Identifying and Quantifying Coordinated Manipulation of Upvotes and Downvotes in *Naver News* Comments

Jiwan Jeong*
School of Computing
KAIST
jiwanjeong@gmail.com

Jeong-han Kang
Department of Sociology
Yonsei University
jhk55@yonsei.ac.kr

Sue Moon
School of Computing
KAIST
sbmoon@kaist.edu

Abstract

Today, many news sites let users write comments on news articles, rate others' comments by upvoting and downvoting, and order the comments by the rating. Top-rated comments are placed right below the news article and read widely, reaching a large audience and wielding great influence. As their importance grew, upvotes and downvotes are increasingly manipulated by coordinated efforts to hide existing top comments and push certain comments to the top. In this paper, we analyze comment sections of articles targeted by coordinated efforts and identify a trace of vote manipulation. Based on the findings, we propose a parameterized classifier that distinguishes comment threads affected by coordinated voting. The classifier only uses the number of upvotes and downvotes of comments. Therefore it is widely applicable to general vote-based curation systems where contents are sorted by the difference of upvotes and downvotes. Using the classifier and our choice of parameters, we have examined six years of the entire commenting history on a leading news portal in South Korea. Manual inspection in partisan online communities could only identify a few hundreds of targeted articles. With our classifier, we have identified more than ten thousand comment threads with a high likelihood of manipulation. We also observe a significant increase in coordinated manipulation in recent years.

Introduction

Manipulation of public opinion over the Internet has emerged as a critical threat to our society across the globe (Woolley and Howard 2018). For the past decade or so, malicious actors have developed numerous techniques to utilize online platforms as a tool for scheming propaganda (Marwick and Lewis 2017). Fake news (Lazer et al. 2018) and social bots (Ferrara et al. 2016) are already getting great attention, but a vast spectrum of malicious efforts remain unaccounted for.

In this paper, we focus on coordinated manipulation of upvotes and downvotes on news comment sections. Today, many sites allow users to write comments, rate others' comments by either upvoting or downvoting, and order the comments by the votes (Stroud et al. 2017), as shown in Figure 1.

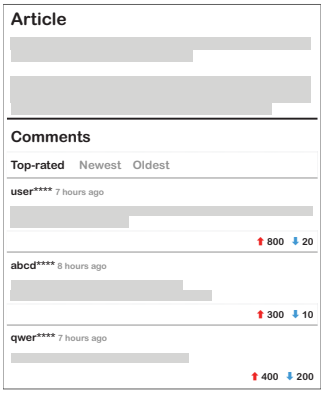

Figure 1: Simplified view of today's online news pages. The comment thread follows right below the news article. Readers can add a comment and upvote or downvote other comments. Many sites, by default, sort comments by the votes.

Readers choose to promote *good* comments by upvoting and punish *bad* comments by downvoting. Ranked by the votes, high-quality comments stay on top of the comment thread, whereas offensive, abusive, or malicious comments become hardly visible. Top-rated comments are placed right below the news article and read widely, reaching a large audience and wielding influence.

As the importance of top comments grew, people began to look at them as an opportunity for their online propaganda. A group of people flock to the same article, massively downvote unwanted comments, if any, and upvote comments of their choice to the top. Unlike conventional opinion trolls posting offensive partisan contents repeatedly (Zelenkauskaite and Niezgoda 2017; Mihaylov et al. 2018), they focus on a small number of well-designed comments and boost their visibility by vote manipulation (Carman et al. 2018). The affected comment threads represent one-sided opinions and distort readers' perception of public opinion.

In this paper, we study the coordinated manipulation of upvotes and downvotes using the data from Korea's largest news site, Naver News. The main goal of this paper is to classify comment threads affected by coordinated voting, and to measure the prevalence of such manipulation on Naver News in recent years.

*Also with Data Science Group, Center for Mathematical and Computational Sciences, Institute for Basic Science (IBS)

We have collected the entire articles and their comments published through Naver News from July 2013 to June 2019, and call-for-action posts from three partisan online communities targeting hundreds of those articles. Comparing the targeted articles against other articles, we have found out that highly upvoted comments with near-zero ranking scores–namely *zeroed comments*–are convincing evidence of coordinated voting. Based on the finding, we propose a parameterized classifier that distinguishes coordinated efforts. This classifier only uses the number of upvotes and downvotes of comments. Thus it can be widely applied to similar vote-based systems that sort contents by the difference of upvotes and downvotes, regardless of languages or data limitation.

Using the classifier and our choice of parameters, we examine the prevalence of such manipulation in Naver News comments. We have classified more than ten thousand comment threads as manipulated. We report that this type of coordinated manipulation has increased significantly in recent years and accounts for more than 30% of the most read articles in the political news category in certain periods.

News sites and other vote-based curating services, as well as the users, must heed to the growth of this type of manipulation and could make use of our classifier.

Background

In this paper, we study the manipulation of upvotes and downvotes in comment sections. While this paper focuses on a single platform, Naver News, many platforms are undergoing similar manipulative attacks called *vote brigading* (Wikipedia 2019b). This section describes recent online trends and environments that foster this type of misbehavior.

Wide adoption of vote-based curation

User-generated content is an essential resource in online services today. While high-quality or widely-agreeable content brings valuable user experience, abusive and deceptive content has emerged as a critical problem (Tsikerdekis and Zeadally 2014). To make good content visible as well as to hide problematic content at scale, many online services have adopted vote-based ranking (Momeni, Cardie, and Diakopoulos 2016). User comments on Yahoo! News and AOL News, posts on Reddit, and answers on StackExchange are a few such examples. In a vote-based ranking, the number of upvotes and downvotes determines the position of the content, thus affects its visibility (Lerman and Hogg 2014). It also indicates the degree of agreement or approval of the readers, and people gauge public reaction based on the vote counts. Votes are an efficient quality assessment mechanism (Goodman 2013; Ghosh and Hummel 2014) but, at the same time, a prime target for opinion manipulation (Li et al. 2019; Jeong et al. 2020).

Voluntary mobilization of partisan subcultures

Until recently, online manipulations have been mainly performed by bots (Ferrara et al. 2016) or crowd-sourced workers (Wang et al. 2012). However, recent studies report that such campaigns are increasingly operated by voluntary participation from partisan online subcultures. For example, Reddit's /The_Donald subreddit users mobilize information campaigns in 2016 election (Flores-Saviaga, Keegan, and Savage 2018), 4chan's /pol/ board users attack YouTube comment sections (Hine et al. 2017; Mariconti et al. 2019), and many other Internet subcultures develop their own way of spreading propaganda (Marwick and Lewis 2017). Their coordinated activities distort information flows and online discourse.

Public opinion perception in comment section

Comment sections in online news sites are where the readers express opinions and see others' perspectives (Diakopoulos and Naaman 2011; Springer, Engelmann, and Pfaffinger 2015). Although comments are written by a relatively smaller number of people than their readers (Stroud, Van Duyn, and Peacock 2016; Kim, Oh, and Choi 2016), readers tend to estimate public opinion based on those comments (Lee 2012; Neubaum and Krämer 2016) and change their attitude about the topic when they are exposed to others' comments (Lee and Jang 2010; Lee, Kim, and Cho 2016). Comment sections have long been a main target of opinion trolls spreading offensive and partisan ideas (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015; Mihaylov et al. 2018). To moderate such malicious comments, many news sites rely on user-contributed upvotes and downvotes (Stroud et al. 2017).

Weaponization of upvotes and downvotes

Manipulative actors have developed strategic use of upvotes and downvotes as a tool to control information flows in today's online news environment (Carman et al. 2018). The motivations for vote manipulation are deeply rooted in social theories such as the mere-exposure effect (Zajonc 1968), the bandwagon effect (Leibenstein 1950), and the spiral of silence theory (Noelle-Neumann 1974). Also, recent randomized experiments have shown that a small number of vote manipulation often affect the final rating significantly (Muchnik, Aral, and Taylor 2013; Glenski and Weninger 2017; Carman et al. 2018).

Vote Manipulation on Naver News

While Facebook and Twitter are often the common ground for public opinion manipulation in many countries, Naver News is the number one news portal site in Korea and its user comment section is the prime target.

Naver News, the target platform

Naver is the biggest online portal in South Korea and Naver News is the number one news site, where 65% of Korean adults use as their major news consumption channel (Kim and Kim 2018). As many news sites do, Naver News lets users write comments, upvote or downvote others' comments, and order the comments by the ranking score of the votes.

Since July 2013, Naver News displayed comments by the ranking score calculated as $u - d$, where u and d are the

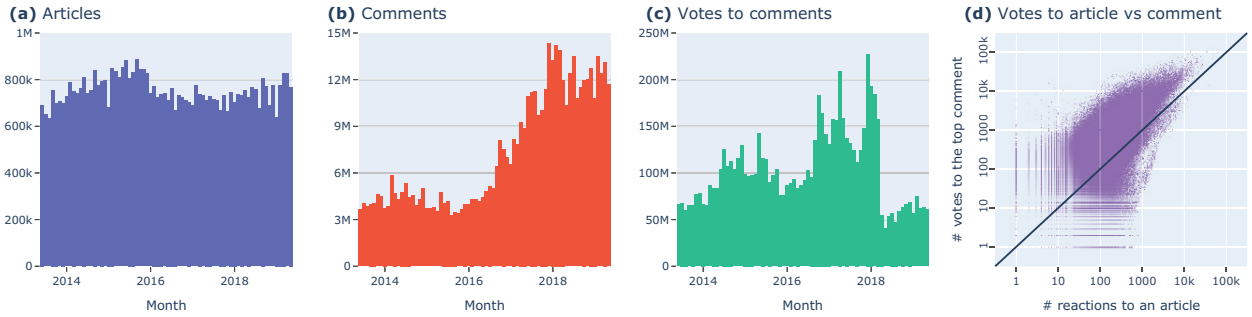

Figure 2: Volume of Naver News. (a) Monthly number of articles published through Naver News. (b) Number of comments on the articles. (c) Total number of upvotes and downvotes to the comments. (d) Comparison of the volume of votes to the most voted comment against that to the article itself.

numbers of upvotes and downvotes a comment received, respectively (Naver News 2013). We have collected the entire articles published through Naver News between July 2013 and June 2019 and all comments on the articles. As Figure 2 shows, writing and rating comments are immensely popular in Naver News. In any month, users wrote millions of comments, and those comments received hundreds of millions of upvotes and downvotes.

In particular, the top-rated comments in Naver wield great influence. A straightforward way to estimate the reach of top-rated comments is to compare the volume of votes to its top comment and the article itself. We compare the total number of votes of the most voted comment against that of the article itself in Figure 2(d). For easy reference, we draw a line of equal reactions to the most voted comment and the article. For most articles, users react more to the top comment than the article itself.

In December 2015, Naver News changed the ranking score to $u - 3d$ (Naver News 2015) to penalize abusive comments efficiently. However, this change made it easy to downvote any comments and get them off the top rank; that is, only a third of effort was needed than before to downgrade a comment. Accordingly, people began to use voting as a tool for online propaganda, and soon it became a critical social problem (Kim and Oh 2018).

Naver News recognized this source of vulnerability for relatively easy manipulation and, in December 2017, switched the ranking score back to $u-d$ (Naver News 2017). Despite the change, the vote manipulation problem continued to 2018. In April 2018, as the 2018 opinion rigging scandal erupted (Wikipedia 2019a), Naver News began to restrict users' voting activities heavily (Naver News 2018a; 2018b; 2018c), resulting in a significant drop in the volume of votes, as seen in Figure 2(c).

We put additional details on Naver News in detail in Appendix.

The way of mobilization

Over more than two years, we have observed in a selected set of ideology-driven online communities people mobilizing for semi-organized vote manipulations. They share target articles by posting on the forum with specific *call-for-action* keywords. The community-specific jargon allows the members to easily find and participate in ongoing operations, as well as prevent outsiders from tracing the manipulation via search engines.

In closed communities, such manipulations were performed more systematically. In February 2018, a Google Docs link of a vote manipulation manual, presumed to be from a political support group, was leaked (Jeong 2018). The manual contains detailed instructions, including monitoring schedules, news topics to target, recommended tones of arguments, and how to use a secret browser installed on their encrypted USB flash drives.

Whether semi-organized or systematic, the way to mobilize collective action is practically the same. Once a few members discover a target news article, they share the link to the article through community forums, social media, or chat rooms. Then others flock to the article, massively downvote the unwanted comments, if any, and upvote comments of their choice to the top. They do not necessarily write comments. Often they choose among existing comments and upvote them. The boosted comments are often neutral and well-written, ingeniously diverting the flow of unfavorable conversations.

Naver News does not provide a direct link to a comment, so the only way to access any comment is to scroll down from the top of the thread to its position. To click votes to a comment, one must log in with an authenticated Naver account. Fake accounts are impossible to make, so the participants were encouraged to exploit their family members' accounts (Lee 2017). It is not easy to automate voting on Naver News because Naver News does not provide an open API. Therefore, many of the operations were done manually by the participants (Jeong 2018; Park and Lee 2018; Kim 2018).

Questions, challenges and our design decision

We have seen some online communities mobilizing vote manipulation, but they are just a tip of the iceberg. Because many of such manipulation efforts are conducted via unobservable channels such as membership-only forums or private chatrooms (Lee 2017; Park and Lee 2018; Jeong 2018).

Then, how many articles were affected by such coordinated efforts? When did people begin this type of vote manipulation? How big was the impact?

Identifying and quantifying vote manipulation is an essential step to answer the questions, but is a serious challenge because the ground truth dataset is almost impossible to build. In this study, we have collected a limited number of call-for-action posts, but we do not know whether they have successfully mobilized coordinated action or not. For a given comment thread that we could not find any corresponding call-for-action, are we sure that it has not been targeted by a closed community?

Also, the information available for votes is very limited. Naver News displays the numbers of upvotes and downvotes per comment, but there is no information about who the voters are and when the votes were clicked. Given the limited information, we only use the numbers of upvotes and downvotes per comment in our analysis. We demonstrate later that only with the numbers of upvotes and downvotes we could still infer the presence of coordinated actions.

On the lack of ground truth, we choose not to attempt labeling data and not to rely on supervised machine learning. Instead, we focus on quantitative characteristics that are common in targeted articles but almost nonexistent in general articles and explain why it is convincing evidence of coordinated voting.

In summary, considering the challenges and decisions above, we define our problem as follows. *Given a snapshot of a comment thread, can we estimate the degree of coordinated voting and identify manipulation using only the number of upvotes and downvotes for each comment?*

Dataset

Naver News publishes daily lists of the entire articles published through the platform.[1] We have collected all articles published from July 17th, 2013 to June 30th, 2019, and the entire comments on those articles. To collect the data, we built a Python package for crawling Naver News, now available on PyPI.[2]

The comment threads were collected at least 1 week later from the article publication, thus the dataset we have is a stabilized snapshot of each comment thread. Naver News anonymizes the author of a comment, thus we cannot identify who the writer is. For each comment, the precise numbers of upvotes and downvotes are provided, but no information is available for who the voters are and when the votes were clicked.

During our collection period, Nave News changed the ranking score equation twice. Accordingly, we divided the dataset into three periods as in Table 1. In this paper, we mainly use data from Period II when the comments were sorted by $u - 3d$, because our collection of call-for-action posts is concentrated on that period. The targeted and comparative articles described in the following subsections all belong to Period II. The data from other periods will be used in the longitudinal analysis.

[1]https://news.naver.com/main/list.nhn

[2]https://pypi.org/project/portalnews

Table 1: Naver News Dataset Statistics

	Period I	Period II	Period III
From	2013-07-17	2015-12-08	2017-12-01
To	2015-12-07	2017-11-30	2019-06-30
Ranking Eq.	$u - d$	$u - 3d$	$u - d$
# Articles	22.6M	17.4M	14.1M
# Comments	122M	166M	235M
# Votes	2.74B	2.92B	1.74B

Targeted articles on Naver News

We looked for call-for-action posts in three open, anonymous, and yet partisan online communities: a far-right, a female chauvinism, and a political fandom community. These communities developed their own jargon for vote manipulation. By looking at words frequently appearing on these sites, we have built a list of call-for-action words per site.[3]

From the three online communities, we have collected posts that include such call-for-action words in the title and a link to a Naver News article in the body. Then, we excluded the posts with fewer than 10 comments on each community because we think that the exposure was not enough to bring coordinated efforts. Finally, we manually checked the titles and sorted out valid call-for-action posts. In total, we collected 316 call-for-action posts and 281 targeted articles. The targeted articles have 566,000 comments in total.

Matched articles for comparison

To compare to the targeted articles, we constructed a set of articles of similar characteristics. For each targeted article, we randomly selected an article from the same category, having a similar number of comments within a 5% error range. Thus, the matched set of articles, compared to the targeted articles, has exactly the same composition of categories and almost identical distribution of comment counts. In total, the 281 matched articles have about 564,000 comments. We presume that these articles are far less likely to have been targeted.

Weather forecasting articles

In addition to the matched articles, we have chosen weather forecasting articles. We assume it is very likely that people visit weather forecasting articles individually without any preference or coordinated efforts. The headlines of important weather forecasting articles are prefixed with "[Weather]" in Naver News. We collected all articles with the prefix in the title. Again, to control the reaction size, we randomly selected a weather article for each targeted article to have a similar number of comments within a 5% error range. The resulting set contains 281 weather forecasting articles and about 567,000 comments.

Identifying the Trace of Vote Manipulation

The first part of this study is to identify the trace of coordinated manipulation. What combinations of upvotes and

[3]We will provide information about the communities and the call-for-action words upon requests.

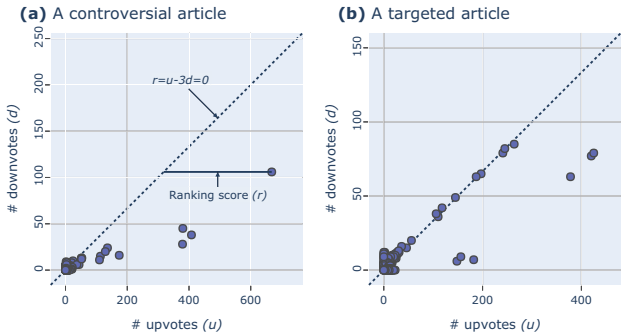

Figure 3: Scatter plots of upvotes and downvotes to comments to (a) a controversial article and (b) a targeted article.

downvotes of comments are prevalent on targeted articles but rare on general articles? If any, how can we explain that it is a trace of coordinated voting?

Prevalence of zeroed comments in targeted articles

We begin our comparison with simple scatter plots of upvotes and downvotes of comment threads on two sample articles. In Figure 3, each point represents a comment, where the horizontal coordinate is the number of upvotes u, and the vertical coordinate that of downvotes d. Recall that during Period II, Naver News displayed comments by the ranking score, $r = u - 3d$. In Figure 3 we put a dotted line to mark where the ranking score is zero, $r = 0$. The horizontal vector from the dotted line to a point is the final ranking score, r, of the corresponding comment. That is, the comments above the dotted line have negative ranking scores.

Figure 3(a) represents a comment thread to a news article about a controversial politician. We see a few comments with more than 300 upvotes and a few more comments with more than 100 votes, and most comments are clustered at the origin. The top-ranked comment has received almost 700 upvotes and about 100 downvotes. From the outlook of the comment thread alone, we gather that the top-ranked comments were favorably received but not without disagreement. Only a small number of comments receive many upvotes and remain top-ranked, while most comments remain lowly voted.

In Figure 3(b) of a targeted political article, we see similar points with large numbers of upvotes and smaller numbers of downvotes below the zero-score line. Yet, there are many points on or near the zero-score line, which have not been observed in Figure 3(a). There are only 4 comments with more than 200 upvotes in Figure 3(a) and they are all top-ranked accordingly. In Figure 3(b), there are at least 5 comments that have received more than 200 upvotes but, in the end, received enough downvotes to reach a ranking score of near 0. We call such comments with large numbers of votes but near-zero ranking scores, *zeroed comments*.

Such zeroed comments are prevalent in the targeted articles, but rare in the general articles. To show the difference, we compare the conditional probability distribution of the upvote proportion of comments grouped by the logged number of total votes, $Pr\{\frac{u}{u+d} | \log(u+d)\}$, in Figure 4. Here, each vertical grid represents the probability distribution of the corresponding number of total votes, with the probability represented as a color. Thus the sum of the color values of each vertical grid is equal to 1.

Figure 4(a) shows the distribution for all comments on the entire articles. Most of the highly-voted comments received a dominant proportion of upvotes than downvotes. On the other hand, in targeted articles as shown in Figure 4(b), a significant amount of highly-voted comments have an upvote proportion near 75%, which are zeroed comments. However, such zeroed comments are barely observed in the matched articles and the weather articles in Figures 4(c) and (d). That is, the news category or the volume of comments on a thread do not show correlation with zeroed comments.

These zeroed comments are not easy to understand nor explain and we dig further in the next subsection.

Growth dynamics of zeroed comments

The existence of zeroed comments drew our attention to its unlikely growth dynamics. Unless reaching a high ranking score, a comment would not be exposed at the top of the thread, and later readers are less likely to see and vote. Once a comment reaches a negative score, it is placed below most comments with near-zero scores, thus loses visibility. This is the reason why controversial comments that attracted a comparable ratio of upvotes and downvotes do not have high exposure and remain near the origin, as in Figure 3(a).

However, we observed that such comments with a large number of votes were prevalent in most of the targeted articles. In order to explain why zeroed comments are rare in general articles but prevalent in targeted articles and to claim that such comments are a trace of coordinated activity, we posit the following premise. *When all users individually visit a comment thread, they arrive in a random order irrespective of their preferences to the comments.* On the contrary, we characterize sequential visits from a group of users with a biased preference as a coordinated activity. Figure 5 illustrates the difference in arrival patterns.

Contradiction of zeroed comments Under the premise, zeroed comments are unlikely to occur in individual and independent voting. Let us take an example. Consider a comment with 300 upvotes and 100 downvotes. From the total number of votes, we gather it must have remained highly ranked for a long time, because it is almost the only way to be exposed to a large audience. However, if the upvotes and downvotes have arrived randomly, that is, in a well-mixed order, the ranking score of the comment mostly remains near zero and often falls below zero, as shown by the randomized growth paths in Figure 6(a).

There exists about 2.24×10^{96} possible growth paths from (0,0) to (300,100).[4] Among all the candidate paths, 99.7% reached a negative score before receiving all votes. If a comment already reached a negative score and lost visibility, how could it receive remaining votes? Also, 98.5% never

[4]The numbers are calculated by dynamic programming equations in the Appendix.

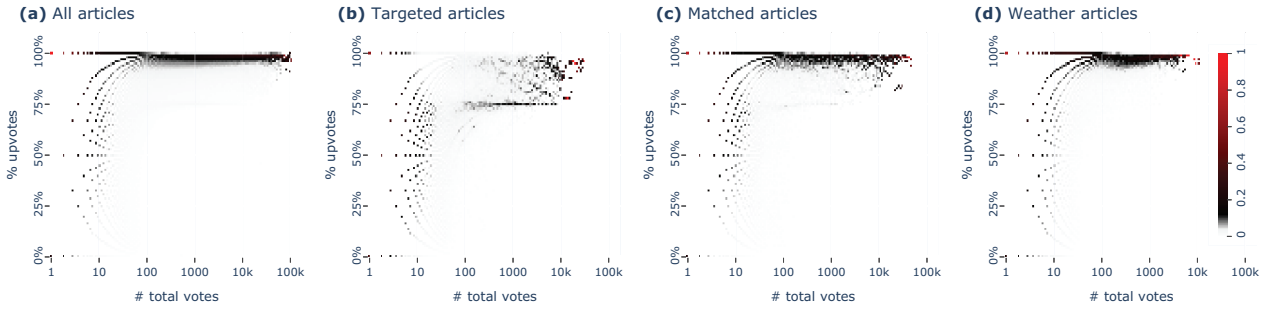

Figure 4: Conditional probability distribution of the upvote proportion of comments given the logged number of total votes, $Pr\{\frac{u}{u+d}|\log(u+d)\}$, aggregating all comments on (a) the entire articles, (b) targeted articles, (c) matched articles, and (d) weather articles.

Figure 5: Illustration of user arrival patterns in (a) individual behavior and (b) coordinated behavior.

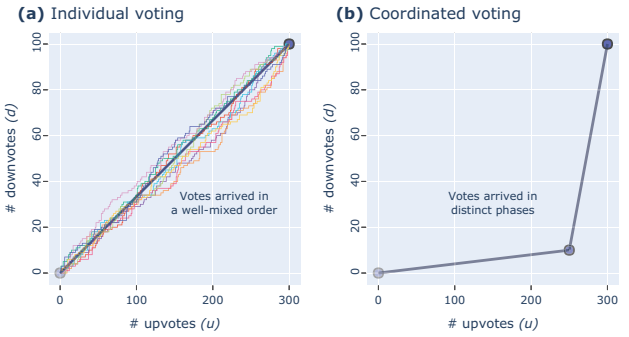

Figure 6: Expected growth dynamics of zeroed comments in (a) individual voting and (b) coordinated voting.

reached a ranking score of 50 even for a moment. If a comment has never kept a high ranking score for a while and never been placed at the top, how could it reach hundreds of voters? It is unlikely.

Then, how do such zeroed comments occur? It is reasonable to assume that the votes have arrived in two contrary phases. In Figure 6(b), the comment first gets highly upvoted and exposed at the top, then downvoted by later readers of the opposite preference. Once a comment gets downvoted enough, it becomes buried under the majority of other comments on the thread and ends up on the zero-score line. However, the growth paths like Figure 6(b) is extremely unlikely with random arrivals. Among the all possible permutations of the vote arrivals, only 0.00000384% ever reach a ranking score 100, and 0.00000000000000000000000000000135% reach 200. Therefore, we argue that the votes are not cumulated by random arrivals of individual voting.

Coordinated voting explains zeroed comments A plausible cause for such growth dynamics is coordinated vote manipulation. When *coordinated downvoting* takes place, actors look over tens or even hundreds of comments from the top of a thread, identify targets, and downvote them. Once the targeted comments get downvoted enough and become zeroed, they lose exposure even to the actors, thus end up on the zero-score line. Conversely, comments boosted by *coordinated upvoting* get downvoted later by the general public end up near the zero-score line.

Our conceptualization of coordinated voting is based on the result of collective behavior rather than the intention. The growth dynamics in Figure 6(b) may take place without a particular motivation. For example, an article published during the work hour might be first read by teenagers and later by the employed, resulting in two distinctive voting phases. We consider such cases as *unintentionally coordinated* behavior, because such dynamics possibly distort others' perception of social consensus as well as intentionally coordinated voting.

Here, we distinguish individual voting and coordinated voting in terms of the arrival order of votes. Highly-voted zeroed comments are almost impossible to occur if the upvotes and downvotes arrive in a well-mixed order. Therefore, we argue that zeroed comments are a convincing trace of impactful coordinated vote manipulation.

Quantifying Zeroed Comments and Articles

Then, how can we make the use of zeroed comments in classifying coordinated manipulation? Here, we present how we quantify zeroed comments and how we apply this in our classifier.

Quantitive definition of zeroed comments

Before moving to the quantitative analysis, we introduce the following notation to refer to comments satisfying certain conditions concisely.

Definition 1 *A comment is a [condition]-comment if the comment satisfies the condition.*[5]

[5]For example, [u≥50]-comments refer to the comments with 50 or more upvotes.

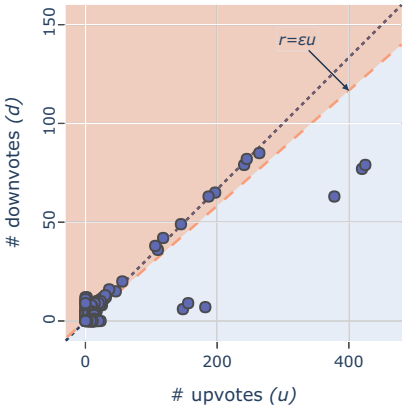

Figure 7: Quantitative definition of zeroed comments.

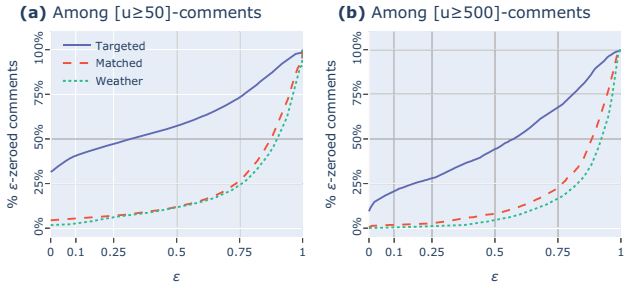

Figure 8: Proportion of ϵ-zeroed comments among (a) [u$\geq$50]-comments and (b) [u$\geq$500]-comments.

In the previous section, we have described the unlikelihood of zeroed comments based on the randomized growth dynamics in individual and independent voting. That is, a comment with a non-dominant ratio of upvotes to downvotes does not have the growth momentum to be exposed at the top of the thread, thus it should remain near the origin. To that extent, we define the border of zeroed comments and non-zeroed comments as a straight line passing through the origin on the upvote-downvote scatter plot as the following.

Definition 2 *A comment is an ϵ-zeroed comment if $r \leq \epsilon u$ given $r = u - 3d$, where u, d and r are the number of upvotes, the number of downvotes and the ranking score, respectively.*

Figure 7 illustrates our quantitative definition of zeroed comments. Then, what value of ϵ should we take? We compare the proportion of ϵ-zeroed comments among highly-upvoted comments by article groups, varying ϵ from 0 to 1 in Figure 8.

When $\epsilon = 1$, all comments are classified as zeroed comments by definition. If $\epsilon = 0$, only the comments with a non-positive score are classified as zeroed comments. For example, in Figure 8(a), more than 30% of [u$\geq$50]-comments on targeted articles have non-positive ranking scores, but the proportion is less than 5% in the matched articles and 2% in weather articles.

In Figure 8, any value of ϵ not close to 1 distinguishes targeted articles well, but we want to set the value conservatively in order to avoid misclassifying non-coordinated action as manipulation. Therefore, we choose to use $\epsilon = 0.1$ for the rest of our analysis and refer them as *zeroed comments*, omitting ϵ, for brevity.

Definition 3 *A comment is a zeroed comment if it is a 0.1-zeroed comment, i.e. $r \leq 0.1 \times u$.*

Human evaluation of zeroed comments

Downvoting a comment with high upvotes to a near-zero score requires many down-voters. Likewise, boosting a comment needs up-voters. Thus the higher the number of upvotes a zeroed comment received, the more likely it was affected by coordinated voting. Also, one or two zeroed comments may occur by chance, but multiple zeroed comments on a single article increase the likelihood of coordinated efforts.

To check the above assumptions, we manually examine hundreds of randomly chosen articles with zeroed comments and categorize them either manipulated or not. Note that, this human examination does not aim at evaluating the precision of zeroed comments in identifying coordinated manipulation. Such validation is not feasible as we lack the ground truth. Rather, our goal is to demonstrate how the increase in zeroed comments aligns with the number of upvotes and, eventually, the likelihood of vote manipulation.

Per comment A comment with 10 upvotes can easily become zeroed without much effort, but a comment with 100 upvotes would not. To that extent, we grouped zeroed comments on the entire articles by the number of upvotes—[20,50), [50,100), [100,200), and [200,500)—and randomly picked up 50 comments in each group.

Factors we took into consideration in our manual inspection are text contents, the order of creation, timestamps, and vote counts. We decided a zeroed comment was more likely affected by coordinated efforts if the comment had a partisan argument or when the later highly-voted comments were written consecutively and have a one-sided opinion. Also, when there exists a comment that reported a coordinated attack with a proof of the corresponding call-for-action.

On the contrary, we considered a zeroed comment was less likely affected by coordinated manipulation in the following cases: if the zeroed comment was one of the earliest created on the thread because it is well placed in oldest-first sorting regardless of the score; if most of the other comments have a lower ratio of upvotes to downvotes than the selected zeroed comment. In cases we were not sure about the existence of coordination, we labeled the comments as ambiguous.

Figure 9(a) presents the inspection result. As we have expected, the higher the number of upvotes, the more the zeroed comments seem to be affected by coordinated efforts. For [20$\leq$u<50]-zeroed comments, we validated 58% were affected by coordinated efforts. The number increases to 76% in [50$\leq$u<100]-, 88% in [100$\leq$u<200]-, and 94% in [200$\leq$u<500]-zeroed comments.

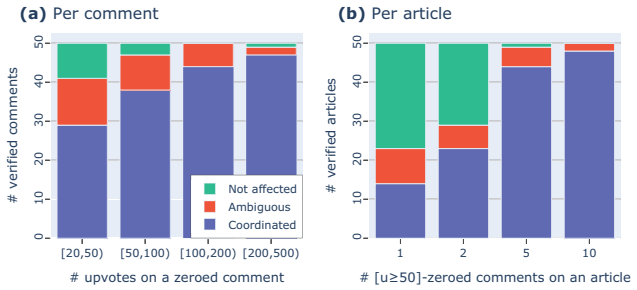

Figure 9: Human validation result of zeroed comments at (a) comment-level and (b) article-level.

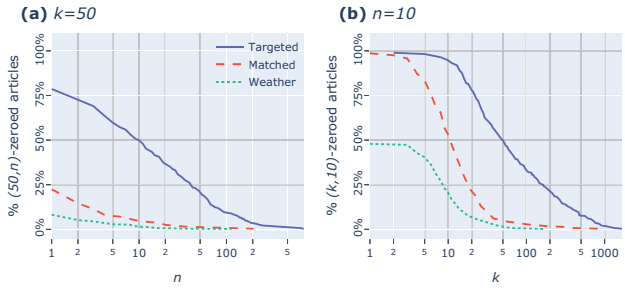

Figure 10: Proportion of *(k,n)*-zeroed articles by article groups, when (a) *k=50* and (b) *n=10*.

Per article Next, we conducted a similar validation by the article. Here, we focused on the number of zeroed comments per article. We posit that a large number of zeroed comments on a single article strongly imply coordinated vote manipulation. Based on this assumption, we grouped the entire articles by the exact number of [u≥50]-zeroed comments on their comment threads. For each group with 1, 2, 5, and 10 of [u≥50]-zeroed comments, we randomly selected 50 articles. Again, we inspected each article's entire comments and manually inspected existence of coordinated activity.

We plot the result in Figure 9(b). In articles with only one or two zeroed comments, the majority of the zeroed comments were the earliest created ones. However, when there are more zeroed comments in an article, the zeroed comments are less from the earliest. When there are 5 or 10 zeroed comments on an article, we observed that those zeroed comments often have similar partisanship and were replaced by later top comments of the opposite view. We validated that 88% of articles with exactly 5 of [u≥50]-zeroed comments were affected by coordinated manipulation. The proportion was 96% in articles with exactly 10 of [u≥50]-zeroed comments.

In summary, from our manual inspection of zeroed comments, we conclude that the higher the upvotes and the higher the co-occurrence of zeroed comments, the more likely that the article has been manipulated.

Quantitative definition of zeroed articles

In targeted articles, many comments with high upvotes become zeroed. We refer to such articles as *zeroed articles* and define as the following.

Definition 4 *An article is a (k,n)-zeroed article if it has n or more of [u≥k]-zeroed comments.*

For example, the comment thread in Figure 7 has 10 of [u≥50]-zeroed comments. Thus, this article is a (50,10)-zeroed article. By definition, any (k,n)-zeroed article is also a (k-1,n)-zeroed article and a (k,n-1)-zeroed article. That is, the article in Figure 7 is also a (50,9)-zeroed article and a (49,10)-zeroed article.

Interpretation of the parameters The result of Figure 9 suggests that the higher the values of k and n, the more likely (k,n)-zeroed articles are affected by coordinated vote manipulation. For example, (50,10)-zeroed articles are expected to classify coordinated efforts with precision close to 96%.

The values of k and n not only affect the precision but also reflect the degree of manipulation in the classified articles. Downvoting a comment with higher upvotes needs more down-voters, and boosting a comment to higher upvotes needs more up-voters. That is, k is related to the *size* of coordinated actors. Also, the number of zeroed comments in an article reflects the number of affected comments. Therefore, n is related to the *breadth* of the coordinated voting.

By modifying the value of k and n, we can detect comment threads that beyond a certain degree of coordinated efforts. Choosing higher values for k and n capture higher degree of coordinated efforts with high precision, but do not capture manipulations of the smaller degree thus lower the recall. The choice of the parameters is up to the users by their purpose.

Proportion (k,n)-zeroed articles In Figure 10(a), we plot the proportion of (50,n)-zeroed articles by the article groups fixing $k = 50$ and varying n. About 80% of the targeted articles belong to (50,1)-zeroed articles. That is, 80% of the targeted articles have at least one [u≥50]-zeroed comment. The proportion of such articles is about 20% in the matched articles and less than 10% in the weather articles.

The proportion of (50,10)-zeroed articles is 50% in targeted articles, but 5% in matched articles and 1% in weather articles. As we mentioned earlier, we expect that about 96% of the (50,10)-zeroed articles are affected by coordinated manipulation. In targeted articles, a significant proportion of articles have hundreds of [u≥50]-zeroed comments, implying massive breadth of the coordinated voting.

Next, we compare the proportion of zeroed articles by fixing $n = 10$ and varying k in Figure 10(b). The targeted articles often have 10 or more zeroed comments with hundreds of upvotes, by successfully inviting a large number of participants. However, such articles are very rare in the other articles.

Longitudinal Analysis

In the previous section, we have introduced new measures, zeroed comments and zeroed articles. In practice, most cases of concerted efforts are impossible to excavate, as numerous

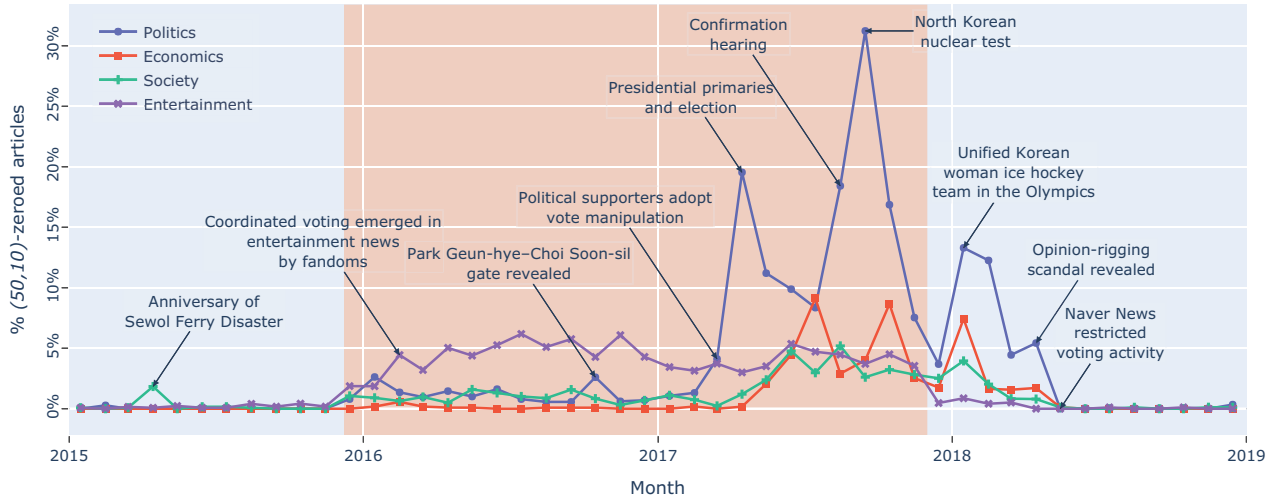

Figure 11: Proportion of *(50,10)-zeroed articles* in Ranking News—the daily lists of 30 most read articles on Naver News per category—binned by month. The shade marks the period when Naver News sorted comments by $u - 3d$. At other times, comments were sorted by $u - d$.

side channels are possible, such as membership-only communities, private chat rooms, or offline coordination.

Then, how many articles were affected by such coordinated efforts? We counted (50,10)-zeroed articles among the entire articles published from July 2013 to June 2019, and the number was 11,586. Note that, for the article published when Naver News sorts comments by $r = u - d$, we define comments satisfying $r \leq 0.1 \times u$ as zeroed comments.

Zeroed articles among *Ranking News* articles

In particular, we focus on the most read articles on Naver News in order to gauge the impact of vote manipulation. Naver News publishes *Ranking News* lists: daily lists of 30 most read articles for the following 7 categories: politics, economy, society, life/culture, world, IT/science, and entertainment.[6] We collect all the Ranking News articles and analyze the commenting trends. In total, the dataset includes about 470,000 articles and 400 million comments.

Among the 11,586 of all (50,10)-zeroed articles, 5,111 (44%) belong to the Ranking News. It implies that the manipulative actors mainly targeted the most influential articles.

For each category, we group the articles by the month of publication and calculate the proportion of (50,10)-zeroed articles. Figure 11 shows the trends in the politics, economics, society and entertainment categories. Other categories had relative less proportion of (50,10)-zeroed articles, thus we omit them in the plot. Also, we omit the period before 2015 and after 2019, because there were only a small number of (50,10)-zeroed articles in those periods.

In Figure 11, the coordinated vote manipulation began to emerge in December 2015. At that time, Naver News changed the ranking score equation from $u - d$ to $u - 3d$ for efficient moderation of abusive comments (Naver News 2015). In 2016, the coordinated vote manipulation was the most prevalent in entertainment news section. We examine the comments of these articles and find out that they were mostly fandoms of musicians, movies, or TV shows. The fandoms follow their celebrities or objects very closely and mob any article about what they support. Also, some articles in the society category are detected, which are mainly about gender issues and mobilized by radical feminists or anti-feminists.

Since 2017, such coordinated voting became immensely prevalent in politics news articles and also increased in society and economy categories, which are closely related to politics. In particular, 2017 was politically a very dynamic and turbulent period in the modern history of South Korea. In December 2016, a call for the impeachment of then-president Park Geun-Hye passed in the National Assembly and was confirmed by the Constitutional Court in March 2017. A presidential election was held in May 2017, and President Moon Jae-In was elected. In September 2017, North Korea conducted its 6th nuclear test and a missile test over Japan. These events resulted in heated discussions about the policy towards North Korea in political news comment sections. Accordingly, the mobilization of vote manipulation from political partisans peaked in this period.

In December 2017, on the prevalence of vote manipulation, Naver News changed the ranking score back to $u - d$ (Naver News 2017). However, the coordinated voting still remained and affected comment threads in early 2018. In April 2018, an opinion-rigging scandal broke out. Naver News began to restrict heavily users' commenting and voting activities. For example, since April 25, Naver News has limited the total number of upvotes and downvotes a user can make a day to 50 and enforced the 10 seconds interval between consecutive votes (Naver News 2018a). In addition,

[6] https://m.news.naver.com/rankingList.nhn

since May 15, comments on articles in the politics category can only be sorted by the newest-first (Naver News 2018b). Once the new restriction set up, the proportion of zeroed article dropped to almost zero.

Discussions

Fake consensus

Today's media ecosystem is highly participatory. News articles are published through the Internet, and the readers share their opinions by leaving comments and clicking upvotes or downvotes. Sorted by the difference between upvotes and downvotes, the top-rated comments reach a large audience, and the readers often estimate the social consensus by the vote counts.

However, in this paper, we show that the upvotes and downvotes are severely manipulated to hide existing top comments or boost certain comments to the top. The focus is on manipulating the visibility and popularity of opinions, whether or not the content is true. We argue that such vote manipulation is a type of misinformation, namely *'fake consensus.'*

Similar manipulations are also increasing on other platforms. For example, fake likes on Facebook (De Cristofaro et al. 2014; Badri Satya et al. 2016), fake views on YouTube (Chen, Zhou, and Chiu 2015; Keller 2018), or fake streams on Spotify (Leight 2019) are a few examples. Addressing these manipulations requires novel approaches because publicly available data is very limited.

Limitations of (k,n)-zeroed article

We have introduced (k,n)-zeroed articles as a parameterized classifier for the detection of coordinated vote manipulation. The design of (k,n)-zeroed article is based on the occurrence of zeroed comments in coordinated voting. Accordingly, the method has a limitation in detecting the cases where coordinated upvoting took place, but the boosted comments were not downvoted enough in the end, or the cases where coordinated downvoting took place but fail to drag down the targets to near-zero scores.

However, such non-zeroed comments are common in most articles irrespective of vote manipulation and from random arrivals of votes. That is, non-zeroed comments are likely to occur even if the arrival order is randomized. Thus the impact of such mobilization is limited and not enough to be a game changer.

Toward user interface augmentation

One controversial issue on the coordinated vote manipulation is the voluntariness of the participants. Should we consider their mobilization as freedom of expression? Even so, should we still inform other readers of such systematic and coordinated efforts?

The countermeasure of Naver News was to put heavy limitations on voting activity. The coordinated vote manipulation on Naver News, measured by (50,10)-zeroed article, has finally decreased after the restriction. However, by doing so, even ordinary users came to lose their freedom of commenting and voting activity.

This paper presents and analyzes the prevalence of coordinated behavior on a news portal site. News sites and other vote-based curating services, as well as users, must heed to the growth of this type of manipulation. We believe users would benefit by being knowledgeable about the likelihood of concerted manipulation and we recommend the user interface of web sites and services include such information.

Conclusion

Across the globe, Internet manipulation has emerged as a critical problem. Fake identities–social bots (Ferrara et al. 2016), sockpuppets (Kumar et al. 2017; Jen, Nuland, and Stamos 2017), paid trolls (Keller et al. 2017; Mihaylov et al. 2018) and crowd-workers (Wang et al. 2012; Lee, Tamilarasan, and Caverlee 2013; Fayazi et al. 2015)–mimic grassroots activities in various online platforms. Fake contents–rumors (Kwon et al. 2013), misinformation (Del Vicario et al. 2016) and fake/false news (Lazer et al. 2018; Vosoughi, Roy, and Aral 2018)–spread quickly and widely through social media deceiving the readers.

In addition, we are increasingly facing artificially boosted popularity of certain contents on the Internet, what we call *fake consensus*. Recently, we observed one of such cases on comment sections in online news sites. Facilitating recent commenting interfaces where comments are ordered by user voting, people attempt to place their comments at the top by concerted effort. Herding to a news article, they skim through the comment thread, upvoting their comments and downvoting the opponents'. The aim is to make their views regarded as public opinion by illusional popularity.

In this paper, we examined the coordinated comment section manipulation problem. We collected manipulation cases from partisan online communities and characterized the commenting dynamics. Based on the observation, we introduced zeroed comments and zeroed articles which highly distinguish targeted articles from general articles. Using these measures, we examined comment sections on Naver News spanning several years. The proposed method is cheap to calculate and only uses the number of upvotes and downvotes, thus would be widely applicable to other rank-order systems regardless of the language or anonymity.

This study not only reports a media manipulation case but also provides understanding about how people abuse user voting systems. The Internet has empowered user participation in the last decade, but also participatory misbehavior. We believe that the next step of the Internet should be moderating such participatory misbehavior.

Acknowledgments

This work was partially supported by Barun ICT Research Center at Yonsei University and the Institute for Basic Science (IBS-R029-C2). The data collection and interpretation was possible because Naver News archives all the previous articles and comments, and provides their policy modifications to the public. We thank Naver News for their data availability and open policy.

Appendix

Detailed Description about Naver News

News category News articles on Naver News are classified into the seven categories: politics, economy, society, life/culture, world, IT/science, and entertainment. (Sports news is separated.) The categorization is not edited by Naver News. The categorization scheme is de facto standard in South Korea, and press companies assign the categories before they provide the articles to Naver News.

Authentication and anonymity Creating a Naver ID requires valid authentication with the social security number or a registered cell phone number. To write a comment on Naver News, users need to sign in with their Naver, Facebook, or Twitter ID. However, to upvote or downvote a comment, users must sign in with a Naver ID. On the comment section, only the first 4 letters of a user ID are visible and followed by 4 asterisks regardless of its length, like 'abcd****.'

Votes to an article In Naver News, users can express emotions to articles just as we do on Facebook posts. Until recent, users could express only sympathy to an article, similar to like in Facebook. In March 2017, Naver News updated the vote feature enabling users to express more emotions; users can choose one of five pre-defined emotions, including "like," "warm," "sad," "angry," and "want follow-up stories."

Commenting interface Like many other news sites, Naver News lets users to write comments and rate them by upvoting or downvoting. The top 10 comments are displayed right below each news article, and users can load next comments by clicking 'read more' button. Since July 17th, 2013, Naver ordered the comments by $u - d$. On December 8th, 2015, Naver changed the rating score to $u - 3d$. On November 30th, 2017, Naver changed the rating score back to $u-d$. In addition, users have the choice in sorting the comments in newest-first, oldest-first, or highest proportion of upvotes first. Naver News does not provide a direct link to a comment, so the only way to access any comment is to scroll down from the top of the thread to its position.

Counting growth paths of upvotes and downvotes

To count the number of all possible growth paths of a comment with u upvotes and d downvotes, we use the following dynamic programming equation. The number of paths from $(0, 0)$ to (u, d) that never have greater/less score than s is,

$$F(u,d,s) = \begin{cases} 0, & \text{if } r \lessgtr s \\ 1, & \text{else if } u = 0 \text{ or } d = 0 \\ F(u\text{-}1, d, s) + F(u, d\text{-}1, s), & \text{else} \end{cases}$$

where $r = u - 3d$ or $r = u - d$ in Naver News.

References

Badri Satya, P. R.; Lee, K.; Lee, D.; Tran, T.; and Zhang, J. J. 2016. Uncovering Fake Likers in Online Social Networks. In *ACM International Conference on Information and Knowledge Management*, 2365–2370.

Carman, M.; Koerber, M.; Li, J.; Choo, K.-K. R.; and Ashman, H. 2018. Manipulating Visibility of Political and Apolitical Threads on Reddit via Score Boosting. In *IEEE International Conference on Trust, Security And Privacy in Computing and Communications*, 184–190.

Chen, L.; Zhou, Y.; and Chiu, D. M. 2015. Analysis and Detection of Fake Views in Online Video Services. *ACM Transactions on Multimedia Computing, Communications, and Applications* 11(2s):1–20.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial Behavior in Online Discussion Communities. In *International AAAI Conference on Web and Social Media*, 61–70.

De Cristofaro, E.; Friedman, A.; Jourjon, G.; Kaafar, M. A.; and Shafiq, M. Z. 2014. Paying for Likes? Understanding Facebook Like Fraud Using Honeypots. In *ACM Internet Measurement Conference*, 129–136.

Del Vicario, M.; Bessi, A.; Zollo, F.; Petroni, F.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America* 113(3):554–559.

Diakopoulos, N., and Naaman, M. 2011. Towards quality discourse in online news comments. In *ACM Conference on Computer-Supported Cooperative Work*, 133–142.

Fayazi, A.; Lee, K.; Caverlee, J.; and Squicciarini, A. 2015. Uncovering Crowdsourced Manipulation of Online Reviews. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 233–242.

Ferrara, E.; Varol, O.; Davis, C. A.; Menczer, F.; and Flammini, A. 2016. The Rise of Social Bots. *Communications of the ACM* 59(7):96–104.

Flores-Saviaga, C.; Keegan, B. C.; and Savage, S. 2018. Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community. In *International AAAI Conference on Web and Social Media*, 82–91.

Ghosh, A., and Hummel, P. 2014. A game-theoretic analysis of rank-order mechanisms for user-generated content. *Journal of Economic Theory* 154:349–374.

Glenski, M., and Weninger, T. 2017. Rating Effects on Social News Posts and Comments. *ACM Transactions on Intelligent Systems and Technology* 8(6):78:1–19.

Goodman, E. 2013. *Online comment moderation: emerging best practices*. World Association of Newspapers and News Publishers (WAN-IFRA).

Hine, G. E.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web. In *International AAAI Conference on Web and Social Media*, 1–10.

Jen, W.; Nuland, W.; and Stamos, A. 2017. *Information Operations and Facebook*.

Jeong, D.; Han, S.-P.; Park, S.; and Lee, S. K. 2020. Fighting Abuse while Promoting Free Speech: Policies to Reduce Opinion Manipulation in Online Platforms. In *Proceedings of the Hawaii International Conference on System Sciences*, 3981–3990.

Jeong, E. 2018. https://news.joins.com/article/22348244. *Joongang Daily*.

Keller, F. B.; Schoch, D.; Stier, S.; and Yang, J. 2017. How to Manipulate Social Media: Analyzing Political Astroturfing Using Ground Truth Data from South Korea. In *International AAAI Conference on Web and Social Media*, 564–567.

Keller, M. H. 2018. The Flourishing Business of Fake YouTube Views. *The New York Times*.

Kim, S., and Kim, W.-G. 2018. A Survey on Internet Users' Awareness for Portal News Service and Comments [http://www.kpf.or.kr/site/kpf/research/selectMediaPdsView.do?seq=574592]. *Media Issue* 4(5):1–13.

Kim, S., and Oh, S. 2018. *Operational Status and Improvement Direction of News Comments [*http://www.kpf.or.kr/site/kpf/research/selectMediaPdsView.do?seq=574912*]*. Korea Press Foundation.

Kim, S.; Oh, S.; and Choi, M. 2016. Analysis of Commenting Culture [http://www.kpf.or.kr/site/kpf/research/selectMediaPdsView.do?seq=573889]. *Media Issue* 2(10):1–16.

Kim, J. 2018. http://www.hani.co.kr/arti/economy/it/828980.html. *The Hankyoreh* (2018.01.23):A20.

Kumar, S.; Cheng, J.; Leskovec, J.; and Subrahmanian, V. S. 2017. An Army of Me: Sockpuppets in Online Discussion Communities. In *International World Wide Web Conference*, 857–866.

Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Prominent Features of Rumor Propagation in Online Social Media. In *IEEE International Conference on Data Mining*, 1103–1108.

Lazer, D. M. J.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; and Zittrain, J. L. 2018. The science of fake news. *Science* 359(6380):1094–1096.

Lee, E.-J., and Jang, Y. J. 2010. What Do Others' Reactions to News on Internet Portal Sites Tell Us? Effects of Presentation Format and Readers' Need for Cognition on Reality Perception. *Communication Research* 37(6):825–846.

Lee, E.-J.; Kim, H. S.; and Cho, J. 2016. How user comments affect news processing and reality perception: Activation and refutation of regional prejudice. *Communication Monographs* 84(1):75–93.

Lee, K.; Tamilarasan, P.; and Caverlee, J. 2013. Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media. In *International AAAI Conference on Web and Social Media*, 331–340.

Lee, E.-J. 2012. That's Not the Way It Is: How User-Generated Comments on the News Affect Perceived Media Bias. *Journal of Computer-Mediated Communication* 18(1):32–45.

Lee, M. 2017. https://news.hankyung.com/article/2017041742897. *The Korea Economic Daily*.

Leibenstein, H. 1950. Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand. *The Quarterly Journal of Economics* 64(2):183–207.

Leight, E. 2019. Fake Streams Could Be Costing Artists $300 Million a Year. *Rolling Stone*.

Lerman, K., and Hogg, T. 2014. Leveraging Position Bias to Improve Peer Recommendation. *PLOS ONE* 9(6):e98914.

Li, L.; Zheng, H.; Chen, D.; and Zhu, B. 2019. Not Only Online Review but Also Its Helpfulness Is Manipulated: Evidence From Peer to Peer Lending Forum. In *Pacific Asia Conference on Information Systems*.

Mariconti, E.; Suarez-Tangil, G.; Blackburn, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Serrano, J. L.; and Stringhini, G. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW).

Marwick, A., and Lewis, R. 2017. *Media Manipulation and Disinformation Online*. Data & Society Research Institute.

Mihaylov, T.; Mihaylova, T.; Nakov, P.; Màrquez, L.; Georgiev, G. D.; and Koychev, I. K. 2018. The Dark Side of News Community Forums: Opinion Manipulation Trolls. *Internet Research* 28(5):1292–1312.

Momeni, E.; Cardie, C.; and Diakopoulos, N. 2016. A Survey on Assessment and Ranking Methodologies for User-Generated Content on the Web. *ACM Computing Surveys* 48(3):1–49.

Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social Influence Bias: A Randomized Experiment. *Science* 341(6146):647–651.

Naver News. 2013. https://news.naver.com/main/ombudsman/readView.nhn?notiId=283.

Naver News. 2015. https://news.naver.com/main/ombudsman/readView.nhn?notiId=333.

Naver News. 2017. https://news.naver.com/main/ombudsman/readView.nhn?notiId=393.

Naver News. 2018a. https://news.naver.com/main/ombudsman/readView.nhn?notiId=412.

Naver News. 2018b. https://news.naver.com/main/ombudsman/readView.nhn?notiId=415.

Naver News. 2018c. https://news.naver.com/main/ombudsman/readView.nhn?notiId=424.

Neubaum, G., and Krämer, N. C. 2016. Monitoring the Opinion of the Crowd: Psychological Mechanisms Underlying Public Opinion Perceptions on Social Media. *Media Psychology* 20(3):502–531.

Noelle-Neumann, E. 1974. The Spiral of Silence: A Theory of Public Opinion. *Journal of Communication* 24(2):43–51.

Park, S., and Lee, Y. 2018. https://news.joins.com/article/22303380. *Joongang Sunday* (567):1.

Springer, N.; Engelmann, I.; and Pfaffinger, C. 2015. User comments: motives and inhibitors to write and read. *Information, Communication & Society* 18(7):798–815.

Stroud, N. J.; Van Duyn, E.; Alizor, A.; Alibhai, A.; and Lang, C. 2017. Comment Section Survey Across 20 News Sites.

Stroud, N. J.; Van Duyn, E.; and Peacock, C. 2016. News Commenters and Comment Readers.

Tsikerdekis, M., and Zeadally, S. 2014. Online deception in social media. *Communications of the ACM* 57(9):72–80.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

Wang, G.; Wilson, C.; Zhao, X.; Zhu, Y.; Mohanlal, M.; Zheng, H.; and Zhao, B. Y. 2012. Serf and Turf: Crowdturfing for Fun and Profit. In *International World Wide Web Conference*, 679–688.

Wikipedia. 2019a. 2018 opinion rigging scandal in South Korea — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=2018_opinion_rigging_scandal_in_South_Korea&oldid=875161784. [Online; accessed 15-September-2019].

Wikipedia. 2019b. Vote brigading — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Vote_brigading&oldid=901507024. [Online; accessed 15-September-2019].

Woolley, S. C., and Howard, P. N., eds. 2018. *Computational propaganda*. Oxford studies in digital politics. Oxford University Press.

Zajonc, R. B. 1968. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology* 9(2, Part 2):1–27.

Zelenkauskaite, A., and Niezgoda, B. 2017. "Stop Kremlin trolls:" Ideological trolling as calling out, rebuttal, and reactions on online news portal commenting. *First Monday* 22(5).