention: Evidence from Trump's Tweets and the 2020 Election". In: *Disinformation in Open Online Media*. Ed. by Davide Ceolin, Tommaso Caselli, and Marina Tulin. Cham: Springer Nature Switzerland, pp. 88–102. ISBN: 978-3-031-47896-3. DOI: 10.1007/978-3-031-47896-3_7.

The first chapter is a joint work with Jian Cao, Nicholas Adams-cohen, and R.Michael Alvarez.

The second chapter is a joint work with the Activision research group. List of co-authors includes Rafal Kocielnik (co-first), Mitchell Linegar, Deshawn Sambrano, Fereshteh Soltani, MJ (Min) Kim, Nabiha Naqvie, Grant Cahill, Animashree Anandkumar, and R.Michael Alvarez.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

THE EFFECT OF MISINFORMATION INTERVENTION: EVIDENCE FROM TRUMP'S TWEETS AND THE 2020 ELECTION

Li, Zhuofang. et al. (2023). "The Effect of Misinformation Intervention: Evidence from Trump's Tweets and the 2020 Election". In: *Disinformation in Open Online Media*. Ed. by Davide Ceolin, Tommaso Caselli, and Marina Tulin. Cham: Springer Nature Switzerland, pp. 88–102. ISBN: 978-3-031-47896-3. DOI: 10.1007/978-3-031-47896-3_7.

1.1 Introduction

Research shows that people use social media platforms like Twitter, Facebook, and YouTube to spread misinformation and conspiracy theories about many different subjects (Allington et al., 2021). Recognizing this problem, these platforms have engaged in different approaches to protect their users from misinformation and platform manipulation, for example Twitter's Platform Manipulation efforts.[1] However, recently some states like Florida and Texas have developed policies to block social media platforms from moderating conversations online, especially those that might involve constitutionally-protected political speech.

Much of the concern about the role of social media platforms in the rapid and viral spread of misinformation and conspiratorial ideas has roots in the 2016 American presidential election, with allegations of foreign interference on social media (U.S. Senate Select Committee on Intelligence, 2020). Other studies showed that the spread and consumption of fake news on social media was widespread among Americans in the 2016 election cycle (Allcott and Gentzkow, 2017). Social media platforms developed monitoring and intervention policies in the aftermath of the 2016 election, often with limited public transparency and unknown efficacy.

Detecting misinformation and other undesirable behavior on social media in real-time is difficult, in particular when well-resourced and strategic agents are conducting the behavior (Srikanth et al., 2021). They engage in many strategies to avoid

[1] See https://transparency.twitter.com/en/reports/platform-manipulation.html.

detection, and have strong incentives to hide their activities and identities. In response, social media platforms use many approaches to detect, mitigate, and prevent the spread of false and misleading information. However, research is mixed about whether the strategies used by social media platforms are effective at preventing the spread of misinformation (Thèro and Vincent, 2022), (Carey et al., 2022), (Clayton et al., 2020), (Porter and Wood, 2021), (Sanderson et al., 2021), (Vosoughi, Roy, and Aral, 2018), (Pennycook and Rand, 2019).

In this paper, we use a unique set of natural experiments that occurred during the 2020 presidential election, employing three unique datasets described below in the Data section. In 2020 (as we discuss in the next section), Twitter used various tools to prevent the spread of information in a series of tweets that President Donald Trump posted. These tweets were deemed to violate Twitter's policies about spreading electoral misinformation. We use a synthetic control methodology to develop counterfactuals that allow us to test the efficacy of Twitter's actions on Trump's tweets, allowing us to make causal inferences from the real-world observational data from the 2020 election. Research demonstrates that the synthetic control methodology is a powerful tool for causal inference (Alberto Abadie and Gardeazabal, 2003), (Abadie Abadie, Diamond, and Hainmueller, 2010), (Alberto Abadie, 2021). This is one of the important contributions of our work—showing how synthetic control can help researchers make causal inferences about interventions in social media.

Using this methodology we produce important causal estimates that allow us to study whether Twitter's content moderation actions in the 2020 presidential election were effective. Our results indicate that for the Trump tweets we studied, Twitter's actions can reduce their dissemination. This is not universally the case, as there are situations where misinformation continues to flow after Twitter's content moderation efforts—and where there seems to be little change (one way or the other) after the platform used restrictions or warnings to slow the spread of misinformation. Our results have implications for the current debates about social media platform content moderation which we consider in the paper's Discussion.

1.2 Twitter's Moderation of Trump's Tweets in 2020

In October 2020, Twitter applied a "Civic Integrity Policy"[2] to prevent use of their platform for electoral or civic interference. Policy violations included misleading

[2] https://help.twitter.com/en/rules-and-policies/election-integrity-policy.

(a) (b)

Figure 1.1: Examples of Twitter-labeled Trump tweets. Source `https://factba.se/topic/deleted-tweets`

information about how to participate in the election, voter suppression or intimidation, and false details about electoral outcomes. Depending on the severity of the violation, Twitter could engage in several actions, including labeling the tweet as misinformation, deleting the message entirely, or locking or permanently suspending the offending account.

One of the most prominent uses of the Civic Integrity Policy in 2020 was for Twitter to use warnings or restrictions on then-President Trump's tweets, as he was disputing the integrity of the election and disseminating misinformation about election fraud. During the period of time we focus on in this study, Twitter mainly applied two types of treatments to Trump's tweets that were determined as violations of the "Civic Integrity Policy":

- Disputed (restricted): Content could be hidden or deleted; the user's ability to reply, retweet, and like the tweet could be turned off; or a label/warning message could be applied to the tweet before it was shared or liked. This treatment was applied frequently between November 4th, 2020 to November 7th, 2020.
- Disputed (not restricted): Content was visible, and users can reply, retweet, or like the tweet; and a warning message was applied to the tweet. This treatment appears throughout the study period.

On January 8, 2021, Twitter suspended @realDonaldTrump, at which time the account had approximately 88.7 million followers.

In Figure 1.1 we provide two examples of Trump's tweets, one for each measure. The example in Panel 1a regards allegations being made regarding the election tabulation and post-election auditing in Georgia from November 18, 2020. The example in Panel 1b shows an example of restricted tweet, which was posted on November 4th, 2020.

Twitter's decision to censor and label Trump's tweets in an attempt to prevent the spread of misinformation was highly controversial. Many people, particularly those within the Republican party, launched a backlash against Twitter following their decision to label Trump's tweets as misinformation. Crucially, it remains uncertain whether Twitter's actions worked as intended: did censoring and labeling these tweets in the 2020 election prevent the subsequent spread of election misinformation on Twitter?

We are interested in the effects of Twitter's actions. Labeling Trump's tweets could have two different consequences: it could have operated as (we assume) Twitter desired: suppressing the further spread of misleading information. Or, given the backlash towards Twitter's policy, it could have amplified the spread of misleading information.

In this paper, we use three novel datasets to study the question. The first one is a unique dataset of over 15 million tweets about the election, a real-time collection that started before the November 2020 general election and ended after Twitter suspended Trump's account. The second one is the $ElectionMisinfo2020$ dataset which consists of tweets directly linked to confirmed misinformation stories in the 2020 election. The third one is from Trump's tweet archive, which collected Trump's tweets and showed whether Twitter took action regarding each of those tweets.

Our main findings are nuanced. There is evidence that for some of Trump's tweets, Twitter's actions reduced misinformation. We find that this is in particular the case for a set of tweets that Twitter placed restrictions on early in election 2020. But we also show that the content moderation efforts generally worked in many cases, but did not work in others. In the set of social media conversations about election fraud in the 2020 election that form the basis of our study, we find that Twitter's actions reduced the subsequent flow of election misinformation by approximately 15%.

In the next section of the paper, we connect our research to the theory about how the public receives and processes information, and what happens when attempts are made to suppress the dissemination of political information. These theories guide

and shape our hypotheses. Following this, we delve into our data sources, detailing both the collection and preprocessing of tweets, and then outline the methodologies employed to test our hypotheses. We then present our results and conclude by discussing the implications and limitations of our analysis.

1.3 Does Labeling and Limiting Misinformation Work?

We use public opinion and censorship theory to guide our research. Public opinion theory regards how the public receives, accepts, and processes political information. Assuming that the public acts in a rational manner, they will use information shortcuts to reduce information costs (Downs, 1957). Rational citizens will not obtain and process all available information, as argued in the theory of public opinion (Zaller, 1992), and applied to the reception and processing of social media information. We assume that citizens will follow and process incoming social media information following the "receive-accept-sample" (or RAS) model (Zaller, 1992), (Adams-Cohen, 2019), (Adams-Cohen, 2020).

In the RAS model, the citizen receives information (usually from elites), accepts the information (usually filtering it ideologically or by partisanship), and then samples from recently received information when needed (say answering a survey or voting on a ballot measure). The RAS model provides a theoretical framework in which citizens will be selective about information; partisan citizens will receive and accept information from elites with whom they share partisan affiliations. Partisanship is an important heuristic or information shortcut used by citizens (Lupia, 1992), (Lupia, 1994), (Popkin, 1991).

However, on social media platforms like Twitter, information is not necessarily passed from a partisan elite to a partisan citizen directly—the platform uses algorithms that can alter the flow of information. Furthermore, as was made clear with many of Trump's tweets concerning the 2020 election, the platform can intervene directly by blocking or impeding the ability of an elite to tweet, labeling the elite's messages as misinformation, or making it difficult or impossible for those who view the elite's post to redistribute the message. While Twitter, and similar social media platforms, are private companies, like governments they can control the flow of information on their platforms.

Next, we draw upon the theory of censorship (Roberts, 2018). That theory argues that three mechanisms can be used to censor information online: fear, friction, and flooding. Censoring information through fear means using tools like financial sanc-

tions or the threat of imprisonment to coerce citizens and elites to not disseminate information. Friction regards efforts to slow or make more difficult the dissemination of information. Flooding involves disseminating large quantities of competing information, which serves to make it more difficult and costly to find the information that the government aims to censor.

As (Roberts, 2018) points out, introducing friction works in situations where "the cost added by censorship to the information is enough to offset the benefits of consuming or disseminating information" (p. 72). Recall that the RAS model notes that citizens use heuristics like partisanship to determine which elites they follow and whether they receive information from those elites. In situations where Twitter imposes no friction on Trump's tweets, the RAS model should apply: Republicans should be more likely to receive and accept Trump's tweets, most likely in the form of additional conversation about the topics of Trump's tweets online.

This theoretical foundation allows us to formulate the following two hypotheses:

- Hypothesis 1: Trump's tweets steer the direction of conversation, resulting in a higher volume of tweets concerning the topics that Trump discusses.
- Hypothesis 2: Actions taken by Twitter (restrictions, warnings) lessen the influence of Trump's tweets. Intervening on Trump's tweets will reduce the subsequent discussions, mitigating the effect of our first hypothesis. Consequently, these measures decrease the number of election fraud tweets by Republicans relative to unrestricted tweets.

In the next section we describe our data and methods, as well as how we test these hypotheses.

1.4 Data

We used three datasets in this study: Trump's tweets obtained from the Trump Twitter Archive[3] and Factba.se[4]; 2020 general election tweets that we collected using the Twitter API; and election misinformation tweets dataset *ElectionMisinfo2020*(Kennedy et al., 2022).[5]

[3]https://www.thetrumparchive.com

[4]https://factba.se/topic/flagged-tweets

[5]The data and code used in this paper is available at `https://https://github.com/jian-frank-cao/Disinformation-Intervention`.

Since Twitter suspended Trump's account, we could not directly obtain his tweets from the Twitter API. Therefore, we used the Trump Twitter Archive to obtain all tweets posted by Trump from September 1, 2020, to December 15, 2020. This dataset contains tweet IDs, times, retweet counts, and texts. Additionally, we used Factba.se to identify the tweets that were labeled. Since Factba.se does not differentiate between restricted and warned tweets (documenting both types as "flagged" tweets), we marked "flagged" tweets with zero retweet counts as restricted tweets, and the remaining "flagged" tweets as warned tweets.

We collected the 2020 general election tweets dataset using the Twitter API from June 2020 to January 2021. We utilized the long-term Twitter monitor developed by (Cao, Adams-Cohen, and Alvarez, 2021) and keywords related to election fraud, remote voting, polling places, and other election topics. We used this dataset to study how Twitter's restrictions influenced the retweeting of Trump's tweets.

The election misinformation tweets dataset (Kennedy et al., 2022) is at the core of this study. It comprises tweets identified in 456 distinct misinformation stories from September through December 2020. For each tweet, the dataset displays the misinformation story it is part of, its identification number[6], the identification numbers of the tweets it retweeted/quoted/replied to, and its partisan lean (left, right, unknown). We used this dataset to construct time series of misinformation counts and study how Trump's tweets and Twitter's labeling impacted these time series.

We find 576 tweets of Trump directly appear in the dataset. Among the 576 tweets, there are 10 restricted tweets, 108 warned tweets, and 458 unrestricted/unwarned tweets. Fifty-nine tweets are directly labeled as misinformation and the summary of the fifty-nine tweets can be found in Table 1.1.

1.5 Effects of Trump's Tweets

The first question we are interested in is the effects of Trump's tweets on the spread of misinformation. Our hypothesis is that Trump's tweets would directly lead to an increasing spread of the corresponding misinformation. To investigate this, we take each of Trump's tweets that appear in the ElectionMisinfo2020 dataset, plot the volumes of the corresponding misinformation story around the posting time of Trump, and look at the direct effect of Trump's posting on the time series.[7] Note

[6]The tweet ID can uniquely identify a message on Twitter, including tweet, reply, quote, and retweet

[7]Out of all Trump's tweets, there are two that were posted close enough in time that their active periods overlap. In this specific instance, we study the combined effect of these tweets, using the

Table 1.1: Summary of Trump's Tweets in the ElecMisinfo2020 Dataset

Story Number	Description	Count	Hard	Soft	Unrestricted	Retweet
Story 1	ballot harvesting: Ilhan Omar Project Veritas Video	3	0	0	3	1
Story 2	tech: dominion	34	0	23	11	12
Story 3	Late:Extended Ballots	1	0	0	1	0
Story 4	dead voters: general ticket	5	0	5	0	1
Story 5	Digital dumps: Michigan 128000 votes	2	2	0	0	0
Story 6	partisan vcr: Nevada whistleblower	1	0	1	0	1
Story 7	Physical Mail Mistakes: Deceased and Inactive CA	1	0	0	1	1
Story 8	Physical Mail Mistakes: MI Misprints for Troops	2	0	0	2	2
Story 9	poll watchers: Philly no entry list	1	0	1	0	0
Story 10	Physical Mail Fraud: Democratic TX Mayor	1	0	0	1	0
Story 11	Other: Stop The Steal Pushed	1	0	0	1	0
Story 12	Other: Candidate Fraud Biden Fraud Quote	2	0	0	2	0
Story 13	protests: stop the steal rallies	1	0	1	0	1
Story 14	Physical Mail Fraud: PA Misprinted Corrections	2	0	0	2	0
Story 15	Statistics: Math Video	1	0	0	1	0
Story 16	Physical Mail Mistakes: NYPost Ballot Typo	1	0	0	1	0

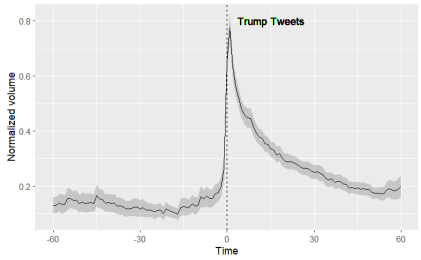
that we take all of Trump's tweets, regardless of Twitter's actions, which could bias the result downward. Therefore, the effect we discuss here might be a lower bound.

We find very similar patterns among almost all the 120-min windows around Trump's posting time: the tweet volume rises sharply, and then gradually decreases, eventually equilibrating at a stable volume that is higher than the level before the posting event. To estimate the average effect, we first normalize each 120-min window by applying the following transformation to each time window.

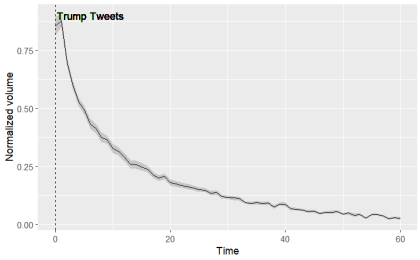
$$\hat{Y}_i = \frac{Y_i - min(Y)}{max(Y) - min(Y)}$$

Figure 1.2a shows the average time series and the confidence interval among all the time windows. We also independently count retweets of Trump's tweets. The normalized average time series and the confidence interval among all windows are shown in Figure 1.2b. Additional, we also plot the average normalized volume for left-lean and right-lean tweets separately in Figure 1.2c and Figure 1.2d. This shows

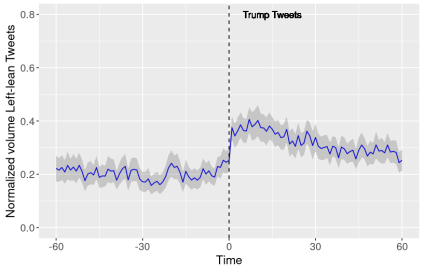
timestamp of the first tweet as the reference point for our analysis.



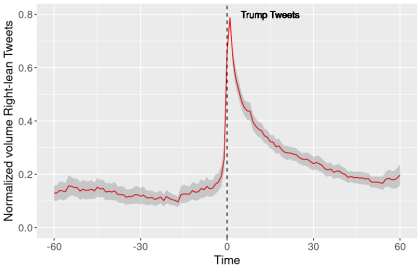
(a) Average tweet time series



(b) Average retweet time series



(c) Average left-lean tweet time series



(d) Average right-lean tweet time series

Figure 1.2: Average normalized time series among all 120-min time intervals affected by Trump's tweet

that the overall effect on volume is much larger for right-leaning tweets, which is consistent with Hypothesis 1.

To quantify the effect of Trump's tweets and the heterogeneity across different topics, we perform a t-test for each story on the average column in three time periods. Period 1 spans 30 minutes before Trump's tweet ($T = -30$ to $T = -1$). Period 2 spans 30 minutes after Trump's tweet ($T = 0$ to $T = 29$). Lastly, Period 3 is 30 minutes to 60 minutes after Trump's tweet ($T = 30$ to $T = 59$).

We compare the data from Periods 1 and 2 to see the immediate effect of the tweet, and from Periods 1 and 3 for the longer-term impact. The volume per minute comparisons before and after each tweet, along with the t-test results, are displayed in Figure 1.3. The graph indicates that Trump's tweet has a heterogeneous effect across different topics. We can observe that for most of the topics, there is an increase in volume either immediately or after 30 minutes. The volume does not immediately increase for some topics with Twitter's intervention like "Dead voters", "Nevada Whistleblower", and "Poll Watcher", which provides evidence in support of Hypothesis 2.

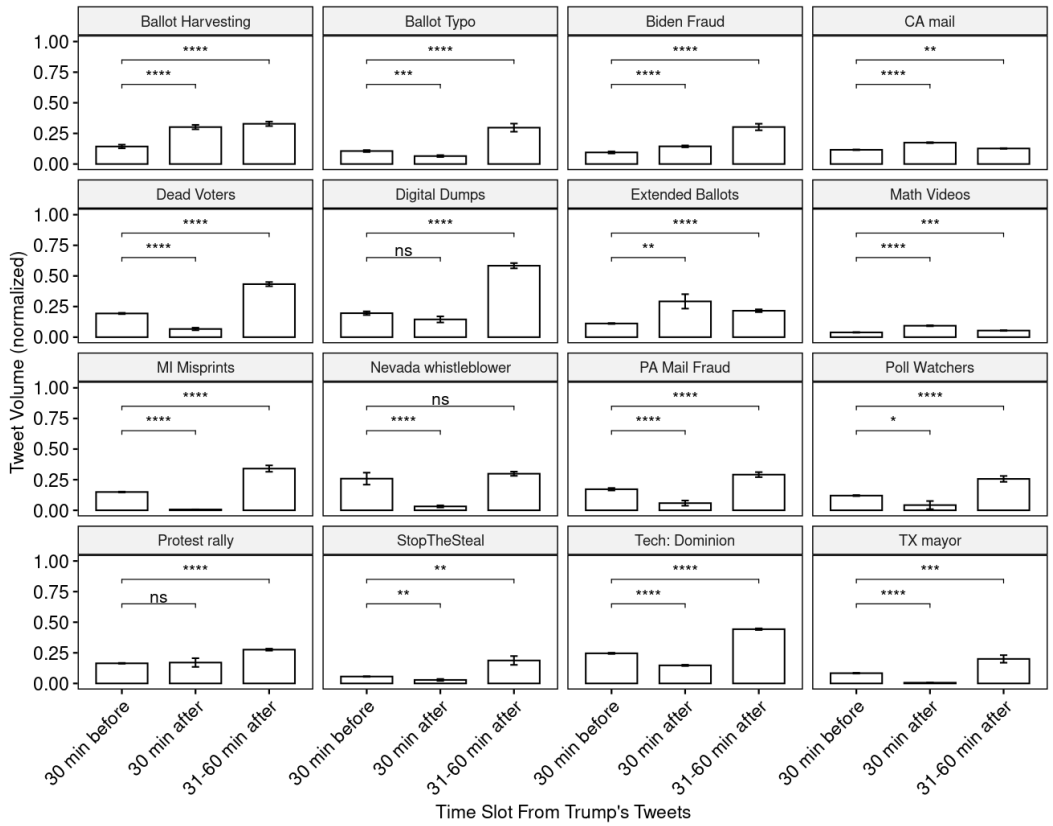


Figure 1.3: Effect of Trump's tweets by topic and time. Note: ns: $p > 0.05$; *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$; ****: $p \leq 0.0001$

1.6 Effects of Twitter's Actions

To study the effects of labeling, we first estimate the time Twitter applied the label. We then derive time series of misinformation related to Trump's tweets. We compare the time series of messages where a label was applied against counterfactuals derived from messages that were not subject to any restriction. Using this data, we estimate the effects of labeling.

To study the effects of labeling, it is necessary to know when the labeling became effective, i.e., the treatment time. However, Twitter has not disclosed the exact timing of the labels, only stating that they applied labels between 5 to 30 minutes after Trump posted the tweets.[8] Fortunately, our 2020 general election data contains real-time retweets of 7 out of 26 of Trump's restricted tweets. Each retweet contains a retweet status object that points to Trump's original tweet and shows its latest retweet count by the time the retweet was collected by our Twitter monitor. The time series of cumulative retweets are shown in Figure 1.4. We can see that the time series

[8] https://www.youtube.com/watch?v=ONYuLP7sHFQ&t=4701s

stopped around 20 to 240 minutes after Trump tweeted because Twitter restricted users' ability to retweet, and no more new retweets were collected. The stopping points (red) are our estimates of labeling time. Notice that labeling took around 1.5 to 4 hours in September, while it only took around 30 minutes in November. Twitter expedited its labeling, likely because election misinformation was spreading fast and the potential damage to society was great. Since we cannot directly estimate the labeling time of warned tweets as retweeting was not restricted, we assume it is similar to that of the restricted tweets.

Next, we derive the time series of misinformation tweets related to Trump's tweets. For each of Trump's tweets included in the misinformation data set, we find that tweet's corresponding story.[9] We then compute the number of tweets posted per minute, across the entire misinformation dataset, from the associated misinformation story. We focus on the time series from T to $T + 120$, where T is the timestamp of Trump's tweet. Thus, each of Trump's tweets produces a misinformation time series. When the Trump tweet that produces this time series is "labeled" by Twitter, we refer to this as a "labeled misinformation time series", and if the Trump tweet is "unlabeled", an "unlabeled misinformation time series".[10].

We use synthetic control to construct counterfactuals of the labeled time series. For each labeled misinformation time series, if there are more than five unlabeled misinformation time series in the same story, we use them to estimate the synthetic control. Otherwise, we disregard the stories and use all unlabeled time series. Based on the estimates of labeling time in Figure 1.4, assuming most labeling was imposed after $T + 20$, we estimate the synthetic control using the T to $T + 19$ sub-series to ensure that it closely resembles the labeled time series in the first 20 minutes.

We show synthetic controls for all of Trump's restricted tweets in Figure 1.5a and 15 out of 201 warned tweets in Figure 1.5b. The synthetic control, i.e., the estimated tweets if there was no restriction, is shown in red, and the observed tweets are shown in blue. The area between the red and blue curves are the estimated effects of labeling. If the red curve is above the blue curve, then the effect of labeling is negative, which means labeling reduces the spreading of misinformation. For example, those

[9] In some cases, the Trump tweet is not directly in the misinformation dataset, but we do find the tweet's associated retweets, quotes, and replies. In all cases where we find more than ten examples of retweets, quotes, or replies with a story in the misinformation dataset, we define the Trump tweet's misinformation story as the most common across this set of retweets, quotes, and replies. If we find fewer than ten examples, we drop this Trump tweet from our analysis.

[10] If any Trump tweet did not lead to significant corresponding misinformation time series from T to $T + 120$, i.e., fewer than 100 tweets per minute on average, we dropped it from our analysis

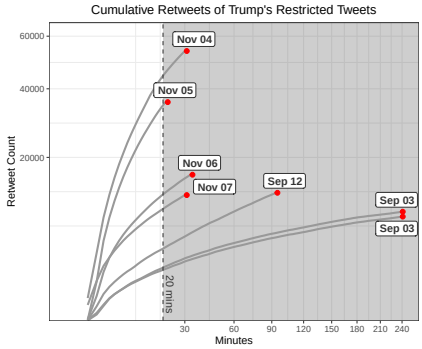


Figure 1.4: How fast did Twitter apply the restriction?

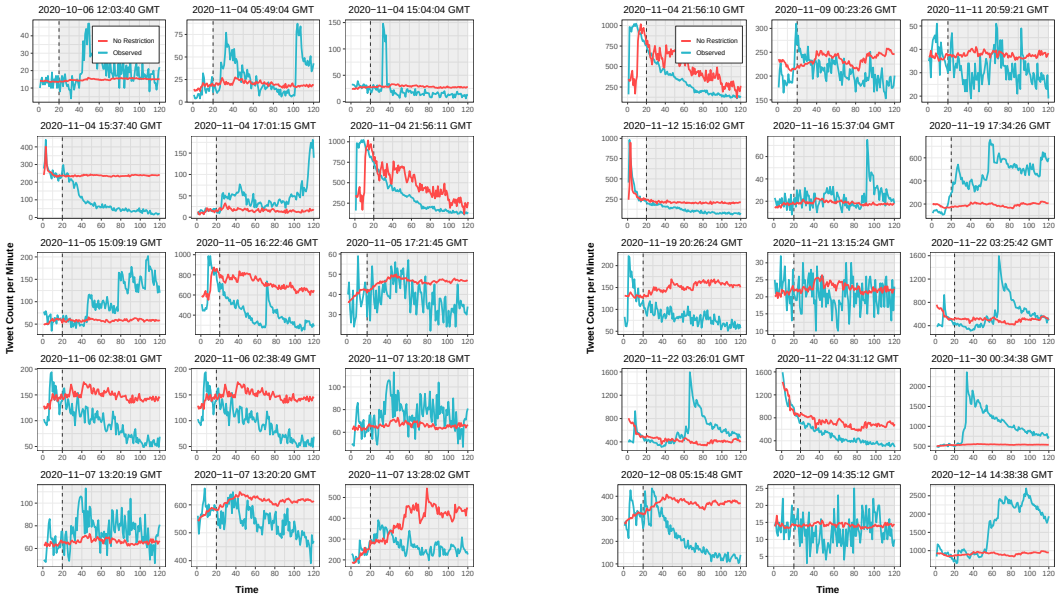
where there is solid evidence from the synthetic control methodology that Twitter's content moderation reduced misinformation are Trump's that Twitter restricted on Nov 04 15:37:40, Nov 04 21:56:11, and Nov 05 16:22:46. Additionally the synthetic control methodology indicates that Twitter's content moderation reduced misinformation in the instances where they placed warnings on Trump's tweets about the election on Nov 04 21:56:10, Nov 09 00:23:26, and Nov 12 15:16:02. On the other hand, if the blue curve is above the red curve, this indicates positive treatment effects, in which Twitter's labeling stimulates more discussion about misinformation. For example, the synthetic control method indicates positive treatment effects when Twitter restricted Trump's tweets on Nov 05 15:09:19 and when Twitter placed warning labels on his election-related tweets on Nov 19 17:34:26, Nov 30 00:34:38, and Dec 14 14:38:38.

Overall, the synthetic control results shown for the thirty examples in Figure 1.5a and Figure 1.5b provide a nuanced perspective on Twitter's attempts in 2020 to slow or stop the spread of election misinformation by restricting or placing warning labels on Trump's tweets. Among the restricted tweets (Figure 1.5a) we see relatively clear evidence in 8 of the 15 instances for restriction reducing the subsequent spread of misinformation. Similarly, among the Trump tweets where warning labels were used, 6 of the 15 examples show that the subsequent spread of misinformation was slowed.

With these synthetic controls, we quantify the labeling effect using ratios of average tweets in the second hour:

$$\phi_i = \frac{\frac{\sum_{t=61}^{120}(Observed)_{i,t}}{60}}{\frac{\sum_{t=61}^{120}(Estimated\ No\ Restriction)_{i,t}}{60}} \tag{1.1}$$

The ratio is less than one if the average observed tweets from $T + 61$ to $T + 120$



(a) Estimated No Restriction Tweets vs. Observed Tweets (Restricted)

(b) Estimated No Restriction Tweets vs. Observed Tweets (Warned) (Sample)

Figure 1.5: Effects of Restriction

is smaller than the average estimated tweets if there was no restriction, i.e., the blue curve is above the red curve, and it is larger than one otherwise. Since this study is interested in analyzing how Twitter's labeling reduces the spreading of misinformation, we focus on Trump's tweets that are associated with a large number of misinformation tweets and exclude time series that have on average fewer than 100 tweets per minute.

The ratios are shown in Figure 1.6 and Table 1.2. We use shapes to distinguish Trump's restricted and warned tweets and use colors to show stories. We see that the majority of the ratios are less than one, i.e., in the green area, as fifty-five of the tweets in this analysis are in the green area. Importantly we note that of the six tweets in this sample that were restricted, five of the restricted instances were ones where the subsequent flow of misinformation were reduced, and in only one of those instances was the subsequent flow of misinformation not reduced by the restriction of Trump's tweets. It is also important to note that these six restricted tweets were in the immediate aftermath of the 2020 presidential election, at a time when mitigating the spread of misinformation might have been most influential. We also must note, however, that thirty of the tweets in this analysis (the vast majority of which were those with warning labels) show positive treatment effects, meaning that misinformation increased after the warning labels were used. Some of the

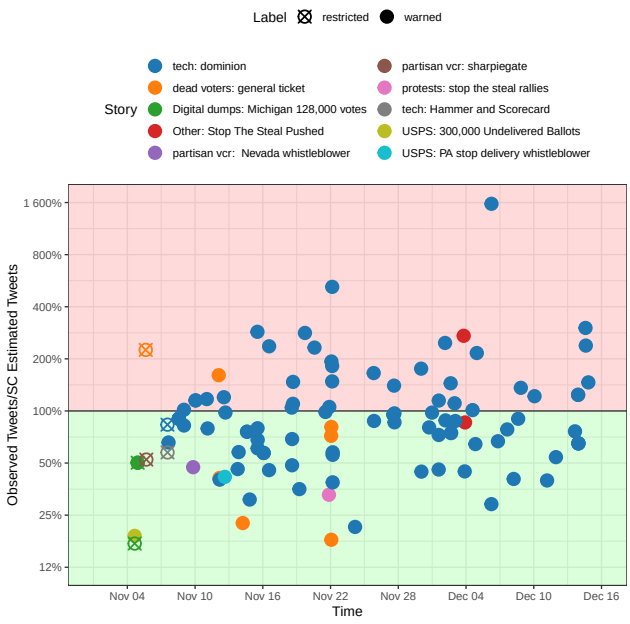


Figure 1.6: Quantification of the estimated treatment effect

tweets with warning labels have sizable increases in post-moderation spread, one of the tweets with a warning label in early December 2020 saw a 1600% increase in post-treatment misinformation spread.

	Restricted	**Warned**
(12.5%, 25%]	1(16.67%)	4(3.92%)
(25%, 50%]	0(0%)	17(16.67%)
(50%, 100%]	4(66.67%)	45(44.12%)
(100%, 200%]	0(0%)	25(24.51%)
(200%, 400%]	1(16.67%)	9(8.82%)
(400%, 800%]	0(0%)	1(0.98%)
(800%, 1600%]	0(0%)	1(0.98%)
Total	6	102

Table 1.2: Tweets in Estimated Ratio Ranges

Finally, for interested readers, we also show the distribution of ϕ_i in Table 1.2.A t-test of $log(\phi_i)$ yields a t-value of -2.4737 and a p-value of 0.0149, which means labeling significantly ($P < 0.05$) reduces the volume of misinformation tweets in the testing period $[T+61, T+120]$. The mean of the log effect $\overline{log(\phi_i)} = -0.1673$ indicates that, on average, Twitter's labeling reduces $1 - e^{-0.1673} = 15.41\%$ of misinformation tweets.

1.7 Discussion

Existing literature presents conflicting findings on the ability of social media platforms to mitigate the spread of misinformation effectively. Our study, however, takes a more targeted approach, examining a particular facet of platform moderation. We utilize a unique dataset and adopt a sophisticated causal inference methodology to increase the validity of our conclusions. Our findings suggest that actions taken by social media platforms can mitigate the subsequent spread of misinformation. We call for further research to better understand the conditions under which moderation is possible and which interventions are the most effective.

In particular, the next stage of research needs to tackle the conditions when content moderation has the desired treatment effect. Is restriction more effective than labeling (we see intriguing evidence that the answer may be yes in Figure 1.6)? Does it matter when a platform applies restrictions or labels? Does the speed at which moderation is carried out affect its effectiveness? Is the wording of the warning label important for restricting subsequent spread? There are many additional questions that researchers and social media companies should tackle.

It is important to view our conclusions in through lens of the current moment, wherein some social media channels opt for less moderation, ostensibly to champion free speech. Discussions surrounding the policies being implemented by states such as Florida and Texas, in conjunction with legal debates about the moderation of certain social media dialogues, highlight potential restrictions on content moderation. While Constitutionally-protected political speech might be an area where content moderation is problematic, that should not imply that social media platforms should stop efforts to prevent the spread of child pornography, voting disenfranchisement, sexual and racial harassment, or the use of their platforms by terrorists organizations. The research community needs to step up our involvement in these debates, and provide research that can help social media platforms develop appropriate content moderation policies that protect rights while preventing illegal behavior and social harm.

References

Abadie, Abadie, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program". In: *Journal of the American Statistical Association* 105.490, pp. 493–505.

Abadie, Alberto (June 2021). "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects". In: *Journal of Economic Literature* 59.2, pp. 391–425.

Abadie, Alberto and Javier Gardeazabal (2003). "The Economic Costs of Conflict: A Case Study of the Basque Country". In: *The American Economic Review* 93 (1), pp. 113–32.

Adams-Cohen, Nicholas (2019). "New Perspectives in Political Communication." California Institute of Technology, doi:10.7907/7TDG-4R42. PhD thesis.

– (2020). "Policy change and public opinion: Measuring shifting political sentiment with social media data". In: *American Politics Research* 48.5, pp. 612–621.

Allcott, Hunt and Matthew Gentzkow (Mar. 2017). "Social media and fake news in the 2016 election". In: *Journal of Economic Perspectives* 31 (2), pp. 211–236. ISSN: 08953309. DOI: 10.1257/JEP.31.2.211.

Allington, Daniel et al. (2021). "Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency". In: *Psychological Medicine* 51.10, pp. 1763–1769. DOI: 10.1017/S003329172000224X.

Cao, Jian, Nicholas Adams-Cohen, and R Michael Alvarez (2021). "Reliable and efficient long-term social media monitoring". In: *J. Comput. Commun.* 09.10, pp. 97–109.

Carey, John M et al. (2022). "The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada". In: *Nature Human Behavior*. DOI: 10.1038/s41562-021-01278-3. URL: https://doi.org/10.1038/s41562-021-01278-3.

Clayton, Katherine et al. (Dec. 2020). "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media". In: *Political Behavior* 42 (4), pp. 1073–1095. ISSN: 15736687. DOI: 10.1007/S11109-019-09533-0.

Downs, Anthony (1957). "An Economic Theory of Democracy". In.

Kennedy, Ian et al. (2022). "Repeat Spreaders and Election Delegitimization: A Comprehensive Dataset of Misinformation Tweets from the 2020 US Election". In: *Journal of Quantitative Description: Digital Media* 2.

Lupia, Arthur (1992). "Busy voters, agenda control, and the power of information". In: *American Political Science Review* 86.2, pp. 390–403.

– (1994). "Shortcuts versus encyclopedias: Information and voting behavior in California insurance reform elections". In: *American Political Science Review* 88.1, pp. 63–76.

Pennycook, Gordon and David G. Rand (Feb. 2019). "Fighting misinformation on social media using crowdsourced judgments of news source quality". In: *Proceedings of the National Academy of Sciences of the United States of America* 116 (7), pp. 2521–2526. ISSN: 10916490. DOI: 10.1073/PNAS.1806781116.

Popkin, Samuel L. (1991). *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. University of Chicago Press.

Porter, Ethan and Thomas J. Wood (Sept. 2021). "The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom". In: *Proceedings of the National Academy of Sciences of the United States of America* 118 (37), e2104235118. ISSN: 10916490. DOI: 10.1073/PNAS.2104235118/SUPPL_FILE/PNAS.2104235118.SAPP.PDF. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2104235118.

Roberts, Margaret E. (2018). *Censored*. Princeton University Press.

Sanderson, Zeve et al. (2021). "Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform". In: *Harvard Kennedy School Misinformation Review* 2 (4).

Srikanth, Maya et al. (2021). "Dynamic Social Media Monitoring for Fast-Evolving Online Discussions". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. Virtual Event, Singapore: Association for Computing Machinery, pp. 3576–3584. ISBN: 9781450383325. DOI: 10.1145/3447548.3467171. URL: https://doi.org/10.1145/3447548.3467171.

Thèro, Hélo ise and Emmanuel M Vincent (2022). "Investigating Facebook's interventions against accounts that repeatedly share misinformation". In: *Information Processing and Management* 59, p. 102804. DOI: 10.1016/j.ipm.2021.102804. URL: https://doi.org/10.1016/j.ipm.2021.102804.

U.S. Senate Select Committee on Intelligence (2020). *Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Volume 2: Russia's Use of Social Media*. U.S. Senate Select Committee on Intelligence Report 116-XX. Washington D.C. URL: https://www.intelligence.senate.gov/sites/default/files/documents/Report%5C_Volume2.pdf.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (Mar. 2018). "The spread of true and false news online". In: *Science* 359 (6380), pp. 1146–1151. ISSN: 10959203. DOI: 10.1126/SCIENCE.AAP9559.

Zaller, John R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Chapter 2

ONLINE MODERATION IN GAMES: HOW INTERVENTION AFFECTS PLAYER BEHAVIOR

2.1 Introduction

The proliferation of online social interactions and competitive action games has enriched the virtual landscape, providing opportunities for entertainment, improved well-being, and social connections (Kriz, 2020; Bourgonjon et al., 2016). However, this digital frontier is not without challenges. While the majority of players engage in respectful and enjoyable gameplay, a small percentage have leveraged these platforms to exhibit disruptive behaviors such as cheating, trolling, and offensive speech (C. Cook et al., 2019). Moderation, or the regulation of user behavior by platforms, has thus emerged as an essential component of online gaming. The process demands tangible resources, including employees, infrastructure, and time, particularly when human review is required to mitigate complex challenges that cannot easily be handled via automation alone, such as domain shift (changes in expression of disruptive behavior or its definition) (Srikanth et al., 2021) and strategic classification (when the players strategically alter their behavior to circumvent automated systems and avoid detection) (Frommel and R. Mandryk, 2022). Platforms face tangible constraints in moderating user behavior. Though algorithmic tools can lessen the content requiring review, the human element is vital (Levkovitz, 2023). This is due to aspects challenging to automate such as interpretation ambiguity (Beres et al., 2021), the need for broader contextual understanding (Frommel and R. Mandryk, 2022), and the need for common-sense judgment. Optimal moderation efforts will likely require collaboration between human judgment and technological tools (Rieder and Skop, 2021; Link, Hellingrath, and Ling, 2016). Unfortunately, the ratio of human moderators to the volume of content requiring moderation leads to bottlenecks in the review process (Gorwa, Binns, and Katzenbach, 2020). Thus many important scientific challenges exist for moderating distributive player behavior (Kocielnik et al., 2023).

Prior work: Despite the importance of moderation, empirical work exploring the causal effects of this practice within the gaming community remains sparse (Wijkstra et al., 2023), with most work focusing more on developing novel data-driven

approaches to detecting toxic behaviors (Canossa et al., 2021; Weld et al., 2021), testing theory-informed hypotheses related to the emergence of toxicity (Kwak, Blackburn, and Han, 2015), or studying the toxicity in various social communication platforms rather than in games themselves (Ghosh, 2021). These studies, however, do not examine the effects that moderation of toxic behavior has on players in real-world gaming situations. The few prior studies that do examine the impact of moderation relied only on small sample survey-based examinations around the self-reported perceptions of players (Ma, Li, and Kou, 2023; Kou and Gui, 2021; Kordyaka and Kruse, 2021) or moderators (Cullen and Kairam, 2022; Aguerri, Santisteban, and Miró-Llinares, 2023). Both of these lack the scale to draw conclusions about the effectiveness of different types and properties of moderation at scale in real-world gaming titles. Indeed a recent review of intervention systems for toxicity highlighted that only a few interventions are evaluated with players and in commercial settings (Wijkstra et al., 2023), highlighting the potential for more research with higher external validity. Our study fills this gap by examining real-world large-scale moderation data from one of the more renowned titles in the industry—2022's *Call of Duty®: Modern Warfare®II*.

Our work: In this paper, we utilize a quasi-experimental design and the latest causal machine learning methods (causalML)(Kaddour et al., 2022) to analyze the impact of moderation on player behavior. We specifically examine and compare the behavior of players that were moderated as compared to those that were not in terms of the impact on subsequent offensive behavior (repeated offenses) and the number of days with matches played (player experience). We focus on players who were eventually subjected to human moderation and control for consistent types of behavior. Taking into account the principle of immediacy we also examine the effect of delayed consequences by examining the player's behavioral measures in the context of delayed versus immediate moderation. Finally, we evaluate the impact the severity of applied moderation actions has on immediate post-moderation player behavior.

Findings: Our results reveal a nuanced relationship between moderation and player experience and illustrate trade-offs between immediate and delayed moderation as well as varying severity of punishment. Specifically, our analysis of player behavior shows that moderation effectively lowers disruptive behavior by up to 70% but can lead to up to 20% fewer matches played per day. Immediate moderation as compared to moderation with delay is more effective at reducing toxicity. Cheaters seem to

respond differently to moderation in that moderation results in a larger reduction of participation (days with matches) for these players than for toxic players. These results offer valuable insights into how moderation affects player experience and disruptive behavior, including the unique response of cheaters.

Contributions: This study offers several significant contributions:

1. We present one of the first studies examining large-scale real-world moderation efforts from one of the most popular gaming franchises.

2. We uncover an important tradeoff between moderation effectiveness and impact on player experience, as well as the importance and impact of immediate versus delayed moderation.

3. Our findings lead to a rethinking of moderation practices, and we provide a discussion of the real-world implications, setting the agenda for comprehensive analysis and mitigation opportunities in a domain affecting millions daily.

4. Our study sheds critical light on the strategies for mitigating disruptive behaviors in online environments, a topic of increasing relevance given the widespread nature of such behaviors across various digital platforms. The insights gleaned from our research are particularly pertinent for online and gaming platforms that strive to foster user connections and maintain a healthy digital ecosystem. By addressing the nuances of how different moderation techniques influence user behavior, our findings offer valuable guidance for platform administrators and community managers in their ongoing efforts to cultivate safe, engaging, and respectful online communities. Therefore, this study extends beyond its immediate context to provide actionable knowledge in the broader realm of digital interaction and community management.

2.2 Research Questions and Hypotheses

Understanding Moderation Impact on Disruptive Player Behavior

Common practices in moderation rely on the punishment of undesirable behaviors via various interventions (Ma, Li, and Kou, 2023). Such punitive actions are naturally expected to discourage the repetition of the same offense in the future (GGWP, 2023). *Deterrence Theory* seems particularly relevant in understanding how such actions can affect player behavior (Pratt et al., 2017). This theory introduced the

concepts of *General Deterrence* and *Specific Deterrence*. The first concept aims to prevent the general population from committing offenses by making examples out of those who are caught and punished. The second concept focuses on preventing an individual who has already committed an offense from reoffending. In this context, moderation actions can be seen as leveraging both these mechanisms, especially if announced publicly.

Secondly, from the behavior change perspective, the *Theory of Planned Behavior* (Ajzen, 2020) identifies a subjective norm component that impacts behavior. This component represents an individual's perception of social normative pressures, or the beliefs of relevant others about what behaviors should or should not be performed. Moderation actions can be seen as a mechanisms for shaping such subjective norm, especially if publicly known (i.e., via code of conduct (ABK, 2023)) and accepted by the community.

Finally, the Cognitive Behavioral Theory suggests that understanding and changing negative thoughts can lead to changes in behavior (Lochman, 1992). In the context of online gaming, moderation interventions, particularly focused on helping the players understand what they are doing wrong, could aim to alter harmful thought patterns that lead to disruptive behavior.

Understanding Moderation Impact on Player Experience

Moderation can also impact player experience in various ways. First, some players may consciously and purposefully engage in disruptive behavior. Such players are fully aware that their behavior is violating the code of conduct (ABK, 2024) or social rules and yet they still choose to engage in it (Lee, Jeong, and Jeon, 2019). These players may have a higher tolerance for toxic behavior or may even derive enjoyment from it. In such cases, consistent moderation can deter these players if they perceive their attempts at being disruptive as ultimately unsuccessful. Some of these players may choose to simply stop playing rather than change their behavior.

Second, if moderation is seen as a form of negative feedback, then some players may simply be more sensitive to punishment (Kim et al., 2015). Negative feedback has been shown to lead to defensiveness, anger, and repudiation (Fong et al., 2016) which may negatively affect player experience.

Third, the players might disagree with the moderation's determination that their behavior was disruptive, This is common, as there are substantial disagreements on what is considered toxic even among seasoned players (Beres et al., 2021).

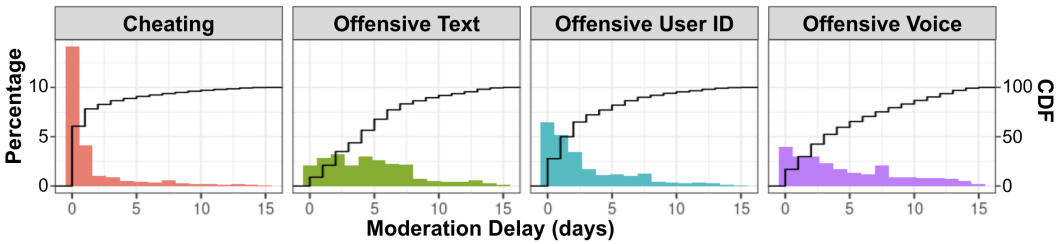


Figure 2.1: Distribution of delay in moderation; Figure showing both PDF and CDF of its distribution. Data generated from Feb 1st, 2023 to April 12th, 2023.

Individual tolerances for disruptive behavior vary, leading to diverse perceptions of what crosses the line. Consequently, if a player feels unfairly punished, this can lead to decreased trust in the moderation fairness (Steinkuehler, 2023) and negatively impact their gaming experience. Recent work has shown the importance of explanations in increasing players' perceptions of fairness of moderation in games (Ma, Li, and Kou, 2023).

Understanding the Impact of Moderation Delay

Moderation actions, especially in high-volume production systems may come at a delay. In Figure 2.1 we show that such delay indeed happens in our dataset. A delay introduces a period of time during which a player's disruptive behavior is not met with any reaction. Delays in moderation within online gaming platforms like Call of Duty®: Modern Warfare®II often occur due to operational constraints, such as the need to reach a specific threshold of reports before taking action, or reliance on human moderators for case review. While sometimes unintentional, these delays may inadvertently allow for a period of self-regulation among players (Liau et al., 2015). Understanding the effects of these delays is crucial for optimizing moderation processes.

The deterrence theory, rooted in criminology, suggests that the certainty, severity, and celerity (swiftness) of punishment are key determinants in preventing undesirable behavior (Pratt et al., 2017). Similarly, from a behavioral economics perspective, the phenomenon of "loss aversion" can be applied to understand how players might react to the potential loss of access or privileges within a game due to moderation (Kahneman and Tversky, 2013). The immediate and certain prospect of punishment can have a more significant deterrent effect than the severity of the punishment itself.

Research in educational contexts indicates that postponing disciplinary actions can

diminish their impact (Abramowitz and O'Leary, 1990). Likewise, in the realm of social media, it has been observed that delayed measures in moderating content, such as the removal of inappropriate postings, tend to be less effective (Srinivasan et al., 2019). When responses to negative behaviors are delayed, there's a decreased probability of adherence from the individuals involved. Additionally, from the standpoint of learning theories, the *timing* of feedback is crucial in determining the effectiveness of such feedback (Thurlings et al., 2013). Studies have shown that the action-effect delay can diminish an individual's sense of agency (Wen, 2019), which means a player might disassociate their disruptive behavior from the punishment and from the ability to control the behavior that leads to such punishment. As a result, moderation might be less effective if given with substantial delay.

Understanding the Impact of Severity of Moderation Actions

Another aspect of moderation is the severity of the applied punishment. The natural expectation is that a more severe punishment for disruptive behavior should result in more effective deterrence (Klepper and Nagin, 1989). Existing moderation approaches in online gaming rely heavily on this assumption in implementing escalating punishment severity (GGWP, 2023). More severe punishment should have pronounced effects on reducing repeated offenses due to the increased costs of being disruptive. At the same time, it can also drive away players who feel unfairly punished, are sensitive to negative feedback, or engage in disruptive behavior on purpose. In that case, we would expect that the severity of moderation actions correlates with the impact on player repeated offenses and player experience.

All these indications lead to two primary research questions and their associated hypotheses:

RQ1: What is the impact of moderation on players' disruptive behavior?

Hypothesis 1.1: Moderation on online gaming platforms leads to decreased disruptive behaviors among moderated players.

Hypothesis 1.2: Delayed moderation is less effective at reducing disruptive behaviors compared to immediate moderation.

Hypothesis 1.3: More severe moderation on online gaming platforms leads to more decrease in disruptive behaviors among moderated players.

Hypothesis 1.4: The effectiveness of severe moderation interventions in reducing disruptive behavior is enhanced when those interventions are implemented imme-

diately.

RQ2: What is the impact of moderation on player experience?

Hypothesis 2.1: Moderation on online gaming platforms leads to decreased player experience among moderated players.

Hypothesis 2.2: The negative impact of moderation actions on player gaming experience is amplified when those actions are implemented with a delay, as compared to immediate implementation.

Hypothesis 2.3: More severe actions result in more reduction in player experience.

2.3 Background and Related Work

Call of Duty®: Modern Warfare®II, released in 2022, is a first-person action game created by Infinity Ward and brought to market by Activision. Set across various continents, the game immerses players in a contemporary and warfare environment. It features a narrative-driven single-player campaign along with a variety of multiplayer online modes.

Online multiplayer video games like Call of Duty®: Modern Warfare®II, enjoyed in real time with others (Kordyaka and Kruse, 2021), offer a plethora of benefits (Kriz, 2020). These range from the enjoyment and satisfaction of basic psychological needs (Raith et al., 2021) to the facilitation of social relationships and aiding in coping and recovery mechanisms (Trepte, Reinecke, and Juechems, 2012). The positive impacts on user experience, such as enhanced fun and additional social exchange, are significant (Bourgonjon et al., 2016). However, the presence of in-game toxicity can undermine these benefits(C. Cook et al., 2019).

Over the past decade, research in video games has included a substantial focus on toxicity (Märtens et al., 2015; Wijkstra et al., 2023; Frommel and Regan L. Mandryk, 2023). The increasing popularity of online gaming (Paschke et al., 2021) has unfortunately led to a rise in the frequency of toxic behavior (Kordyaka, Laato, et al., 2023; Kordyaka, Park, et al., 2023). However, this behavior is not universal among players. Estimates suggest that only about 1% of a player base might consistently exhibit toxicity (Stoop et al., 2019). This small percentage of players disproportionately contributes to the overall toxicity in a game. For example, in League of Legends, it is estimated that this 1% of the player base is responsible for approximately 5% of the toxic speech (Stoop et al., 2019). Still, such

disruptive content, over time, can affect a disproportionately large percentage of the player population (Frommel, Regan L Mandryk, and Klarkowski, 2022). Therefore, strategies aimed at moderating toxic behavior, particularly those targeting these consistently toxic players, could significantly impact the reduction of overall toxic behavior on these online platforms. However, recent work highlighted substantial challenges involved in moderating toxic content in real-world high-volume systems (Kocielnik et al., 2023).

Studies Around Perception and Impact of Moderation

The examination of user experiences and policies around content moderation across various platforms reveals a nuanced understanding of the challenges and strategies in combating online toxicity. Research in this area often relies on self-reported data, highlighting both the strategies employed and the perceptions of their effectiveness. For instance, C. L. Cook, Patel, and Wohn (2021) utilizing surveys from over 900 users of commercially-moderated platforms, illustrates a paradox where user-moderated platforms, despite their greater transparency, are not perceived as less toxic compared to their commercially moderated counterparts. Similarly, Patel, C. L. Cook, and Wohn (2021) engaged 902 users across six social media platforms to solicit opinions on effective countermeasures against toxicity, proposing strategies that remain largely untested for their efficacy.

Specific user groups and platform contexts have also been explored, as seen in Lyu et al. (2024) focus on the moderation experiences of blind users on TikTok, and Cai, Wohn, and Almoqbel (2021) insights from volunteer moderators on Twitch. Particularly notable is Fox and W. Y. Tang (2017) exploration of harassment experiences among women in online video games, emphasizing the significant impact of both general and sexual harassment on women's participation and highlighting the pivotal role of the video game industry in addressing these challenges. These studies underscore the contextual and demographic specificity of moderation challenges but suffer from low ecological validity, limiting the generalizability of their findings.

In addition to these, others have identified strategies to maximize the efficacy of moderation. Among the most effective were having high transparency for when and why players are being moderated for toxic or disruptive behavior and reinforce good behavior (Lapolla, 2020). Furthermore PRADEL et al. (2024) and Ma, Li, and Kou (2023) delve into the preferences for content moderation through controlled survey studies and player experiences in online multiplayer games, respectively.

Both highlight a significant demand for transparency and fairness in moderation processes, albeit within constrained research designs that question their broader applicability.

The exploration of content moderation in the context of specific games, as in Kou (2021), Kou and Gui (2017), and Kou and Gui (2021), provides a deeper, yet narrow, view into the practices and perceptions within distinct gaming communities. These studies reveal complex dynamics around the use of permanent bans, automated moderation systems, and user reporting mechanisms, particularly within the League of Legends community. However, their focus on singular platforms or punitive measures, alongside methodological limitations, underscores the need for more comprehensive and scalable research approaches.

Overall, these studies highlight a critical tension in content moderation research: the need to balance effective moderation with preserving community autonomy and freedom of expression, all while grappling with the methodological challenge of ensuring ecological validity and generalizability.

Intervention Approaches to Combating Toxicity

In contrast to the empirical observations of moderation experiences, a body of work proposes theoretical frameworks and interventions aimed at mitigating toxic behavior, particularly within gaming environments (Kordyaka and Kruse, 2021; Kordyaka and Kruse, 2021; Kordyaka, Jahn, and Niehaves, 2020). These theoretical propositions try to address toxic behavior through game design and psychological frameworks, suggesting that interventions need to be deeply integrated into the design and operation of online platforms. However, these studies largely rely on theoretical propositions and survey data, highlighting a gap between conceptual models and their practical implementation.

Systematic reviews and surveys of existing moderation approaches (Wijkstra et al., 2023; Frommel, Johnson, and Regan L Mandryk, 2023; Frommel and R. Mandryk, 2022) further contribute to understanding the landscape of interventions against toxicity. They underscore the scarcity of evaluated interventions within commercial settings and the complex relationship between perceived toxicity, social connectedness, and player retention. These insights suggest a critical need for applied research that transcends theoretical discussions to implement and evaluate interventions within real-world gaming environments.

Despite these contributions, the field remains nascent in terms of practical imple-

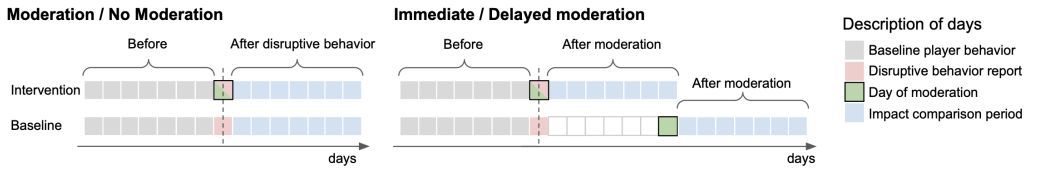


Figure 2.2: Overview of the quasi-experimental setup used for answering our research questions.

mentation and long-term evaluation of proposed interventions. The limitations of current studies, including their low ecological validity and the reliance on self-reported measures, underscore the urgent need for more robust, empirical research that can inform the development of effective, scalable solutions to combat online toxicity particularly in the gaming context.

Estimating Causal Impact from Observational Data

It is difficult to study the effects of moderation in real-world systems actively used by millions of users (Kohavi, D. Tang, and Xu, 2020). In part, this is because the random assignment of moderation by platforms implies not policing some users' bad behavior, thereby exposing other users. This paper joins a burgeoning literature examining the causal effects of moderation on subsequent user behavior (Kaddour et al., 2022; Kreif and DiazOrdaz, 2019). Some approaches to estimating the effect of moderation include fuzzy regression discontinuity designs (Horta Ribeiro, Cheng, and West, 2023), presenting personalized messages to users before they share misinformation (Wojcik et al., 2022), differences-in-differences based on the timing of legislation (Jiménez-Durán, Müller, and Schwarz, 2023), propensity score matching (Zhang et al., 2023), and experimentally assigning reports of harmful content (Jiménez-Durán, 2022).

This paper also engages with the extensive body of work at the nexus of the computer science (CS) and human-computer interaction (HCI) literature. However, a significant portion of this literature is based on correlational evidence from observational data or limited lab studies, which makes it difficult to isolate the assignment of treatment from its effects (see (Jiménez-Durán, 2022) for further discussion on this point).

2.4 Methodology

To examine the causal effect of moderation timing on player behavior in online gaming, we use causal machine learning methodologies (Künzel et al., 2019) to discern causal relationships from observational data. The goal is to estimate the Conditional Average Treatment Effect (CATE) and the Average Treatment Effect (ATE) of moderation and delay in moderation. For CATE, the estimator provides insights into how the treatment effect varies among different sub-groups of players, based on their individual characteristics or covariates. This level of granularity is crucial for understanding if the treatment (moderation in this case) has different impacts on players with varying types, such as by different total scores, damage taken, and damage taken, for example. By contrast, the ATE reflects the overall average impact of the treatment across all players.

Unlike traditional predictive models, these techniques aim to estimate the effect of interventions or treatments in a non-experimental setting. Table 2.1 discusses some commonly used learners in causal ML to estimate CATE and ATE.

Quasi-Experimental Design

The ideal examination of moderation effects as well as the impact of moderation delay would involve the random assignment of players into different experimental groups.

Unfortunately, such an experiment would necessitate allowing some players to engage in disruptive behavior without any consequences. This is naturally not ideal in real-world systems affecting a large number of players every day. We therefore resort to the analysis of observational data where certain moderation actions have already been applied. Following a quasi-experimental setup as depicted in Figure 2.2, we designate synthetic control groups.

Denote the report date as T and the moderation date as M. For estimating the effect of moderation, we select players who were reported for an offense and were moderated on the same day as treatment ($M = T$) and players who were reported for an offense and moderated with a substantial delay ($M >= T+7$) as control. We need to include players that were eventually moderated as our control, due to the known noise in the reporting of cheating and toxic behavior by other players (i.e., players may be reported unfairly (Beres et al., 2021)). As such we only include players with human moderation team verified reports of offensive behavior.

For estimating the effect of delay, we analyze the impact of delayed moderation

by comparing players moderated quickly in the first week ($M \leq T + 3$) against those moderated with substantial delay in the second week ($M >= T + 7$). This method encounters challenges due to possible inherent disparities between players moderated early and later. Players moderated in the first week may exhibit more severe infractions, while delays for others may be caused by factors like moderator backlog. This discrepancy could undermine the reliability of using the quasi-random timing of moderation actions for comparison. To mitigate this, we utilize propensity score estimation (Rosenbaum and Rubin, 1983), following approaches used in observational data studies for causal inference (refer to (Lechner, 2010) for an exhaustive review). This technique enables us to balance the player groups based on observed confounders for a more accurate comparison, essentially creating a controlled virtual experiment to evaluate the effects of moderation timing.

Learner	**Setup**	**Mathematical Formulation**
T Learner	Two separate models: one for treatment and one for control.	$\hat{\tau}(X) = \hat{Y}_T(X) - \hat{Y}_C(X)$, where $\hat{Y}_T(X) = f_T(X)$ and $\hat{Y}_C(X) = f_C(X)$.
X Learner	Designed for imbalance in treatment and control groups, crossing information.	Estimates treatment effects for both groups, then uses a weighting scheme to learn from the opposite group.
R Learner	Based on Robinson transformation; semi-parametric approach.	Regresses outcome on covariates and treatment to get residuals, then regresses residuals on treatment.
S Learner	Single model approach with treatment indicator as a feature.	$\hat{Y}(X, W) = f(X, W)$. CATE estimated by comparing predictions for $W = 1$ and $W = 0$.
DR Learner	Combines propensity score weighting and regression adjustment.	Adjusts for propensity score $e(X) = P(W = 1\|X)$ and outcome regression $\mu(W, X)$.

Table 2.1: Summary of Causal ML Techniques

Causal Machine Learning Modeling

We select the estimation method by comparing several different learners and the distribution of CATE. FigureS2 shows the results estimated by other learners. In this study, both the CATE and the ATE are estimated using the Doubly Robust (DR) Learner estimator. The DR Learner combines the propensity score model (PSM) and the outcome regression model (ORM) to offer robustness against potential model misspecifications (Sant'Anna and Zhao, 2020) : the DR approach can still yield unbiased estimates even if only one of these models is correctly specified. This feature is crucial in situations where model assumptions may not be strictly met, as

is often the case in observational studies.

Additionally, the DR Learner is superior in handling heterogeneous treatment effects, allowing for a more nuanced understanding of how treatment impacts different subgroups. It can also incorporate non-linear relationships and interaction terms more effectively than linear regression. Another significant advantage is the flexibility in choosing different Meta Learners for both components of the model, enabling the use of more sophisticated approaches like machine learning algorithms when necessary. This flexibility and robustness make the DR Learner a valuable tool for treatment-effect estimation, especially in complex datasets where traditional methods like linear regression may not be sufficient. When deploying the DR Learner, we also perform a two-stage model selection by first selecting the Meta Learner, which is linked to estimating potential outcomes among the treatment and control groups, and then selecting the Treatment Effect Learner which is linked to estimating the treatment effects using potential outcomes among two groups. We show the model selection detail in the appendix.

Consider players subjected to moderation (treatment group, $W = 1$) or not (control group, $W = 0$). Let X represent the covariates (player characteristics) and Y be the outcome of interest (e.g., repeat report rate). Using this notation, we can define the CATE as follows.

$$\widehat{CATE}(X) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i(Y_i - \mu(1, X_i))}{e(X_i)} + \mu(1, X_i)\right) - \left(\frac{(1 - W_i)(Y_i - \mu(0, X_i))}{1 - e(X_i)} + \mu(0, X_i)\right) \tag{2.1}$$

again, where W_i indicates if player i is in the treatment group, Y_i is the observed outcome, $\mu(W, X)$ is the outcome predicted by the ORM, and $e(X)$ is the propensity score estimated by PSM.

The PSM estimates the probability of receiving treatment (moderation) given covariates:

$$e(X) = P(W = 1|X) \tag{2.2}$$

But the ORM predicts the potential outcome for both treatment and control conditions given the covariates:

$$\mu(W, X) = E[Y|W, X] \tag{2.3}$$

Simultaneously, the DR estimator also allows for the estimation of ATE. This gives a broader view of the effectiveness of the moderation. The ATE is calculated as the

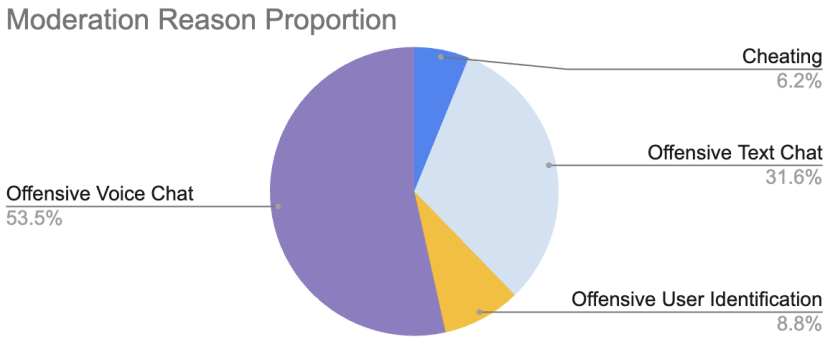


Figure 2.3: Breakdown of unique player proportion by moderation reason in our dataset. Data generated from Feb 1st, 2023 to April 12th, 2023.

average of the estimated CATEs:

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N}\widehat{CATE}(X_i) \tag{2.4}$$

Metric Selection

To assess the impact of moderation on players, we utilize two key metrics to quantify disruptive behavior and player experience: the report and participation rates. The report rate, calculated as the ratio of reports received to matches played each week, allows us to directly track the frequency of disruptive behavior as perceived by the community. We also restrict to reports of the same report reason as the moderation reason to more accurately capture the effect of moderation.

The participation rate measures the proportion of days within a week where a player actively engages in at least one match. We have chosen the participation rate as a metric for player experience because it directly reflects a player's active choice to engage with the game. Unlike passive metrics (e.g., logins), participation in matches indicates a conscious decision to invest time and effort.

2.5 Data

In Call of Duty®: Modern Warfare®II, if players suspect other players are cheating or feel a player is using language or a name that they find offensive, players have the opportunity to report other players with the option of selecting corresponding report reasons. This might happen during or outside of a match. Moderation review is made after the report is generated and moderation actions can be made case by case. If the disruptive behavior is confirmed, certain actions could be implemented based

on the given category of moderation reason.[1] There are four possible moderation reasons: Cheater, Offensive Voice Chat, Offensive Text Chat, and Offensive User ID. Figure 2.3 summarizes the proportion of all possible moderation reasons in our dataset. Table 2.3 summarizes possible actions that could be taken given each moderation reason. Among each category of offensive type, there are multiple actions being taken. These actions can be further broken down into moderate actions and severe actions, according to the severity of these actions. Table 2.2 shows the correspondence of actions and moderation reason, as well as the assigned severity of each action.

Offense Type	**List of Moderation Actions Taken**	**Severity**
Cheater	Remove From Leaderboard (1)	N/A
Offensive Text Chat	Warning Notice, Ban Feature (3)	Mild
Offensive Text Chat	**Penalty Notice**, Ban Feature (2)	Severe
Offensive User ID	Rename User, Adjust Rename Token, Update Clantag, Penalty Notice (4)	Mild
Offensive User ID	Rename User, Adjust Rename Token, Update Clantag, Penalty Notice, **Ban Feature** (5)	Severe
Offensive Voice Chat	Ban Feature (6)	Mild
Offensive Voice Chat	Ban Feature, **Ban Feature, Penalty Notice** (7)	Severe

Table 2.2: Classification of Moderation Actions Based on Severity. Several actions can be taken in response to particular disruptive player behavior. We make the additional actions taken in the severe category bold.

Data Collection

To answer the research questions, we pull the whole player report data generated from Feb 1st, 2023 to April 12th, 2023.[2] The database contains reported players, reporting players, and the reason for reporting. We also join the moderation records with the report data. The moderation record contains reported players, moderation date, reporting players associated with the action, and the specific actions taken. Note that one moderation action could target multiple reports and be directly linked to multiple reporting players. We then join these two tables to get reports with moderation information. We want to focus on players who are verified to have had

[1] A detailed explanation on moderation in COD can be found in https://www.callofduty.com/blog/2024/01/call-of-duty-ricochet-modern-warfare-iii-warzone-anti-cheat-progress-report

[2] For voice chat reports, we use March 21st to April 12th, for the voice chat moderation system was not in place before that.

Moderation Reason	Actions Taken	Ratio
Cheater	Remove From Leaderboards	97.37%
Cheater	Ranking Service, Remove From Leaderboards	2.63%
Offensive text chat	Ban Feature	0.09%
Offensive text chat	Penalty Notice	0.05%
Offensive text chat	Penalty Notice, Ban Feature	99.55%
Offensive text chat	Warning Notice, Ban Feature	0.31%
Offensive user identification	Delete Profile, Rename User, Adjust Rename Token	0.01%
Offensive user identification	Delete Profile, Rename User, Adjust Rename Token, Remove Clantag	0.09%
Offensive user identification	Delete Profile, Rename User, Adjust Rename Token, Remove Clantag	0.14%
Offensive user identification	Adjust Rename Token, Update Clantag, Penalty Notice, Ban Feature	0.03%
Offensive user identification	Warning Notice	0.19%
offensive user identification	Rename User, Adjust Rename Token	1.15%
Offensive user identification	Rename User, Adjust Rename Token, Penalty Notice, Ban Feature	0.10%
Offensive user identification	Rename User, Adjust Rename Token, Update Clantag, Penalty Notice	5.49%
Offensive user identification	Rename User, Adjust Rename Token, Update Clantag, Penalty Notice, Ban Feature	92.81%
Offensive voice chat	Ban Feature	26.75%
Offensive voice chat	Ban Feature, Ban Feature, Penalty Notice	41.78%
Offensive voice chat	Ban Feature, Penalty Notice	0.01%
Offensive voice chat	Warning Notice	0.01%
Offensive voice chat	Warning Notice, Ban Feature	31.46%

Table 2.3: Moderation Actions and Their Relative Frequency in Our Dataset. In many cases, multiple actions are taken simultaneously to moderate the player.

disruptive behavior to avoid known issues with misreporting (Kou and Gui, 2021). Therefore, for each player who is reported on day T, we only keep the sample if the player is moderated within 14 days and when the moderation can be directly linked to reports. In other words, denote the day they are moderated as M, we only keep those with $M \in [T, T + 14]$. To address cases when one moderation event targets multiple reports, we keep the first record of the report associated with each moderation record as the report date.

Constructing Comparison Groups

Next, to measure the effect and compare activities before and after moderation, we create a time series for each player. Specifically, for each player, we extract the sample period t such that $t \in [T - 7, T + 6]$ and period $t \in [M, M + 6]$. We

calculate the number of reports each player receives on these days with the same report reason the player is moderated, as well as the number of matches they play. To test our hypothesis, we construct two different outcome variables Y_i. The first one is the change in the weekly report rate. The report rate is calculated for each week w as $ReportRate_w = \frac{\text{Num Reports Reveived}_w}{\text{Num Matches Played}_w}$, and the change in report rate is measured as $Y_{report} = ReportRate_{w1} - ReportRate_{w0}$, where $w0 = [T-7, T-1]$, and $w1 = [T, T+6]$. The other outcome variable is the change in engagement rate. The engagement rate is calculated as the proportion of days in a week with any matches played, $Engagement_w = \frac{\text{Num days with Matches Played}_w}{7}$, and the change in engagement rate is measured as $Y_{engagement} = Engagement_{w1} - Engagement_{w0}$, where $w0 = [T-7, T-1]$, and $w1 = [T, T+6]$. [3] Following our design, we calculate the outcome measure for the treatment group for whom $M = T$ and the control group for whom $M >= T + 7$

In Figure 2.1 we report the delay in moderation by different disruptive player behavior categories. This delay is measured as the number of days from when the first report about a player was submitted until the time the moderator examined the reported case. This can be due to the need for manual inspection, or even in an automated system due to the need for accumulation of sufficient evidence to justify taking an action (e.g., the threshold of reports). To test our hypothesis on delayed moderation, we measure the change in report rate and engagement rate using a different time window. Specifically, $Y_{report,delay} = ReportRate_{w1} - ReportRate_{w0}$, where $w0 = [T-7, T-1]$, and $w1 = [M, M+6]$; $Engagement_w = \frac{\text{Num days with Matches Played}_w}{7}$, and the change in engagement rate is measured as $Y_{engagement,delay} = Engagement_{w1} - Engagement_{w0}$, where $w0 = [T-7, T-1]$, and $w1 = [M, M+6]$; Following our design, we calculate the outcome measure for the treatment group for whom $M <= T + 3$ and the control group for whom $M >= T + 7$. A detailed description of the data pre-processing can be found in the Appendix.

[3] For cheaters, there is a shadow-ban system that could forbid some players from joining a match, which could result in over-estimating the impact on engagement rate. We therefore exclude those players from the dataset.

2.6 Results

In the graphs of Figure 2.4, we present the estimated effects of moderation measures on altering player behavior, specifically regarding the recurrence of reports for misconduct or repeated offenses (RQ1). Figure 2.5 report the effect of moderation on user experience. (RQ2). In both graphs, we depict the effects of moderation vs. no moderation, delayed moderation vs. immediate moderation, as well as the effects of different severity.

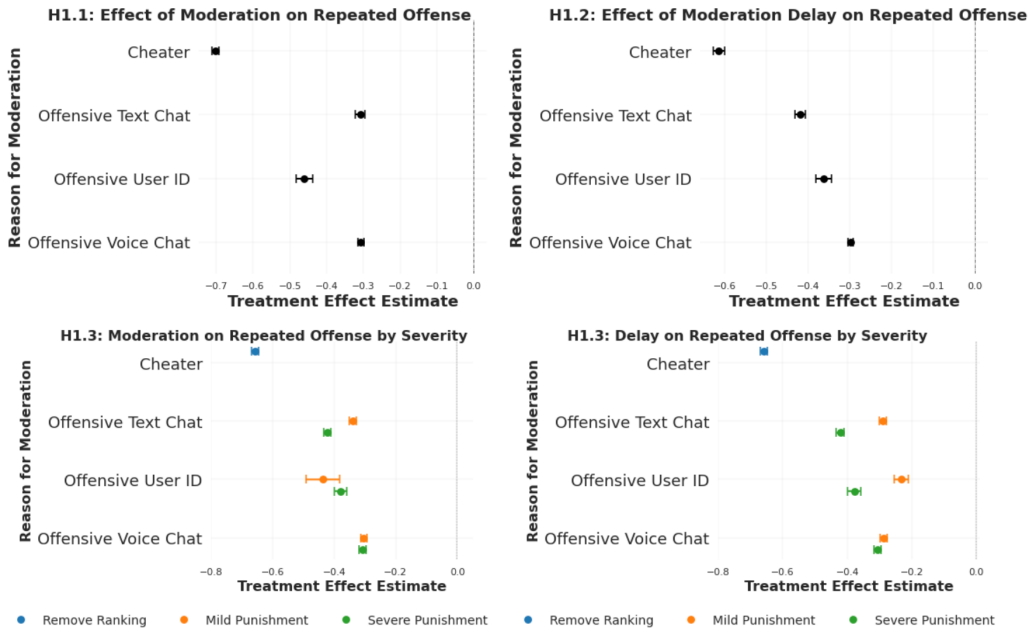


Figure 2.4: Top: Effect of moderation/no moderation, delayed moderation/immediate moderation on repeated offense rates (a measure of toxicity). Bottom: Breakdown by action severity (RQ1).

Impact of Moderation on Repeated Disruptive Behavior (RQ1)

Testing Hypothesis 1.1 : Moderation on online gaming platforms leads to decreased disruptive behaviors among moderated players.

The evidence supports Hypothesis 1, demonstrating a clear decrease in disruptive behaviors among players who were moderated for various infractions. For instance, moderation targeted at cheating resulted in a substantial -70.33% (95% CI: -71.25%, -69.41%) reduction in subsequent report rates, indicating a significant deterrence effect. Similarly, moderation for offensive text chat and Offensive User ID led to reductions in report rates by -30.84% (95% CI: -32.08%, -20.60%) and -46.05% (95% CI: -48.29%, -43.81%), respectively. These figures suggest that moderation

effectively curtails repeated offenses, confirming the hypothesis that moderation leads to decreased disruptive behaviors among moderated players (Hypothesis 1.1 supported).

Testing Hypothesis 1.2: Delayed moderation is less effective at reducing disruptive behaviors compared to immediate moderation.

The evidence suggests a significant deterrent effect of immediate moderation on repeat offenses: a -61.44% (95% CI: -62.78%, -60.10%) decrease in cheating, a -41.89% (95% CI: -43.05%, -40.72%) reduction in Offensive Text Chat incidents, a -36.36% (95% CI: -38.23%, -34.49%) decline in Offensive User ID cases, and a -29.81% (95% CI: -30.51%, -29.12%) drop in Offensive Voice Chat offenses. These findings lend robust support to hypothesis 1.2, positing that immediate moderation is more effective than delayed actions in curtailing disruptive behaviors on gaming platforms (Hypothesis 1.2 supported).

Testing Hypothesis 1.3 : More severe moderation on online gaming platforms leads to more decrease in disruptive behaviors among moderated players.

Figure 2.4 further breaks down the effects of different actions among each moderation reason group according to Table 2.2. We can better understand the impact of these moderation strategies by looking at the variation in severity and the effect size. The evidence does not support Hypothesis 1.3. Severe actions for Offensive Text Chat result in a reduction of -30.58%, 95% CI (-33.02%, -28.14%) while mild action leads to a decrease of -33.91%, with 95% CI (-36.38%, -31.44%); Severe action for Offensive Voice Chat results in a reduction of -25.48% with 95% CI (-28.14%, -22.82%) as compared to mild actions -30.35% with 95% CI (-32.44%, -28.26%). Severe actions for Offensive User ID results in a reduction of -43.69% with 95% CI (-54.45%, -32.92%) as compared to mild action at -47.11% with 95% CI (-51.93%, -42.29%). Evidence indicates that severe and mild actions have at least similar effects in reducing disruptive behavior, with mild actions being marginally better (Hypothesis 1.3 rejected).

Testing Hypothesis 1.4: The effectiveness of severe moderation interventions in reducing disruptive behavior is enhanced when those interventions are implemented immediately.

Evidence in the figure also partially supports Hypothesis 1.4. In terms of effects on subsequent disruptive behavior, the differences between immediate and delayed actions are larger for severe actions compared to mild actions. For Offensive Text

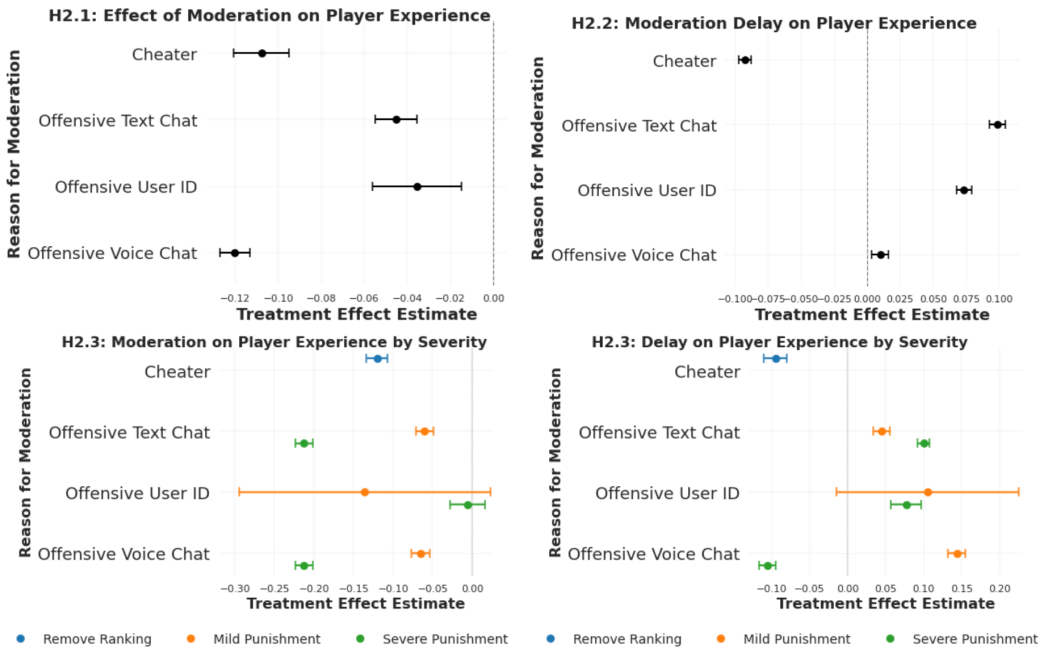


Figure 2.5: Top: Effect of moderation/no moderation, delayed moderation/immediate moderation on player experience (proportion of days with participation). Bottom: Breakdown by action severity (RQ2).

Chat, the difference is -42.21% (95% CI: -43.38%, -41.04%) for severe actions versus -29.04% (95% CI: -30.12%, -27.96%) for mild actions; for Offensive User ID, the difference is -37.83% (95% CI: -39.80%, -35.85%) for severe actions versus -23.28% (95% CI: -25.54%, -21.03%) for mild actions; for Offensive Voice Chat, the difference is -30.68% (95% CI: -31.81%, -29.54%) for severe actions versus -28.72% (95% CI: -29.87%, -27.57%) for mild actions. We can see the severe actions taken immediately are more effective for Offensive Text Chat and Offensive User ID, but not for Offensive Voice Chat. However, we can see that all the actions are more effective when taken immediately, regardless of the severity (Hypothesis 1.4 partially supported).

Impact of Moderation on Player Experience (RQ2)

The data from our study provides insights not only into the effectiveness of moderation in reducing disruptive behavior, but also into the impact it has on the experience of moderated players. These results are depicted in Fig 2.5 suggest the following:

Testing Hypothesis 2.1: Moderation on online gaming platforms leads to decreased player experience among moderated players.

The provided evidence also supports Hypothesis 2.1, showing an inverse relationship between moderation actions and the experience of the moderated player. Moderation for Cheating was associated with a -10.80% decline in participation, while moderation for Offensive Text Chat and Offensive User ID resulted in decreases of -4.53% (95% CI: -5.51%, -3.55%) and -3.57% (95% CI: -5.64%, -1.50%) in participation, respectively. Moreover, the most significant drop in participation was observed following moderation for Offensive Voice Chat, with a -12.03% (95% CI: -12.73%, -11.32%) reduction in participation. These findings indicate that while moderation is effective in reducing disruptive behavior, it also leads to a decrease in player experience levels among those subjected to moderation actions (Hypothesis 2.1 supported).

Testing Hypothesis 2.2: The negative impact of moderation actions on player gaming experience is amplified when those actions are implemented with a delay, as compared to immediate implementation.

While immediate moderation in response to Cheating led to a -9.28% (95% CI: -9.75%, -8.81%) decrease in player experience, suggesting a potential overreach deterring not only negative behaviors but also reducing general activity, immediate actions against Offensive Text Chat and User ID were associated with increases in player experience for the moderated player by 9.87% (95% CI: 9.25%, 10.49%) and 7.35% (95% CI: 6.79%, 7.91%), respectively). This indicates a positive reception among players to swift actions that foster a respectful community environment. The minimal impact (0.99%(95% CI: 0.37%, 1.62%) increase) observed in the context of Offensive Voice Chat moderation further suggests that the nature of the offense and player perceptions of moderation fairness significantly influence user experience outcomes. Considering these mixed impacts on experience—with immediate moderation leading to both increases and decreases in participation depending on the offense type (Hypothesis 2.2 partially supported).

Testing Hypothesis 2.3: More severe actions result in more reduction in player experience.

Evidence in Figure 2.5 only partly supports this hypothesis. Specifically for Offensive Voice Chat, severe action results in a -21.2% (95% CI (-0.22.35%, -20.19%) decrease in player experience compared to mild action at -6.57%(-7.70%, -5.45%). Also, immediate actions for Offensive Voice Chat result in -10.60% more (95% CI (-12.83%,-8.38%)) reduction in player experience compared to delayed actions. However, the result does not hold for other actions. (Hypothesis 2.3 partially sup-

ported).

Investigating the Heterogeneity of Moderation Effects

To further investigate the heterogeneity of treatment effects across different types of players, we explore the relationship between individual CATE and player features. We plot the correlation between some key player features and the CATE on report rate and participation in Figure 2.6. We observe that the treatment effects on both report rate and participation are strongly correlated with key performance/skill indicators, such as average scores in each match, a damage skill indicator (calculated as average damage dealt / average damage taken), and the average kill-death ratio (calculated as average kills / average deaths). This indicates that players with higher skills exhibit a greater reduction in disruptive behavior following moderation when they continue playing the game. However, they are also more likely to leave the game and stop playing. We suggest that the high correlation between skill and moderation effects can be attributed to the following reasons:

Increased Visibility and Accountability: Better players often have a higher profile within the gaming community, making their actions more visible and, consequently, more susceptible to scrutiny. When such players are moderated, the action may serve as a stronger deterrent not just to the individual but also to the broader community, amplifying the perceived consequences of misconduct.

Behavioral Adjustment Sensitivity: Better players, having demonstrated a capacity to adapt and strategize within the game, might be more adept at adjusting their behavior in response to moderation, reflecting a higher treatment effect in reducing misconduct.

Also, the correlation between the effect on player experience and skill can possibly be attributed to the following:

Higher Opportunity Cost: For better players, the opportunity cost of not participating in the game due to moderation (temporary bans or the choice to disengage following a moderation action) is potentially higher. This is because they might lose more in terms of rankings, status, or even potential streaming viewership if they are also content creators. Thus, the impact on player experience could be more significant for these individuals.

Increased Sensitivity to Platform Experience Changes: Players with higher skill levels might be more sensitive to changes in the platform's social environment or to the imposition of restrictions. Such players might perceive moderation actions as a more significant infringement on their gameplay experience, leading to a greater reduction in player experience.

2.7 Discussion

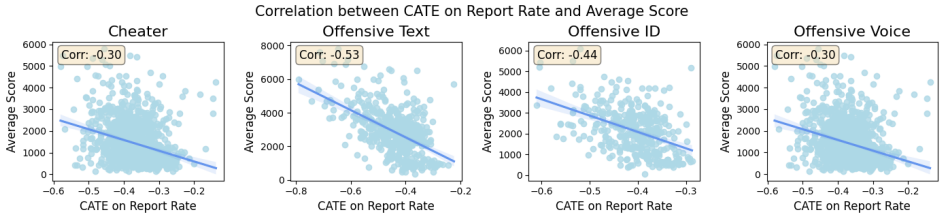
Several observations from our analysis warrant further in-depth interpretation and considerations in future research on the moderation of disruptive player behavior.

Different Behavior of Cheaters and Toxic Players: Our study distinctly illustrates the differential impact of moderation on two forms of misconduct within online platforms: cheating and toxicity. Moderation measures targeting cheating achieved a remarkable reduction in subsequent report rates, highlighting the effectiveness of such interventions in deterring this specific type of misconduct. In contrast, actions against toxicity, manifested through Offensive Text Chat and user IDs, also resulted in significant decreases in report rates, though the magnitude of reduction varied, suggesting that while moderation is universally effective, its impact is nuanced by the nature of the offense. This emphasizes the necessity of adopting targeted moderation strategies that address the specific challenges posed by different forms of misconduct, from direct and tangible cheating to the more subjective and varied instances of toxicity. The varied reductions in report rates between cheating and toxic behaviors also reflect the complexity of moderating diverse forms of misconduct, each requiring a tailored approach to effectively mitigate their occurrence and ensure a healthy online environment.

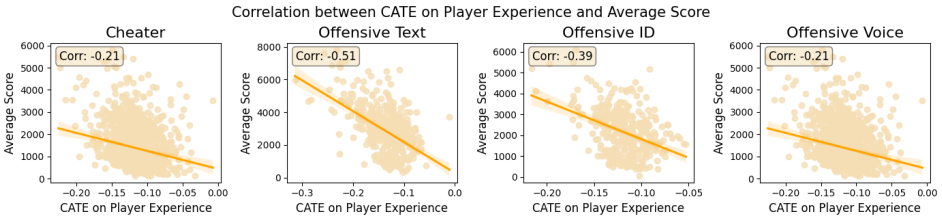
Importance of Moderation Swiftness: Immediate moderation proves more effective in reducing disruptive behaviors than delayed actions, emphasizing the importance of timely intervention. The effects on player experience vary, with some immediate actions enhancing participation, suggesting player appreciation for efforts to maintain a respectful community. Yet, the impact differs based on the offense's nature and the moderation's severity, pointing to the complex relationship between moderation practices and their outcomes on player behavior and platform dynamics. These insights underline the necessity for tailored moderation strategies that consider the timing and severity of actions to optimally balance community standards enforcement with player experience.

Figure 2.6: Correlation of CATE and game features.

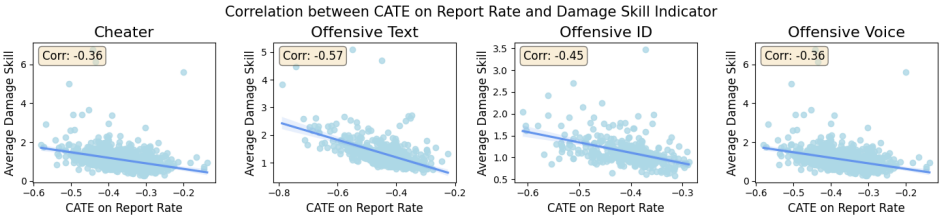
(a)



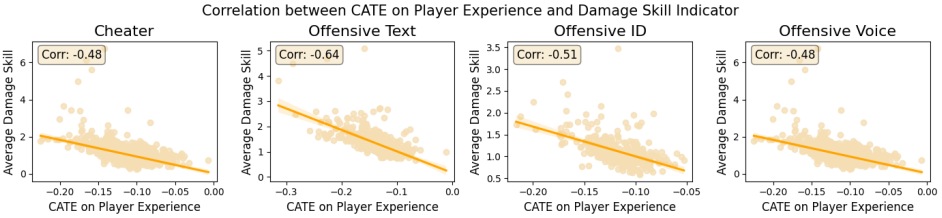
(b)



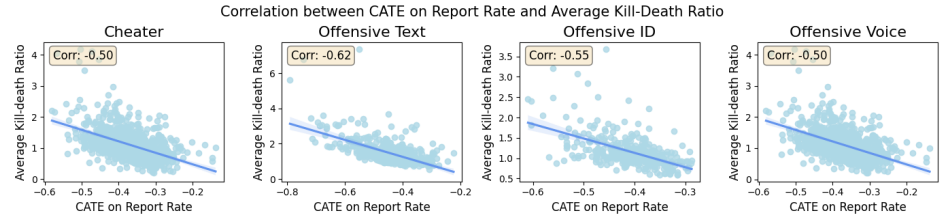
(c)



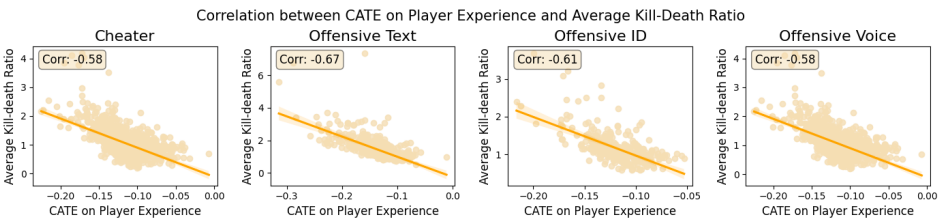
(d)



(e)



(f)



Our analysis suggests that prompt moderation plays a crucial role in mitigating repeat offenses, effectively upholding community standards, especially in the face of severe violations like cheating. By implementing swift moderation actions, we not only deter negative behavior but also foster a culture of accountability within the gaming community. It is important to note, however, that moderation's impact on player experience varies. For less-severe offenses, a refined approach to moderation is advocated—one that aims to correct behavior without necessarily reducing the days players engage with the platform. This strategy may lead to two outcomes: either players adjust their behavior to comply with community standards and continue to maintain a good and respectful experience on the platform, or those resistant to change and unwilling to alter their disruptive behavior may choose to leave the platform. We note that this also highlights the effectiveness of moderation, as there is no room for such purposefully disruptive players on the platform to begin with. This dual outcome underscores the importance of developing tailored moderation strategies that differentiate between the severity and nature of offenses, ensuring both a healthy community environment and high player experience when being subjected to moderation.

Potential for Tailoring Interventions: The data also provides valuable insights for designing more targeted intervention strategies. By understanding which types of offenses are more responsive to immediate versus delayed moderation, platforms can develop more effective approaches. Designing interventions that result in effective behavior change while keeping the players engaged in a prosocial way is also vital. This could involve incorporating feedback mechanisms, transparency mechanisms (Ma, Li, and Kou, 2023), or response systems to be more effective at promoting prosocial play while preserving a high gaming experience for the moderated players. This is, of course, as long as such players can be effectively motivated to act prosocially and they are respectful of community values..

Threat of Spill-Over from Moderation: Moreover, the timing of moderation actions for specific behaviors, such as offensive chat, can lead to spill-over effects on other forms of toxic behavior. Monitoring changes in these ancillary behaviors in response to moderation actions offers a more comprehensive view of the health of the gaming community. For instance, strict immediate moderation for a particular offense might lead to an increase in different types of toxic behavior as a reactionary response from the player base.

Future Directions

Looking to the future, research and practical applications could explore the use of bandit algorithms to dynamically adjust moderation strategies based on real-time results. Such an approach could continuously optimize the approach so that it maximizes effectiveness in enforcing community standards while preserving the highest possible player experience for the players that require moderation.

2.8 Conclusion

In this study, we employed a unique dataset from the Call of Duty®: Modern Warfare®II gaming community which contains real-world data on gameplay and moderation activities. We used this data to estimate the causal impact of moderation actions on disruptive player behaviors, a topic of growing importance in the digital era. Our comprehensive analysis, grounded in a real-world and widely popular online gaming context, provides novel insights into the dynamics of player interactions and the efficacy of moderation strategies.

Our results reveal the differential impact of moderation resulting in effectively changing the behavior of some players while driving those unwilling to change away from the platform. We further underscore the significance of timely moderation interventions. We observed that immediate moderation is in general, more effective compared to a delayed one, especially for the cheating behavior. In addition, we revealed some complex relationships between the moderation effect, moderation time, and severity of actions. This finding provides much-needed empirical evidence for the necessity of prompt and efficient moderation processes in online gaming spaces.

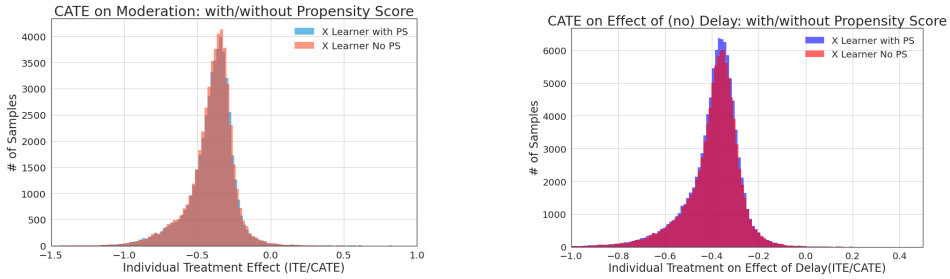
In conclusion, our research makes a substantial contribution to the field by leveraging a unique dataset to make a causal claim about the effectiveness of moderation in online gaming. The findings not only enrich the academic discourse on digital behavior management but also provide practical guidelines for the design of more effective and nuanced moderation strategies in online communities. This study, therefore, serves as a pivotal reference point for future research and practical applications in the rapidly evolving domain of online behavior management.

Acknowledgements We thank Gary Quan, Jonathan Lane, and Michael Vance for helpful comments. Due to the player confidentiality and data privacy policies of ABK, the data and code used in this paper cannot be made available by the researchers.

S1 Supplementary Information

Estimating CATE With and Without the Propensity Score

Fig S1 shows the impact of propensity score correction on the CATE estimated under X-Learner. We note the DR Learner we us for the main results, always uses propensity score, hence we plot the CATE estimates to the X Learner.



(a) X Learner with/without Pscore (b) X Learner with/without Pscore

Figure S1: Robustness Checking: Comparing learners with/without Pscore.

Data Pre-Processing

Our analysis focused on users moderated for cheating, offensive user IDs, or offensive behavior in text or voice chat from 2023-02-01 to 2023-04-19. We considered only the first moderation instance per user. In cases with multiple moderations, the longest delay before moderation was used so as to ensure a pessimistic bound on the effect of a shorter moderation delay.

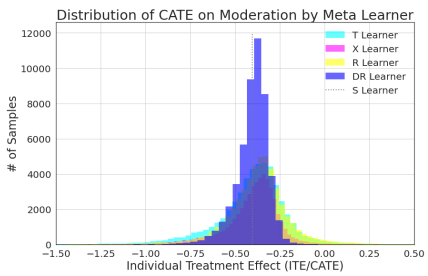
- **Gather Report Data:** Collect data from three different types of player reports (Cheating, Offensive Text Chat, and Offensive User Identification) over a specified date range.
- **Process Moderation Data:** Combine the reports with data on moderation actions, categorizing each report based on the type of moderation action taken (if any).
- **Aggregate Data:** Aggregate the combined data by calculating number of reports by each moderation date, reasons, and actions, for each moderated players, calculating the time delay between each report and the corresponding moderation action. Join the gameplay (number of matches by player-date) records to the report records.
- **Refine for Minimum Lag:** Filter the data to keep only those records where the delay between the report and the moderation action is the minimum for

each reported player, based on the moderation date and reason. Ensure each player has only one record left to eliminate duplicates.

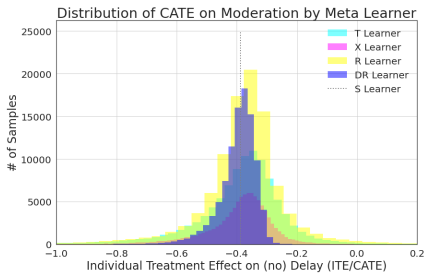
- **Create Datasets for effect of moderation:** Label players by lag of reports, separate players in treatment group (lag = 0) and control group (lag >= 7). Calculate average participation and report rate in the 7-day window before and after receiving reports.
- **Create Datasets for effect of delay:** Label players by lag of reports, separate players in treatment group (lag <= 3) and control group (lag >= 7). Calculate average participation and report rate in the 7-day window before receiving reports and 7-day window after receiving moderation.

Model Selection and Modeling Robustness Evaluation

Selection of Meta Learner: In Figure S2, we display the CATE distribution of multiple Learners, including T Learner, S Learner, X Learner, R Learner, and DR Learner. The two graphs indicate that all the Learners give relatively close results in the range of individual treatment effects (x-axis), providing evidence of the robustness of the result. In particular, the mean and mode of the estimates are close and significantly different from 0. Among all the learners, DR Learner gives the highest concentration among all learners, providing the most consistent estimates. Therefore, we proceed to use the DR Learner for our main results.



(a) Distribution of Conditional Average Treatment Effect (CATE) by Meta-Learner for impact of moderation vs. no moderation.



(b) Distribution of Conditional Average Treatment Effect (CATE) by Meta-Learner for impact of immediate vs. delayed moderation.

Figure S2: Robustness Checking: Comparing different Meta Learners.

Machine Learning Estimator Selection Under Doubly Robust (DR) Meta Learner: We evaluate different base models (machine learning estimators) under the best DR Meta Learner. Figure S3 shows the distribution of CATE treatment effects using DR

Meta Learner while varying the type of the Base Learner. In the first graph, we observe the performance of single models. The Random Forest Learner's distribution has the widest spread, suggesting a higher variance in its treatment effect estimates. The XGB Learner appears to have the narrowest and most symmetric distribution around zero, which suggests less variability and more conservative estimates of the treatment-effect.

Aside from the Base Learner used to estimate outcomes and treatment effects in both the control and treatment groups, we can also select a Treatment-Effect Learner. This model is used to estimate the treatment effects in the treatment group. Based on the best-performing Base Learner (XGB in our case), we further vary the machine learning estimator used for the treatment effect Learner. The second graph shows Combined Learners. Here, the XGB+Linear Learner's distribution is the narrowest, implying highly consistent treatment effect estimates. The XGB+RF Learner and XGB+XGB Learner have wider distributions, indicating less consistency in their predictions.

Similarly to effect of moderation, we also perform the same two-stage Learner selection for the effect of delay. In the third graph for single models, the Linear Learner has the highest peak with the smallest spread, indicating the least variance in estimation. Conditional on the Linear-Base Learner, we further select Linear Learner as the treatment effect estimator according to the fourth graph since it offers the most consistent treatment-effect estimates

Propensity Modeling: To correct for the potential systematic difference in participant assignment to treatment vs. control, we calculate the propensity score (balancing score) and estimate its impact on outcome estimates. Figure S4 gives the distribution of propensity score among treatment and control groups. The graphs indicate that in both propensity estimation, propensity score distribution for treatment and control group are in a similar range. The difference in peak is due to the treatment/control group size being more balanced in the data we use for the effect of moderation, where the number of players who are moderated and not are roughly the same. For the effect of delay, since majority of the players are moderated within 3 days, we have fewer players in the control group.

In both graphs, treatment and control groups overlap significantly, implying that for any given propensity score, there are individuals from both the treatment and control groups.

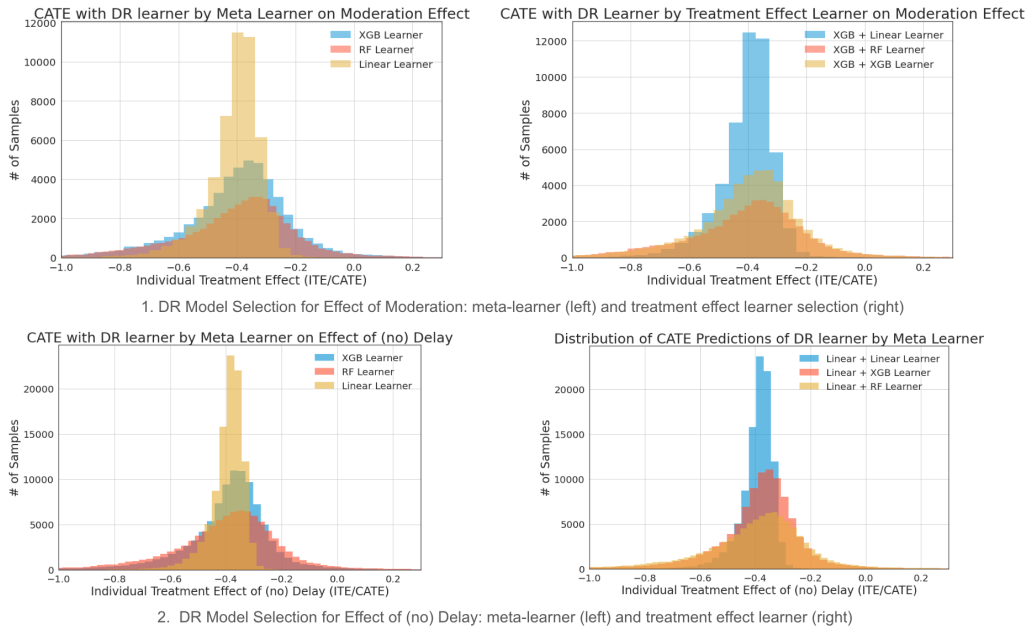


Figure S3: Based model selection (Machine Learning Estimator) among DR model class. Top left & bottom Left: selecting general Meta-Learner. Top right & bottom Right: selecting Treatment Effect Learners.

Figure S4 also shows the contribution of the each feature to the propensity score after we do propensity score matching with KNN. The graph shows that most features become more balanced after matching with a smaller effect size. which provides evidence on the validity of the propensity score. The graphs provide reassurance that the propensity scores are capturing information on the likelihood of treatment.

In Table S1 we summarize the gaming features we use to estimate the propensity score and display the mean value of each group.

References

ABK (Apr. 2023). *Call of Duty Code of Conduct | FPS Game Terms*. `https://www.callofduty.com/values`. (Accessed on 04/19/2023).

– (Jan. 2024). *Anti-Toxicity / Disruptive Behavior Progress Report – Modern Warfare III Season 2*. `https://www.callofduty.com/blog/2024/01/call-of-duty-ricochet-modern-warfare-iii-warzone-anti-cheat-progress-report`. (Accessed on 02/12/2024).

Abramowitz, Ann J and Susan G O'Leary (1990). "Effectiveness of delayed punishment in an applied setting". In: *Behavior Therapy* 21.2, pp. 231–239.

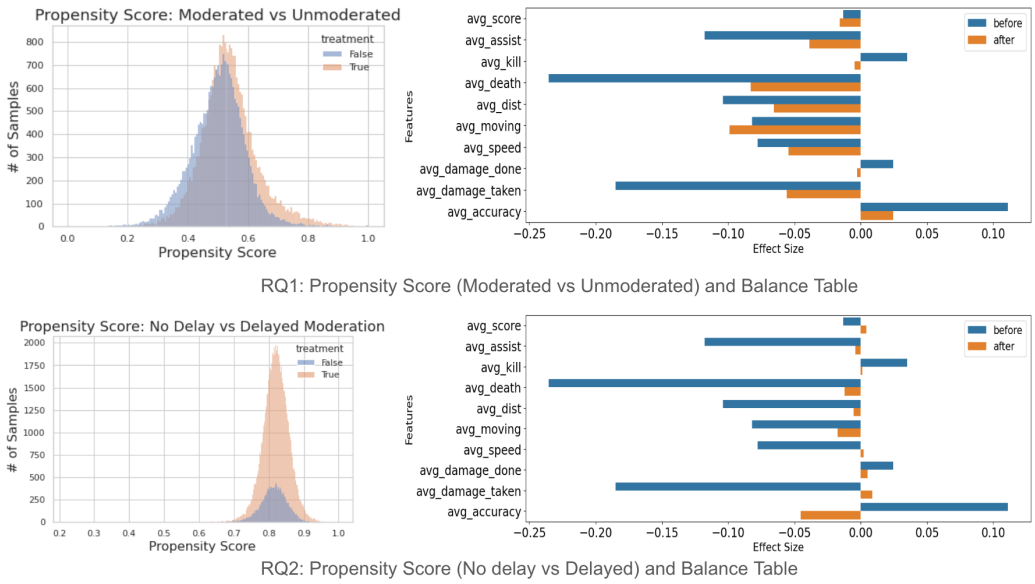


Figure S4: Propensity Score distribution and feature balance. Top left & bottom left: Distribution propensity scores. Top right & bottom right: standardized mean differences across covariates before and after matching.

Features (mean)	**Unmoderated**	**Moderated**	**Delayed Moderation**	**Immediate Moderation**
Score	2156.4	2137.5	2250.1	2257.6
Assist	3.3	3.0	3.4	3.3
Kill	15.1	15.4	15.7	16.1
Death	13.2	11.7	13.4	12.6
Distance	42646.9	41099.2	43232.9	42612.5
Moving	77.9	76.8	78.4	78.1
Speed	135.7	132.7	137.3	136.6
Damage done	1619.5	1649.8	1680.4	1711.2
Damage taken	1447.5	1300.2	1472.7	1396.4
Accuracy	20.6	21.1	20.8	21.1

Table S1: Feature Balance Table

Aguerri, Jesús C, Mario Santisteban, and Fernando Miró-Llinares (2023). "The Enemy Hates Best? Toxicity in League of Legends and Its Content Moderation Implications". In: *European Journal on Criminal Policy and Research*, pp. 1–20.

Ajzen, Icek (2020). "The theory of planned behavior: Frequently asked questions". In: *Human Behavior and Emerging Technologies* 2.4, pp. 314–324.

Beres, Nicole A et al. (2021). “Don’t you know that you’re toxic: Normalization of toxicity in online gaming”. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–15.

Bourgonjon, Jeroen et al. (2016). “Players’ perspectives on the positive impact of video games: A qualitative content analysis of online forum discussions”. In: *New Media & Society* 18.8, pp. 1732–1749.

Cai, Jie, Donghee Yvette Wohn, and Mashael Almoqbel (2021). “Moderation visibility: Mapping the strategies of volunteer moderators in live streaming micro communities”. In: *ACM International Conference on Interactive Media Experiences*, pp. 61–72.

Canossa, Alessandro et al. (2021). “For honor, for toxicity: Detecting toxic behavior through gameplay”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CHI PLAY, pp. 1–29.

Cook, Christine et al. (2019). “For whom the gamer trolls: A study of trolling interactions in the online gaming context”. In: *Journal of Computer-Mediated Communication* 24.6, pp. 293–318.

Cook, Christine L, Aashka Patel, and Donghee Yvette Wohn (2021). “Commercial versus volunteer: Comparing user perceptions of toxicity and transparency in content moderation across social media platforms”. In: *Frontiers in Human Dynamics* 3, p. 626409.

Cullen, Amanda LL and Sanjay R Kairam (2022). “Practicing moderation: Community moderation as reflective practice”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW1, pp. 1–32.

Fong, Carlton J et al. (2016). “Deconstructing constructive criticism: The nature of academic emotions associated with constructive, positive, and negative feedback”. In: *Learning and Individual Differences* 49, pp. 393–399.

Fox, Jesse and Wai Yen Tang (2017). “Women’s experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies”. In: *New media & society* 19.8, pp. 1290–1307.

Frommel, Julian, Daniel Johnson, and Regan L Mandryk (2023). “How perceived toxicity of gaming communities is associated with social capital, satisfaction of relatedness, and loneliness”. In: *Computers in Human Behavior Reports* 10, p. 100302.

Frommel, Julian and Regan Mandryk (2022). “Effective Toxicity Prediction in Online Multiplayer Gaming: Four Obstacles to Making Approaches Usable”. In: *Mensch und Computer 2022-Workshopband*.

Frommel, Julian, Regan L Mandryk, and Madison Klarkowski (2022). “Combating Game-based Toxicity and Harassment: Challenges to Ensuring Safe and Healthy Spaces in Esports”. In: *BOOK OF ABSTRACTS*, p. 87.

Frommel, Julian and Regan L. Mandryk (Dec. 2023). "Individual Control over Exposure to Combat Toxicity in Games". In: *ACM Games* 1.4. DOI: 10.1145/3633768. URL: https://doi.org/10.1145/3633768.

GGWP (Nov. 2023). *Game Changer: Impact of Chat Sanctions on Toxicity–GGWP–the first AI-powered game moderation platform*. https://www.ggwp.com/blog/game-changer-impact-of-chat-sanctions-on-toxicity/. (Accessed on 02/11/2024).

Ghosh, Ayushi (2021). "Analyzing Toxicity in Online Gaming Communities". In: *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.10, pp. 4448–4455.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach (2020). "Algorithmic content moderation: Technical and political challenges in the automation of platform governance". In: *Big Data & Society* 7.1, p. 2053951719897945.

Horta Ribeiro, Manoel, Justin Cheng, and Robert West (2023). "Automated Content Moderation Increases Adherence to Community Guidelines". In: WWW '23, pp. 2666–2676. DOI: 10.1145/3543507.3583275. URL: https://doi.org/10.1145/3543507.3583275.

Jiménez-Durán, Rafael (2022). *The economics of content moderation: Theory and experimental evidence from hate speech on Twitter*. Tech. rep. 324. Chicago, IL: University of Chicago Booth School of Business, Stigler Center for the Study of the Economy and the State.

Jiménez-Durán, Rafael, Karsten Müller, and Carlo Schwarz (2023). "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG". In: URL: http://dx.doi.org/10.2139/ssrn.4230296.

Kaddour, Jean et al. (2022). "Causal machine learning: A survey and open problems". In: *arXiv preprint arXiv:2206.15475*.

Kahneman, Daniel and Amos Tversky (2013). "Prospect theory: An analysis of decision under risk". In: *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, pp. 99–127.

Kim, Sang Hee et al. (2015). "Individual differences in sensitivity to reward and punishment and neural activity during reward and avoidance learning". In: *Social cognitive and affective neuroscience* 10.9, pp. 1219–1227.

Klepper, Steven and Daniel Nagin (1989). "The deterrent effect of perceived certainty and severity of punishment revisited". In: *Criminology* 27.4, pp. 721–746.

Kocielnik, Rafal et al. (2023). "Challenges in Moderating Disruptive Player Behavior in Online Competitive Action Games". In: *Frontiers in Computer Science* 6, p. 1283735.

Kohavi, Ron, Diane Tang, and Ya Xu (2020). *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press.

Kordyaka, Bastian, Katharina Jahn, and Bjoern Niehaves (2020). "Towards a unified theory of toxic behavior in video games". In: *Internet Research* 30.4, pp. 1081–1102.

Kordyaka, Bastian and Björn Kruse (2021). "Curing toxicity–developing design principles to buffer toxic behaviour in massive multiplayer online games". In: *Safer communities* 20.3, pp. 133–149.

Kordyaka, Bastian, Samuli Laato, et al. (2023). "The Cycle of Toxicity: Exploring Relationships between Personality and Player Roles in Toxic Behavior in Multiplayer Online Battle Arena Games". In: *Proceedings of the ACM on Human-Computer Interaction* 7.CHI PLAY, pp. 611–641.

Kordyaka, Bastian, Solip Park, et al. (2023). "Exploring the relationship between offline cultural environments and toxic behavior tendencies in multiplayer online games". In: *ACM Transactions on Social Computing*.

Kou, Yubo (2021). "Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2, pp. 1–21.

Kou, Yubo and Xinning Gui (2017). "When code governs community". In.

– (2021). "Flag and flaggability in automated moderation: The case of reporting toxic behavior in an online game community". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–12.

Kreif, Noemi and Karla DiazOrdaz (2019). "Machine learning in policy evaluation: new tools for causal inference". In: *arXiv preprint arXiv:1903.00402*.

Kriz, Willy C (2020). "Gaming in the Time of COVID-19". In: *Simulation & Gaming* 51.4, pp. 403–410.

Künzel, Sören R et al. (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10, pp. 4156–4165.

Kwak, Haewoon, Jeremy Blackburn, and Seungyeop Han (2015). "Exploring cyberbullying and other toxic behavior in team competition online games". In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 3739–3748.

Lapolla, Matthew (2020). "Tackling Toxicity: Identifying and Addressing Toxic Behavior in Online Video Games". In.

Lechner, Michael (2010). "The Estimation of Causal Effects by Difference-in-Difference Methods". In: *Foundations and Trends® in Econometrics* 4.3, pp. 165–224. DOI: 10.1561/0800000014.

Lee, Sung Je, Eui Jun Jeong, and Joon Hyun Jeon (2019). "Disruptive behaviors in online games: effects of moral positioning, competitive motivation, and aggression in "League of Legends"". In: *Social Behavior and Personality: an international journal* 47.2, pp. 1–9.

Levkovitz, Zohar (Apr. 2023). *Content moderators alone can't clean up our toxic internet*. https://www.fastcompany.com/90515733/content-moderators-alone-cant-clean-up-our-toxic-internet. (Accessed on 04/20/2023).

Liau, Albert K et al. (2015). "Impulsivity, self-regulation, and pathological video gaming among youth: Testing a mediation model". In: *Asia Pacific Journal of Public Health* 27.2, NP2188–NP2196.

Link, Daniel, Bernd Hellingrath, and Jie Ling (2016). "A Human-is-the-Loop Approach for Semi-Automated Content Moderation." In: *ISCRAM*.

Lochman, John E (1992). "Cognitive-behavioral intervention with aggressive boys: three-year follow-up and preventive effects." In: *Journal of consulting and clinical psychology* 60.3, p. 426.

Lyu, Yao et al. (2024). ""' I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.": A Study of Blind TikTokers' Content Moderation Experiences". In: *arXiv preprint arXiv:2401.11663*.

Ma, Renkai, Yao Li, and Yubo Kou (2023). "Transparency, Fairness, and Coping: How Players Experience Moderation in Multiplayer Online Games". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21.

Märtens, Marcus et al. (2015). "Toxicity detection in multiplayer online games". In: *2015 International Workshop on Network and Systems Support for Games (NetGames)*. IEEE, pp. 1–6.

Paschke, Kerstin et al. (2021). "Adolescent gaming and social media usage before and during the COVID-19 pandemic". In: *Sucht*.

Patel, Aashka, Christine L Cook, and Donghee Yvette Wohn (2021). "User Opinions on Effective Strategies Against Social Media Toxicity". In: *Proceedings of the Annual Hawaii International Conference on System Sciences*.

PRADEL, FRANZISKA et al. (2024). "Toxic Speech and Limited Demand for Content Moderation on Social Media". In: *American Political Science Review*, pp. 1–18.

Pratt, Travis C et al. (2017). "The empirical status of deterrence theory: A meta-analysis". In: *Taking stock*. Routledge, pp. 367–395.

Raith, Lisa et al. (2021). "Massively multiplayer online games and well-being: A systematic literature review". In: *Frontiers in Psychology* 12, p. 698799.

Rieder, Bernhard and Yarden Skop (2021). "The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API". In: *Big Data & Society* 8.2, p. 20539517211046181.

Rosenbaum, Paul R and Donald B Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55.

Sant'Anna, Pedro HC and Jun Zhao (2020). "Doubly robust difference-in-differences estimators". In: *Journal of Econometrics* 219.1, pp. 101–122.

Srikanth, Maya et al. (2021). "Dynamic social media monitoring for fast-evolving online discussions". In: *arXiv preprint arXiv:2102.12596*.

Srinivasan, Kumar Bhargav et al. (2019). "Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–21.

Steinkuehler, Constance (2023). "Games as social platforms". In: *ACM Games: Research and Practice* 1.1, pp. 1–2.

Stoop, Wessel et al. (2019). "Detecting harassment in real-time as conversations develop". In: *Proceedings of the Third Workshop on Abusive Language Online*, pp. 19–24.

Thurlings, Marieke et al. (2013). "Understanding feedback: A learning theory perspective". In: *Educational Research Review* 9, pp. 1–15.

Trepte, Sabine, Leonard Reinecke, and Keno Juechems (2012). "The social side of gaming: How playing online computer games creates online and offline social support". In: *Computers in Human behavior* 28.3, pp. 832–839.

Weld, Henry et al. (2021). "Conda: a contextual dual-annotated dataset for in-game toxicity understanding and detection". In: *arXiv preprint arXiv:2106.06213*.

Wen, Wen (2019). "Does Delay in Feedback Diminish Sense of Agency? A Review". In: *Consciousness and Cognition* 73, p. 102759. DOI: 10.1016/j.concog.2019.05.007.

Wijkstra, Michel et al. (2023). "Help, My Game Is Toxic! First Insights from a Systematic Literature Review on Intervention Systems for Toxic Behaviors in Online Video Games". In: *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pp. 3–9.

Wojcik, Stefan et al. (2022). "Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation". In: *arXiv preprint arXiv:2210.15723*.

Zhang, Xiaohui et al. (2023). "Social Media Moderation and Content Generation: Evidence from User Bans". In: Available at SSRN: https://ssrn.com/abstract=4089011 or http://dx.doi.org/10.2139/ssrn.4089011.

Chapter 3

HIGH DIMENSIONAL TOPIC MODELING

3.1 Introduction

We are living in an information age when oceans of text data get generated every day—millions of social media posts, blogs, and news articles. As social scientists, when we get access to those gigantic datasets, the first thing we want to do is to see what content is in these datasets, i.e., to understand the major issues discussed by these documents. One common tool used today for this purpose is Latent Dirichlet Allocation (LDA) topic modeling, which extracts the key topics and common patterns from large corpora.

At the same time, we should be aware that topics are dynamic by nature—conversations are changing all the time. Suppose an economist wants to know the topics people discuss on social media regarding the U.S. macroeconomy in the first and second quarters of 2022. One might observe that the topic patterns are completely different across different time periods. We might see the discussion of stimulating the economy in the first quarter. In the second quarter, however, the major concerns switched to skyrocketing inflation and overheated economy. Therefore, the distribution of topics should be changing by time, with probably most of the topics occurring only for short periods of time.

We should also be aware that we live in an age when conversations differ by the groups the speakers belong to. Topics are different in different spatial locations. Democrats and Republicans have different opinions on many important social issues like abortion, climate change, and gun control. Men and women have different topics of for various subjects. Many social scientists are interested in learning the differences between groups. It would be natural to assume that there is a group effect in topic modeling.

However, time and group effects have long been ignored by the state-of-the-art feature extraction methods on text data. Most existing topic modeling approaches social scientists use today fail to utilize any of the additional information, even if we have access to them. Accounting for these types of information during the stage of topic modeling could significantly improve the topic modeling results. A large body of literature on topic modeling explore how to incorporate the additional information

the text data carries, examples including the structural topic model (STM), correlated topic models, joint sentiment topic models, and dynamic topic modeling. Those studies typically extend LDA to make more complicated parametric assumptions on how text data are generated under the influence of say, time or sentiment. In that sense, those alternative approaches are not ideal enough in their flexibility. They also rely on algorithms like EM/Gibbs Sampling, which are slow and make it inefficient to use for researchers. Social scientists, therefore, lack appropriate, easy-to-use tools to analyze text data with metadata that is potentially helpful.

In this paper, I introduce a novel high-dimensional topic modeling (HDTM)framework for social scientists to address all of these issues using just one model. The approach is flexible, easy-to-use, and can be parallelized. The method is based on non-negative tensor rank decomposition (also known as CANDECOMP/PARAFAC/CP decomposition), which is capable of handling high-dimensional text data with time and group labels. Identification comes from the structure of the high-dimensional data itself, instead of parametric modeling. This helps us avoid the caveats of over-parameterizing, and makes it possible to estimate a higher-dimensional topic model quickly with parallel computing. Given enough metadata, the model can potentially handle an arbitrary number of dimensions. This means that the method can provide not only topics and topic evolution over time, but also topic descriptions in other dimensions, like sentiment, gender, and partisanship, among others. The flexibility makes the approach an extremely powerful method that helps social scientists in virtually any field to make inferences from their text data.

I test my model on multiple synthetic and real-world datasets and show its capability of recovering information on higher dimensions. I also take a closer look at LDA and show that it performs much worse than my approach in handling those datasets. In fact, LDA gives wrong results and is unable to discover the right topics given dynamic datasets. This indicates that many previous studies social scientists did using LDA may be problematic and the results may be unreliable.

The remaining sections are organized as follows: section 2 presents a review of previous topic modeling approaches; section 3 introduces the model; section 4 establishes the theoretical foundation; section 5 presents the experimental results on synthetic data and comparison to existing approaches; section 6 applies the method to several real-world datasets. Section 7 and 8 discuss and conclude.

3.2 Related Work

This work is related to the large strand of literature on topic modeling. Topic modeling is an unsupervised machine learning technique that can discover the underlying "topics" from a set of documents, where topics are defined as a set of word and phrase patterns that summarize the word groups and similar expressions that can characterize a set of documents. The canonical approach people use for topic modeling is LDA(Blei, Ng, and Jordan, 2003). It has broad applications in many fields in social science, like communication research (Maier et al., 2018) evolution of economic research (Ambrosino et al., 2018), and financial time series analysis (Kanungsukkasem and Leelanupab, 2019). Note that many studies here actually involve high-dimensional information like time series, and we will see the caveats of that later.

LDA model documents as distributions of topics and topics as a distribution of words. Specifically, it assumes that the data-generating process for each document w in a corpus D is as follows:

- choose size $N \sim Poisson(\psi)$
- choose the word distribution $\theta \sim Dir(\alpha)$
- For each of the N words w_n: choose a topic $z_n \sim Multinomial(\theta)$; choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial distribution conditional on the topic z_n

The underlying assumptions are: topics are invariant over time, and each topic is independently drawn from the Dirichlet distribution, and therefore independent of each other. These restrictions make it difficult to model topics that are correlated, and also make temporal interpretation of topics unclear.

There are several existing studies that attempt to model the topic dynamics over time explicitly, one is dynamic topic modeling (Blei and Lafferty, 2006). It's a refinement based on LDA with additional assumptions on how the parameters α and β change over time, namely

$$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I) \tag{3.1}$$

$$\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I) \tag{3.2}$$

The other commonly used one is topic over time (Wang and McCallum, 2006). The refinement they make is to add a generation process of timestamp. Specifically, they make an additional assumption on the distribution of timestamp:

$$t_n \sim \text{Beta}(\psi_{z_n})$$

which means that the timestamp associated with each token is distributed according to $\text{Beta}(\psi_z)$ where z is the topic associated with that token. Then, one can obtain the joint distribution $p(\theta, z, t \mid \alpha)$ of word distribution θ, topic z, time t conditional on parameter α and thereby obtain the distribution of time label $p(t|\alpha)$ by integration.

There are also studies that attempt to incorporate metadata in the topic modeling process. A well-known example is STM(Roberts et al., 2013). The approach is also one of the LDA-like models. One of the key assumptions is that topics follow a logistic normal distribution of metadata over documents, i.e.,

$$\eta \sim \mathcal{N}(X_\gamma, \Sigma)$$

The other key assumption they make is that topic contents are sparse deviations from a word-specific baseline i.e., topic content

$$\beta_{k,g} \propto \exp(m + \mathcal{K}^{(k)} + \mathcal{K}^{(g)} + \mathcal{K}^{(k,g)})$$

Those two assumptions bring more unknown parameters into the model, making it computationally less tractable. The algorithm they use is variational EM, which is slow and inefficient. In addition, STM only gives us some weights on how each attribute affects the distribution, and it is hard to tell from the weights how groups are different in terms of topics they discuss. The ability of STM to handle time-series labels is also limited.

The three methods above all make additional assumptions based on LDA. There are two caveats to these approaches: first, these assumptions are not very likely to be true. In particular, it is questionable how much using the Dirichlet distribution to model topics fits the real-life situation. The evolution of topics also might not follow a conditional normal distribution. Second, adding more assumptions make it computationally less tractable to estimate the parameters. In fact, all of these approaches rely on approximate inference methods like Kalman Filtering and Gibbs Sampling to estimate the parameters. These approaches are slow and computationally intractable.

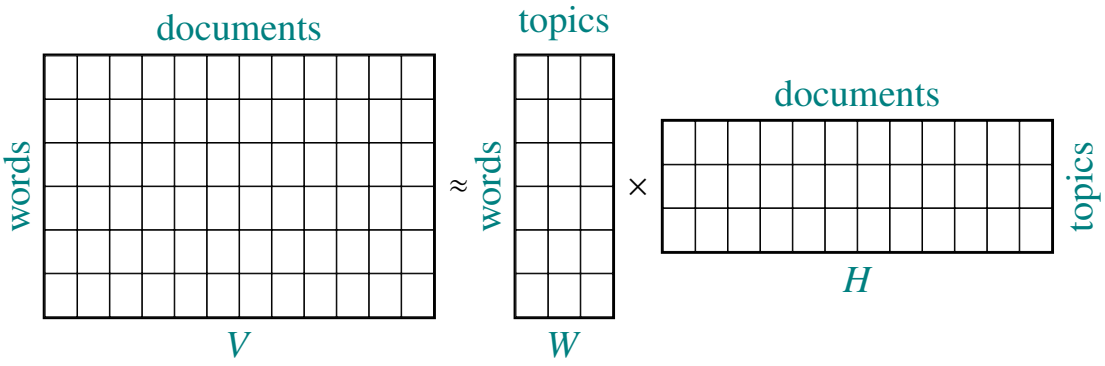


Figure 3.1: NMF-based topic modeling

There are lots of other versions of dynamic models, but a large proportion of them are modeling the dynamics with complicated parametric modeling, which also suffers from the same caveats mentioned above. Our method is different from all existing approaches like STM/topic over time in that we do not make any of these parametric assumptions. Our identification exploits the high-dimensional structure of the data, making the data itself tell us what the group effect/time evolution is. It has the advantage that it is "non-parametric" in the sense that it does not depend on any parametric assumption on how topics are changing over, nor does it make any assumption on how documents are generated, which makes it free from the over-parameterizing problems.

Moreover, the estimation method we focus on is tensor rank decomposition, which is much more computationally tractable than those sampling methods. It also makes it possible to parallelize the process. This leads to the second strand of related literature: the application and computation of non-negative tensor decomposition.

An alternative method to LDA is non-negative matrix factorization-based topic modeling (NMF) (Lee and Seung, 1999). The idea behind it is to learn an object by learning its parts. For the application to text data, it approximates a matrix V (word $\times$ documents) with two non-negative matrices W and H, i.e., $V \approx WH$. W gives the word-topic matrix and H gives the topic-document matrix. Our approach can be seen as an extension of NMF in higher dimensions.

Similar to NMF, non-negative tensor CP decomposition (NCPD) is a non-parametric approach to extracting the lower-dimensional representation of a tensor. It has been applied to many fields, including computer vision and image processing (Shashua and Hazan, 2005), feature extraction, and health data analysis (Anaissi, Suleiman, and Zandavi, 2020). Recently, there are studies that apply this method to text

analysis (Ahn et al., 2020), which focus on modeling topic evolution over time. The study shows that the non-negative tensor rank decomposition of a three-way tensor is able to detect topic evolutions over time. They also show that the decomposition is robust and stable against noise. However, the theoretical foundation from a topic modeling point of view has not been discussed. Also, a more detailed comparison with existing approaches from the topic modeling point of view is lacking.

In recent years, a lot of attention has been drawn to neural-network-based pre-trained language models. Some topic modeling approaches have been developed using the embedding output from those large language models, examples including BERTopic(Grootendorst, 2022) and Top2vec(Angelov, 2020). The key ideas behind these are similar: they transform documents into sentence embedding, and then cluster the resulting embedding to get topics. However, note that the quality of topics would largely depend on the pre-trained language models, which might not perform very well without fine-tuning. Also, those frameworks still do not consider metadata for the modeling part. BERTopic provides “dynamic topic modeling” and ”class-based topic modeling” options that give topic evolution over time or classes. However, they are essentially summarizing topics over the time dimension or over a given class, which is done after the modeling stage. In other words, those methods do not utilize non-text information that are potentially helpful during the modeling stage.

There are several gaps in the literature that we want to address. First, current applications of non-negative CP decomposition on text data mostly stay within 3D. We will show that it can be generalized to even higher dimensions with additional metadata. Second, current applications lack theoretical background from a language model point of view. We fill the gap by providing a theoretical interpretation, which resembles probabilistic latent semantic analysis (PLSA) in high dimensions. Third, we lack a more direct and closer examination of the performance against canonical approaches like LDA. We fill the gap by comparing different approaches by coherence score measures on both synthetic and real-world datasets.

3.3 The Model

Data Structure: The 3D Case

A tensor is a multidimensional array (Kolda and Bader, 2009). A N-way tensor is an element of the tensor product of N vector spaces. A 1-way tensor is an array; a 2-way tensor is a matrix; here, we focus on higher-order tensors, namely, when

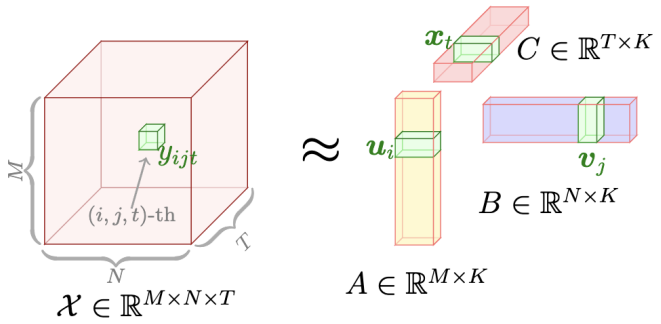


Figure 3.2: 3D Tensor Rank Decomposition

$N \geq 3$. The reason is that our data structure naturally resembles a 3D cube: suppose the vocabulary size is N, then, a document can be seen as a $1D$ array of length N, each element being the probability of appearance of some word. We have T time slices, each of which contains a set of documents M. This gives us a 3-way tensor of dimension $M \times N \times T$.

Non-Negative Tensor Rank Decomposition

The CP decomposition factorizes a tensor into the sum of products of rank-one tensors. The data we deal with would have three dimensions: document, word, and time. Therefore the data we have would be a rank-3 tensor. Specifically, suppose for each period, we have a set of documents $\mathcal{D}$ with size N, and we can construct a vocabulary list with size N, then the M documents can be represented by a matrix of size $M \times N$, with each row representing a document. The entries of each column would be the vocabulary composition of each word. Therefore the slice of documents at time t is represented by $\mathcal{D}_t = [d_{1t}, d_{2t}, d_{3t}...d_{Nt}]$

Piling up the document matrix of T periods would give us a tensor $\mathcal{X}$ of dimension $M \times N \times T$.

Tensor rank decomposition treats the tensor as a whole. It decomposes the tensor into a sum of rank-one tensors. The rank-K CP decomposition of a tensor $\mathcal{X} \in \mathbf{R}_{M \times N \times T}$ would be

$$\mathcal{X} = \sum_{k=1}^{K} a_k \otimes b_k \otimes c_k$$

where $\otimes$ is the tensor product, meaning that

$$x_{ijt} = \sum_{k=1}^{K} a_{ki} b_{kj} c_{kt}$$

In the case of dynamic topic modeling, the three dimensions can be chosen according to one's own analysis purposes. For instance, to track individuals over time, the three dimensions can be word, user, and time; to compare documents in different locations over time, the three dimensions can be word, location, and time.

Uniqueness

The CANDECOMP/PARAFAC decomposition factorizes a tensor into a sum of component rank-one tensors. The factor matrices refer to the combination of the vectors from the rank-one components, i.e.,

$$A = [a_1, a_2, ...a_K]$$
$$B = [b_1, b_2, ...b_K]$$
$$C = [c_1, c_2, ...c_K]$$

We can express each time slice of the tensor as

$$\mathcal{X}_k = AD^{(k)}B^T$$

where $D^{(k)} = diag(c_{k:})$ for $k = 1...K$

The tensor decomposition of any rank-R three-way tensor is unique under weak conditions. Kruskal's results say that a sufficient condition for the uniqueness for the CP decomposition is

$$k_A + k_B + k_C \geq 2R + 2$$

where k_A is the k-rank of matrix A.

Algorithm

The non-negative tensor decomposition problem can be formalized as the minimization of the following regularized KL-divergence loss function:

$$\min D(\mathcal{X}||\hat{\mathcal{X}}) = \sum_{i,j,t} x_{i,j,t} \log \frac{x_{i,j,t}}{\hat{x}_{i,j,t}} - x_{i,j,t} + \hat{x}_{i,j,t}$$

where

$$\tilde{\mathcal{X}} = \sum_{k=1}^{K} A_{:k} \otimes B_{:k} \otimes C_{:k}$$

A state-of-the-art algorithm for non-negative CP decomposition with the above loss function is multiplicative updates. Many researchers apply multiplicative updates to image processing (Welling and Weber, 2001) and sound source separation (Coyle, 2005). We present the algorithm for a third-order tensor as follows:

Algorithm 1 Multiplicative Update

Input: Tensor $\mathcal{X} \in \mathbf{R}^{M \times N \times T}$, convergence criterion ϵ
Output: matrices A, B, and C

1: Initialize A, B, and C randomly or with SVD
2: **repeat**
3: $\quad A = A \circ X_{(1)}(B \odot C) \circ \frac{1}{A(B^T B)(C^T C)}$
4: $\quad B = B \circ X_{(2)}(A \odot C) \circ \frac{1}{B(A^T A)(C^T C)}$
5: $\quad C = C \circ X_{(3)}(A \odot B) \circ \frac{1}{C(A^T A)(B^T B)}$
6: **until** reconstruction error $L(\mathcal{X}, A, B, C) < \epsilon$

where $\circ$ denotes the element-wise product; $\odot$ denotes the Khatri-Rao product; fractions are done element-wisely. $X_{(i)}$ are the matricized tensor in mode i.

Going Beyond the 3D World

Adding the time dimension does not exploit the whole potential of the model. The nature of high-dimensional tensor decomposition makes it flexible enough that enable us to explore more aspects of the data. For instance, we could add another dimension that indicates the sentiment of the document, which would allow us to discern part of the semantic meaning of each topic—for instance, we could not only identify the climate change topic, we could even separate apart the topics that are positive and negative against it. Similarly, if we have the network information, we would be able to keep track of which communities are the main participants in each topic. Any other pre-labeling information like partisanship would also work. These are natural extensions of the model into higher dimensions, and it provides a good way to analyze an unprecedentedly rich set of information.

Specifically, the approach can be extended to deal with datasets with the following tensor data structure:

$$\begin{aligned} &\mathcal{X}(m, n, t, i_1, i_2, i_3 \ldots) \\ &= Prob(\text{word } m \text{ occurs in document } n \\ &\text{when time} = t, I_1 = i_1, I_2 = i_2, I_3 = i_3 \ldots) \end{aligned}$$

where I_1, I_2, I_3... are a set of attributes that are ordinal variables or categorical variables.

We can easily express $\mathcal{X}$ as a N-way tensor, and perform the same decomposition

$$\mathcal{X} \approx \sum_{k=1}^{K} u_k^1 \otimes u_k^2 \otimes u_k^3 \otimes ... \otimes u_k^n$$

$$\mathcal{X} \approx [[U^1, U^2, U^3, ..., U^n]]$$

The non-negative tensor decomposition would yield a set of factors. If we normalize them, we would get

$$U^1(i_1, p) = Prob(I_1 = i_1 \text{ in topic p})$$
$$U^2(i_2, p) = Prob(I_2 = i_2 \text{ in topic p})$$
$$U^3(i_3, p) = Prob(I_3 = i_3 \text{ in topic p})$$
$$.....$$

For instance, if we add a dummy variable $I_1 \in$ (Republican, Democrats) that indicates the partisanship of the author of each document, the resulting factor would give us the partisanship component of each topic. If we add a categorical variable $I_2 \in$ (positive, negative, neutral) that indicates the sentiment of each document, we would get a factor matrix that tells us the sentiment component of each topic. If we add an indicator variable $I_3 \in$ (community 1, community 2...) that indicates which cluster the document belongs to, which could be generated by some community detection algorithm, we can get a factor matrix that tells use the community-level participant component of each topic. Possible additional attributes include gender, race and ethnicity, social class, or similar types of features. The method has very loose assumptions on the set of categorical variables: they can be ordered or unordered.

The factor matrices we obtain can be used for inferential purposes. These matrices can tell us how topics separate among different groups/dimensions. This methodology will thus be very useful in research efforts that are interested in learning about these types of group separations, for example, gender studies, sociology, or history.

Label Topic Components for Each Document

Going from high dimensions to lower dimensions is easy, since high dimension contains all the lower-order information. We can obtain the topic component or topic labels of each document through unfolding higher-dimensional factor matrices into lower dimension. Specifically, let U^d be the topic-word matrix, $\{U^{-d}\}$ be the rest of the factor matrices. Then, we can obtain the topic-document matrix by calculating the Khatri-Rao product of $\{U^{-d}\}$. Formally,

$$L = U^1 \odot U^2 \odot ... \odot U^n$$
$$\text{for } U^1 ... U^n \in \{U^{-d}\}$$

Let $E_1 \times R$, $E_2 \times R$, ... $E_n \times R$ be the size of matrices U^1, ... $U^n \in \{U^{-d}\}$, then the size of L would be $\prod_{k=1}^{n}(E_k) \times R$. Let $\tilde{L}$ denote the row-wise normalized L , then topic t component of the document with indices e_1, e_2 ... e_n (note that we use zero-indexing) would be:

$$L\Big(\sum_{m=1}^{n-1}(e_m * \prod_{k=m+1}^{n} E_k) + e_n, t\Big)$$
$$= L\Big(e_1 * \prod_{k=2}^{n}(E_k) + e_2 * \prod_{k=3}^{n}(E_k) + ... + e_{n-1} * E_n + e_n, t\Big)$$

Interested readers can see a proof in appendix 3.9

3.4 Theoretical Background

In this section, we will see that the method is a high-dimensional generalization of the PLSA(Hoffman, 1999).

Existing Results

PLSA is a latent variable model for co-occurrence data with an unobserved class variable $z \in \mathcal{Z}$ with each observation. For a given collection of text document $\mathcal{D} = \{d_1, ...d_N\}$ with a set of vocabulary $\mathcal{W} = \{w_1, ...w_M\}$, the PLSA model is defined as

$$P(d, w) = P(d)P(w|d)$$

where

$$P(w|d) = \sum_{z \in \mathcal{Z}} P(w|z)P(z|d)$$

PLSA makes a conditional independence assumption that d and w are independent conditional on the state of the associated latent variable. With that assumption, the model can be further parameterized as

$$P(d, w) = \sum_{z \in \mathcal{Z}} P(z)P(w|z)P(d|z)$$

Studies have shown that PLSA and NMF-based topic modeling are equivalent(Gaussier and Goutte, 2005). Specifically, PLSA is NMF with KL divergence.

Data-Generating process

The language model is formulated as followed. Given a set of vocabulary $\mathcal{W} = \{w_1, ...w_M\}$, a set of document $\mathcal{D} = \{d_1, ...d_N\}$, a series of time $\mathcal{T} = \{t_1, t_2...t_T\}$ and a set of M metadata $\mathcal{I}_1, \mathcal{I}_2, ...\mathcal{I}_M$. Assume the following data-generating process:

1. select a document class d from $\mathcal{D}$ with $P(d)$
2. select a topic z from $\mathcal{Z}$ with $P(z|d)$
3. select a time label t with $P(t|z)$
4. select attributes i_1, i_2, ..., i_M with $P(i_1|z)$, $P(i_2|z)$, ..., $P(i_M|z)$ independently
5. select word w from $\mathcal{W}$ with $P(w|z, t, i_1, ...i_M)$

Additionally, we make the following assumptions:

- A1: document class d and topic z are independent, i.e., $P(z_p|d) = P(d)$
- A2: word w is conditionally independent with t and i given z_p, i.e., $(w \perp\!\!\!\perp t|z_p)$ and $(w \perp\!\!\!\perp i|z_p)$

With Assumption A1, The joint probability of occurrences $(w, d, t, i_1, ...i_M)$ can be then modeled as

$$P(w, d, t, i_1, i_2...i_M)$$
$$= \sum_p P(d)P(z_p|d)P(t|z_p)P(i_1|z_p)...P(i_M|z_p)P(w|z_p, t, i_1, i_2, ..i_M)$$

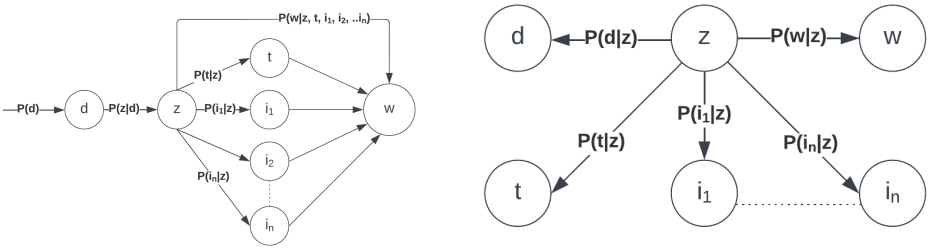
By adding the conditional independence assumption A2 the model above is equivalent to

$$P(w, d, t, i_1, i_2...i_n)$$
$$= \sum_p P(d|z_p)P(w|z_p)P(t|z_p)P(i_1|z_p)...P(i_n|z_p)P(z_p)$$

which is the PLSA model in high dimensions. We are also giving each topic the same weight, i.e., we assume $P(z_p) = P(z)\ \forall\ p$. Therefore, the formula reduces to

$$P(w, d, t, i_1, i_2...i_n)$$
$$\propto \sum_p P(d|z_p)P(w|z_p)P(t|z_p)P(i_1|z_p)...P(i_n|z_p)$$

Therefore, we have just shown the data-generating process is equivalent to PLSA in high dimension. A graph representation of both can be seen in Figure 3.3.



(a) Original Tensor-PLSA (b) With Independence Assumption

Figure 3.3: Data-Generating Process

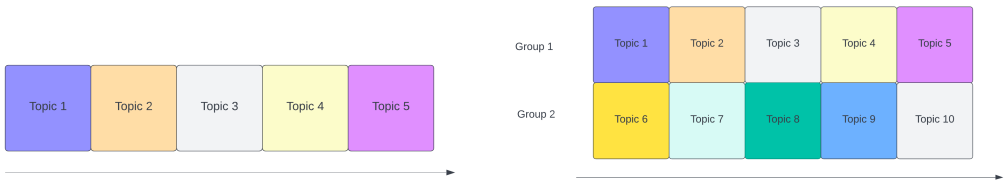
See a proof of the equivalence between the approach and Tensor-PLSA in Appendix 3.9. The result helps to show that the factor matrices U^1, U^1, ... U^N we recover indeed capture the $topic - attribute$ relation that we want, i.e.,

$$U^1(i_1, p) \approx Prob(I_1 = i_1 \text{ in topic p})$$
$$U^2(i_2, p) \approx Prob(I_2 = i_2 \text{ in topic p})$$
$$U^3(i_3, p) \approx Prob(I_3 = i_3 \text{ in topic p})$$
$$.....$$

3.5 Some Simulations

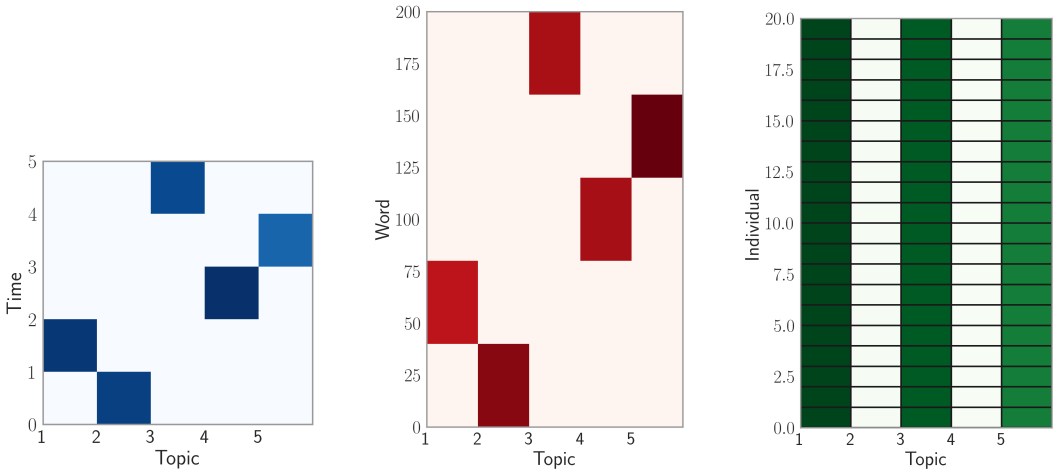
Fully Synthetic Data

We first test the model on fully synthetic data. To test the 3-D version, we construct a "dynamic" topic evolution dataset. Suppose there are twenty individuals that post documents from period 1 through 5. Suppose there are five topics and topics differ in every period. In other words, topic 1 is discussed in period 1, topic 2 is discussed in period 2, etc. In addition, suppose the vocabulary size is 200, and each topic takes up $1/5$ of the vocab, which means they are fully orthogonal. We construct a tensor of size $20 \times 200 \times 5$ to model a scenario of "topics changing over time". The visualization of the timeline is in Graph 3.4a We perform non-negative tensor rank decomposition on the dataset. We obtain a set of three factor matrices that can recover the exact tensor with construction error 0. The visualization of each matrix is shown below.



(a) 3D tensor (b) 4D tensor

Figure 3.4: Fully synthetic data timeline



(a) Topic-time matrix (b) Topic-word matrix (c) Topic-individual matrix

Figure 3.5: Factor matrices of 3D fully synthetic data

We can see that the evolution of topics, word component of each topic, and topic

component on the individual level are all fully captured by the three factor matrices. In particular, the topic-time matrix indicates that each topic appears in one period, and that each period has only one topic. The topic-word matrix indicates that topic 1 covers word $40-80$, topic 2 covers word $1-39$ ... and the topic-individual matrix indicates that every topic is contributed equally by every individual. These are the same as the data-generating process up to the permutation of topic labels. We are also interested in seeing whether the model recovers higher dimensions. So we add a dimension "group" (e.g., partisanship/sentiment) to model the case when topics could differ by which group each individual posts in. Suppose each individual participates in two groups. From period 1 to 5, every individual discusses topic 1 to 5 in each period, respectively, in group 1, and at the same time, they discuss topic 6 to 10 in group 2. The timeline can be seen in Graph 3.4b.

The results show that we can recover all the dimensions correctly: the topic-individual matrix shows that each topic is contributed equally by each individual; the topic-word matrix shows that each topic covers $1/10$ of the vocab size; the topic-time matrix shows that each topic appears once on the timeline; the topic-group matrix shows that each topic appears in one group and identifies the group correctly.

We also run LDA on the synthetic dataset. Surprisingly, LDA is not able to identify all of the topics and generates duplicate topics in both 5-topic and 10-topic cases. The topic-word matrices are shown in Figure 3.7a and Figure 3.7b. Interestingly, this pattern is very similar to what we observe in our previous implementation of LDA on the real-world dataset. LDA on the election fraud tweet dataset also generates duplicate topics, with a bunch of other topics almost identical. Our test of LDA on synthetic datasets here tells us it might not be

Semi-Synthetic Data

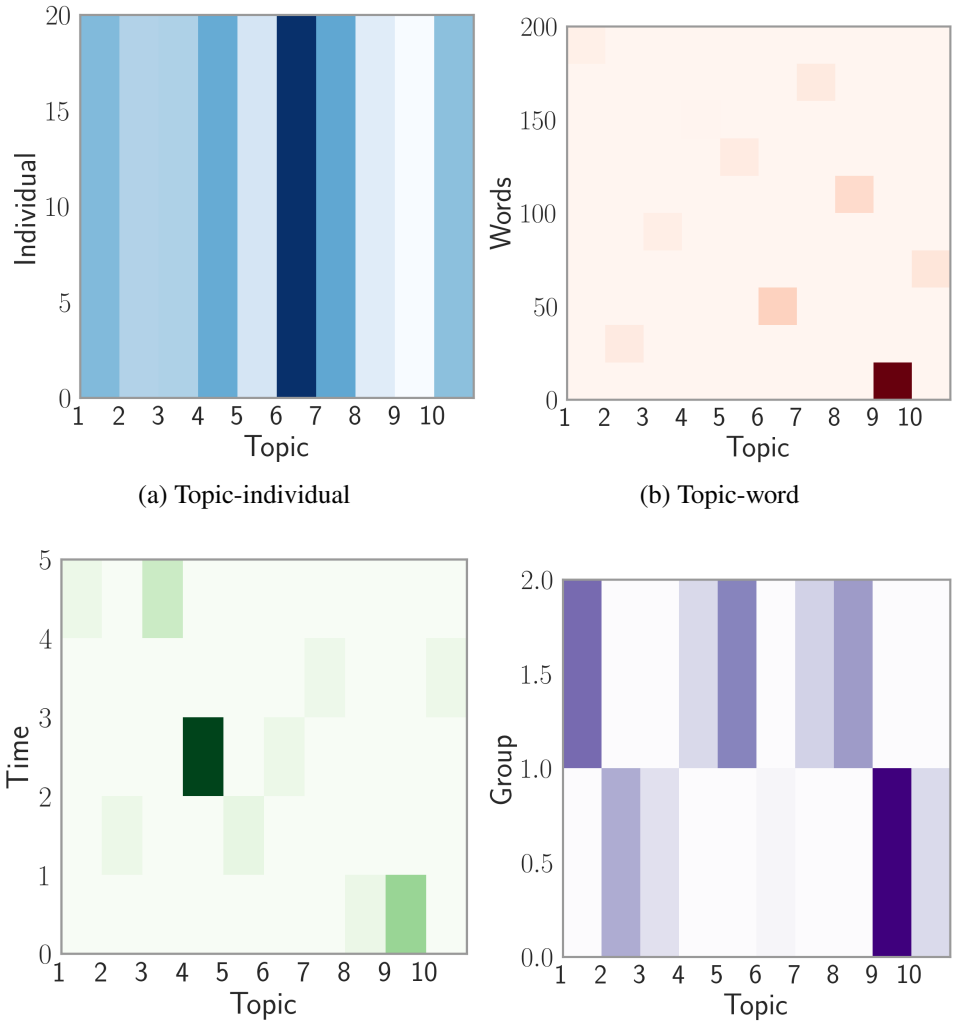
In this section, we test the 3D and 4D versions of the model using some real-world labeled text data.

3D Data

We first test the existing 3D version of the model, using a similar idea as in (Ahn et al., 2020). They construct a 3-D dataset by pulling different topics from labeled datasets and putting them together to form a tensor.

We use the sci-kit learn newsgroup dataset[1], which comprises around 18000 news-

[1]Link for the Scikit-learn dataset



(a) Topic-individual

(b) Topic-word

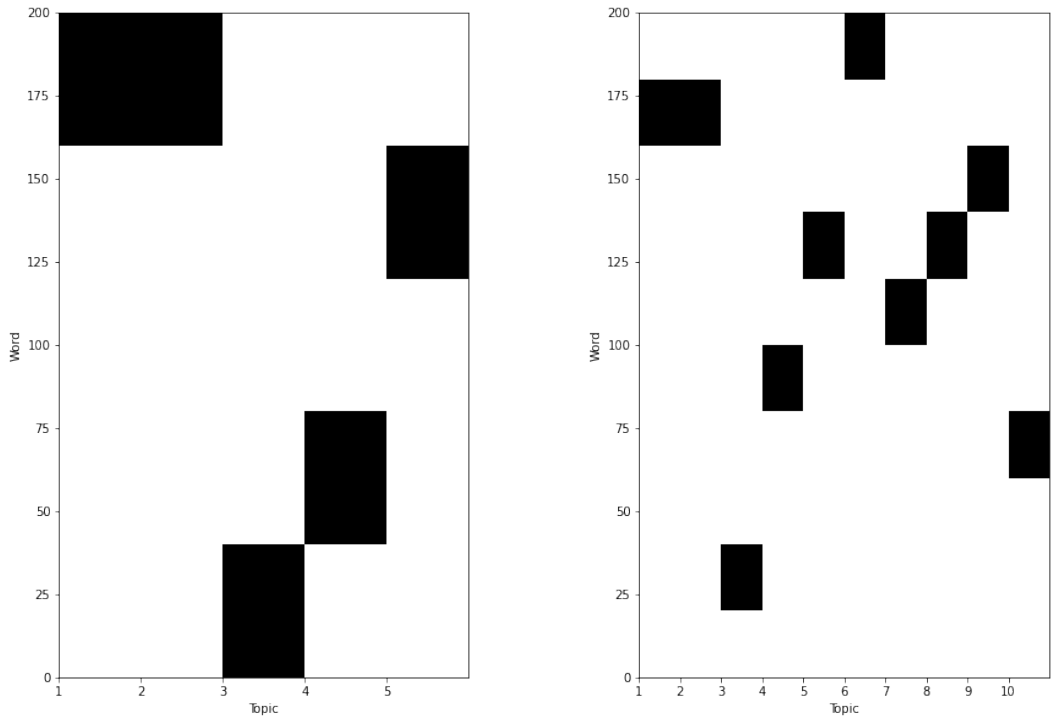
(c) Topic-time

(d) Topic-group

Figure 3.6: Factor matrices of 4D fully synthetic dataset

groups posts on twenty topics. We pull data from five categories of sklearn dataset. Assume there are five categories: religion, space, sales, baseball, and windows. Extract 3000 vocabs.

To construct the dynamic dataset, for every time period, we randomly sample 26 documents from one of the five chosen categories and regard that as a slide (26×3000). This can be seen as if twenty-six individuals discuss the same topic in every period. We sample 40 periods in total and put all slides from different time periods together as a tensor. Our tensor size is therefore (individual, word, time) = $(26, 3000, 40)$. Some topics cover longer terms, some cover shorter terms. The evolution of topic is defined in Figure 3.8.



(a) LDA: 5 topics (b) LDA: 10 topics

Figure 3.7: Topic-word of LDA on fully synthetic data

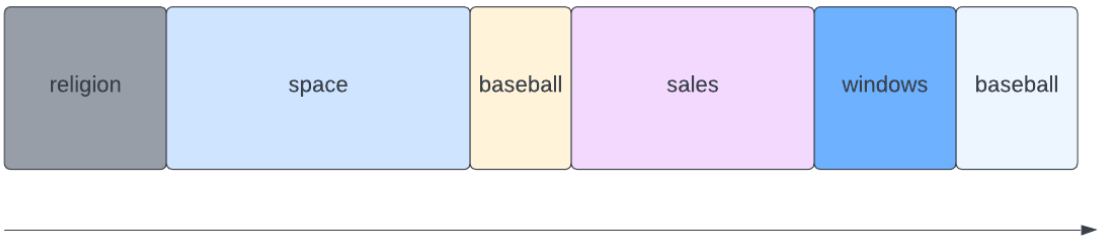


Figure 3.8: Semi-synthetic data

Table 3.1 and Table 3.2 give the topics discovered by both models. We can see that LDA is unable to identify half of the topics. The identified ones are also noisy with some irrelevant words. HDTM, by contrast, identify all topics correctly and gives much better topics in terms of quality.

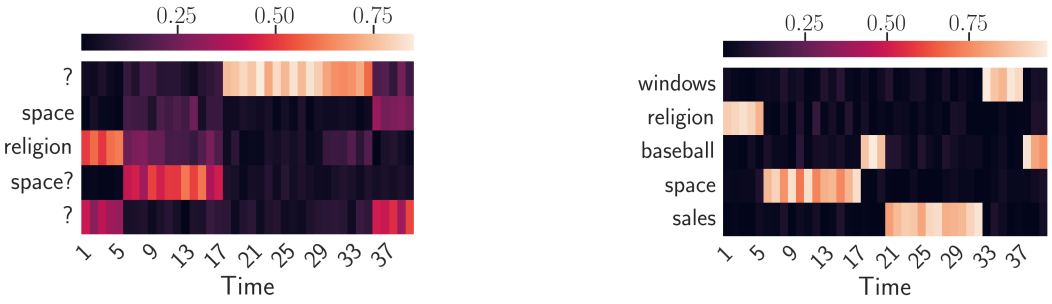
Figure 3.16a and Figure 3.16b give the topic evolution generated by each model. The color in the heatmap stands for the intensity of each topic at any given time period. We can see from the topics and time evolution that HDTM is able to capture the right topics, giving the same time evolution as the data-generating process. While LDA is unable to capture some of the topics, giving a much more noisy topic revolution

Table 3.1: 3D Semi-synthetic Data: LDA

Topics	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Top 10 words	would	earth	god	space	would
	one	sun	time	launch	one
	new	edu	would	would	god
	like	space	one	like	appears
	sale	system	people	system	window
	please	solar	church	much	art
	get	planet	know	one	widget
	edu	spacecraft	good	shuttle	man
	drive	use	see	satellite	like
	year	also	think	us	use
Label	?	space	religion	space?	?

Table 3.2: 3D semi-synthetic data: HDTM

Topics	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Top 10 words	server	god	year	space	sale
	window	church	games	would	new
	widget	jesus	game	launch	drive
	motif	would	team	like	offer
	use	christians	baseball	nasa	shipping
	windows	christian	hit	orbit	interested
	please	christ	runs	much	please
	application	marriage	pitching	shuttle	condition
	xterm	people	last	earth	email
	x11r5	one	fan	one	asking
Label	windows	religion	baseball	space	sales



(a) LDA

(b) HDTM

Figure 3.9: 3D semi-synthetic data: LDA vs. HDTM

Table 3.3: 3D Semi-synthetic Data: Coherence Scores

Coherence Measure	u_mass	c_v	c_uci	c_npmi
LDA	-2.128	0.416	-1.306	-0.026
HDTM	-1.852	0.643	-0.134	0.081

than HDTM.

Table 3.3 gives the comparison of coherence scores. We can see that HDTM dominates LDA in all four different coherence measures, meaning that HDTM gives more semantically interpretable topics compared to LDA.

4D Data

We add another dimension "group" for the semi-synthetic data by drawing documents from ten categories and splitting them into two groups. It is simulating the case when individuals generate documents each period in different groups, and that topics differ by group. Figure 3.10 shows the timeline for both groups.

Table 3.4 and Table 3.5 present the topics we discover for the 4D dataset. We can correctly identify all ten topics using 4-dimensional topic modeling, while we fail to identify half of the topics using LDA.

Figure 3.11b and Figure 3.11a show the topic evolution for labeled topics of both models. Figure 3.11c provides the extra dimension HDTM gives us, which correctly identifies the group membership of each topic.

We also provide the coherence score measures in Table 3.6. Again, HDTM excels in every measure of coherence score.

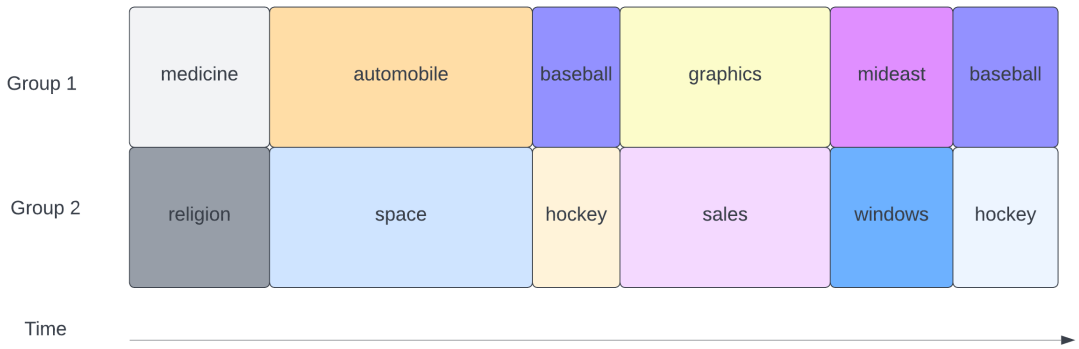


Figure 3.10: Semi-synthetic Data

Table 3.4: 4D Semi-synthetic data: 4D Modeling

Topics	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Top 10 words	car	year	server	god	israel	space	would	sale	graphics	game
	cars	games	widget	church	jews	would	medical	offer	image	hockey
	engine	game	window	jesus	armenian	launch	doctor	new	thanks	team
	like	team	motif	christians	people	orbit	patients	drive	would	nhl
	would	hit	use	christian	armenians	nasa	think	shipping	know	games
	get	baseball	xterm	would	turkish	shuttle	one	condition	format	players
	one	pitching	windows	christ	arab	like	know	interested	program	season
	dealer	runs	x11r5	marriage	muslim	moon	disease	please	anyone	espn
	oil	last	please	bible	israeli	earth	edu	asking	please	play
	also	first	application	sin	peace	much	treatment	email	files	would
Label	automobile	baseball	windows	religion	mideast	space	medical	sales	graphics	hockey

Table 3.5: 4D Semi-synthetic Data: LDA

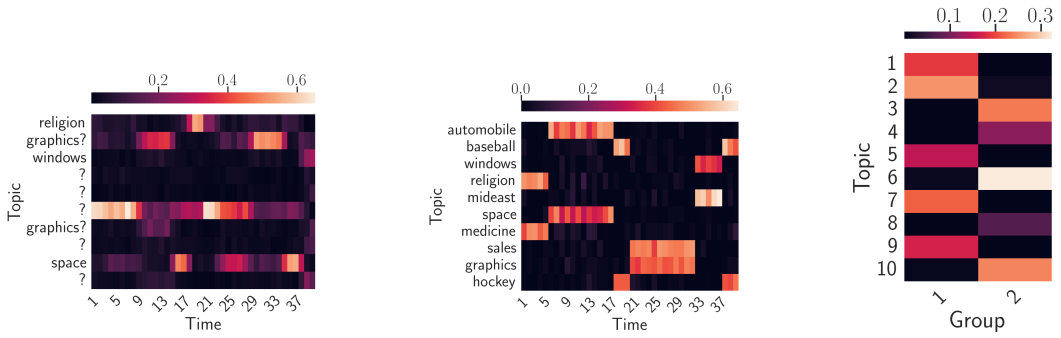
Topics	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Top 10 words	people	edu	server	please	key	would	jpeg	doug	space	use
	said	data	sco	files	xterm	one	gif	could	year	color
	us	image	widget	comp	map	car	file	symbol	launch	using
	god	please	biz	postscript	define	may	image	lib	game	would
	turkish	available	display	put	line	like	format	undefined	one	colors
	one	software	window	event	keys	get	use	one	would	display
	jews	also	x11	command	motif	time	get	server	good	program
	say	graphics	events	could	ms	think	files	even	two	frame
	armenian	information	screen	broken	string	also	think	read	like	bit
	armenians	pub	static	page	end	see	images	get	first	window
Label	religion	graphics?	windows	?	?	?	graphics?	?	space	?

Table 3.6: 4D Semi-synthetic Data: Coherence Scores

Coherence Measure	u_mass	c_v	c_uci	c_npmi
LDA	-2.658	0.458	-2.067	-0.018
HDTM	-1.843	0.604	-0.152	0.082

3.6 Empirical Application

In this section, we provide several applications on real-world datasets. The first one is a comparison of the model in 3D(HDTM) and LDA on the Reuters-21578 dataset, one of the most commonly used data collections for text categorization studies. It is also one of the datasets used in Blei, et al (2013). The second application applies 4-dimensional and 5-dimensional versions of the model to a real-world dataset that consists of political actors' tweets.



(a) LDA (b) 4D: Topic-time (c) 4D: Topic-group

Figure 3.11: 4D semi-synthetic data: LDA vs. 4D Modeling

Reuters-21578

Reuters-21578[2] is the dataset Blei et al (2003) use to showcase the application of LDA to document classification. The dataset is collected from the Reuters financial newswire service in 1987 and contains the date label, which makes it feasible to implement dynamic topic modeling.

Here, we test the same dataset with tensor-based dynamic topic modeling method and add the time dimension. We focus on the "ModLewis" subset, which contains 12449 documents, and obtain a subset with 8418 documents that span 23 days. Each time slide consists of 366 documents.

We observe fairly different topic over time patterns in Figure 3.17 and Figure 3.18. With HDTM, we get topics that mostly span a short time. Some of the topics disappear and reoccur later on. By contrast, LDA gives much more noisy topic evolution patterns, with most topics spanning a long period. We also obtain a higher coherence score -2.2487 compared to LDA(-1.693)[3].

Note that our approach is potentially beneficial to finance researchers who use financial reports as a data source. The approach gives a time series for each topic, and it is handy to use them as predictors of important economics/finance variables.

The Congressional Tweet Dataset

The congressional tweet dataset consists of tweets from legislators in congress from 2017 to 2022. The dataset provides the name of each legislator, the party they belong to, and the date of each tweet. It makes the data a great fit for the analytic framework

[2]https://huggingface.co/datasets/reuters21578

[3]The measure we use here is c_umass. The other three measures fail to converge.

of our HDTM since it provides more categorical information on the tweets. We obtain the sentiment label from the existing joint sentiment topic modeling of the dataset (Ebanks, 2021). The sentiment label tells us whether the tweet is positive, negative, or neutral. Using this, together with the time labels and the partisanship labels, we create two high-order tensors:

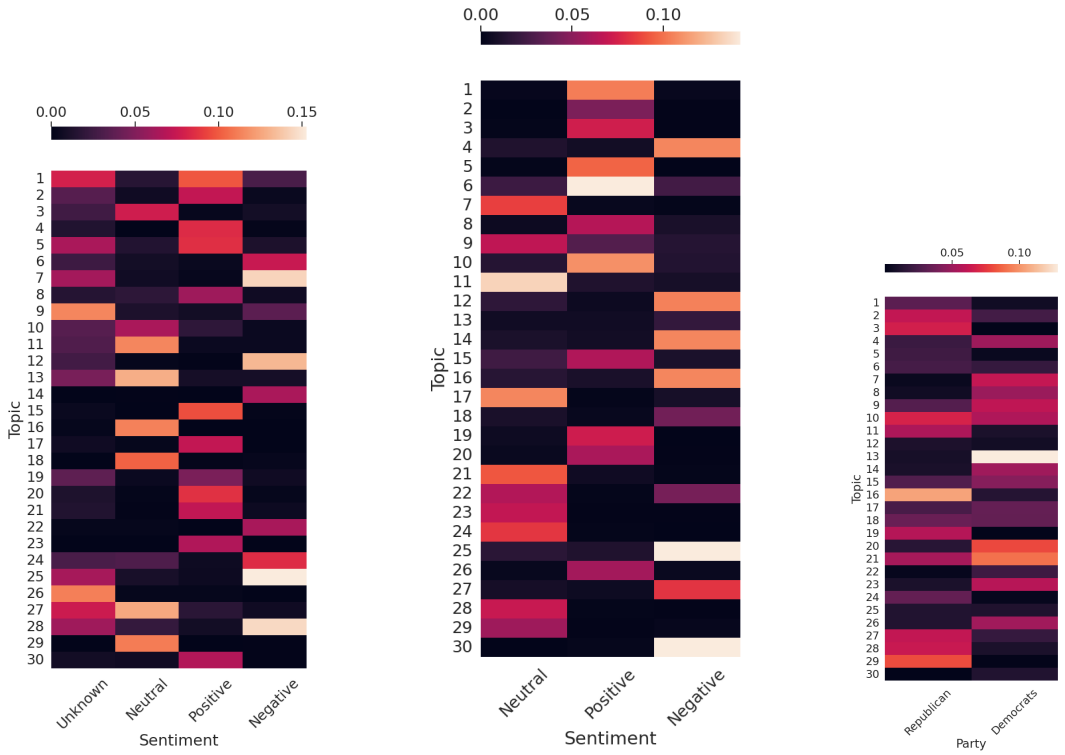
- 4-D: (word, document, time, sentiment)
- 5-D: (word, document, time, sentiment, partisanship)

We test the 4-D and 5-D versions of our model on the congress tweet dataset with thirty topics. For comparison purposes, we also run LDA on the same datasets, without the extra dimensions. And then we evaluate the topics obtained by the two models.

Coherence measures are shown in Table 3.7 and Table 3.15. Overall, HDTM achieves higher coherence scores than LDA in all cases. The qualities of the topics are significantly better. Note that we are able to obtain both long-lasting and short-lasting topics(Christmas).

Figure 3.19 and Figure 3.20 shows the evolution of topics obtained from 4D modeling and LDA. Figure 3.21 and Figure 3.22 show the evolution of topics obtained from 5D modeling and the LDA counterpart. Figure 3.12 shows the sentiment component and partisanship component of each topic obtained from 4D and 5D modeling.

We observe that most topics have a clear sentiment and partisanship. Using the extra dimensions, we can make inferences on the semantic meaning of each topic. For instance, the factor matrix tells us that topic 14 is mainly about Trump. We can also see it is mostly Democrats that are tweeting about Trump, and that most of the tweets are negative. We can then understand that it is a "criticizing Trump" topic. The time evolution shows clearly how several peaks of the topic align with some key social events, like BLM, COVID, the Supreme Court nomination of Barrett, and the election. A visualization based on multi-dimensional factor matrix is displayed in Figure 3.13. As another example, topic 3 displayed in Figure 3.14 features "happy birthday Trump". Almost 100% of the participants are Republican. The sentiment is also almost all positive. The time evolution graph clearly shows that it is a topic that emerged in the third week of June 2020 and faded quickly. This corresponds to Trump's birthday, which was on June 14th, 2020. The extra dimensions help



(a) 4D: Topic-sentiment (b) 5D: Topic-sentiment (c) 5D: Topic-party

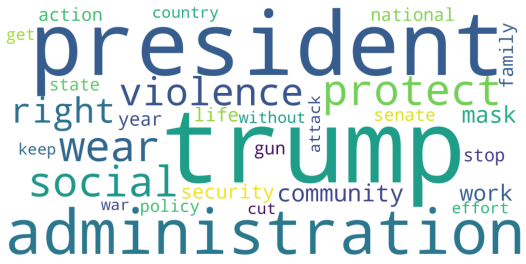
Figure 3.12: Extra dimensions from high-dimensional topic modeling

us clearly understand the context of the topic. These examples also show that both long-term and short-term topics can be detected with our approach.

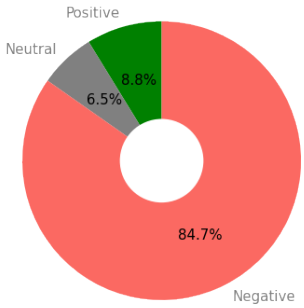
Note that with LDA, researchers can only get one out of the four subgraphs that we show. Therefore, we are unable to directly make inferences like that using LDA—LDA only gives the topic-word dimension and it is hard to tell the exact meaning of each topic. To make inferences, researchers have to label the topics of each document and summarize them using the metadata. However, the metadata information is ignored during the modeling process. And thus LDA performs much worse than our model.

With the help of the additional dimensions, we are able to dive deeper into each topic and get more information out of it, which might help us to get the actual semantic meaning of each topic. In comparison, LDA can only give us word clouds, and it cannot tell us more than the word component of each topic, which makes it difficult to make meaningful inferences.

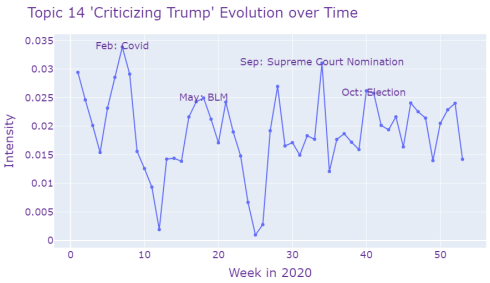
Figure 3.15 shows a comparison of different approaches on the congressional tweet dataset. Four different measures of coherence scores are reported. Overall, HDTM



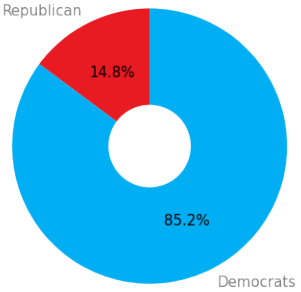
(a) 5D: Topic-word



(b) 5D: Topic-sentiment



(c) 5D: Topic-time



(d) 5D: Topic-party

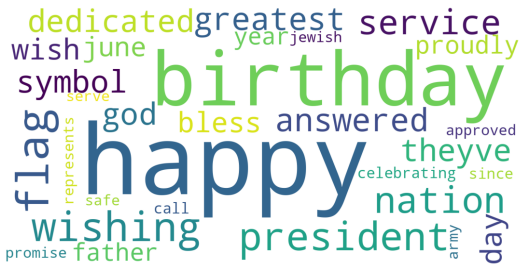
Figure 3.13: Topic 14 ("Criticism of President Trump" description with 5D modeling on the congress dataset

Table 3.7: Coherence Score: 4D Modeling vs. LDA

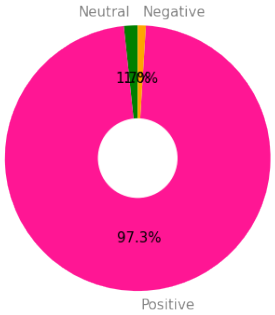
Coherence Measure	u_mass	c_v	c_uci	c_npmi
LDA	-7.198	0.371	-3.171	-0.076
4-D modeling	-3.840	0.601	0.345	0.125

Table 3.8: Coherence Score: 5D Modeling vs. LDA

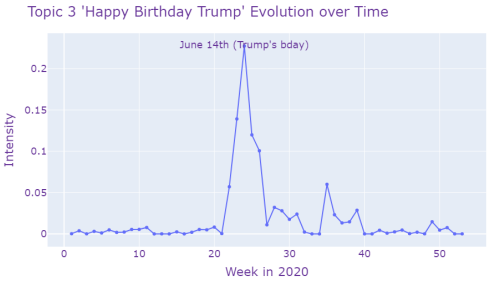
Coherence Measure	u_mass	c_v	c_uci	c_npmi
LDA	-5.724	0.396	-1.582	-0.019
5-D modeling	-3.633	0.645	0.772	0.149



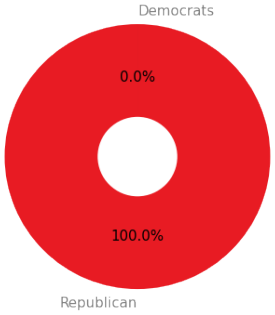
(a) 5D: Topic-word



(b) 5D: Topic-sentiment



(c) 5D: Topic-time



(d) 5D: Topic-party

Figure 3.14: Topic 3 (“Happy Birthday Trump" description with 5D modeling on the congress dataset

always achieves significantly better results than LDA. The comparison with the 2D version HDTM suggests that we indeed have gain in performance through adding the extra dimensions. HDTM is on par with BERTopic in three out of the four coherence measures. However, note that BERTopic utilizes some “void” topics that contain “outliers” of the corpus. Those void topics can help improve coherence while making a large number of documents unclassified. Using our model, all the documents contribute to the topics and can be clearly labeled. Therefore, the tiny difference in coherence is not too much loss considering the benefit of getting more information and more informative labels.

Next, we show the runtime comparison of the application of 4D and 5D modeling and LDA in gensim, a commonly used python-based library. Since the two methods converge to different metrics, we use coherence score as the convergence criterion.

Overall, the plot of LDA is steeper, indicating that it takes LDA shorter time to converge. The curves for 4D and 5D modeling keep increasing progressively as the time goes, eventually converging to a higher coherence measure. Comparing the time they converge to the same coherence score, the runtime loss is not significant.

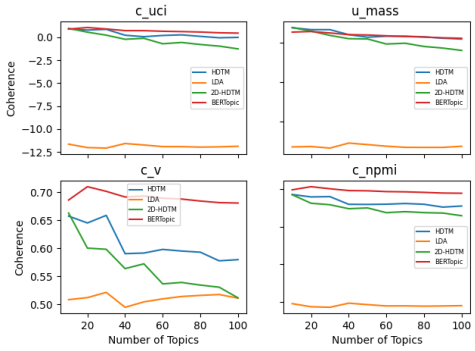
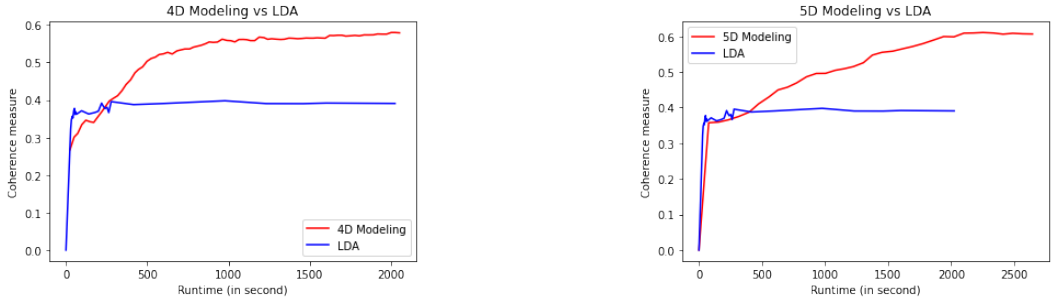


Figure 3.15: Topic coherence comparison between HDTM, LDA, 2D-HDTM, and BERTopic



(a) 4D HDTM vs. LDA

(b) 5D HDTM vs. LDA

Figure 3.16: Runtime comparison

3.7 Discussion

Reflection on Social Science

The framework developed in this paper is meaningful for social science researchers in a number of ways.

First, it is a framework that is general and flexible enough that can be applied to virtually any social science field that needs analytic tools to conduct inferences on text data. It is especially helpful when researchers want to make inferences on high-dimensional metadata. While in the paper I have focused on use cases that come from political science, there are potential use cases in other social sciences. For instance, researchers in economic history might need to analyze historical textual data and see how certain key issues evolved over time; gender studies scholars may want to make inferences on how males and females are different in generating content on social media; legal researchers might want to sift through datasets of cases and identify those related to certain times-locations-topics. My method is capable of fulfilling all tasks mentioned above, among others.

Second, the methodology developed in this paper relies on very few assumptions about how the data is generated. In that sense, it is a non-parametric method that is free from over-parameterization/over-fitting caveats suffered from existing state-of-the-art topic modeling methods.

Third, it is easy and fast to implement. Researchers can rely on any off-the-shelf tensor rank decomposition package to implement my method. In this paper, we also provide a way to parallelize the algorithm and make it thousands of times faster than existing approaches. It significantly saves time and energy. I will make the code open-source so that researchers have access to it.

Future Work

First, the current pipeline does not account for the fact that the data is sparse. We could save a lot of memory if we are able to store only the non-zero entries and the indexes with a sparse matrix representation. Suppose the vocabulary size is W, the number of documents is D, and each document contains h proportion of the vocabulary. Then for a n-way tensor, we only need to store $(n+1)h * WD$ data points using the sparse representation. While for a regular tensor, we need at least to store WD number of data points. When the data is sparse, $h << 1$, and thus $(n+1)hWD << WD$, saving $\frac{1}{(n+1)h}$ times memory. Some efficient ways to deal with sparse tensors can be further explored.

Second, the model can recover each of the higher dimension, but it does not tell us how much each dimension contribute to the topic identification. An interesting avenue for future work will use ablation studies to look closer on how important each dimension contributes to the model performance.

In addition, another direction we hope to explore is PARAFAC2. It is different from CP decomposition in that it does not enforce the constraint that each slice of the tensor has to be of the same dimension. In other words, our tensor has to be a perfect "block" that is balanced. PARAFAC2, on the other hand, allows for a tensor that consists of matrices of different sizes in one mode. In the real world, the document-word matrix of each time period could contain different numbers of documents. In other words, we could have a series of matrices across time with different dimensions. This occurs in the Twitter datasets, for instance, when different numbers of documents are generated across time. In that situation, PARAFAC2 would be an alternative candidate to the PARAFAC decomposition we adopt.

3.8 Conclusion

In this paper, I propose a novel analytic framework for analyzing social science text data. The framework is based on high-order tensor decomposition. It enables researchers to incorporate metadata of arbitrary dimensions into one model. The identification comes from exploiting the high-dimensional structure of the data to recover each dimension, which is “non-parametric" and thus free from the caveats of over-parameterization that many existing approaches suffer from. With extra dimensions, the method gives a topic summary in many more aspects—as many aspects as the researchers have information on, not just the topic-word summary. This makes it possible for social scientists to get a more comprehensive understanding of each topic and make proper inferences on group differences and semantic meaning.

Simulation on synthetic data suggests the method is much more capable of detecting topics varying over time/group than the state-of-the-art approach people use—my approach can recover the right topics while LDA gives wrong topics. It also beats LDA in coherence scores by a large margin. Applications on several real-world datasets also demonstrate how powerful the method is in extracting key information from large datasets.

3.9 Appendix

Topic Component Matrix

Suppose the tensor rank decomposition of a n-way tensor $\mathcal{X}$ is $\mathcal{X} = U^1 \otimes U^2 \otimes U^3 ... \otimes U^n$. Suppose U^d is the topic-word matrix. We want to show the Khatri-Rao product of matrices other than U^d

$$L = U^1 \odot U^2 \odot ... \odot U^n$$

for $U^1, ...U^n \in \{U^{-d}\}$ is the topic-document label matrix. i.e.,

$$\text{Prob}(w, e_1, e_2, ...e_n)) = L(\sum_{m=1}^{n-1}(e_m * \prod_{k=m+1}^{n} E_k) + e_n, :) \cdot U^{d^T}(:, w)$$

Proof. By definition of the Khatri-Rao product,

$$L\big(\sum_{m=1}^{n-1}(e_m * \prod_{k=m+1}^{n} E_k) + e_n, t\big) = U^1(e_1, t) * U^2(e_2, t) * ... * U^n(e_n, t)$$

By definition of tensor CP decomposition,

$$\text{Prob}(w, e_1, e_2, ...e_n)) = \mathcal{X}(e_1, e_2...w, ...e_n) = \sum_{t=1}^{T}\prod_{k=1}^{N} U^k(e_k, t)$$

Substitute the expression of L, we obtain

$$\begin{aligned}\text{Prob}(w, e_1, e_2, ...e_n) &= \sum_{t=1}^{T}[L(\sum_{m=1}^{n-1}(e_m * \prod_{k=m+1}^{n} E_k) + e_n, t) * U^d(w, t)] \\ &= L(\sum_{m=1}^{n-1}(e_m * \prod_{k=m+1}^{n} E_k) + e_n, :) \cdot U^{d\intercal}(:, w) \\ &= \sum_{t=1}^{T} \text{Prob}(t|e_1, e_2, ..e_n) * \text{Prob}(w|t)\end{aligned}$$

This means that L multiplied by U^{d^T} gives us the word-document matrix, which is the exact matrix we get when we "flatten" the tensor in 2-dimension.

□

The Equivalence with Tensor-PLSA

Proof. Let $x_{w,d,t,i_1,...i_n}$ be the data point in tensor $\mathcal{X}$. A PLSA model can be described as approximating $x_{w,d,t,i_1,...i_n}$ with

$$P(w, d, t, i_1, i_2...i_n) = \sum_{p} P(d|z_p)P(w|z_p)P(i_1|z_p)...P(i_n|z_p)P(z_p)$$

where $z_p \in \mathcal{Z}$ is a set of topics.

Let $c_{w,d,t,i_1,...i_n} = \sum_p U_{w,p} \cdot U_{d,p} \cdot U_{t,p} \cdot U^1_{i_1,p}... \cdot U^n_{i_n,p}$ is the reconstructed data.

Non-negative CP decomposition with KL divergence minimizes the following loss function:

$$L_{NNCP} = \sum_{w,d,t,i_1,..i_n} x_{w,d,t,i_1,i_2...i_n} log \frac{x_{w,d,t,i_1,...i_n}}{c_{w,d,t,i_1,...i_n}}$$

The tensor-PLSA model minimizes

$$L_{PLSA} = -\sum_{w,d,t,i_1,..i_n} x_{w,d,t,i_1,i_2...i_n} log\big(P(w, d, t, i_1, i_2...i_n)\big)$$

Notice that

$$L_{NNCP} = L_{PLSA} + \sum_{w,d,t,i_1,..i_n} x_{w,d,t,i_1,i_2...i_n} log x_{w,d,t,i_1,i_2...i_n}$$

The second term can be seen as a constant. Therefore, the two minimization problems are equivalent. □

Visualization

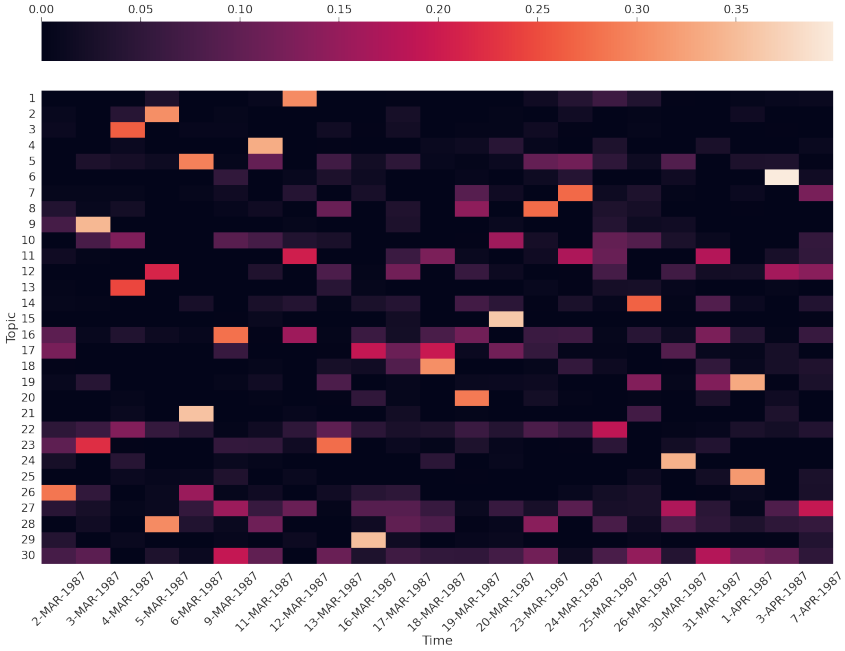


Figure 3.17: Topic-time: HDTM on Reuters-21578

Table 3.9: Congress Dataset Topics: 5D Modeling

Topic 1	barrett	coney	judge	amy	supreme	justice	confirmation	court	barretts	confirm
Topic 2	merry	christmas	wishing	holiday	family	wish	safe	happy	joy	hope
Topic 3	happy	birthday	flag	wishing	president	nation	service	answered	greatest	symbol
Topic 4	police	racial	officer	change	end	community	policing	hold	brutality	george
Topic 5	luther	martin	king	life	legacy	remember	equality	man	nation	changed
Topic 6	business	small	care	health	worker	relief	local	community	hospital	keep
Topic 7	enrollment	open	health	december	plan	insurance	visit	coverage	sign	affordable
Topic 8	postal	postmaster	general	usps	service	dejoy	mail	change	trump	delivery
Topic 9	relief	covid	vaccine	get	family	package	unemployment	small	stimulus	pas
Topic 10	veteran	woman	service	work	year	state	happy	congratulation	men	nation
Topic 11	job	economy	trade	tune	back	business	joining	discus	million	economic
Topic 12	impeachment	president	senate	trial	Democrat	witness	manager	partisan	case	evidence
Topic 13	senate	passed	act	pas	trump	black	year	Republican	step	stop
Topic 14	trump	president	administration	protect	wear	social	violence	right	community	mask
Topic 15	health	coronavirus	public	response	paid	family	resource	emergency	free	working
Topic 16	Democrat	china	law	right	chinese	communist	america	president	officer	political
Topic 17	question	information	town	hall	telephone	business	coronavirus	testing	payment	loan
Topic 18	election	every	biden	count	ballot	legal	democracy	joe	cast	voter
Topic 19	happy	wishing	wish	family	healthy	year	blessed	blessing	safe	prosperous
Topic 20	happy	wishing	year	safe	healthy	wish	wonderful	prosperous	birthday	hoping
Topic 21	de	la	office	community	school	federal	student	el	census	visit
Topic 22	Republican	direct	payment	senate	check	struggling	increase	pas	relief	voted
Topic 23	early	voting	location	day	open	plan	week	voice	sure	election
Topic 24	business	small	loan	program	protection	paycheck	application	relief	forgivable	assistance
Topic 25	spread	social	distancing	stay	home	coronavirus	slow	practice	follow	avoid
Topic 26	health	care	affordable	supreme	public	court	million	worker	access	fighting
Topic 27	iran	soleimani	terrorist	attack	qassem	responsible	hundred	military	terror	iranian
Topic 28	vaccine	operation	warp	effective	safe	speed	first	news	dos	fda
Topic 29	paycheck	protection	program	enhancement	small	business	care	additional	funding	health
Topic 30	voting	right	lewis	john	act	pas	restore	anniversary	ballot	access

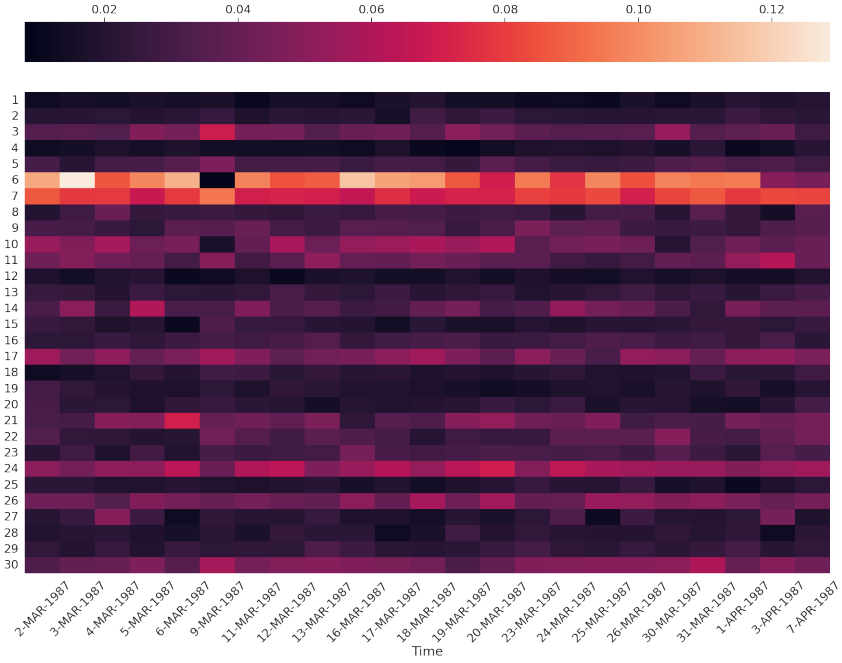


Figure 3.18: Topic-time: LDA on Reuters-21578

Table 3.10: Congress Dataset Topics: LDA

Topic 1	woman	call	resource	information	please	visit	town	issue	men	hall
Topic 2	national	day	happy	mask	security	year	hope	always	deserve	force
Topic 3	even	officer	fair	general	check	share	role	victim	individual	believe
Topic 4	senate	passed	bipartisan	legislation	act	bring	attack	since	protecting	create
Topic 5	police	stay	part	hand	follow	moment	keep	ready	story	north
Topic 6	every	better	way	affordable	death	day	lead	place	spread	middle
Topic 7	food	celebrate	emergency	assistance	disaster	supporting	damage	apply	weve	despite
Topic 8	Republican	stop	Democrat	political	senate	order	letter	money	payment	trying
Topic 9	life	working	work	together	forward	look	save	hard	prevent	worked
Topic 10	live	watch	democracy	war	question	tune	answer	long	ill	joining
Topic 11	service	student	school	military	congressional	team	city	last	high	budget
Topic 12	health	care	public	worker	crisis	medical	system	cost	protect	face
Topic 13	two	month	past	another	package	strong	free	using	took	prescription
Topic 14	voting	big	biden	defense	passing	voice	voter	thought	ballot	demand
Topic 15	wear	case	challenge	virus	virtual	folk	resolution	begin	generation	practice
Topic 16	de	la	el	vulnerable	en	le ¶	que	forced	los	
Topic 17	action	veteran	violence	hold	essential	step	leadership	taking	real	others
Topic 18	get	back	social	voted	covid	getting	hero	doe	direct	working
Topic 19	community	justice	child	leader	safety	keep	safe	serve	family	decision
Topic 20	lost	testing	find	open	senior	news	honored	plan	secretary	youre
Topic 21	fighting	america	world	making	ever	china	deal	lot	responsibility	reminder
Topic 22	trump	president	administration	white	history	show	away	mark	donald	entire
Topic 23	state	end	government	united	human	right	federal	local	judge	recognize
Topic 24	never	remember	million	job	made	still	done	cut	water	clean
Topic 25	election	read	full	black	office	statement	mental	resident	equal	contact
Topic 26	power	report	official	global	everything	admin	abuse	front	restore	party
Topic 27	policy	future	change	address	climate	increase	meet	work	commitment	pleased
Topic 28	family	friend	loved	home	hearing	year	ago	enough	freedom	thousand
Topic 29	protect	law	act	court	access	introduced	supreme	rule	line	allow
Topic 30	business	small	relief	pas	funding	provide	program	fund	federal	benefit

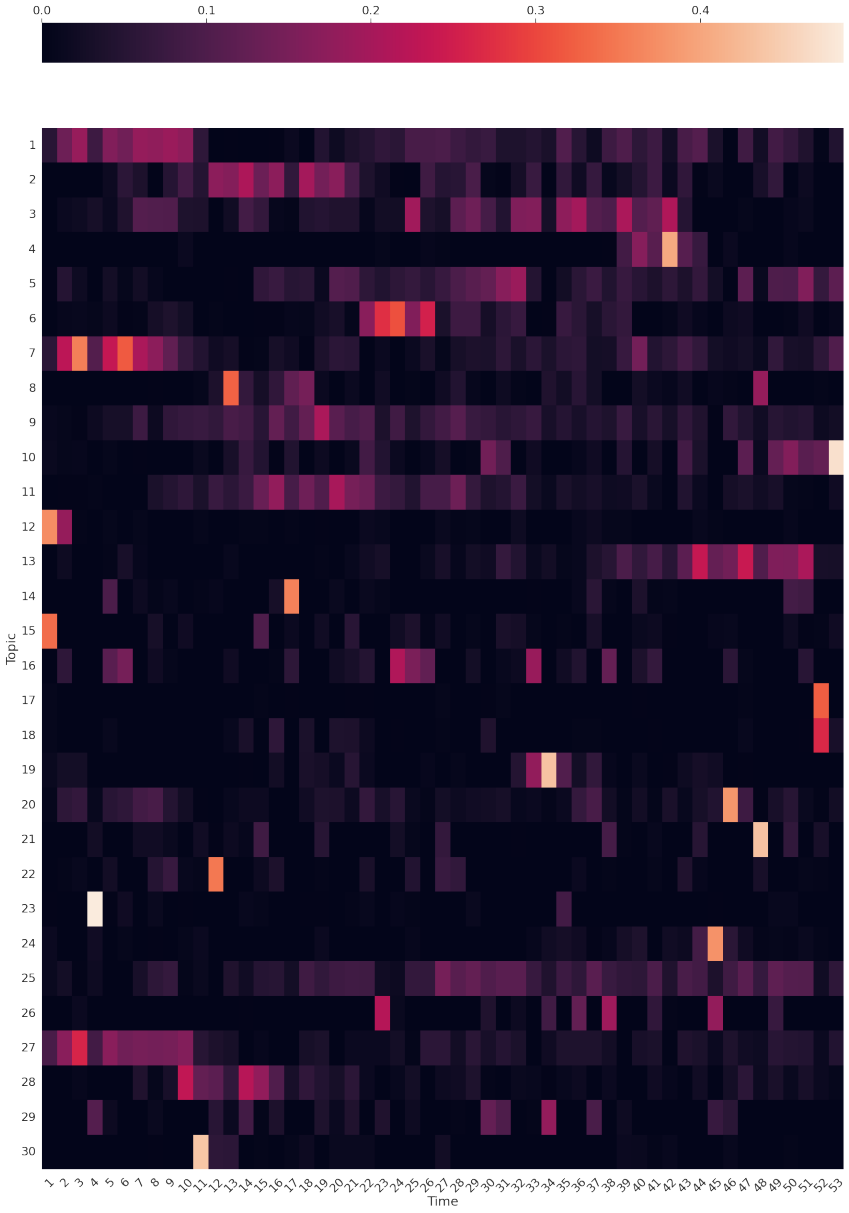


Figure 3.19: 4D topic over time

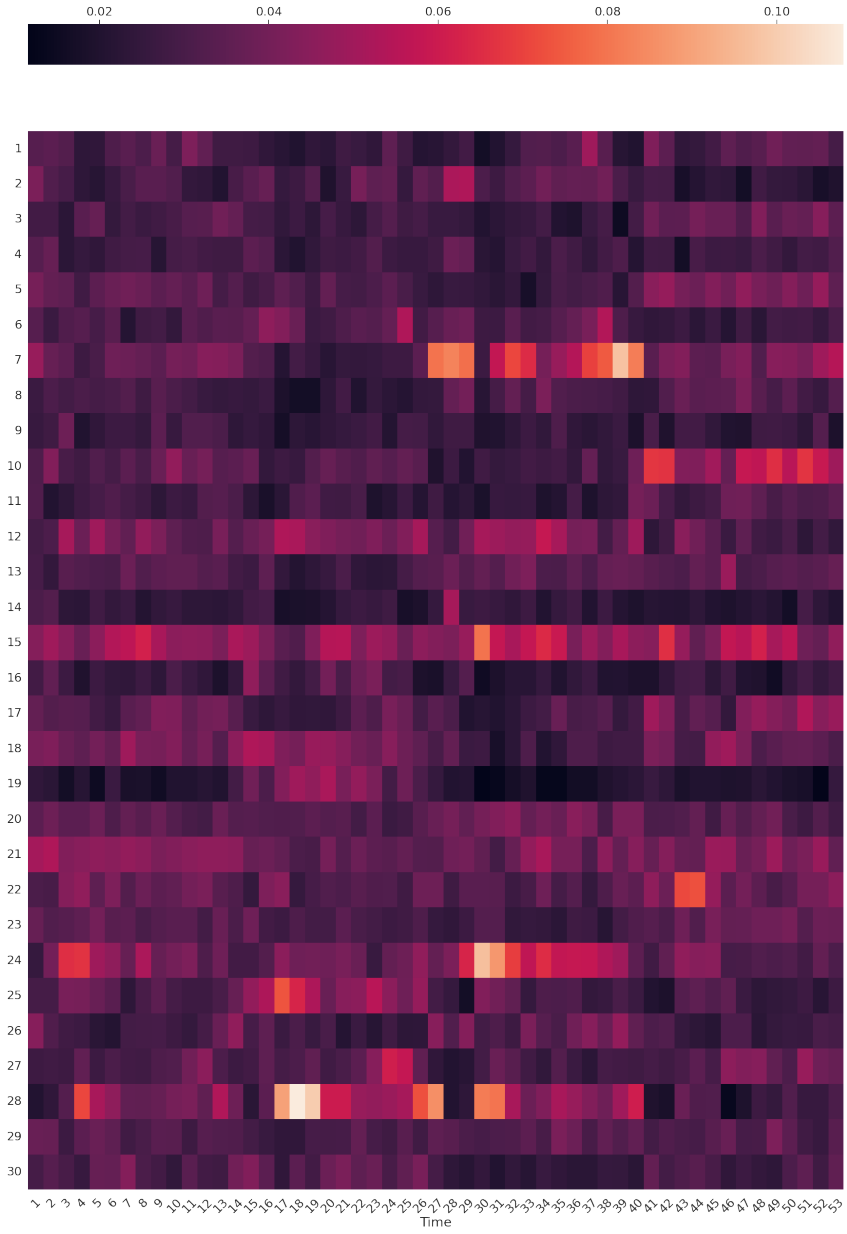


Figure 3.20: 4D topic over time(LDA)

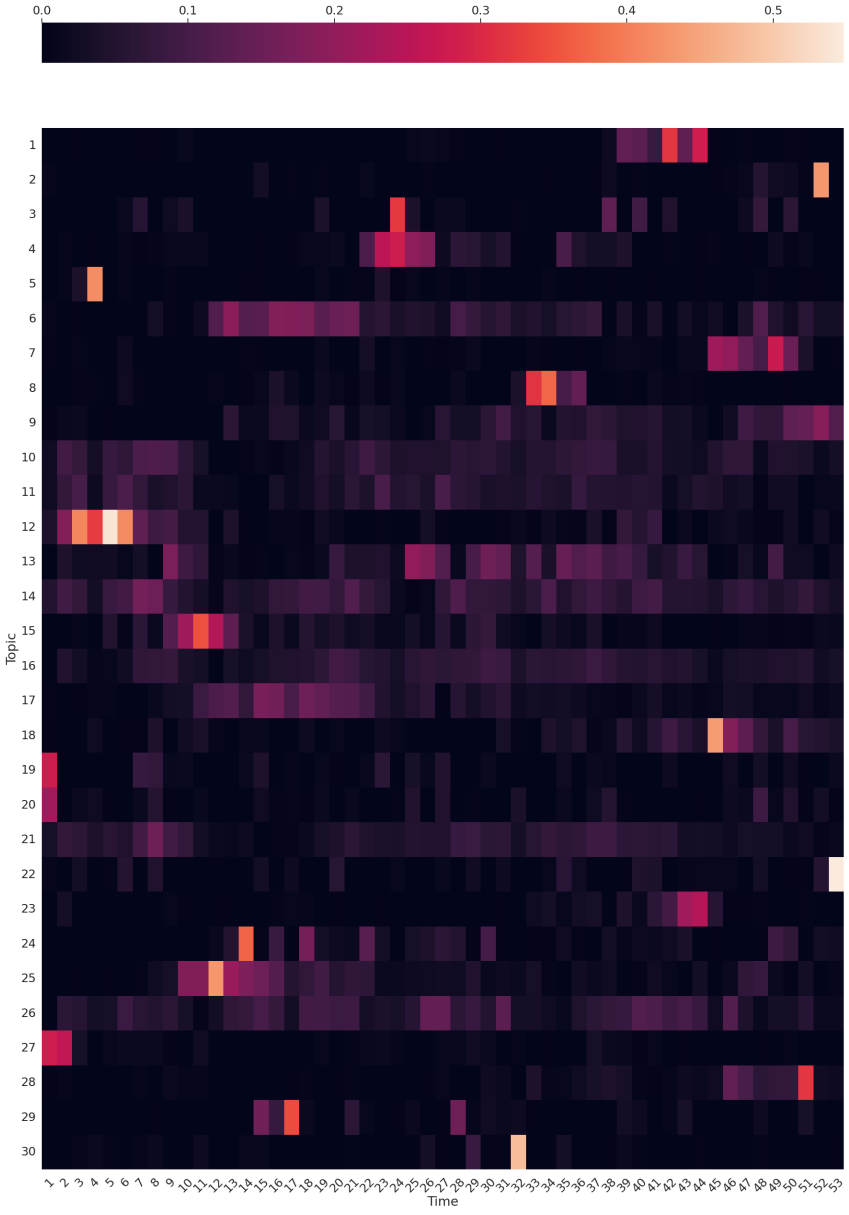


Figure 3.21: 5D topic over time

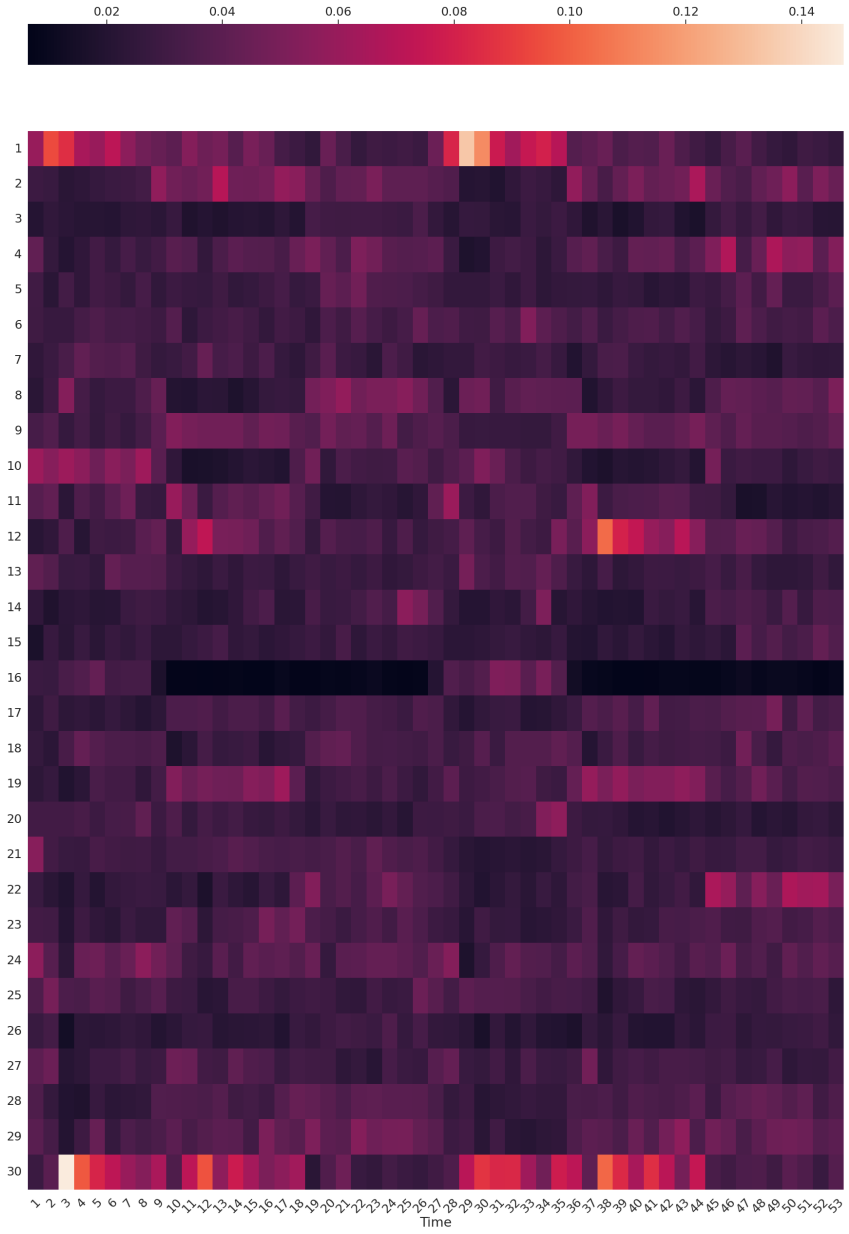


Figure 3.22: 5D topic over time (LDA)

References

Ahn, Miju et al. (2020). "On Large-Scale Dynamic Topic Modeling with Non-negative CP Tensor Decomposition". In: *CoRR* abs/2001.00631. arXiv: 2001.00631. URL: http://arxiv.org/abs/2001.00631.

Ambrosino, Angela et al. (2018). "What topic modeling could reveal about the evolution of economics". In: *Journal of Economic Methodology* 25.4, pp. 329–348. DOI: 10.1080/1350178X.2018.1529215. eprint: https://doi.org/10.1080/1350178X.2018.1529215. URL: https://doi.org/10.1080/1350178X.2018.1529215.

Anaissi, Ali, Basem Suleiman, and Seid Miad Zandavi (2020). *NeCPD: An Online Tensor Decomposition with Optimal Stochastic Gradient Descent*. DOI: 10.48550/ARXIV.2003.08844. URL: https://arxiv.org/abs/2003.08844.

Angelov, Dimo (2020). *Top2Vec: Distributed Representations of Topics*. DOI: 10.48550/ARXIV.2008.09470. URL: https://arxiv.org/abs/2008.09470.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 113–120. ISBN: 1595933832. DOI: 10.1145/1143844.1143859. URL: https://doi.org/10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (Mar. 2003). "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3.null, pp. 993–1022. ISSN: 1532-4435.

Coyle, E. (Jan. 2005). "Non-negative tensor factorisation for sound source separation". English. In: *IET Conference Proceedings*, 8–12(4). URL: https://digital-library.theiet.org/content/conferences/10.1049/cp_20050279.

Ebanks, D. (2021). "Tensor-Base Implementation for Joint Sentiment Topic Modeling: Methods for Large Scale Text Data". English. In.

Gaussier, Eric and Cyril Goutte (2005). "Relation between PLSA and NMF and Implications". In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '05. Salvador, Brazil: Association for Computing Machinery, pp. 601–602. ISBN: 1595930345. DOI: 10.1145/1076034.1076148. URL: https://doi.org/10.1145/1076034.1076148.

Grootendorst, Maarten (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. DOI: 10.48550/ARXIV.2203.05794. URL: https://arxiv.org/abs/2203.05794.

Hoffman, Thomas (1999). "Probabilistic latent semantic analysis". In: *In UAI'99*, pp. 289–296.

Kanungsukkasem, Nont and Teerapong Leelanupab (2019). "Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction". In: *IEEE Access* 7, pp. 71645–71664. DOI: 10.1109/ACCESS.2019.2919993.

Kolda, Tamara G. and Brett W. Bader (Sept. 2009). "Tensor Decompositions and Applications". In: *SIAM Review* 51.3, pp. 455–500. DOI: 10.1137/07070111X.

Lee, Daniel D. and H. Sebastian Seung (1999). "Learning the parts of objects by nonnegative matrix factorization". In: *Nature* 401, pp. 788–791.

Maier, Daniel et al. (2018). "Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology". In: *Communication Methods and Measures* 12, pp. 93–118. DOI: 10.1080/19312458.2018.1430754.

Roberts, Margaret et al. (2013). *The structural topic model and applied social science*. Working Paper. Harvard University OpenScholar. URL: https://EconPapers.repec.org/RePEc:qsh:wpaper:132666.

Shashua, Amnon and Tamir Hazan (2005). "Non-Negative Tensor Factorization with Applications to Statistics and Computer Vision". In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML '05. Bonn, Germany: Association for Computing Machinery, pp. 792–799. ISBN: 1595931805. DOI: 10.1145/1102351.1102451. URL: https://doi.org/10.1145/1102351.1102451.

Wang, Xuerui and Andrew McCallum (2006). "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. Philadelphia, PA, USA: Association for Computing Machinery, pp. 424–433. ISBN: 1595933395. DOI: 10.1145/1150402.1150450. URL: https://doi.org/10.1145/1150402.1150450.

Welling, Max and Markus Weber (2001). "Positive tensor factorization". In: *Pattern Recognition Letters* 22.12. Selected Papers from the 11th Portuguese Conference on Pattern Recognition - RECPAD2000, pp. 1255–1261. ISSN: 0167-8655. DOI: https://doi.org/10.1016/S0167-8655(01)00070-8. URL: https://www.sciencedirect.com/science/article/pii/S0167865501000708.