MIS Quarterly

RESEARCH ARTICLE

DISINFORMATION SPILLOVER: UNCOVERING THE RIPPLE EFFECT OF BOT-ASSISTED FAKE SOCIAL ENGAGEMENT ON PUBLIC ATTENTION[1]

Sanghak Lee
W. P. Carey School of Business, Arizona State University
Tempe, AZ, U.S.A. {sanghak.lee@asu.edu}

Donghyuk Shin
College of Business, Korea Advanced Institute of Science and Technology (KAIST)
Seoul, REPUBLIC OF KOREA {dhs@kaist.ac.kr}

K. Hazel Kwon
Walter Cronkite School of Journalism and Mass Communication, Arizona State University
Phoenix, AZ, U.S.A. {khkwon@asu.edu}

Sang Pil Han
W. P. Carey School of Business, Arizona State University
Tempe, AZ, U.S.A. {shan73@asu.edu}

Seok Kee Lee
College of IT Engineering, Hansung University
Seoul, REPUBLIC OF KOREA {seelee@hansung.ac.kr}

Disinformation activities that aim to manipulate public opinion pose serious challenges to managing online platforms. One of the most widely used disinformation techniques is bot-assisted fake social engagement, which is used to falsely and quickly amplify the salience of information at scale. Based on agenda-setting theory, we hypothesize that bot-assisted fake social engagement boosts public attention in the manner intended by the manipulator. Leveraging a proven case of bot-assisted fake social engagement operation in a highly trafficked news portal, this study examines the impact of fake social engagement on the digital public's news consumption, search activities, and political sentiment. For that purpose, we used ground-truth labels of the manipulator's bot accounts, as well as real-time clickstream logs generated by ordinary public users. Results show that bot-assisted fake social engagement operations disproportionately increase the digital public's attention to not only the topical domain of the manipulator's interest (i.e., political news) but also to specific attributes of the topic (i.e., political keywords and sentiment) that align with the manipulator's intention. We discuss managerial and policy implications for increasingly cluttered online platforms.

Keywords: Opinion manipulation, disinformation, fake social engagement, political bots, online platforms, econometrics, machine learning

[1] Likoebe M. Maruping was the accepting senior editor for this paper.
Ofer Arazy served as the associate editor.

Introduction

Disinformation, also known as adversarial information operation (Weedon et al., 2017) or network or computational propaganda (Benkler et al., 2018; Bradshaw & Howard, 2018), refers to deceptive informational activities that aim to manipulate public opinion (Freelon & Wells, 2020). Recent cases of disinformation operations have shown that employing programmable bots to disrupt digital information commons has become common globally (Bradshaw et al., 2021).

The current study aims to explore whether *bot-assisted fake social engagement*, a widely used technique of disinformation operation, influences public attention to information beyond the manipulated context. Fake social engagement operations falsely amplify the salience of given information by manipulating the volume of user engagement with it (e.g., up/down voting, liking, sharing). Fake social engagement operations are integral to false amplification in today's digital environment because user engagement metrics serve as fundamental signals for many digital platforms' content curation algorithms to determine what to prioritize for display.[2] By deploying bots, manipulators can boost engagement metrics at scale with rapidity (Salge et al., 2022). Despite being a major prong of today's disinformation operation, how bot-assisted fake social engagement affects the networked public's information consumption patterns has not yet been empirically explored. This inquiry should be of particular interest to information systems (IS) scholars, as it suggests that bots can pollute information commons by abusing the platform's content curation policy (Mindel et al., 2018).

In addition to focusing on bot-assisted fake social engagement, this study broadens the understanding of disinformation influence by scoping the sphere of influence *beyond* the immediate context that disinformation actors disrupt. Existing studies have predominantly focused on users' direct engagement with manipulated content, for example, on how users accept (Effron & Raj, 2020) or share (or intend to share) fake information (e.g., Pennycook et al., 2021; Weidner et al., 2020). Few studies have evaluated the extent to which the manipulation effect continues even *after* users leave the context in which manipulation occurs. By exploring how falsely amplified messages insinuate themselves into organic user choices of nonmanipulated information, this study reflects the reality that disinformation operates not in isolation but in a broader information ecosystem.

To examine the spillover effect of bot-assisted fake social engagement on manufacturing users' attention to information, this study adopts a media theory, agenda-setting theory (McCombs & Valenzuela, 2020). Building on this theoretical framework, we maintain that bot-assisted fake social engagement amplifies the salience of not only the targeted topic (or issue) itself but also its specific traits. The amplified salience of the targeted topic and its traits, in turn, influence the distribution of public users' attention in a broader information consumption context. This theoretical framework is applied to a specific empirical case of South Korea's "Druking" scandal, one of the most infamous opinion rigging scandals in this country.[3] The Druking scandal is the epitome of a bot-assisted fake social engagement, as described more later. Bot-assisted fake social engagement operation is not particular to the Druking case but is widely observed across various global contexts, for example, politicians creating fake likes to inflate the popularity of their posts on Facebook (Wong & Ernest, 2021).

To broaden the scope of disinformation influence, we look at how bot-assisted fake social engagement operations change *general users'* informational behaviors, regardless of their political affinity. Previous studies have illustrated the ways in which disinformation actors (and their contents) mobilize a small, niche group of ideologically likeminded users (e.g., Bail et al., 2020; Bastos & Mercea, 2019; Freelon et al., 2022), or disloyal heavy internet users (Nelson & Taneja, 2018). However, no study, to our knowledge, has focused on general public users, except one survey study that found no effect of disinformation campaigns on them on Twitter—currently known as X (Bail et al., 2020). Based on large-scale clickstream data, the current study shows how bot-assisted fake social engagement influences the general public's informational behaviors in terms of what types of news they subsequently view, what keywords they use to search, and what they actually click on. Empirical research about the effects of disinformation on the general public's broader information consumption patterns is scant, due to the rarity of data. Such research requires a natural experiment setting that captures general users' real-time access to information while contrasting between users who are exposed to a disinformation operation and nonexposed users. Our data source meets both conditions, offering a unique opportunity to examine disinformation effects on general users at scale.

In the following sections, we first review the current bot-assisted disinformation research to point out two gaps in the existing literature. We then introduce agenda-setting theory as a theoretical framework to explain how disinformation operation, particularly fake social engagement, can influence public attention to information, followed by the presentation of our empirical study on the Druking scandal, an exemplary case of bot-assisted fake social engagement operation.

[2] For example, Facebook (https://blog.hootsuite.com/facebook-algorithm) and Twitter algorithms (https://blog.hootsuite.com/twitter-algorithm) are centered around engagement metrics.

[3] https://en.wikipedia.org/wiki/2018_opinion_rigging_scandal_in_South_Korea

Literature Review

Disinformation Research: Connecting Bot-Centered and User-Centered Approaches

Digital opinion manipulations, largely known as disinformation, have evolved into a serious problem of cybersecurity, presenting substantial challenges to public communication and information systems. Increased academic attention has focused on understanding bots and their roles in disinformation diffusion. For example, Stella et al. (2018) examined how social bots (e.g., software-controlled social media accounts) maneuvered political opinion dynamics on Twitter during the 2017 Catalan referendum. Also, Gorodnichenko et al. (2021) described the diffusion of information in social media and the role of bots in shaping public opinion based on their analysis of Twitter data on Brexit and the U.S. presidential election in 2016. Other studies have inferred political bot activities based on the bot-ness measure or ephemerality of accounts to understand the impact of bot-like accounts on amplifying political messages (Bastos & Mercea, 2019; Boichak et al., 2021).

Whereas most literature on disinformation bots is based on the network-structural view, Salge et al. (2022) importantly suggested taking "a processual view of diffusion" (p. 230) based on the concept of algorithmic conduit brokerage. Algorithmic conduit brokerage refers to the deliberate design and programming of bots as information brokers. Bots are programmed to play the role of information broker in multiple ways, ranging from social information alerts to rearranging shapes, forms, and structures of information (reconfiguration), to adding/inserting new information (embellishment) and transmitting information (Salge et al., 2022). Considering that Twitter has been a dominant platform for bot research, the algorithmic conduit brokerage perspective is perhaps best suited to Twitter-like platforms. For example, Salge et al. (2022) emphasized the role of bots in "actually transferring information between parties" (p. 230). While such actual transmission can be observed on Twitter in a relatively obvious form (i.e., retweets), it may not be apparent in other types of platforms—for example, online comment sections.

Nevertheless, the algorithmic conduit brokerage perspective offers overarching insights into the theorizing of disinformation bots. For example, Salge et al. (2022) suggest "algorithmic social alertness" as the first step for bot activity through which bots are programmed to search and discover already existing content and curate it in the programmer's (i.e., human manipulator behind the bot operation) favor. Algorithmic social alertness is performed on a variety of web platforms, not just on platforms with dynamic social feeds, such as Twitter but also in rather linearly designed platforms, such as discussion forums or comment sections. More importantly, the ability of algorithmic conduit brokerage for "rapid scaling" is integral to a wide spectrum of platforms wherein the main goal of bot activities is amplification. The "outcome [of bot actions] … is always *high volume* and not necessarily *high reach*, although both are certainly possible" (Salge et al., 2022, p. 247, italics original). In online comment sections, for example, bot-assisted fake social engagement may not necessarily increase audience reach but will generate high volumes of clicks or votes, which can help amplify the targeted information's visibility by rearranging the display of information. In this case, bot-assisted fake social engagement operations in online comment spaces exploit the bot's capacity for rapid scaling.

Whereas algorithmic conduit brokerage theory and related bot research have offered insights into the mechanism of bot-driven information diffusion, bot-centered disinformation research has been disconnected from another main branch of disinformation research that centers on the effects of disinformation on users' perceptions, attitudes, and behaviors. This "user-centered" research has revealed conditions in which users become vulnerable to falsehoods and evaluated how users interact with fake messages. For example, Pennycook et al.'s experimental study (2018) highlighted the "illusory truth effect" of fake news, one type of disinformation, showing that even a single prior exposure could enhance the (falsely) perceived accuracy of fake news. In the context of science communication, Scheufele and Krause (2019) examined the processes through which citizens become subject to scientific disinformation, concluding that vulnerability to scientific falsehoods should be determined not only by individual-level characteristics, such as the person's ability and motivation to detect falsehoods, but also by group-level and societal factors that facilitate access to correct(ive) information. Other studies (Carnahan & Garrett, 2020; Kahan et al., 2017) have shown that users not only accept deceptive information but also contribute to its propagation when the message affirms their cultural or political identity; conversely, users resist the correcting message if it is identity-threatening.

Overall, the user-centered disinformation research suggests that it is not the general population but specific audience groups that are prone to engaging with fake content. For example, Chen et al. (2021) showed that COVID-19 misinformation about the inefficacy of wearing masks and an election conspiracy theory of voter fraud was pushed by a "small but dense cluster of conservative users" on Twitter (p. 2). Nelson and Taneja (2018) analyzed audience visitation data of fake and real news sites during the 2016 U.S. presidential election campaigns, finding that heavy internet users who were not loyal to a mainstream news outlet were the main fake news consumers.

While existing user-centered studies have offered lessons on what makes users engage with or react to disinformation messages, these studies have predominantly examined message characteristics, providing little explanation about what happens to users when their attention is "hacked" by bots' false amplification (Marwick & Lewis, 2017, p. 19). To summarize, the two branches of disinformation research, bot-based information diffusion studies and user effect studies, have rarely been integrated, leaving the question of how bots' amplifying activities alter the digital public's information consumption patterns open to further exploration.

Missing Pieces: Influence Spillover and Fake Social Engagement

To fill in this gap, we expand on two aspects of a real-world disinformation operation that have not yet been thoroughly explored by empirical research. First, in reality, disinformation never occurs in an isolated dyad between the manipulator (or manipulated content) and a user. On the contrary, the immediate context in which users are exposed to a perpetrator's action is a subset of a larger information ecosystem. Therefore, the effect of a successful disinformation campaign is likely to extend beyond the direct interaction between the manipulator (or manipulated content) and the user and spill into other settings of information consumption. Several qualitative case studies have alluded to this point by describing how disinformation perpetrators work not in isolation but exploit existing media networks. For example, successful disinformation content created in an online troll community does not stay within the community but is picked up by mainstream media attention, reaching broad audiences (Phillips, 2015; Marwick & Lewis, 2017).

That said, systematic empirical analyses of disinformation effects have mostly focused only on direct interactions between manipulative content and users. This is understandable because it is rare to obtain data that represent the spillover of disinformation influence. Nevertheless, previous findings that disinformation is engaged only by niche audience groups are based on such limited measures, resulting in an incomplete representation of disinformation's sphere of influence. For example, Nelson and Taneja (2018) measured fake news consumption by using site-visitation data, one of the most proactive measures of audience engagement. Based on this, they argued that broad users were seldom vulnerable to fake news. Similarly, Bail et al.'s Twitter study (2020) found that Russia's disinformation accounts were engaged mostly by highly partisan users with high-frequency usage of Twitter, concluding that "Russian trolls might have failed to sow discord because they mostly interacted with those who were already highly polarized" (p. 243). However, the Bail et al. study was based on non-representative survey data matched with the metrics of direct engagement with the troll accounts or their messages. The findings of these studies are a partial snapshot of disinformation reality because disinformation influence can also be *indirect*: Users may be surreptitiously exposed to the manipulator's intention, and even the simplest exposure could result in a ripple effect on the consumption of other informational sources without further direct interactions with the manipulators or their content.

Second, disinformation operations entail not only the creation of fake content but also the creation of fake engagement with existing content in the manipulator's favor. Thus far, a considerable body of literature has focused on the effects of the former—for example, by examining what makes fake content persuasive, how it is propagated, and how it is detected (e.g., Vosoughi et al., 2018; Cresci, 2020), with little attention paid to the disruption of the digital information commons caused by fake social engagement. Thus far, disinformation studies that have examined social engagement have mainly focused on organic social engagement with fake content. For example, Edelson et al. (2021) found that about 70% of all user engagements across far-right news pages on Facebook were made with misinformation content. In another study, Freelon et al. (2022) showed that user engagement with disinformation tweets became disproportionately large when the tweets originated from fake accounts pretending to be Black activists. A handful of studies have paid attention to fake (mostly bot-assisted) social engagement activities (e.g., Boichak et al., 2021); however, to our knowledge, no study has taken a user-centered approach to examine how bot-assisted fake social engagement affects individual users' informational behaviors.

The reasons for the dearth of user-centered studies in the bot literature are twofold. First, it is difficult to detect bot activities unless ground-truth labels are available. As a result, developing detection techniques is a complex scientific problem that demands considerable effort (e.g., Varol et al., 2017; Cresci, 2020). Second, the primary bot activities have thus far functioned as information brokers (i.e., algorithmic conduit brokerage) rather than original content creators. Since the consequence of conduit brokerage is more nuanced than content creation, it is difficult to empirically differentiate between users who are exposed to bot activities and those who are not.

Despite the challenges, understanding the effects of bot-assisted fake social engagement on individual users is imperative because of the phenomenon's prevalence and significance. Bot-assisted fake social engagement is prevalent due to its cost-effectiveness (e.g., Jeong et al., 2020; Carman et al., 2018; Schäfer et al., 2017; Rossi et al., 2020). Also, it is a powerful tactic because social engagement metrics are pivotal indicators of content popularity fed into a platform's content curation algorithms. Accordingly, we first ask the following question:

RQ: *Does bot-assisted fake social engagement have spillover effects on public attention to information beyond the manipulated context?*

Bot-Assisted Fake Social Engagement and Public Attention: An Agenda-Setting Theoretical Framework

Bot-assisted fake social engagement operations center on the interplay among human perpetrators, automation (bots), and platform algorithms to "manufacture consensus or to otherwise give the illusion of general support for a (perhaps controversial) political idea or policy, with the goal of creating *a bandwagon effect*" (Woolley & Howard, 2016, p.4, emphasis added). In information consumption contexts, the bandwagon effect is manifest in the shift of public attention to certain types of information.

Agenda-setting theory (McCombs & Valenzuela, 2020) is a useful theoretical framework for explaining how fake social engagement influences public attention. It suggests that the media has the ability to influence audiences in terms of which issue to pay attention to as an important public agenda and which attributes of the issue to pay attention to in order to make sense of the issue (McCombs & Valenzuela, 2020). The agenda-setting effect of news media on the public's mind has been well documented in the media and journalism literature since the seminal evidence of news media's agenda-setting function a half-century ago. McCombs and Shaw (1972) found a significant association between the amount of news coverage of political agendas during an election campaign and the public ranking of the importance of agendas for the election. Numerous studies have since then confirmed that the public's understanding of political reality is influenced by the salience of issues emphasized in news coverage.

Provided that the media's agenda-setting effect occurs by increasing the salience of information, disinformation actors may also play the role of agenda setters by amplifying the salience of the information that conveys their preferences. A few studies have alluded to this point. For example, Guo and Vargo (2020) showed that fake news stories exaggerated politician attributes, such as moral quality, leadership quality, and intellectual ability, to affect public attitudes toward political candidates. Rojecki and Meraz (2016) examined conspiratorial information transmissions during the 2004 U.S. presidential campaigns. They found that while the visibility of conspiracies on the Google search results was not directly associated with users' overall search trend—an indicator of the naturally occurring volume of online public attention—the visibility of conspiratorial information on the search results influenced traditional media's coverage of it, which in turn was associated with users' overall search trend. Vargo et al. (2018) analyzed big data from news archives, demonstrating that fake news sites had a stronger "intermedia" agenda-setting effect (p. 2030) on legitimate news coverage than fact-checking sites, particularly by transferring their agendas to partisan news outlets (e.g., Fox News).

While extant studies have alluded to the agenda-setting potential of disinformation, they have focused only on fake news sites and the transporting of their narratives to other mainstream media outlets. To our knowledge, no study has examined disinformation's agenda-setting effect on general public users, particularly in terms of bot-assisted fake social engagement operations.

Bot-assisted fake social engagement operations facilitate a bandwagon of public attention through the mechanism of rapid scaling (Salge et al., 2022). The rapidly inflated engagement volume makes it look like a large number of "real" users are interested in the (falsely) amplified topic, which can, in turn, increase organic public attention to the topic. Technically speaking, social engagements can be manipulated solely by human workers. However, fake engagement operations would have little impact on rearranging the salience of information unless the metric is rapidly fabricated at scale. In other words, bots' "rapid scaling" (Salge et al., 2022) of social engagement and the subsequent bandwagon effect resonates with the tenet of agenda-setting theory.

Agenda-setting theory includes two levels of media effects on shaping public attention to news agendas (McCombs & Valenzuela, 2020). The first-level agenda-setting effect, also known as "issue agenda setting" (Kim et al., 2002), refers to the media's ability to determine the hierarchy of public agendas by informing the audience *what* topic (issue or object) it should pay more attention to. The frequency of topics in news articles influence how the audience prioritizes the importance of these topics (McCombs & Valenzuela, 2020). For example, if the media covers news about Samsung Galaxy smartphones more frequently than Apple iPhones, the audience will pay more attention to Samsung's smartphones than Apple's. The first-level agenda-setting effect can occur on an even more abstract topic domain. For example, if the media reports on foreign affairs more frequently than on the domestic economy, the audience will be likely to consider international politics to be a more important current issue than domestic economic conditions.

In other words, first-level agenda setting is about the media's influence on public attention to a topic. In this study's empirical context, where disinformation was related to a political issue, we posit a hypothesis that suggests the first-level agenda-setting effect of bot-assisted fake social engagement on public attention to a political topic. That is, when a bot-assisted fake social engagement operation targets political content, the intensity of exposure to the operation predicts a relative increase in public attention to political news compared to non-political news.

H1: *The salience of bot-assisted fake social engagement predicts an increase in public attention to political news compared to non-political news.*

Meanwhile, second-level agenda setting, also known as "attribute agenda setting" (Kim et al., 2002), focuses on the presentation of attributes, qualities, or characteristics of a certain topic and its effect on *how* the audience will subsequently perceive or feel about that topic (Kiousis, 2005; McCombs & Valenzuela, 2020). For example, if the media frequently focuses on Samsung Galaxy's foldable design when reporting on smartphone features, the audience will begin to prioritize foldable design as the smartphone's most important attribute and pay more attention to information related to this feature when they think about Samsung Galaxy smartphones. On the contrary, if the media frequently reports on Samsung Galaxy's alleged benchmark manipulation with regard to the speed, battery life, and overall performance,[4] the audience will prioritize benchmark manipulation as the smartphone's most important attribute and will pay attention to information related to this feature when they think about Samsung Galaxy smartphones. In other words, the second-level agenda-setting effect is about the media's ability to prime the audience's attitudinal or emotional reaction to a topic, because the selective presentation of attributes transmits sentiment, whether intended or not, and subsequently influences the audience's attitude toward the topic (Coleman & Wu, 2010; Kim et al., 2002).

Likewise, bot-assisted fake social engagement operations may engender the second-level agenda-setting effect by inflating the salience of certain attributes of the targeted topic, which may, in turn, increase public attention to these attributes. Recent advances in agenda-setting theory suggest that the agenda-setting effect occurs not only in the context of single attributes but impacts bundles of mental associations in a so-called network agenda-setting effect (e.g., Guo & Vargo, 2015; Vu et al., 2014). Drawing upon the recent theoretical development of the network agenda-setting model, we posit a second-level agenda-setting hypothesis based on associative keywords and texts resonating with the manipulator's intention. Given this study's empirical context, featuring a disinformation operation directed toward a political issue, we posit a second-level agenda-setting hypothesis that centers on political attributes:

H2: *The salience of bot-assisted fake social engagement predicts an increase in public attention to manipulator-promoted political attributes compared to manipulator-demoted political attributes (e.g., political keywords and sentiment).*

Figure 1 conceptually illustrates the fake social engagement-driven agenda-setting effect thesis that we propose. The lower part of Figure 1 illustrates that bot-assisted fake social engagement distorts the salience of topics and their attributes. In this example, Topic 1 is the target of manipulation in which Attribute A3 is promoted and Attribute A2 is demoted through bot-assisted fake social engagement. This manipulated salience then influences the distribution of public attention, as illustrated by the upper-right part of Figure 1, where Topic 1 becomes the most dominant topic and Attribute A3 emerges as the most salient attribute while A2 becomes the least salient attribute of Topic 1.

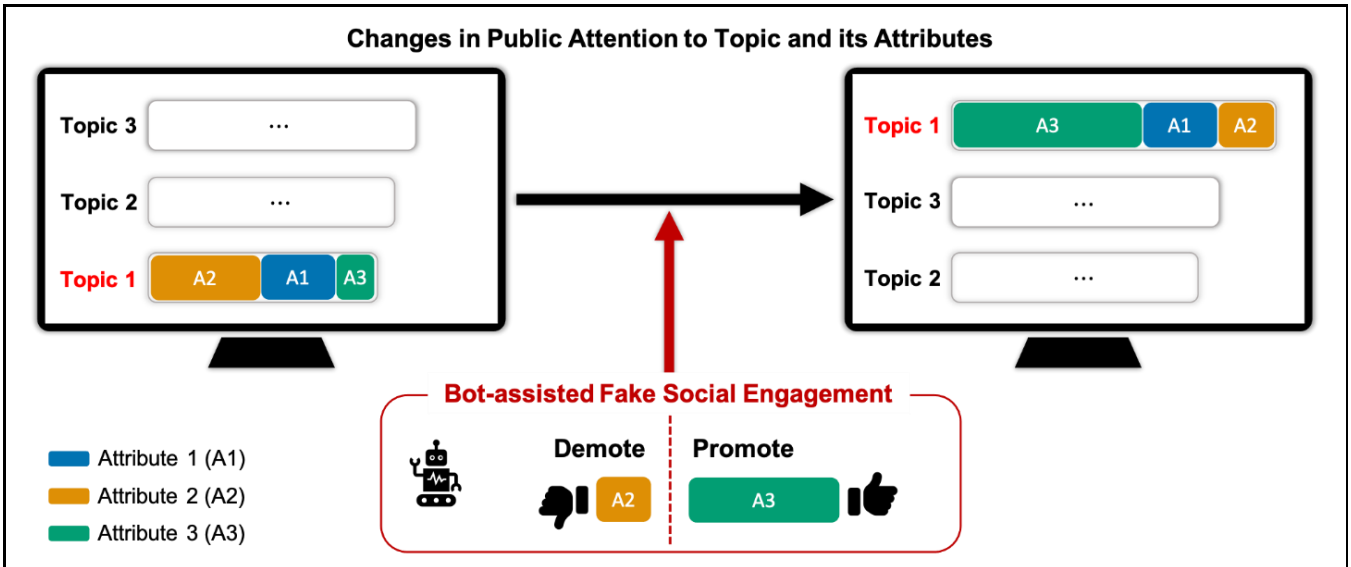


Figure 1. Agenda-Setting Effect in Bot-Assisted Fake Social Engagement Context

[4] https://www.classaction.org/blog/samsung-phone-lagging-class-action-alleges-the-company-misled-consumers-on-speed-and-performance

Research Context

Digital Opinion Manipulation: The 2018 Druking Scandal

We focused on an online opinion-rigging scandal that occurred in South Korea. In 2018, South Korea experienced a major disinformation activity, widely referred to as the Druking scandal (Choe, 2018). "Druking" was the screen name for the disinformation operation team's leader, who had been a popular blogger while secretly founding a shadow company that ran illegitimate internet political campaigns utilizing political trolls. The company operated during the 2017 South Korean presidential election campaign to influence public opinion. While its initial political position was aligned with the Democratic Party (the then ruling party), in 2018 it assumed an anti-government stance. In 2018, the Druking team was indicted for rigging online comments. The main locations where the Druking team operated were spaces for news comments on major Korean portal sites. Given that South Korea has a 96% internet penetration rate, with the vast majority of online news consumption occurring via portal sites and an active presence of news comment culture, such digital opinion manipulations can have substantial ramifications.[5,6]

One of the primary activities of the Druking team was to manipulate the ranking of comments on a news site. To this end, they used a programmable code called "KingCrab," a macro-based bot that cast a massive number of up/down votes for certain targeted comments. The ranking of comments was important because the top-ranked comments achieved a higher degree of visibility than the rest of the comments. The Druking operation team's key action involved the selection of target comments and the manipulation of their rankings by pushing their favored (disapproved) comments to the top (bottom) of the list.

As with Reddit and other online news aggregators, the focal platform determined the ranking of comments on a particular news page based on their net vote count (i.e., total upvotes minus total downvotes per comment). The platform used phone verification to authenticate users' identities during the account registration process and permitted only one upvote or downvote per comment. Nonetheless, the Druking team managed to circumvent this by obtaining and leveraging thousands of legitimately created user accounts to create a large number of upvotes and downvotes in an attempt to manipulate news comment spaces in its favor.

Ground Truth of Opinion Manipulation

The Druking accounts were established based on the verified documents issued by the law enforcement department and the details pertinent to general users are fully de-identified. The subject of the focal article gained traction: 39,827 comments were posted within 24 hours after the story was first published on January 17, 2018, at 9:35 a.m. Korea Standard Time (KST). It is worth noting that none of the comments came from the Druking accounts. That is, they were all posted by authentic users. As a result, the manipulator's primary goal was to affect the popularity of comments created by others rather than to create its own comments. Druking's operation was clearly directed at altering the upvote/downvote counts of existing comments. Over the 24-hour time span, some 2,300 Druking accounts were used approximately 1.2 million times to cast upvotes or downvotes in order to alter the ranking of the current comments targeted by the manipulator.

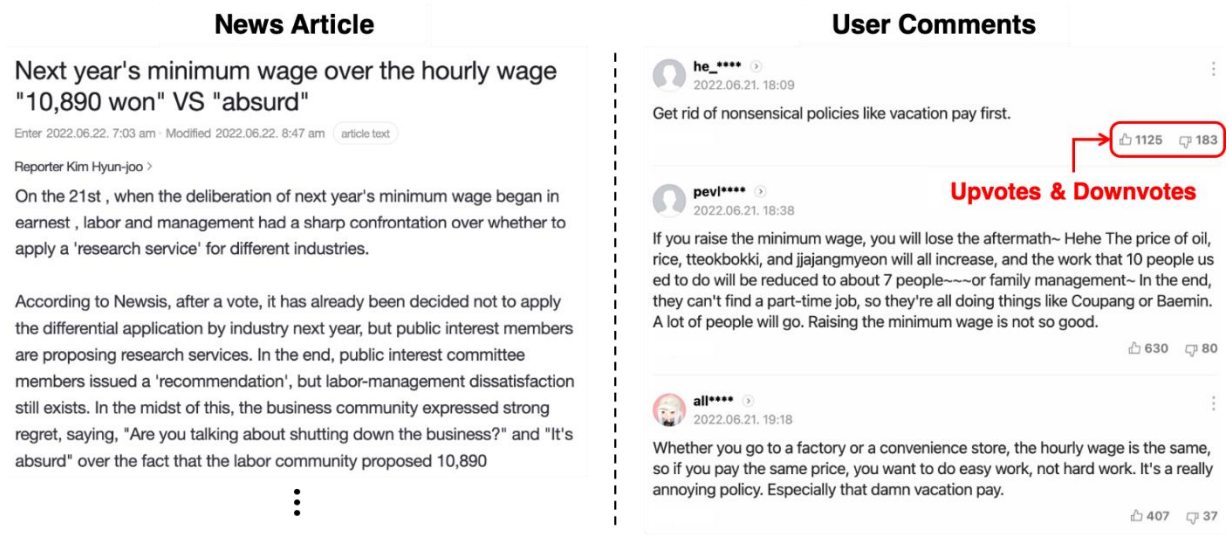
Data

We examined one of the leading online news platforms in South Korea. On this platform, each news article's page was composed of the main article and the user comment space dedicated to the main article. The ranking of comments was determined by their popularity, measured as upvotes (i.e., the number of thumbs-ups it received) compared to downvotes. Therefore, manipulators could escalate (decrease) the ranking of comments they wanted to promote (demote) by generating a large number of upvotes (downvotes) on them using a programmable bot. An example screenshot of a news article and its user comments from the focal platform is shown in Figure 2.

In partnership with the platform's company, we obtained access to its proprietary data on user behaviors and clickstream information, amounting to more than 108 million raw user log entries. The granularity of the data enabled us to observe how the user-generated comment section embedded in a news article's page was shaped over time and how users' news searching and viewing activities across the platform changed after the consumption of a news article's page. In the following sections, we first describe how a bot-assisted fake social engagement operation influenced the real-time process through which the user-generated comment section of the focal article was created. Then we shift our focus to the behaviors of organic users and provide the descriptive statistics that illustrate their activities on the platform.

[5] https://www.digitalnewsreport.org/survey/2020/south-korea-2020

[6] Korea Press Foundation (2018, Media Issue 5): User survey about portal news and comments. https://bit.ly/3KkdnVK



Note: Contents of the article and user comments were translated from Korean to English using Google Translate.

Figure 2. Example Screenshot of the Focal Platform

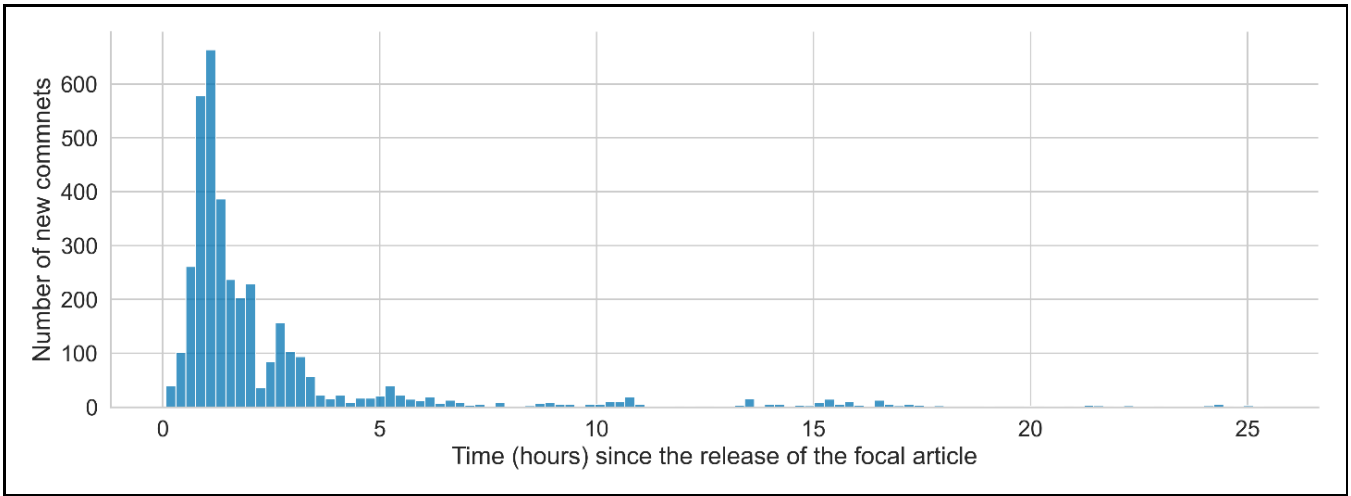


Figure 3. Dynamics of Comment Postings

Formation of a User-Generated Comment Section

When the focal article was published at 9:35 a.m. on January 17, 2018 (KST), the audience began to immediately utilize the comment section. Some commented directly on the article, others indicated their agreement with an existing comment with an upvote or their disagreement with a downvote, while still others viewed the comments passively without responding. The first part of our data showed how the top 1,000 comments on the focal article changed over time, recording the cumulative counts of upvotes and downvotes on each comment. Given the deterministic ranking mechanism of comments, this dataset enabled us to recover the entire process of the formation of the comment section.

A total of 3,775 comments were ranked in the top 1,000 at least once during our sample span. The first comment appeared four minutes after the news article was published, and the last comment was posted 25 hours later. Despite the fact that the news was published on a weekday morning, approximately 80% of the comments were generated within the first three hours, showing that users reacted to the news article quickly (see Figure 3).

The platform ranked user-generated comments in the order of popularity, as determined by the number of upvotes received minus the number of downvotes. As a result, either upvoting or downvoting on a particular comment would influence the comment's salience by changing its relative position. Knowing this, the manipulator used programmable bots to increase the number of upvotes on comments he endorsed while producing downvotes on comments he wanted to suppress. Importantly, however, the manipulator did not have full control because a large portion of votes were generated by organic users with diverse viewpoints.

Figure 4 depicts the number of upvotes and downvotes created by bots, as well as organic users, over time. There was a total of 953,578 votes cast on 3,775 comments, with 719,609 upvotes (75.46%) and 233,969 downvotes (24.54%). The manipulator was responsible for 31.77% of the total upvotes and 20.92% of the total downvotes, and its voting activities increased two hours after the focal article was published. Organic users' votes, on the other hand, appeared more quickly and had a longer tail than the manipulator's votes. This suggests that the manipulator took some time to identify his target news page and the comments on it and prepare for the attack. Then, he ceased the operation when the effect of the votes became muted due to the large volume of accumulated votes.

Comment rankings fluctuated over time and eventually converged to a final rank, as illustrated in Figure 5, which presents 10 different comment convergence trends. The rankings fluctuated significantly within the first three hours and then steadily converged to their final positions at around five hours, which was to be anticipated given that the rankings were decided by the cumulative number of upvotes and downvotes. Although some comments shifted upward or held their status over time, others moved down due to downvotes, a surge of other comments, or the introduction of new comments.

More importantly, the manipulator's vote distribution was clearly distinct from that of organic votes. Figure 6 shows the source of votes for the top 10 comments as of the last time point in our data, ordered by their total net upvotes, including votes from both manipulator and organic users. The manipulator created upvotes to promote six comments (C1, C3, C4, C6, C7, C8, C9) and suppress four (C2, C4, C5, C10), demonstrating his goal-directed behavior. That said, the manipulator did not have complete control of the opinion landscape: While the operation succeeded in positioning six of his preferred comments among the top 10, he was unable to overcome the organic popularity of the other four comments that he disapproved of. Nonetheless, manipulative votes appeared to have a significant impact on the final ranking of the comments, even when organic votes outnumbered them. For example, if the manipulator had voted against comment C1 while supporting comment C2, the relative positions of the two comments would have been reversed. In total, the manipulator promoted 998 comments and suppressed 247 comments.

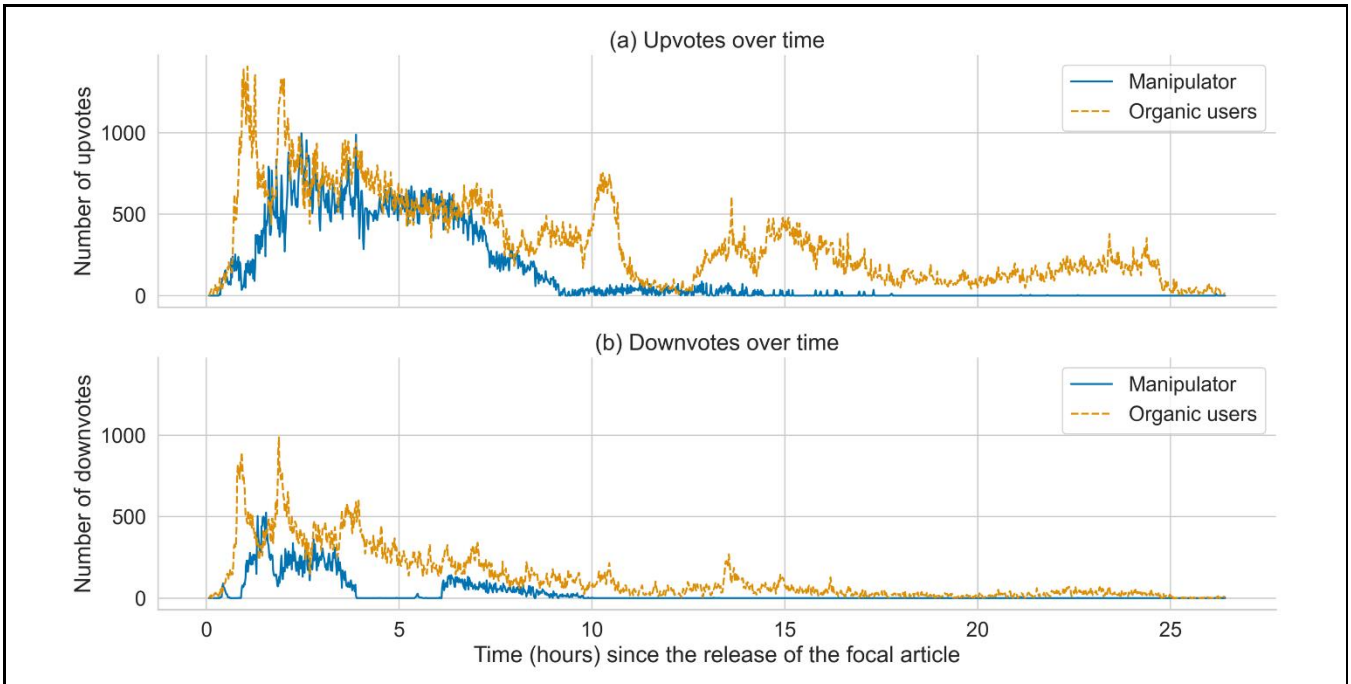


Figure 4. Trends of Upvotes and Downvotes by Manipulators and Organic Users

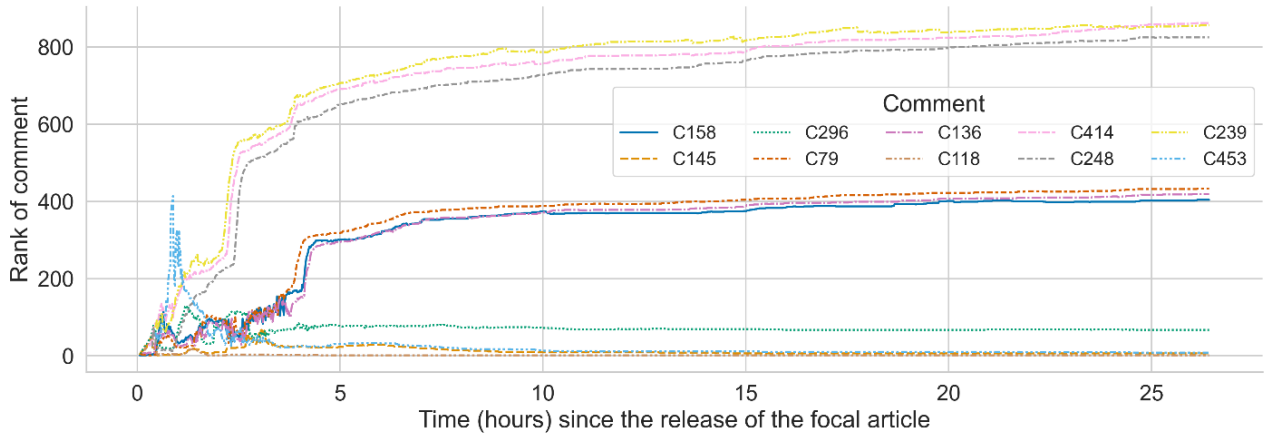
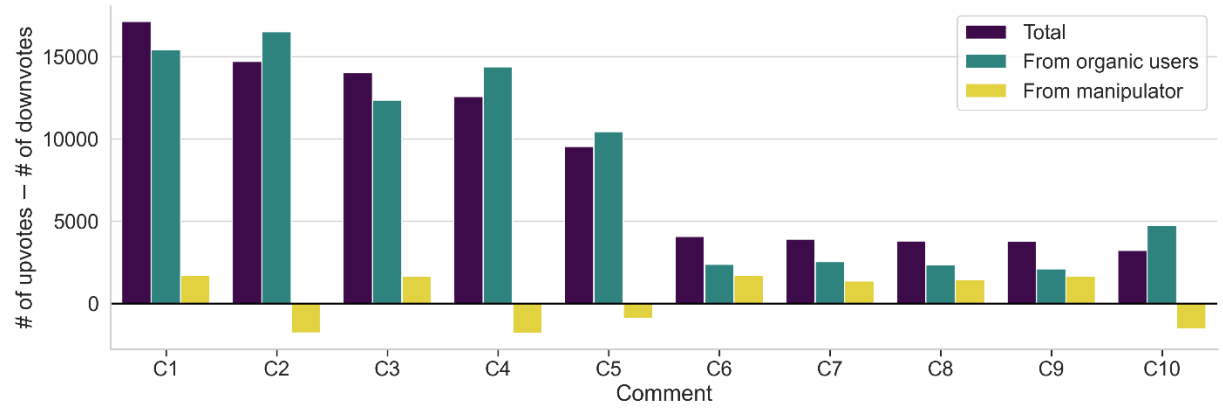


Figure 5. Dynamics of Rankings for Ten Selected Comments



Note: Comments are ordered by their total net upvotes (upvotes minus downvotes), including those from both manipulator and organic users, as of the last time point in our data.

Figure 6. Sources of Votes for the Top 10 Comments

Organic User Activities

The organic user activity dataset was made up of the full server log details of 23,735 general users from January 15, 2018, to January 18, 2018. We removed 384 users who had no activity during the period, were younger than eight years of age, or had missing demographic data. The samples were then divided into one of two categories. The first group, the treatment group, consisted of 17,335 users who visited the focal news article with a manipulator-targeted comment section. The second group, which we called the control group, consisted of 3,868 users who visited one of 34 articles that contained highly similar content to the focal article but had not been attacked by the manipulator.

To identify articles that were extremely similar to the focal article, we used the doc2vec model (Le & Mikolov, 2014), which has rapidly gained popularity in the IS literature (e.g., Qiao et al., 2020; Shin et al., 2020) due to its state-of-the-art performance in various natural language processing tasks. We fine-tuned a large-scale doc2vec model pre-trained on more than 6.3 GB of text data[7] using a total of 342,567 articles that users visited during the sample period. Based on our doc2vec model, we first represented each article by its embedding vector in a latent feature space.[8] Then, using the cosine similarity between their embedding vectors, we computed the similarity between the focal article and all other articles. Along with manual verification, we selected a total of 34 articles that were

[7] https://ko-nlp.github.io/Korpora

[8] We use $d = 300$ for the dimension of the embedding vectors. Other reasonable values of the dimension ($100 \leq d \leq 500$) yield similar empirical results.

most similar to the focal article as control group articles.[9] To avoid cross-contamination, we eliminated 2,148 users who visited both focal and control articles from our sample. As a result, the valid sample included 21,203 organic users.

Our data enabled us to examine both pre- and post-visit log files for each user account since the focal news article was published at 9:35 a.m. on January 17, 2018 (KST). An average user in both the treatment and control groups visited 153.2 pages and spent 3.1 hours per day on the platform over the course of four days (see Table 1). The two groups were comparable in terms of platform engagement, with no statistically significant differences in the number of logs, page views, votes, or amount of time spent during the sample period. However, there was a high degree of heterogeneity among users, which is indicated by large sample standard deviations. In addition, the user activities exhibited a clear pattern of temporal variation. Figure 7 depicts how an average user's page views per hour changed over time. Users were more active in the afternoon and evening than late at night and early in the morning. The first three days contained a higher number of user page views than the remaining days.

Variables

Salience of Bot-Assisted Fake Social Engagement

To falsely amplify the visibility of preferred comments, the manipulator promoted some comments by upvoting them and demoted others by downvoting them. Accordingly, we measured the salience of bot-assisted fake social engagement (FSE henceforth) using the visibility difference between manipulator-promoted and demoted comments at time t:

$$FSE_t = \sum_{\forall k \in K_{promote}} \frac{1}{ranking_{kt}} - \sum_{\forall k \in K_{demote}} \frac{1}{ranking_{kt}}, \quad (1)$$

where $ranking_{kt}$ denotes the ranking of comment k at time t, and $K_{promote}$ and K_{demote} represent the set of manipulator-promoted comments and the set of manipulator-demoted comments, respectively. That is, we subtracted the sum of the inverse rankings of manipulator-demoted comments from the sum of the inverse rankings of manipulator-promoted comments. By using this metric, we not only weighed higher-ranking comments (i.e., comments with a higher rank are more likely to be viewed and hence more salient than comments with a lower rank), but we also accounted for the volume of comments at t.

Figure 8(a) shows the temporal variation of the salience of FSE. It fluctuated between -4 and 4 for the first five hours following the publication of the focal article and then steadied at a positive value of around 3.5 as the comment ranks stabilized. Figure 8(b) depicts the arrival time of organic users at the focal article and its comment section. Visitors to the focal article within the first five hours accounted for 25% of overall viewers of the focal article. If we extended the time window to the first ten hours, the percentage rose to 62%. The variation in users' arrival times at the focal article resulted in a variation in the composition of the comment section to which each user was exposed.

Public Attention

We operationalized public attention by using page views. That is, we measured the total amount of user i's attention to news in time t by the number of news pages that user i viewed during the corresponding window of one hour, which was denoted by PV_{it}. Further, we decomposed public attention to news by topical categories to investigate the shift in public attention caused by FSE (i.e., the first-level agenda-setting effect). In order to test the first-level agenda-setting effect (H1) given the political nature of the focal news article in our empirical context, we compared user attention to political and non-political news articles using the platform's preset news categories. The non-political news sections included sports (PV_{it}^{sports}), entertainment (PV_{it}^{enter}), and other news[10] (PV_{it}^{other}).

Political Attributes

To test the second-level agenda-setting effect (H2), we examined changes in page views within the political news section. Given that Druking promoted and demoted certain political messages, we operationalized public attention to manipulator-intended political attributes by calculating the net difference in page views between the articles whose content matched manipulator-promoted attributes and articles whose content matched manipulator-demoted attributes. Specifically, we constructed two variables: (1) proactive public attention and (2) passive public attention.

Proactive public attention (keyword search and search-induced page views): One way for a user to proactively find news articles is through the use of search keywords. Some search keywords may resonate with the manipulator's intention. According to court proceedings, Druking automated FSE operations by selecting target comments using a list of keywords that aligned with his goal and then setting the desired number of upvotes and downvotes for the selected comments. Thus, the keywords used by Druking should be indicative of the political attributes that he either promoted or demoted.

[9] All control articles have a cosine similarity larger than 0.85 with the focal article, which is at the top 0.0001% of all similarity scores.

[10] Note that the portal site bundles the rest of other topics into a single category called "(other) news."

Table 1. Platform Activity Statistics of Treatment and Control Groups

	Treatment group (users who visited the focal article)	Control group (users who visited the control articles)	Total
Number of users	17335	3868	21203
Number of logs per day	332.0 (332.3)	341.2 (324.7)	333.7 (330.9)
Hours spent on platform per day	3.1 (2.2)	3.3 (2.2)	3.1 (2.2)
Number of page views per day	151.0 (121.2)	163.2 (122.9)	153.2 (121.6)
Number of news page views per day	18.4 (19.6)	25.8 (28.5)	19.7 (21.5)
Number of upvotes per day	14.8 (71.4)	14.0 (62.3)	14.7 (69.8)
Number of downvotes per day	4.9 (36.1)	5.2 (34.7)	4.9 (35.9)

Note: Sample standard deviations are in parentheses.

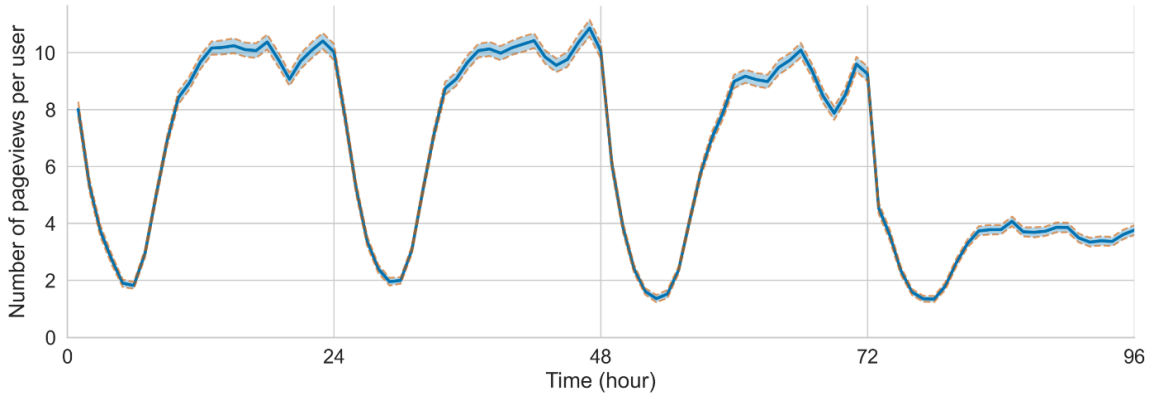


Figure 7. Page Views over Time

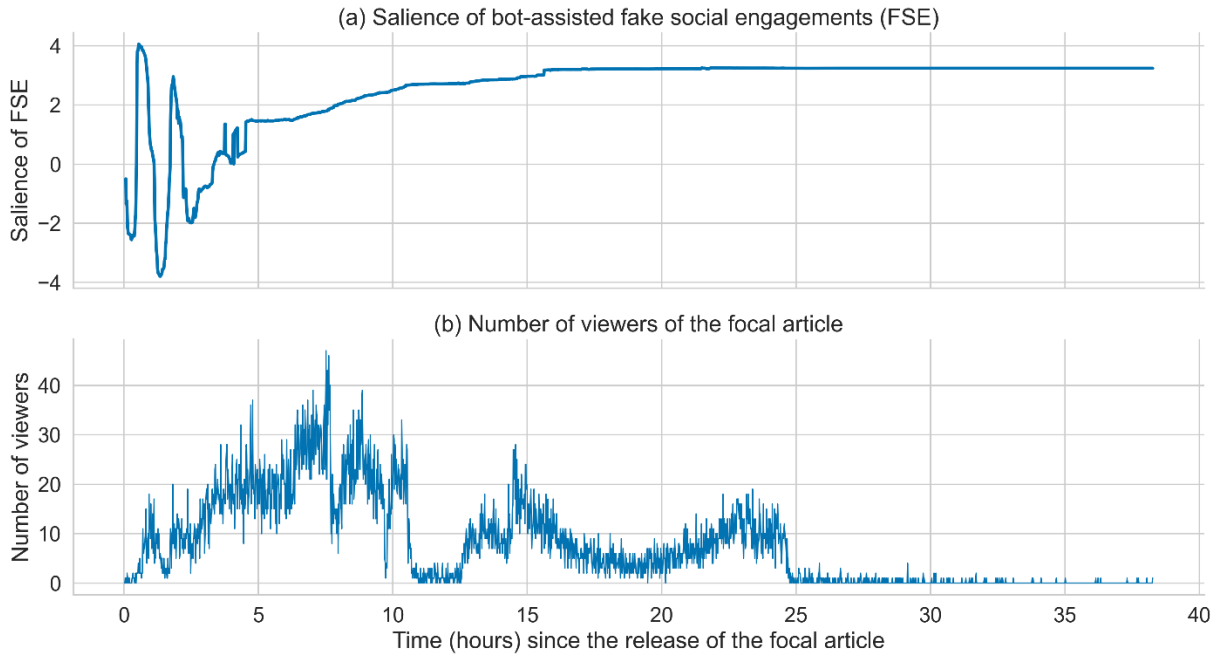


Figure 8. Salience of Bot-Assisted Fake Social Engagement and the Number of Viewers of the Focal Article

To infer Druking's keywords, we first computed the term frequency-inverse document frequency (TF-IDF) scores across all comments. Then, using the ground-truth labels for Druking's bot accounts, we found 43 promoted keywords and 21 demoted keywords with the largest TF-IDF score differences across comments with and without fake social engagement from Druking accounts. The majority of upvoted (or promoted) keywords contained anti-government sentiments, such as innuendo, mockery, or insinuation about the government and president at the time, whereas the majority of downvoted (or demoted) keywords contained terms referring to opinion manipulation operations or investigations into harmful/fake comments. The identified upvoted (downvoted) keywords appeared in 90.5% (96.7%) of promoted (demoted) comments but not in any unmanipulated comments.

Given that general users may not always use search keywords that are identical to manipulator-promoted/-demoted keywords, we included other keywords that were deemed highly associative with the manipulator's keyword list. To do this, we represented each keyword by its word embedding vector using the doc2vec model (explained in the Organic User Activities section). Then, we identified the top-100 most associative and semantically similar search keywords from over seven million distinct search queries by calculating their cosine similarity to the manipulator's keyword list. The final keyword list had 143 keywords associated with manipulator-promoted attributes and 121 keywords associated with manipulator-demoted attributes. Using the list, for each user and time period, we counted the number of search activities containing manipulator-promoted keywords (KS_{it}^{promo}) and the number of searches containing manipulator-demoted keywords (KS_{it}^{demo}). We next operationalized the manipulator-intended political attribute by calculating the net count difference between the aforementioned two types of search activities.

Further, we examined a user's subsequent viewing of news articles following the keyword search. Because a keyword search returns a list of relevant news articles, a user may select one or more of those articles from the list to get more insights, which we refer to as search-induced page views. To measure search-induced page views, we first identified news articles with a headline that contained the search keywords of interest. Then, we counted the number of the identified news articles viewed, contingent on the user viewing them within an hour of the keyword search. Specifically, we used the embedding vectors obtained from our doc2vec model to compute the cosine similarity between search keywords and articles and counted the number of views of the top-n% most similar articles. Based on the similarity measures, we counted news page views driven by the manipulator-promoted search keywords ($PV_{it}^{promo-search}$) and by the manipulator-demoted search keywords ($PV_{it}^{demo-search}$). The difference between these two page views offers additional operationalization of the manipulator-intended political attribute based on search-induced page views.

Passive public attention (unsearched page views of articles with similar headlines): Users do not always navigate news articles by conducting proactive keyword searches; rather, they frequently choose what to read due to incidental exposure to the headline of an article. Indeed, 54% of our sample did not perform any keyword searches during the sample period. To ascertain the effect of FSE on those who "passively" consumed news articles, we used our doc2vec model to examine the unsearched page views of all articles with headlines that were semantically similar to the aforementioned Druking's keywords. Following that, page views were calculated based on the top-10% most similar articles. Finally, we operationalized passive attention to manipulator-intended political attributes by computing the difference in page views between the news articles with headlines that resonated with the manipulator's promoting keywords (PV_{it}^{promo}) and news articles with headlines that were consonant with the manipulator's demoting keywords (PV_{it}^{demo}).

Political sentiment (pro-government vs. anti-government): Additionally, we examined the second-level agenda-setting effect from the perspective of political position. Recall that Druking's goal was to undermine the then-ruling party by promoting anti-government comments while limiting pro-government comments. Therefore, we hypothesize that the salience of FSE would predict a relative increase in public attention to news with anti-government sentiment compared to news with pro-government sentiment.

A crucial step for the analysis of political sentiment was determining the political leanings of news articles that users viewed following their exposure to the focal news page. To identify the political orientation, we adopted an advanced semi-supervised machine learning (ML) approach called label propagation (LP) (see Appendix A for details of our LP model). Semi-supervised learning is best suited for scenarios in which only a small number of labeled samples are available, whereas most of the data are unlabeled (Zhou et al., 2003; Fujiwara & Irie, 2014). The LP model has been shown to achieve considerably more accurate performances for various applications by combining both labeled and unlabeled samples together during training compared to supervised ML models that utilize only labeled samples (e.g., Tarvainen & Valpola, 2017; Iscen et al., 2019). Notably, semi-supervised learning is becoming increasingly popular due to the high cost of expert data labeling along with the increasing need for large-scale training data. In the IS literature, while both supervised and unsupervised ML models have been extensively studied and employed, the investigation of semi-supervised learning has been extremely limited, with the exception of the work by Abbasi et al. (2012) on financial fraud detection.

In our setting, we used the well-known political bias of partisan Korean news media (Lim et al., 2019) as the initial labels of articles, resulting in less than 18% of articles being labeled as either pro- or anti-government. A similar approach was used in David et al. (2016) to predict the political orientation of Facebook users based on posts from the pages of political parties. We note that our LP model achieved the best accuracy (F1-score of 0.913), compared to other representative ML models (see Appendix A). Finally, using the political orientations of articles identified by our LP model, we operationalized public attention to manipulator-intended political sentiment by the difference in the page views between the news articles with pro-government sentiment (PV_{it}^{progov}) and the news articles with anti-government sentiment ($PV_{it}^{antigov}$).

Spillover Effects of Bot-Assisted Fake Social Engagement on Public Attention

RQ: A Spillover Effect of Bot-Assisted FSE on Public Attention to News

Bot-assisted FSE operations target a comment section within a news article page. Hence, the effect of FSE cannot be accurately estimated unless the model accounts for the variance due to the exposure to the news article's content. Furthermore, not all users who visit the attacked article's comment space would be exposed to the same level of manipulation: Depending on when a user visits the article, the salience of FSE is different, as is its effect on the exposed user. Accordingly, we estimated the following two-way fixed effects regression model that controlled the content effect and the exposure effect, along with time and individual fixed effects:

$$PV_{it} = \beta_0 Post_{it} + \beta_1 Post_{it} Focal_i + \beta_2 Post_{it} Focal_i FSE_i + u_i + v_t + \varepsilon_{it}, \quad (2)$$

where $Post_{it}$ is an indicator variable that indicates whether time t occurred after the exposure to the news content or before (1: after, 0: before) in either the treatment or the control group. It is unique to each user since the user visits the focal news article or control news articles at various time periods. $Focal_i$ is an indicator variable that indicates whether user i is in the treatment group (i.e., who visited the focal news article with FSE) or the control group, which was not affected by the manipulator (1: treatment group, 0: control group); u_i is a fixed effect for user i; v_t captures a fixed effect for time t; and ε_{it} represents an idiosyncratic error term that follows a standard normal distribution. Note that the salience of FSE has subscript i instead of subscript t. FSE_i is the salience of fake social engagements that user i was exposed to at the time of her visit to the focal news article (i.e., $FSE_i \equiv FSE_{t=i\text{'s arrival time}}$).

The beta parameters, namely, β_0, β_1, and β_2, measured the change in page views following a visit to the focal news article or a control news article. That is, β_0 captured the change in page views induced by the news content, β_1 represented the baseline effect of the exposure to the manipulated comment section, and β_2 denoted the moderating effect of the salience of FSE.

One challenge in estimating the effect of FSE on public attention is that comment visibility (i.e., comment rankings) is a function of both FSE and organic user engagement. That is, the FSE variable computed by Equation (1) would be affected not only by manipulated votes but also by organic votes cast by general users. Econometrically, this gives rise to an issue of endogeneity due to the correlation between the FSE variable and the idiosyncratic error term in Equation (2). The virality of the focal news, for example, might affect both the number of organic votes and public attention to news simultaneously. Another source of correlation might be reverse causation in that organic users' news consumption might affect the FSE variable by increasing the number of organic votes. Thus, we use two-stage least squares (2SLS) estimation (Greene, 2017; Angrist & Pischke, 2008), using the following first-stage equation, to attribute FSE only to the effect of the salience of manipulation operations across comments:

$$FSE_i = \delta_0 + \delta_1 UV_i + \delta_2 DV_i + \xi_i, \quad (3)$$

where UV_i is the number of the manipulator's upvotes for the comments to which user i was exposed, DV_i is the number of the manipulator's downvotes for the comments to which user i was exposed, and ξ_i is a random error term.

Manipulative votes cast by a bot-assisted manipulator serve as valid instrumental variables for identifying the effect of FSE on organic users' attention and news consumption. First, they are clearly correlated with the FSE variable measured by Equation (1), due to their direct influence on the ranking of comments according to the platform's comment-ranking algorithm, satisfying the inclusion restriction. Second, they are independent of the error term in the organic users' news consumption model (i.e., Equation 2). Because the manipulative votes were generated by bots that cast a vast number of upvotes and downvotes for targeted comments that match Druking's keyword list, public users' attention to news has no bearing on the generation of manipulative votes. In addition, public users are unable to detect or distinguish the presence of manipulative votes from organic votes, further confirming the independence between manipulative votes and the idiosyncratic error term for users' attention to news, satisfying the exclusion restriction.

Table 2. Impact of FSE on News Consumption

	Parameter	Variable	Estimate	*SE*
The first-stage equation	δ_0	Intercept	0.219***	(0.015)
	δ_1	UV_i	0.646***	(0.006)
	δ_2	DV_i	-0.482***	(0.005)
The second-stage equation	β_0	$Post_{it}$	0.059***	(0.009)
	β_1	$Post_{it} \times Focal_i$	0.162***	(0.008)
	β_2	$Post_{it} \times Focal_i \times \widehat{FSE}_i$	0.332***	(0.010)

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Estimates of user and time fixed effects are omitted for brevity.

Table 3. Time-varying Pattern of the FSE Effect

Parameter	Variable	Estimate	*SE*
β_0	$Post_{it}$	0.050***	(0.009)
β_1	$Post_{it} \times Focal_i$	0.158***	(0.008)
$\beta_{2,base}$	$Post_{it} \times Focal_i \times \widehat{FSE}_i$	0.195***	(0.010)
$\beta_{3,short-term}$	$Post_{it} \times Focal_i \times \widehat{FSE}_i \times ST_{it}$	1.144***	(0.024)

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Estimates of user and time fixed effects are omitted for brevity.

Table 2 shows the results of the 2SLS estimation. The first-stage estimation results reveal that as expected, the bot-generated upvotes increased the relative salience of FSE while its downvotes decreased it. Furthermore, both the R-squared and F-statistic values of the regression were large (i.e., $R^2 = 0.390, F = 557$), alleviating the concern of weak instruments (Bound et al., 1995). Notably, the second-stage regression results show that the FSE effect was statistically significant and positive. That is, as the salience of FSE increased by one unit, users increased their subsequent news consumption on the platform by 0.332 pages per hour.

The effect of FSE may change over time. To distinguish its short-term and long-term effects, we introduced an interaction term between the main effect and a short-term dummy (ST_{it}) that represented the first three hours after leaving the focal news page.[11] Table 3 reveals that the short-term effect was greater than the long-term effect. For the first three hours, a one-unit increase in the salience of FSE increased a user's subsequent hourly news consumption by 1.339 ($= 0.195 + 1.144$) page views. After the first three hours, its effect was still positive yet reduced to 0.195 page views per hour. Since our data spanned up to 39 hours from the publication of the focal news article, we were unable to empirically measure the effect's longevity after 39 hours.

In addition, the effect of FSE might vary by demographic group. We investigated the effect's user heterogeneity by incorporating interaction terms with a user's gender and age, respectively. Table 4 shows that the magnitude of the FSE effect was smaller for female users than for male users, and its magnitude was larger for younger users (under the age of thirty) than for those in their sixties or older.

H1: First-Level Agenda Setting (Effect of Bot-Assisted FSE on Public Attention to Political News Over Non-Political News)

According to H1, the salience of FSE should draw more public attention to political news than non-political news. We tested this hypothesis by examining the effect of FSE on the difference in page views between political and non-political news articles (i.e., sports, entertainment, and other miscellaneous news):

$$PV_{it}^{poli} - PV_{it}^{nonpoli} = \beta_0 Post_{it} + \beta_1 Post_{it} Focal_i + \beta_2 Post_{it} Focal_i FSE_i + u_i + v_t + \varepsilon_{it} \quad (4)$$

Overall, the results support H1, showing the positive and significant effect of the salience of FSE on the net difference in page views between the politics section and non-politics sections: $\beta_2 = 0.048$, $p < 0.01$ for the comparison with sports, $\beta_2 = 0.103$, $p < 0.01$ for the comparison with entertainment, $\beta_2 = 0.027$, $p < 0.01$ for the comparison with other news (see Table 5). The results suggest that the rate of increase in news consumption induced by FSE was greater in the political news domain compared to non-political news topics.

[11] The choice of a three-hour window for the short-term period was made empirically by experimenting with different time windows. Although the magnitude of the impact changes according to short-term durations, we observed a consistent pattern in which the impact of the manipulator is temporarily strong but significantly weakens in the long run.

Table 4. Heterogeneity of the FSE Effect

Parameter	Variable	Estimate	*SE*
β_0	$Post_{it}$	0.059***	(0.009)
β_1	$Post_{it} \times Focal_i$	0.162***	(0.008)
$\beta_{2,base}$	$Post_{it} \times Focal_i \times \widehat{FSE}_i$	0.302***	(0.044)
$\beta_{2,\text{female}}$	$Post_{it} \times Focal_i \times \widehat{FSE}_i \times D_{\text{female}}$	-0.043**	(0.021)
$\beta_{2,\text{age0119}}$	$Post_{it} \times Focal_i \times \widehat{FSE}_i \times D_{\text{age0119}}$	0.237***	(0.068)
$\beta_{2,\text{age2029}}$	$Post_{it} \times Focal_i \times \widehat{FSE}_i \times D_{\text{age2029}}$	0.143***	(0.047)
$\beta_{2,\text{age3039}}$	$Post_{it} \times Focal_i \times \widehat{FSE}_i \times D_{\text{age3039}}$	0.001	(0.046)
$\beta_{2,\text{age4049}}$	$Post_{it} \times Focal_i \times \widehat{FSE}_i \times D_{\text{age4049}}$	-0.033	(0.047)
$\beta_{2,\text{age5059}}$	$Post_{it} \times Focal_i \times \widehat{FSE}_i \times D_{\text{age5059}}$	0.033	(0.051)

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Estimates of user and time fixed effects are omitted for brevity.

Table 5. Impact of FSE on Public Attention to Political News over Non-political News

Parameter	Independent variable	Dependent variable		
		Politics vs. sports ($PV_{it}^{poli} - PV_{it}^{sports}$)	Politics vs. entertainment ($PV_{it}^{poli} - PV_{it}^{enter}$)	Politics vs. other news ($PV_{it}^{poli} - PV_{it}^{other}$)
β_0	$Post_{it}$	0.031*** (0.008)	0.042*** (0.007)	0.088*** (0.007)
β_1	$Post_{it} \times Focal_i$	0.059*** (0.006)	0.059*** (0.006)	-0.026*** (0.006)
β_2	$Post_{it} \times Focal_i \times \widehat{FSE}_i$	0.048*** (0.008)	0.103*** (0.008)	0.027*** (0.007)

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are in parentheses. Estimates of user and time fixed effects are omitted for brevity.

H2: Second-Level Agenda Setting (Effect of Bot-Assisted FSE on Public Attention to Manipulator-Promoted Compared to Manipulator-Demoted Political Attributes)

According to H2, the salience of FSE should direct more public attention to manipulator-promoted political attributes than manipulator-demoted political attributes. H2 was tested in three ways: by examining (a) proactive public attention, operationalized by keyword searches and search-induced page views; (b) passive public attention, operationalized by page views of articles whose headlines were semantically similar to manipulator's keywords (no search involved); and (c) political sentiment, operationalized by page views of anti- vs. pro-government news.

First, we regressed the difference in keyword search counts between searches containing manipulator-promoted keywords and manipulator-demoted keywords on the same set of independent variables as in Equation (2). Table 6 shows that the salience of FSE increased the above-mentioned difference in keyword searches ($\beta_2 = 0.040$, $p < 0.01$). Additionally, because a keyword search results in a list of pertinent news articles, we examined the influence of FSE on search-induced page views. We estimated the same fixed effects regression model with the difference in search-induced page views as a new dependent variable. The results show the salience of FSE increased the difference in search-induced page views between articles associative with manipulator-promoted search keywords and articles associative with manipulator-demoted search keywords ($\beta_2 = 0.027$, $p < 0.01$).

Second, we conducted the same fixed effects regression analysis using a different dependent variable: the net difference in page views for articles with and without headlines associated with the manipulator's FSE keywords. The results are consistent with the results for the search-induced page views. That is, the salience of FSE increased the difference in page views between articles with similar headlines to the manipulator's promoting keywords and articles with similar headlines to the manipulator's demoting keywords ($\beta_2 = 0.021$, $p < 0.01$).

Lastly, we tested the second-level agenda-setting effect in terms of political sentiment. The results indicate that the salience of FSE increased the difference in page views between articles with anti-government sentiment and articles with pro-government sentiment ($\beta_2 = 0.007$, $p < 0.01$), which is well-aligned with the manipulator's intention.

To summarize, all results in Table 6 demonstrate that the salience of FSE directed greater public attention to political attributes consistent with the manipulator's goal, thus supporting H2.

Table 6. Impact of FSE on Public Attention to Political Attributes over Non-political Attributes

Parameter	Independent variable	Dependent variable			
		Proactive public attention		Passive public attention	Political sentiment
		Search keywords ($KS_{it}^{promo} - KS_{it}^{demo}$)	Search-induced page views ($PV_{it}^{promo-search} - PV_{it}^{demo-search}$)	Related page views ($PV_{it}^{promo} - PV_{it}^{demo}$)	Political sentiment ($PV_{it}^{antigov} - PV_{it}^{progov}$)
β_0	$Post_{it}$	0.058*** (0.010)	0.040*** (0.004)	0.026*** (0.003)	0.011*** (0.002)
β_1	$Post_{it} \times Focal_i$	-0.041*** (0.008)	-0.025*** (0.003)	0.022*** (0.002)	-0.004** (0.002)
β_2	$Post_{it} \times Focal_i \times \widehat{FSE}_i$	0.040*** (0.010)	0.027*** (0.004)	0.021*** (0.003)	0.007*** (0.003)

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are in parentheses. Estimates of user and time fixed effects are omitted for brevity.

Robustness Checks

To assess the robustness of the empirical findings, we performed a series of robustness checks by operationalizing the model components differently. We first investigated the sensitivity of our results to various ways of measuring FSE (i.e., independent variables) and public attention (i.e., dependent variable). Then, we examined the robustness of the second-level agenda-setting test results by exploring different parameter choices for the machine learning procedure. Lastly, we conducted a Granger causality test by developing a multisite entry, relative time model (Angrist & Pischke, 2008; Autor, 2003).

First, we developed three alternative measures of the salience of FSE. Note that the original FSE was computed by the difference in the sum of the inverse rankings between manipulator-promoted and manipulator-demoted comments (see Equation 1). The first alternative metric was based on averages rather than summation. Since the total number of default comments displayed using a mobile device screen setting was five, while it was 10 for PC users, the average of inverse rankings was not perfectly correlated with the original variable that employed the summation. The second metric summed rankings as they were, rather than taking an inverse and eliminating weights given in the reverse order of their positions. We expected the sign of the effect to be negative since lower values mean a greater salience of manipulator-promoted comments over manipulator-demoted comments. The third metric considered only the top-ranked comment, assigning it a value of 1 if the manipulator supported it, 0 if it was neutral, and -1 if the manipulator opposed it. When each of the three alternative measures of the FSE variable was applied, the results were consistent with the original results, as shown in Tables 7 and 8.

Second, we examined the sensitivity of our empirical findings with respect to the choice of the dependent variable. While the number of page views was used to measure public attention in the main analysis, the time users spent viewing news articles can be used as an alternative proxy for users' attention. When we conducted our analyses with this alternative dependent variable, the results remained consistent. We also tested if our results were driven by a small number of outliers by removing data points whose page views were over the 99th percentiles of the page view distribution. The main results held regardless of the removal of outliers.

Third, we conducted sensitivity analyses with respect to the machine learning models we employed in the process of testing the second-level agenda-setting effect. Because the identification of search keywords associated with the manipulator's intention relied on a parameter of our choice, which determined the number of most aligned keywords, we tried different values (i.e., top 50, top 200) and confirmed that the second-level agenda-setting effects held regardless of the choice of the parameter. Similarly, we also needed to choose a cutoff value for identifying news articles that were in harmony with the manipulator's keywords based on their similarity. We confirmed the robustness of the main findings by exploring different cutoff values (i.e., top 5% and top 20%). In addition, the degree of alignment between the manipulator's intention and news articles can be computed in various ways. While the similarity was measured based on the manipulator's keywords and news titles as in the main analysis, we could use the manipulator-promoted (or demoted) comments instead of the manipulator's keywords or use news content instead of news titles. By employing different combinations in computing similarity, we found consistent support for the second-level agenda-setting effect.

Table 7. Robustness Checks for Baseline and First-Level Agenda-Setting Hypothesis

	Baseline	First-level agenda-setting (H1)		
		Politics vs. sports	Politics vs. entertainment	Politics vs. other news
Original	0.332*** (0.010)	0.048*** (0.008)	0.103*** (0.008)	0.027*** (0.007)
IV: Average of inverse rankings	1.758*** (0.033)	0.431*** (0.027)	0.694*** (0.026)	0.269*** (0.025)
IV: Sum of rankings	-0.007*** (0.001)	-0.006*** (0.001)	-0.006*** (0.001)	-0.004*** (0.001)
IV: Indicator for top ranking	0.435*** (0.008)	0.104*** (0.007)	0.170*** (0.006)	0.065*** (0.006)
DV: Time spent	0.631*** (0.020)	0.095*** (0.017)	0.222*** (0.015)	0.032* (0.018)
DV: Without outliers	0.243*** (0.006)	0.048*** (0.005)	0.090*** (0.004)	0.005 (0.005)
Lead2	-0.020 (0.012)	-0.017 (0.013)	0.005 (0.010)	0.067*** (0.009)
Lag0	0.473*** (0.015)	0.166*** (0.013)	0.198*** (0.012)	0.121*** (0.012)
Lag1	0.251*** (0.017)	0.021 (0.015)	0.073*** (0.014)	0.047*** (0.013)
Lag2	0.074*** (0.021)	0.023 (0.019)	0.031* (0.017)	0.048*** (0.016)

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Reported are the estimates of the parameter of interest (β_2) and their standard errors in parentheses. Estimates of β_0, β_1, user and time fixed effects are omitted for brevity. Following Autor (2003), Lead2 is a relative time dummy that indicates the time span from twelve to twenty-four hours prior to the exposure to the focal article's comment section, and Lag0, Lag1, and Lag2 are relative time dummies for 0~12, 12~24, and 24~36 hours after manipulation exposure, respectively.

Table 8. Robustness Checks for Second-Level Agenda-Setting Hypothesis

	Second-level agenda setting (H2)			
	Search keywords	Search-induced page views	Related page views	Political sentiment
Original	0.040*** (0.010)	0.027*** (0.004)	0.021*** (0.003)	0.007*** (0.003)
IV: Average of inverse rankings	0.236*** (0.034)	0.173*** (0.012)	0.156*** (0.010)	0.022** (0.009)
IV: Sum of rankings	-0.001 (0.001)	-0.001*** (0.000)	-0.002*** (0.000)	0.000 (0.000)
IV: Indicator for top ranking	0.058*** (0.009)	0.042*** (0.003)	0.038*** (0.002)	0.006*** (0.002)
DV: Time spent	NA	0.027*** (0.006)	0.037*** (0.009)	0.020 (0.014)
DV: Without outliers	0.018*** (0.002)	0.009*** (0.001)	0.018*** (0.002)	0.005*** (0.002)
Search keyword level: Top 50	0.038*** (0.011)	0.026*** (0.003)	NA	NA
Search keyword level: Top 200	0.046*** (0.010)	0.027*** (0.004)	NA	NA
Similarity level: Top 5%	NA	0.021*** (0.003)	0.026*** (0.002)	0.004* (0.002)
Similarity level: Top 20%	NA	0.043*** (0.005)	0.019*** (0.004)	0.013*** (0.004)
Article title × Abuser comment	NA	0.030*** (0.003)	0.065*** (0.003)	NA
Article text × Abuser keyword	NA	0.021*** (0.002)	0.023*** (0.002)	NA
Article text × Abuser comment	NA	0.016*** (0.002)	0.009*** (0.002)	NA
Lead2	0.001 (0.006)	0.001 (0.002)	-0.006 (0.004)	0.005 (0.004)
Lag0	0.027*** (0.007)	0.023*** (0.003)	0.071*** (0.004)	0.012*** (0.004)
Lag1	0.015* (0.008)	0.006** (0.003)	0.030*** (0.005)	0.011** (0.005)
Lag2	0.004 (0.010)	0.002 (0.004)	0.026*** (0.006)	0.021*** (0.006)

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Reported are the estimates of the parameter of interest (β_2) and their standard errors in parentheses. Estimates of β_0, β_1, user and time fixed effects are omitted for brevity. Following Autor (2003), Lead2 is a relative time dummy that indicates the time span from twelve to twenty-four hours prior to the exposure to the focal article's comment section, and Lag0, Lag1, and Lag2 are relative time dummies for 0-12, 12-24, and 24-36 hours after manipulation exposure, respectively. NA denotes "not applicable."

Lastly, we developed a multisite entry, relative time model (Angrist & Pischke, 2008; Autor, 2003) to conduct a Granger causality test. If the salience of FSE is a cause, a change in users' news consumption would be predicted by past exposure to FSE (i.e., lag) but not by future exposure to FSE (i.e., lead). The following is the lead-lag regression equation:

$$PV_{it} = \beta_{2,-2} Lag2_{it} Focal_i FSE_i + \beta_{2,-1} Lag1_{it} Focal_i FSE_i + \beta_{2,0} Lag0_{it} Focal_i FSE_i + \beta_{2,+2} Lead2_{it} Focal_i FSE_i + u_i + v_t + \varepsilon_{it}, \quad (5)$$

where $Lead2_{it}$ is a relative time dummy that indicates the time span from twelve to twenty-four hours prior to the exposure to the focal article's comment section, and $Lag0_{it}$, $Lag1_{it}$, and $Lag2_{it}$ are relative time dummies for 0~12, 12~24, and 24~36 hours after manipulation exposure, respectively. Note that a dummy indicating zero to twelve hours before the arrival at the focal article (i.e., $Lag1_{it}$) is omitted as the base group. Tables 7 and 8 shows the null effect of the lead variable. Only after the exposure to the manipulation does the FSE effect become statistically significant, lending support to the causal effect of FSE on

users' news consumption. However, the positive effect does not last long, rapidly diminishing within a day.

In sum, our empirical results hold consistently across various conditions, as summarized in Tables 7 and 8. The robustness check analyses show that our empirical findings were neither driven by nor dependent on a particular choice of dependent variable, independent variable, and parameters for the machine learning procedure.

General Discussion

This study examines the spillover effect of bot-assisted fake social engagement (FSE), a widespread false amplification practice in the global disinformation industry, on manufacturing public attention in a large information ecosystem. Based on the algorithmic conduit brokerage perspective (Salge et al., 2022) and the agenda-setting framework (McCombs & Valenzuela, 2020), we pose a research question of whether FSE produces the spillover effect on public attention to information beyond the immediately manipulated context (RQ) and hypothesize that the salience of FSE shifts public attention in line with the manipulator's intention (H1 and H2). This study advances disinformation research by integrating bot- and user-centered approaches to demonstrate that bots' capacity for the rapid scaling of social engagement elicits a false bandwagon of public attention. We integrate the two approaches by empirically examining the spillover of bot operation effects into a broader information environment. Methodologically, this study leverages a unique large-scale user-behavioral data source and the ground truth of disinformation bot activities, coupled with advanced semi-supervised ML techniques.

Considering that disinformation campaigns have increasingly incorporated automation software, understanding the mechanism of bot-assisted FSE and its effect on the general public's attention may offer theoretical and managerial insights into disinformation's harms on digital information commons. This section discusses the study's theoretical implications, methodological contributions, and managerial implications for scholars, practitioners, and policymakers.

Theoretical Contributions

Theoretically, bot-assisted FSE manifests functions of algorithmic conduit brokerage, particularly in terms of bots' ability for social alerting and rapid scaling (Salge et al., 2022). In addition to an algorithmic conduit brokerage perspective, we use agenda-setting theory to explain a mechanism of how bot-assisted FSE helps the human manipulator (i.e., the programmer behind bots) game (deceptively) the process of public agenda setting on digital platforms. By deploying bots, the manipulator can rapidly amplify social engagement volume at scale, which in turn results in the rearrangement of information positions and eventually elicits a bandwagon of public attention in the manipulator's favor. Further, this study contends that the influence of bot-assisted FSE does not just stay in the immediately manipulated space but leaks into a larger information consumption ecosystem. By adopting agenda-setting theory, this study elaborates a mechanism by which a manipulator plays the role of a public agenda setter by falsely amplifying the salience of selective messages. Importantly, deceptive agenda setting does not necessitate creating one's own fake messages. Manipulators can manufacture public attention by rapidly scaling the visibility of existing genuine content in their favor. Despite its prevalence and significance due to cost-effectiveness, bot-assisted FSE has been largely overlooked in the literature due to the difficulty of obtaining compatible empirical data sources, which should ideally disambiguate inauthentic engagement from organic engagement. In this sense, this study's focus on bot-assisted FSE uniquely advances disinformation research.

In addition to disinformation research, this study contributes to advancing agenda-setting theory by theorizing a deceptive agenda-setting mechanism and developing computational processes to empirically demonstrate it. In particular, our semi-supervised ML modeling approach to detecting and including associative textual cues as compositions of the issue attributes echoes the tenet of the network agenda-setting model, an advanced branch of agenda-setting theory that contends that the audience remember news not only as single issues/attributes but also as a bundle of mental associations (Vu et al., 2014; Guo & Vargo, 2015). To our knowledge, this study is the first attempt to incorporate advanced machine learning techniques to infer associative concepts that represent issue attributes.

The study's findings suggest both first- and second-level agenda-setting effects of bot-assisted FSE on public attention. On the first level, we examined news domain-specific page views by comparing page views for the politics news section to those for non-politics sections. The findings revealed that bot-assisted FSE operations have a first-level agenda-setting effect on how the public allocates its attention, as our findings reveal that the exposed users directed greater attention to political news than to non-political news such as sports and entertainment. On the second level, we compared political attribute-specific news page views between articles that contained manipulator-promoted political attributes and those that contained manipulator-demoted attributes. The findings confirm the second-level agenda-setting effect, as the FSE effect was greater for page views with manipulator-promoted

attributes than for those with manipulator-demoted attributes. The results were consistent for proactive public attention (keyword searches and search-induced page views), passive public attention (other page views that occurred without search), and political sentiment-driven public attention. Altogether, the empirical findings attest to the spillover influence of bot-assisted FSE on the general public's broader information (news) consumption beyond the immediate context targeted by a manipulator. Our findings of disinformation effects on general users' information behaviors add new insights to existing knowledge that has thus far centered around subpopulation groups of ideologically like-minded and/or heavy platform users, based on a somewhat narrow definition of the sphere of disinformation influence within the immediate interaction context.

Methodological Contributions

Disinformation research has employed ML techniques to tackle detection problems. The current study advances this line of research by demonstrating the utility of semi-supervised ML approaches to explore the effects of disinformation on the general public at scale. In particular, given the sheer scale of our data, we note that it is infeasible to manually code all articles, especially as this requires expert domain knowledge and familiarity with the political background. Prior work has primarily utilized supervised ML with carefully engineered features (e.g., Horne et al., 2018; Potthast et al., 2018; Gangula et al., 2019). In practice, the two major drawbacks of such models are that they require (1) vast quantities of curated labeled training data and (2) features unique to the context or characteristics (e.g., lexicon, style) of the focal language, which is usually English. The latter drawback makes it especially difficult to extend these models to other languages (e.g., Korean).

In this paper, we not only used the doc2vec model, an advanced ML technique that has gained popularity in the IS literature, but also demonstrated the utility of the label propagation model, a semi-supervised learning approach, by combining it with representation learning of text embeddings to effectively resolve the aforementioned two issues. Semi-supervised models are not completely new to online data-driven research. For example, studies have successfully used semi-supervised models to classify the political orientations of Twitter users using the retweet network (Badawy et al., 2018; Luceri et al., 2019). That being said, the application case in this study is distinct from previous studies in that we inferred political attributes of news articles using natural language processing.

Following the paradigm shift from manually engineering features to learning representations, we created data-driven text embeddings using the doc2vec model, which facilitated our investigation into the second-level agenda-setting effect. Since text embedding models are not dependent on specific context or language characteristics (Grave et al., 2018), our proposed approach is generalizable to a wide range of languages. Our main approach, semi-supervised learning, is ideally suited for situations with a limited amount of labeled data that is mixed with abundant unlabeled data during training, resulting in substantial performance improvements (Tarvainen & Valpola, 2017; Iscen et al., 2019). Despite its advantages, semi-supervised learning has attracted little attention in the IS literature, with the exception of Abbasi et al. (2012). In this paper, we demonstrate how such an approach can achieve superior accuracy in predicting specific attributes of information (e.g., political orientation of articles). Our study is one of the first in the IS literature to implement semi-supervised learning to empirical research, broadening the ML spectrum beyond the dichotomy of unsupervised and supervised learning.

Managerial and Policy Contributions

This research has practical implications for online platforms and policymakers. First, our results shed new light on the underlying mechanism of bot-assisted disinformation campaigns on online platforms. This knowledge can be particularly helpful in managing digital platforms that battle increasingly complex opinion manipulation by offering guidance in the design and development of manipulation detection algorithms. In particular, we point out that bot-assisted fake social engagements can substantially contribute to changing the visibility of messages by deploying massive engagements simultaneously at a rapid pace. The content visibility may, of course, not be fully controlled by the manipulator yet can nonetheless be altered to some extent. While a keyword-based deployment of bots is a rather simple technique, this disinformation tactic can be easily operated, making the content curation vulnerable to the attack, especially when the curation algorithm is simplistic (as in the case of the net-vote-based rank order used by the studied platform) and no rigorous monitoring protocol exists.

Second, considering the ever-expanding role of digital social conversations in setting the "climate" of public opinion in network societies, it is obvious that the compromised social engagement culture deteriorates the quality of deliberative democracy. Platforms thus must take some social responsibility for the conversational health of society. In particular, bot-assisted manipulation has become increasingly common globally. Intensified bot deployment is deeply problematic because it is scalable and can thus easily generate bandwagon effects (Caldarelli et al., 2020). Our study reiterates the importance of paying managerial attention to bot-assisted false amplification, as well as aspects of human-

crafted false messages, in counteracting disinformation operations. Concerted efforts of online platforms and policy regulators will be necessary, and data-driven empirical insights, such as our findings, can serve as shared intelligence in the process.

Limitations and Future Research

This research is subject to several limitations which in turn highlight potential areas for future research. First, our work relies on observational data of a single event. While empirical analysis of observational data has its own merits (e.g., high external validity), it entails costs such as limited observations, unobservable confounders, and context dependency. Experimental studies where the effect of FSE can be clearly measured under various conditions would complement this research, allowing our findings to be generalized to broader contexts. Second, the context of this research limits us from examining how the effect of FSE might be influenced by social networking. The online news platform studied in this study is a news aggregator, similar to Yahoo News, rather than a social networking service, similar to Twitter or Facebook; thus, it provides limited data on how its users share specific news/information. It would be fascinating to examine the role of social interaction in the context of fake social engagement operations. Third, this research studied a particular type of FSE operated by fake votes on organic comments. Other important contexts of fake engagement, such as sharing articles/ads/posts/videos or paying for targeted ads, paired with relevant data would be very interesting for future work. For example, Bradshaw (2019) studied search engine optimization manipulation by junk news domains that targeted an increase in their discoverability on Google Search. Last, this research measured the short-term effect of FSE, which was manifested by users' news consumption behavior. Therefore, there are remaining questions, such as how persistent the effect would be and whether FSE would affect not only people's information search behavior but also their attitudes or beliefs. We leave these questions to future research.

Conclusion

Despite its limitations, this study has theoretical as well as practical implications for IS researchers, online platforms, and regulators. Many disinformation mechanisms still remain black-boxed, including those related to fake social engagement operations. To our knowledge, this study is the first attempt to unravel the workings of a fake social engagement operation and its broad effect on users. Through the lens of agenda-setting theory, the findings indicate that programmable bots increase the potential for perpetrators to falsely inflate the salience of certain messages and subsequently manufacture public attention to information. This research contributes to the IS literature by broadening our theoretical understanding of a bot-assisted disinformation technique and by demonstrating how a computational and data-driven approach can help quantify its effects on general users' informational behaviors. We hope this study will lead to more IS scholarly attention to the misuse/abuse of digital technologies and their ramifications on cybersocial security.

Acknowledgments

The third author is thankful for the mentoring received through the U.S.-Korea NextGen Scholar program under the sponsorship of the Korea Foundation. The fourth and fifth authors are co-corresponding authors for this paper. The third author's effort was partly supported by DEVCOM Army Research Laboratory-Army Research Office (Award Number: W911NF1910066), MIT-Lincoln Laboratory (Award Number: PO 7000506684), and the National Science Foundation (Award Number: 2210137). The fifth author's effort was financially supported by Hansung University.

References

Abbasi, A., Albrecht, C., Vance, A., & Hansen, J. (2012). MetaFraud: A meta-learning framework for detecting financial fraud. *MIS Quarterly*, *36*(4), 1293-1327. https://doi.org/10.2307/41703508

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Autor, D. H. (2003). Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of Labor Economics*, *21*(1), 1-42. https://doi.org/10.1086/344122

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 258-265). https://doi.org/10.1109/ASONAM.2018.8508646

Bail, C. A., Guay, B., Maloney, E., Combs, A., Hillygus, D. S., Merhout, F., Freelon, D., & Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the National Academy of Sciences*, *117*(1), 243-250. https://doi.org/10.1073/pnas.1906420116

Baly, R., Da San Martino, G., Glass, J., & Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 4982-4991). https://doi.org/10.18653/v1/2020.emnlp-main.404

Bastos, M. T., & Mercea, D. (2019). The Brexit Botnet and user-generated hyperpartisan news. *Social Science Computer Review*, *37*(1), 38-54. https://doi.org/10.1177/0894439317734157

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Boichak, O., Hemsley, J., Jackson, S., Tromble, R., & Tanupabrungsun, S. (2021). Not the bots you are looking for: Patterns and effects of orchestrated interventions in the US and German elections. *International Journal of Communication*, *15*, 814-839. https://ijoc.org/index.php/ijoc/article/view/14866

Bound, J., Jaeger, D. A., Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443-450. https://doi.org/10.2307/2291055

Bradshaw, S. (2019). Disinformation optimised: Gaming search engine algorithms to amplify junk news. *Internet Policy Review*, *8*(4), 1-24. https://doi.org/10.14763/2019.4.1442

Bradshaw, S., Bailey, H., & Howard, P. N. (2021). *Industrialized disinformation: 2020 global inventory of organized social media manipulation*. Computational Propaganda Project at the Oxford Internet Institute. https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation

Bradshaw, S., & Howard, P. N. (2018). *Challenging truth and trust: A global inventory of organized social media manipulation*. Computational Propaganda Project at the Oxford Internet Institute. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/127/2018/07/ct2018.pdf

Caldarelli, G., Nicola, R. D., Vigna, F. D., Petrocchi, M., & Saracco, F. (2020). The role of bot squads in the political propaganda on Twitter. *Communications Physics*, *3*(1), 1-15. https://doi.org/10.1038/s42005-020-0340-4

Carman, M., Koerber, M., Li, J., Choo, K. R., & Ashman, H. (2018). Manipulating visibility of political and apolitical threads on Reddit via score boosting. In *Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 184-190). https://doi.org/10.1109/TrustCom/BigDataSE.2018.00037

Carnahan, D., & Garrett, R. K, (2020). Processing style and responsiveness to corrective information. *International Journal of Public Opinion Research*, *32*(3), 530-546. https://doi.org/10.1093/ijpor/edz037

Chen, E., Chang, H., Rao, A., Lerman, K., Cowan, G., & Ferrara, E. (2021). COVID-19 misinformation and the 2020 US presidential election. *The Harvard Kennedy School Misinformation Review*, *1*, Article 7. https://doi.org/10.37016/mr-2020-57

Choe, S. (2018). Ally of South Korean leader conspired to rig online opinion, inquiry finds. *The New York Times*. https://www.nytimes.com/2018/08/27/world/asia/moon-jae-in-online-scandal.html

Coleman, R., & Wu, H. D. (2010). Proposing emotion as a dimension of affective agenda setting: Separating affect into two components and comparing their second-level effects. *Journalism & Mass Communication Quarterly*, *87*(2), 315-327. http://dx.doi.org/10.1177/107769901008700206

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, *63*(10), 72-83. https://doi.org/10.1145/3409116

David, E., Zhitomirsky-Geffet, M., Koppel, M., & Uzan, H. (2016). Utilizing Facebook pages of the political parties to automatically predict the political orientation of Facebook users. *Online Information Review*, *40*(5), 610-623. http://dx.doi.org/10.1108/OIR-09-2015-0308

Edelson, L., Nguyen, M. K., Goldstein, I., Goga, O., McCoy, D., & Lauinger, T. (2021). Understanding engagement with US (mis)information news sources on Facebook. In *Proceedings of the 21st ACM Internet Measurement Conference* (pp. 444-463). https://doi.org/10.1145/3487552.3487859

Effron, D. A., & Raj, M. (2020). Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science*, *31*(1), 75-87. https://doi.org/10.1177/0956797619887896

Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., & Adams, K. (2022). Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review*, *40*(3), 560-578. https://doi.org/10.1177/0894439320914853

Freelon, D., & Wells, C. (2020) Disinformation as political communication. *Political Communication*, *37*(2), 145-156. https://doi.org/10.1080/10584609.2020.1723755

Fujiwara, Y., & Irie, G. (2014). Efficient label propagation. In *Proceedings of the International Conference on Machine Learning* (pp. 784-792).

Gangula, R. R. R., Duggenpudi, S. R., & Mamidi, R. (2019). Detecting political bias in news articles using headline attention. In *Proceedings of the ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 77-84). https://doi.org/10.18653/v1/W19-4809

Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from #Brexit and #USElection. *European Economic Review*, *136*, Article 103772. https://doi.org/10.1016/j.euroecorev.2021.103772

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation* (pp. 3483-3487).

Greene, W. H. (2017). *Econometric analysis*. Pearson.

Guo, L., & Vargo, C. (2015). The power of message networks: A big-data analysis of the network agenda setting model and issue ownership. *Mass Communication and Society*, *18*(5), 557-576. http://dx.doi.org/10.1080/15205436.2015.1045300

Guo, L., & Vargo, C. (2020). Fake news and emerging online media ecosystem: An integrated intermedia agenda-setting analysis of the 2016 U.S. presidential election. *Communication Research*, *47*(2), 178-200. https://doi.org/10.1177/0093650218777177

Horne, B. D., Dron, W., Khedr, S., & Adali, S. (2018). Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the World Wide Web Conference* (pp. 235-238). https://doi.org/10.1145/3184558.3186987

Iscen, A., Tolias, G., Avrithis, Y., & Chum, O. (2019). Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5070-5079). https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00521

Jeong, J., Kang, J.-H., & Moon, S. (2020). Identifying and quantifying coordinated manipulation of upvotes and downvotes in Naver news comments. In *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 303-314. https://doi.org/10.1609/icwsm.v14i1.7301

Kahan, D. M., Landrum, A., Carpenter, K., Helft, L., & Jamieson, K. H. (2017). Science curiosity and political information processing. *Political Psychology*, *38*(S1), 179-199. http://dx.doi.org/10.1111/pops.12396

Kang, H., & Yang, J. (2020). Quantifying perceived political bias of newspapers through a document classification technique.

Journal of Quantitative Linguistics, *29*(2), 127-150. https://doi.org/10.1080/09296174.2020.1771136

Kim, S.-H., Scheufele, D. A., & Shanahan, J. (2002). Think about it this way: Attribute agenda-setting function of the press and the public's evaluation of a local issue. *Journalism & Mass Communication Quarterly*, *79*(1), 7-25. https://doi.org/10.1177/107769900207900102

Kiousis, S. (2005). Compelling arguments and attitude strength: Exploring the impact of second-level agenda setting on public opinion of presidential candidate images. *Harvard International Journal of Press/Politics*, *10*(2), 3-27. https://doi.org/10.1177/1081180X05276095

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning* (pp. 1188-1196).

Lim, W., Lee, C., & Choi, D. (2019). *Opinion polarization in Korea: Its characteristics and drivers* (Korea Development Institute (KDI) Research Monograph). https://doi.org/10.22740/kdi.rm.2019.03

Luceri, L., Deb, A., Badawy, A., & Ferrara, E. (2019). Red bots do it better: Comparative analysis of social bot partisan behavior. In *Companion Proceedings of the World Wide Web Conference* (pp. 1007-1012). https://doi.org/10.1145/3308560.3316735

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. Data & Society Research Institute. https://datasociety.net/library/media-manipulation-and-disinfo-online

McCombs, M., & Shaw, D. (1972). The agenda setting function of mass media. *Public Opinion Quarterly*, *36*(2), 176-187. https://doi.org/10.1086/267990

McCombs, M., & Valenzuela, S. (2020). *Setting the agenda: Mass media and public opinion*. Polity Press.

Mindel, V., Mathiassen, L., & Rai, A. (2018). The sustainability of polycentric information commons. *MIS Quarterly*, *42*(2), 607-632. http://dx.doi.org/10.25300/MISQ/2018/14015

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society*, *20*(10), 3720-3737. https://doi.org/10.1177/1461444818758715

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865-1880. http://dx.doi.org/10.2139/ssrn.2958246

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590-595. https://doi.org/10.1038/s41586-021-03344-2

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Potthast, M., Kiesel, J., & Reinartz, K. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 3528-3539). https://doi.org/10.18653/v1/P18-1022

Qiao, D., Lee, S.-Y., Whinston, A. B., & Wei, Q. (2020). Financial incentives dampen altruism in online prosocial contributions: A study of online reviews. *Information Systems Research*, *31*(4), 1361-1375. https://doi.org/10.1287/isre.2020.0949

Rojecki, A., & Meraz, S. (2016). Rumors and factitious informational blends: The role of the web in speculative politics. *New Media & Society*, *18*(1), 25-43. https://doi.org/10.1177/1461444814535724

Rossi, S., Rossi, M., Upreti, B., & Liu, Y. (2020). Detecting political bots on Twitter during the 2019 Finnish parliamentary election. In *Proceedings of the Hawaii International Conference on System Sciences* (pp. 2430-2439).

Salge, C., Karahanna, E., & Thatcher, J. B. (2022). Algorithmic processes of social alertness and social transmission: How bots disseminate information on Twitter. *MIS Quarterly*, *46*(1), 229-260. https://doi.org/10.25300/MISQ/2021/15598

Schäfer, F., Evert, S., & Heinrich, P. (2017). Japan's 2014 general election: Political bots, right-wing internet activism, and Prime Minister Shinzō Abe's hidden nationalist agenda. *Big Data*, *5*(4), 294-309. https://doi.org/10.1089/big.2017.0049

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, *116*(16), 7662-7669. https://doi.org/10.1073/pnas.1805871115

Shin, D., He, S., Lee, G. M., Whinston, A. B., Cetintas, S., & Lee, K.-C. (2020). Enhancing social media analysis with visual data analytics: A deep learning approach. *MIS Quarterly*, *44*(4), 1459-1492. https://doi.org/10.25300/misq/2020/14870

Stella, M., Ferrara, E., & De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, *115*(49), 12435-12440. https://doi.org/10.1073/pnas.1803470115

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the Advances in Neural Information Processing Systems* (pp. 1195-1204).

Vargo, C., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, *20*(5), 2028-2049. https://doi.org/10.1177/1461444817712086

Varol, O., Ferrara, E., Davis, C., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 280-289. https://doi.org/10.1609/icwsm.v11i1.14871

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Vu, H. T., Guo, L., & McCombs, M. (2014). Exploring the world outside and the pictures in our heads: A network agenda-setting study. *Journalism & Mass Communication Quarterly*, *91*(4), 669-686. https://doi.org/10.1177/1077699014550090

Weedon, J., Nuland, W., & Stamos, A. (2017). *Information operations and Facebook*. Facebook. https://about.fb.com/wp-content/uploads/2017/04/facebook-and-information-operations-v1.pdf

Weidner, K., Beuk, F., & Bal, A. (2020). Fake news and the willingness to share: A schemer schema and confirmatory bias perspective. *Journal of Product & Brand Management*, *29*(2), 180-187. https://doi.org/10.1108/JPBM-12-2018-2155

Wong, J. C., & Ernst, J. (2021). Facebook knew of Honduran president's manipulation campaign—and let it continue for 11 months. *The Guardian*. https://www.theguardian.com/technology/2021/apr/13/facebook-honduras-juan-orlando-hernandez-fake-engagement

Woolley, S., & Howard, P. (2016). Political communication, computational propaganda, and autonomous agents. *International Journal of Communication*, *10*, 4882-4890.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2003). Learning with local and global consistency. In *Proceedings of the Advanced in Neural Information Processing Systems* (pp. 321-328).

Author Biographies

Sanghak Lee is an associate professor of marketing at the W. P. Carey School of Business, Arizona State University. He holds a B.S. in Chemical Engineering from Seoul National University, an M.S. in Management Engineering from KAIST (Korean Advanced Institute of Science and Technology), and a Ph.D. in Marketing from the Ohio State University. His research primarily focuses on direct utility models, Bayesian econometrics, and choice modeling. His work has been published in prestigious journals, including *Marketing Science* and *Management Science.*

Donghyuk Shin is an associate professor of information systems at the College of Business, Korea Advanced Institute of Science and Technology (KAIST). Before joining KAIST, he held positions as an assistant professor at Arizona State University and a machine learning scientist at Amazon. He earned his Ph.D. in computer science from the University of Texas at Austin. His primary research interest is at the nexus of machine learning and information systems, with a focus on artificial intelligence, digital platforms, and business analytics. His research has been featured in *MIS Quarterly, Management Science*, and at leading machine learning conferences such as NeurIPS, ACM RecSys, and CIKM.

K. Hazel Kwon is a professor of digital audiences and the founder and lead researcher of the Media, Information, Data, and Society (MIDaS) Lab at the Walter Cronkite School of Journalism and Mass Communication, and an affiliate faculty with Global Security Initiative's Center on Narrative, Disinformation, and Strategic Influence at Arizona State University. She has received grants from the DoD, NSF, Social Science Research Council, and the Gates Foundation for her various research projects on social media and participation. She has won multiple awards including the AEJMC Emerging Scholar (2020), Top Faculty Papers from the Broadcast Education Association (2022) and Chinese Communication Association (2021), and the Herbert S. Dordick Dissertation Award (3rd place) from the International Communication Association (2012). In 2020-2021, she was selected as a U.S.-Korea NextGen Scholar (ORCiD Id: 0000-0001-7414-6959).

Sang-Pil Han is an associate professor of information systems at the W. P. Carey School of Business, Arizona State University. His research interests encompass artificial intelligence, digital platforms, and business analytics. Notably, his work has been published in esteemed journals such as *Management Science, MIS Quarterly, Information Systems Research, and Journal of Marketing*. Beyond academia, his insights have been showcased in media outlets like *Harvard Business Review, The Wall Street Journal*, and BBC News. Professor Han's research has garnered support from institutions including the Marketing Science Institute, NET Institute, and Hong Kong General Research Fund, as well as private enterprises. He has held educational leadership roles, notably as co-faculty director for the Master of Science in Business Analytics at ASU. Additionally, he served as an associate editor for *Information Systems Research.* Outside academia, his consultation spans from tech startups such as Mathpresso to nonprofits like Simple Steps.

Seok Kee Lee is a professor in the Department of Computer Engineering at Hansung University in South Korea. He received his Ph.D. in management engineering at KAIST (Korea Advanced Institute of Science and Technology). His current research interests include data analytics and artificial intelligence on consumer behavior. His articles have been published in academic journals including *Information Sciences, International Journal of Consumer Studies,* and *Sustainability*.

Appendix A

We describe our label propagation (LP) model used to infer the political sentiment of articles that users visited. Two main advantages of LP are (1) local consistency: nearby data points are likely to have the same label, and (2) global consistency: data points on the same structure (i.e., manifold or cluster) are likely to have the same label. The core idea of LP is to construct an affinity graph from all labeled and unlabeled samples and then iteratively propagate the known labels to the unlabeled samples according to the graph structure. More formally, the algorithm proceeds as follows:

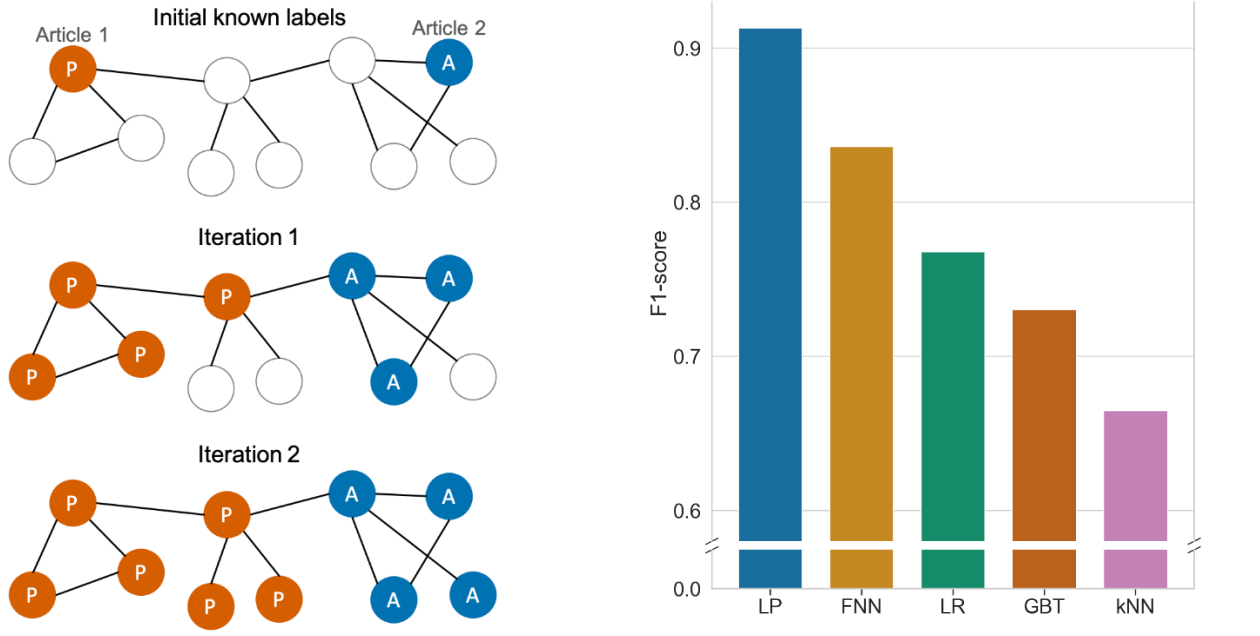
1. Form an affinity graph G and its corresponding adjacency matrix W, where nodes represent samples and edges capture their pairwise similarities (e.g., k-nearest neighbor graph).

2. Construct the normalized Laplacian $L = D^{-1/2}WD^{-1/2}$, where D is the diagonal matrix of node degrees (necessary for convergence).

3. Iterate $F_{t+1} = \lambda LF_t + (1-\lambda)Y$ until convergence, where F_t represents the labels at the t-th iteration, λ is a hyperparameter between 0 and 1 that specifies the relative amount of initial label information to retain, and Y is the vector of initial known labels.

To construct an affinity graph with articles as nodes, we computed pairwise cosine similarities between 342,567 articles using their embedding vectors obtained from our doc2vec model (described in the Organic User Activities section), which has been shown to be accurate in detecting political biases in articles (e.g., Baly et al., 2020; Kang & Yang, 2020). From the pairwise similarities, we formed a sparse k-nearest neighbor graph with $k = 15$ as the affinity graph G. For the iterations in Step 3, we set $\lambda = 0.4$.

Figure A1(a) shows an example of label propagation iterations. Starting from the nodes corresponding to Article 1 (labeled as "P") and Article 2 (labeled as "A"), the initial known labels are propagated to other articles according to the affinity graph at each iteration. We also compare our LP model to other representative supervised ML models, including feed-forward neural network (FNN), logistic regression (LR), gradient boosting trees (GBT), and k-nearest neighbor (kNN) classifiers. Figure A1(b) depicts that the LP model yields the best prediction accuracy (0.913) measured by the F1-score (a standard accuracy metric for classification tasks) averaged over multiple stratified 5-fold cross-validations. We note that hyperparameters of the compared models are tuned with validation sets and F1-scores are reported using separate test sets.



Note: (a) Shows an example of label propagation iterations where the initial known labels of Article 1 ("P") and Article 2 ("A") are spread to other articles (i.e., nodes) according to the affinity graph**. (b)** Presents the prediction accuracies (F1-score) of different ML models showing that label propagation (LP) achieves the best performance.

Figure A1. Example of Label Propagation Iterations and Performance Comparison