



The Effect of Platform Intervention Policies on Fake News Dissemination and Survival: An Empirical Examination

Ka Chung Ng, Jie Tang & Dongwon Lee

To cite this article: Ka Chung Ng, Jie Tang & Dongwon Lee (2021) The Effect of Platform Intervention Policies on Fake News Dissemination and Survival: An Empirical Examination, Journal of Management Information Systems, 38:4, 898-930, DOI: [10.1080/07421222.2021.1990612](https://doi.org/10.1080/07421222.2021.1990612)

To link to this article: <https://doi.org/10.1080/07421222.2021.1990612>



© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 02 Jan 2022.



[Submit your article to this journal](#)



Article views: 3707



[View related articles](#)



[View Crossmark data](#)

The Effect of Platform Intervention Policies on Fake News Dissemination and Survival: An Empirical Examination

Ka Chung Ng ^{a,b}, Jie Tang ^c, and Dongwon Lee ^a

^aDepartment of Information Systems, Business Statistics and Operations Management, School of Business and Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, HONG KONG;

^bDepartment of Management and Marketing, Faculty of Business, Hong Kong Polytechnic University, Hung Hom, Kowloon, HONG KONG; ^cHKU Business School, The University of Hong Kong, Pok Fu Lam, Hong Kong, HONG KONG

ABSTRACT



Fake news on social media has become a serious problem, and social media platforms have started to actively implement various interventions to mitigate its impact. This paper focuses on the effectiveness of two platform interventions, namely a content-level intervention (i.e., a fake news flag that applies to a single post) and an account-level intervention (i.e., a forwarding restriction policy that applies to the entire account). Collecting data from China's largest social media platform, we study the impact of a fake news flag on three fake news dissemination patterns using a propensity score matching method with a difference-in-differences approach. We find that implementing a policy of using fake news flag influences the dissemination of fake news in a more centralized manner via direct forwards and in a less dispersed manner via indirect forwards, and that fake news posts are forwarded more often by influential users. In addition, compared with truthful news, fake news is disseminated in a less centralized and more dispersed manner and survives for a shorter period after a forwarding restriction policy is implemented. This study provides causal empirical evidence of the effect of a fake news flag on fake news dissemination. We also expand the literature on platform interventions to combat fake news by investigating a less studied account-level intervention. We discuss the practical implications of our results for social media platform owners and policymakers.


KEYWORDS

Fake News; Fake News Online; Fake News Flag; Forwarding Restriction Policy; Fake News Dissemination; Quasi-Experiment; Online Disinformation; Platform Policies

Introduction

Online channels such as social media play an important role in information acquisition and dissemination [62]. However, these channels are increasingly affected by the spread of fake news, which should be addressed through substantial efforts, especially during serious social, political, and epidemiological crises like the COVID-19 pandemic of 2020. Unlike content disseminated through traditional channels such as newspapers and broadcasts, social media content can be created, modified, and spread in a much less rigorous way. It can be published by a layperson without sufficient knowledge of a topic, modified, and even distorted during dissemination, ultimately leading to serious and undesirable consequences.

CONTACT Dongwon Lee  dongwon@ust.hk  School of Business and Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, HONG KONG

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

For instance, as reported by CNN,¹ a man in Phoenix, U.S., died of chloroquine phosphate poisoning after taking a product intended for cleaning fish tanks in the hope of recovering from COVID-19, after reading a post on social media advocating this as a treatment. Antonio Guterres, Secretary-General of the United Nations, alerted people to the “dangerous epidemic of fake news” on COVID-19 in the current situation and stressed that social media companies should take responsibility for tackling the spread of fake news.²

In line with this alert, social media platforms have implemented various interventions in recent years, including WhatsApp’s forwarding restriction to slow the spread of fake news,³ Sina Weibo’s launch of its Community Management Center to detect fake news by social reporting,⁴ and Facebook’s fact-checking teams that verify the factuality of news stories.⁵ Although these efforts to protect the credibility of information are recognized, the effectiveness of platform interventions remains unclear [2,41,57]. We believe that it is urgent and important to examine the effectiveness of platform interventions with empirical evidence. This study thus performs a series of analyses to evaluate the effectiveness of platform interventions to limit the spread of fake news. Specifically, we study the effectiveness of platform interventions in terms of fake news dissemination and survival. We further divide fake news dissemination into three patterns to better understand the more nuanced impacts of platform interventions.

Previous studies focused primarily on the content-level platform intervention, which applies to a single piece of information. A fake news flag is a good example of the content-level intervention; it attaches a label to a post to indicate that the post is fake news [44,45,49]. The results of previous studies on its effectiveness mainly focus on the cognitive level. Several studies have shown that flagging fake news can reduce its believability and sharing intentions [17,44]. However, other studies have found that a fake news flag can be ineffective due to confirmation bias [45] and people’s habit of disregarding warnings [55]. These seemingly inconsistent findings based on psychological outcomes motivate us to investigate the influence of a fake news flag on people’s actual behavior in a more generalizable setting. Many social media platforms, such as Twitter, do not prevent the spread of flagged fake news to ensure the practice of free speech, unless the harm caused by such fake news is extremely serious (i.e., a threat to national security).⁶ Besides, fake news may be continuously forwarded even after being flagged as it is most often more novel than real news [69]. Therefore, it is of great interest to understand how a fake news flag works in the real world. Instead of focusing on the psychological outcomes induced by a fake news flag, we take a different approach by using large-scale archival data collected from the field and exploiting a quasi-experiment to establish a causal relationship between a fake news flag and people’s sharing behaviors.

In addition to studying a fake news flag, we identify an important research gap in the relevant literature. As the impact of fake news has become increasingly serious,⁷ platforms have started to implement stricter regulations by imposing activity restrictions on accounts that publish fake news. We refer to this type of restriction as an account-level platform intervention. Unlike a fake news flag, which mitigates fake news by focusing on people’s cognitive processes [25], the restriction intervention directly controls the spread of fake news by limiting people’s engagement with fake news and inducing deterrence among accounts that intend to create and distribute fake news. However, there are concerns about the negative impacts of this intervention, as it may unintentionally restrict freedom of speech and block legitimate contents.⁸ It also takes time for the platform to discern the

legitimacy of a post [12]. Therefore, our work fills this research gap by empirically examining the effectiveness of the account-level intervention on fake news dissemination. As little is known about the impact of the account-level intervention, we also examine its effectiveness in shortening the survival time of fake news. Contrary to popular belief, fake news may not be overwhelmed by a huge amount of information online and disappear quickly (within days) [58,66]. If fake news can survive for an extended period, it is more likely to be spread through likes, sharing, comments, and, more importantly, reading. Exposure to fake news is dangerous, as people may take action without seeking the truth. Therefore, stopping the early spread of fake news is important to minimize its damage and negative social impact. In this regard, in addition to scholarly implications, we believe that understanding the impact of the account-level platform intervention on the survival time of fake news is of great importance for practice.

This study leverages two interventions implemented by Sina Weibo, the largest social media platform in China: a fake news flag as a content-level intervention and a forwarding restriction policy as an account-level intervention. To this end, we empirically examine how these two platform interventions affect fake news and answer the following two research questions:

1. *How does a content-level platform policy, i.e., fake news flags, affect fake news dissemination?*
2. *How does an account-level platform policy, i.e., forwarding restrictions, affect fake news dissemination and fake news survival?*

Using natural language processing and propensity score matching (PSM), we obtain a matched sample of fake news and truthful news to alleviate potential endogeneity issues for empirical analysis. We first study the impact of a fake news flag by using a difference-in-differences (DiD) approach. This specification helps us to identify a causal relationship between a fake news flag and fake news dissemination. We find that a post is distributed through more direct forwards than indirect forwards after being marked as “fake news.” Furthermore, a fake news flag encourages influential users to spread fake news posts to confirm its falsehood. Next, we estimate the impact of implementing a forwarding restriction policy by using the matched sample and controlling for the observable characteristics of the post and the user. Our results show that a forwarding restriction policy affects fake news and truthful news differently. Compared with truthful news, fake news is disseminated in a less centralized but more dispersed manner and has a significantly shorter survival time after the implementation of a forwarding restriction policy.

Overall, we find that a fake news flag and a forwarding restriction policy have different effects on fake news, with the former leading to more centralized and less dispersed dissemination of fake news and the latter yielding the opposite pattern. These results are not contradictory, as fake news flag and forwarding restriction policy are theorized as two different types of platform intervention. Therefore, their different influences on fake news dissemination are expected and can be explained by two mechanisms. The impact of a fake news flag on fake news is explained by the reduction of content ambiguity [35], which affects the weak ties of the fake news publishing account, whereas the impact of a forwarding restriction policy on fake news is explained by relational concerns arising from the strong ties [71]. In practice, these findings can inform social media platforms

about designing interventions to combat the spread of fake news. Although the account-level intervention seems to represent a “one-size-fits-all” policy, our results suggest that it does not affect the normal and desirable dissemination of truthful news.

This study contributes to the literature on platform interventions to combat fake news on social media. We provide empirical evidence based on field data of the causal impact of the content-level intervention (i.e., fake news flag) on fake news dissemination, which extends previous findings based on cognitive outcomes to actual behaviors by examining the practical importance of and capacity for flagging fake news to reduce its harm and social impact. We further investigate a less studied account-level intervention (i.e., forwarding restriction policy) and shed light on its effectiveness in mitigating the spread of fake news.

The rest of this paper is organized as follows. In the next section, we summarize the related literature and identify research gaps. In Section 3, we theorize the impacts of the two platform intervention policies on fake news. In Section 4, we introduce the research context and describe the data collected for the study. In Section 5, we propose our identification strategies. In Section 6, we present and discuss the research results. We discuss the contributions and limitations of this study in Section 7 and conclude our study in Section 8.

Related Literature

Fake News on Social Media

Fake news refers to news posts with deceptive intentions and false content [1,34]. Fake news also strongly overlaps with other deceptive information such as misinformation (false or misleading information) and disinformation (false information that is purposely spread to deceive people) [41]. As social media has changed the way news is created and consumed, such that people typically only read headlines or watch short videos,⁹ we define fake news in a broader sense as any information that is intentionally and verifiably false and could mislead readers.

The issue of fake news on social media has received much attention in previous studies [1,33,45,69], given its huge impact on politics, social crises, and other aspects of social life. One strand of the literature focuses on the empirical analysis of fake news dissemination, using descriptive analyses to examine dissemination patterns in terms of post and user characteristics [42,46,65,69]. For instance, Vosoughi et al. [69] found that fake news spreads farther, faster, deeper, and more broadly than truthful news across various topics, including politics, terrorism, and natural disasters. In addition to these static characteristics, previous studies have adopted a dynamic perspective to study fake news dissemination with informative results [32,65]. For example, Sutton et al. [64] explored how users’ follower-follower networks can influence the transmission of crisis information from a social network perspective. Tang and Ng [66] examined the forwarding behavior of users and found that more forwards are associated with a longer survival time of fake news on social media. The characteristics of fake news recipients have also been examined. For example, in the context of the 2016 U.S. presidential election, studies have shown that people who were older [21,23], politically conservative [21,23], and heavily involved in political news [21] were more likely to engage with fake news.

Another strand of the literature focuses on the psychological mechanisms or consequences of users exposed to fake news on social media. Several studies have posited the existence of confirmation bias, arguing that users tend to believe news that confirms their prior beliefs, regardless of the authenticity of its content [33,34]. When encountering information that does not align with their prior beliefs, individuals experience cognitive dissonance [24] and tend to resolve such dissonance by rejecting new information, as this often requires less effort than changing one's beliefs. Other mechanisms, such as fluency via prior exposure [50], laziness or lack of reasoning [51], and cognitive and affective engagement [42], have also been proposed to explain why people are susceptible to fake news. In terms of outcomes, previous studies have focused on perceived believability [33,34,45], engagement with the news (e.g., read, like, comment, and share) [33,34,43,45,47,49], and fact-checking behavior [68].

In summary, studies have investigated several aspects of fake news, including the characteristics of fake news content, publishers, and receivers; the mechanisms behind people's susceptibility to fake news; and individuals' attitudes and behavioral outcomes when exposed to fake news on social media. However, to the best of our knowledge, relatively few studies have used field data to investigate platform interventions aimed at changing people's behavior toward fake news.

Platform Interventions to Combat Fake News

Aside from understanding the phenomenon of fake news per se, previous studies have focused on platform interventions as mitigation strategies to detect [23,34] and stop [1,2,14,18,33] the spread of fake news. Most empirical studies of platform interventions, as summarized in Table 1, have focused on the content-level intervention, which only regulates one piece of information on social media. A fake news flag is a commonly studied content-level intervention, but the results of previous studies on its effectiveness are mixed.

For instance, Moravec et al. [44] showed that flagging fake news along with training on the meaning of the flag could significantly reduce the believability of fake news. They also showed that conducting flagging interventions to trigger subconscious processing (i.e., by displaying a visual "stop" sign when flagging fake news), deliberative reasoning (i.e., by displaying a text argument when flagging fake news), or a combination of these two approaches can effectively reduce the believability of fake news on social media. Garrett and Poulsen [17] reported that publishers' self-identified flags could reduce people's beliefs and sharing intentions regarding inaccurate messages. However, Moravec et al. [45] found that although a fake news flag can trigger increased cognitive activity in people, it cannot affect their judgments about the truth due to confirmation bias. In the same vein, Ross et al. [55] studied a fake news flag with additional manipulation (either a normal warning message indicating that the focal information was disputed by the third party or a negatively framed risk-handling advice) and found no significant effect of the flag on fake news. Considering the interaction between a fake news flag and the reputation of the information source, Figl et al. [14] found that although the flag may reduce the believability of fake news, this effect is weakened if the source of that fake news has a good reputation. Recently, Pennycook et al. [49] suggested that a fake news flag induces an implied truth effect so that unflagged fake news headlines are considered valid and more accurate by default.

Table 1. Existing Empirical Studies on Platform Intervention against Fake News

Reference	Platform Intervention	Dependent Variable	Data	Fake vs. Truthful News Comparison
This study	Content-level: fake news flag Account-level: forwarding restriction policy	Dissemination pattern and survival time	Field	Yes
Pennycook et al. [49]	Content-level: fake news flag	Accuracy judgment and social media sharing	Behavioral experiment	Yes
Figl et al. [14]	Content-level: fake news flag	News believability	Behavioral experiment	No
Kim and Dennis [33]	Content-level: highlighting source	Engagement with news (read, like, comment, and share)	Behavioral experiment	No
Kim et al. [34]	Content-level: source rating	Engagement with news (read, like, comment, and share)	Behavioral experiment	No
Moravec et al. [45]	Content-level: fake news flag	News believability	Behavioral experiment	Yes
Tang and Ng [66]	Community-level: launch of the social reporting system	Survival time	Field	No
Moravec et al. [44]	Content-level: fake news flag	News believability	Behavioral experiment	No
Ross et al. [55]	Content-level: fake news flag (warning message with/without risk-framed advice)	Number of hits and false alarms identified by a user	Behavioral experiment	No

The literature mainly addresses the effectiveness of a fake news flag based on cognitive and psychological outcomes, such as content believability and sharing intentions. To the best of our knowledge, little research has focused on changes in people's actual behavior in response to a fake news flag. Understanding this effect is crucial for fake news research, as the ultimate goal of any platform intervention is to stop the spread of fake news. In this regard, this study considers a more generalizable setting that exploits field data to investigate the effectiveness of flagging deceptive content by examining changes in people's sharing behavior. In particular, we aim to understand how a post is disseminated after being flagged as fake news.

In light of the huge impact of fake news on society, social media platforms have started to take a proactive approach by restricting the activities of accounts that publish deceptive information. This imposition of restrictions is considered an account-level intervention. Algorithms have been developed to detect and remove malicious and bot accounts created solely to spread fake news [59]. In addition, network-based methods have been proposed to stop the spread of fake news by identifying a set of accounts to monitor [59] or by controlling the flow of information through suspicious accounts [3]. Nevertheless, the effectiveness of this type of platform intervention has been less studied, as shown in Table 1. The account-level intervention is expected to be a better fake news mitigation strategy, as it not only regulates isolated fake news but also prevents the account publishing this fake news from performing activities such as forwarding posts or being followed by other accounts. This intervention should trigger inhibitory emotions such as fear and dread among accounts with the intent to deceive, effectively deterring them from creating and spreading fake news [52]. However, this account-level intervention may be detrimental to the freedom of speech and might inevitably hinder the normal circulation of credible information, i.e., truthful news. Therefore, it is theoretically and practically important to study the account-level intervention and its impact on the dissemination of fake and truthful news.

Hypothesis Development

Conceptualization of Dissemination Characteristics

The main objective of this study is to examine how content-level and account-level platform interventions affect fake news dissemination. We define “the dissemination of a post” as a directed network, with each node representing an account and each link representing a forwarding of the post by the account. We then divide the dissemination of posts into three patterns, namely *centrality*, *dispersibility*, and *influenceability*.

Centrality captures the centralized distribution of a post by counting its direct forwards. This pattern is commonly considered when studying information diffusion [18,21,25]. In our context, high centrality indicates that a post receives more direct forwards than indirect forwards. Dispersibility captures how far and deep a post is distributed in its dissemination network. Vosoughi et al. [69] captured this dissemination characteristic through structural virality [19] and documented that fake news spreads significantly farther, deeper, and more broadly than truthful news. In this study, we propose a similar but more accurate measure than structural virality to represent the dispersibility of fake news, in which a post with high dispersibility indicates that it spreads farther and deeper than a post with low dispersibility.

Finally, influenceability captures whether a post is widely disseminated to other accounts through a few direct forwards. The literature suggests that influential users help to facilitate the cascade and spread of information [11,72]. Therefore, a few direct forwards of a fake news post can also reach many other accounts if it is forwarded by influential users. We represent influenceability as the reach of a post that is distributed through influential users.

Our three proposed dissemination patterns correspond to basic and commonly used measures in social networks, namely degree centrality, closeness centrality, and eigenvector centrality [19,56,73]. These three measures use various concepts of social networks, such as degree [56], shortest path, interconnectedness [40], social influence [73], and power [15], to capture the main aspects of a post dissemination network.

Effect of the Content-Level Intervention: Fake News Flag

To identify the effect of a fake news flag on post dissemination patterns in terms of centrality and dispersibility, we first draw on social tie theory and define two types of social ties for an account: strong ties and weak ties [20]. Strong ties refer to proximate followers who can forward posts directly from the focal account, and weak ties refer to other users with more than one degree of separation from the focal account [71]. Due to the homophily of strong ties, followers are more likely to have the same views and beliefs as the focal account [22,38,48]. Therefore, the forwarding behavior of strong ties is not affected by a fake news flag due to confirmation bias [33,34,45].

In contrast, weak ties are distant followers who are less likely to have the same views and beliefs as the focal account. The forwarding behavior of weak ties is thus affected by a fake news flag that reduces the ambiguity of the post and eliminates the followers' need for information verification. Rumor theory suggests that ambiguity is an important factor that leads to fake news dissemination [35]. For example, Rosnow [54] proposed that uncertainty is a major predictor of rumor generation and transmission. Oh et al. [46] found that the ambiguity of the information source is a significant predictor of rumor dissemination in the context of a social crisis. Therefore, before a post is identified as fake news, its authenticity is ambiguous to its audience. As individuals experience a lack of reliable information in an ambiguous situation, they tend to engage in information seeking, sharing, and elaboration to resolve information uncertainty, incompleteness, or incongruence [32,46]. With a fake news flag, the ambiguity is lifted because the post is verified as fake news. In line with this reasoning, we predict that flagging fake news will reduce its ambiguity, preventing it from spreading farther and more broadly through weak ties. Therefore, we expect that a fake news flag leads to more centralized and less dispersed dissemination of fake news and propose the following hypotheses:

H1a: A fake news flag increases the centrality of the fake news dissemination network.

H1b: A fake news flag decreases the dispersibility of the fake news dissemination network.

Regarding the influenceability of the fake news dissemination network, influential users with many followers are expected to behave more cautiously to protect their authenticity, good reputation, and good public relations [4,13]. They tend to avoid disseminating an

ambiguous post until it has been verified but will forward verified fake news to help dispel it so that their followers are not fooled or confused by fake news posts. As a result, we expect influential users to be more likely to spread a post after it is flagged as fake news, as the ambiguity regarding its authenticity is removed. Accordingly, we propose the following hypothesis:

H1c: A fake news flag increases the influenceability of the fake news dissemination network.

Effect of the Account-Level Intervention: Forwarding Restriction Policy

Unlike the content-level intervention that targets post of questionable reliability, the account-level intervention targets malicious accounts and imposes severe punishment to combat the spread of fake news. It can be a very efficient strategy to fight the wave of fake news, as accounts are completely blocked from posting deceptive information. However, as mentioned above, this intervention can also cause fear and concern among legitimate accounts about publishing trustworthy content and may restrict freedom of speech and the spread of truthful news. Research has also suggested that blocking malicious accounts is problematic if the decision is not transparent and publicly assessable [41]. We thus aim to empirically investigate this less studied platform intervention for better policy design.

To explain the relationship between a forwarding restriction policy and fake news dissemination, we argue that the strong ties and weak ties of an account differ with respect to their relational aspect [71]. Specifically, compared with weak ties, strong ties have a high level of emotional closeness and a strong proximate interpersonal relationship with the focal account [63,71]. When an account publishes a post whose authenticity is uncertain, strong ties tend to avoid forwarding that post by considering that the account may be punished with an activity restriction. For weak ties with less relational consideration, their forwarding behavior is less likely to be affected by a forwarding restriction policy. Taken together, there will be fewer direct forwards made by the strong ties but relatively more indirect forwards made by the weak ties, leading to less centralized and more dispersed dissemination of fake news. Therefore, we propose the following hypotheses:

H2a: Fake news is disseminated in a less centralized manner after the implementation of a forwarding restriction policy.

H2b: Fake news is disseminated in a more dispersed manner after the implementation of a forwarding restriction policy.

In terms of influenceability, as discussed earlier, influential users' forwarding behavior tends to be largely affected by reputational concerns because they feel more accountable for their behavior in the presence of a large audience. In other words, influential users decide to forward a post by considering whether this forwarding will harm their reputation or not. Unlike a fake news flag, which clearly alleviates the problem of sharing fake news by reducing its ambiguity, a forwarding restriction policy does not affect the forwarding behavior of influential users in a predictable direction. On the one hand,

influential users may become more active in forwarding posts to protect the practice of free speech [5]. On the other hand, they may choose to share fewer posts to avoid forwarding fake news that would damage their reputation [4,13]. In line with this reasoning, we consider the effect of a forwarding restriction policy on the influenceability of fake news dissemination as an empirical question, and we do not formally propose a hypothesis here.

Using a forwarding restriction policy is an effective way to stop the spread of fake news, as it restricts the activities of the fake news publishing account instead of just warning others about fake news. Although our above hypotheses posit that fake news will be disseminated in a more dispersed manner with a forwarding restriction policy, the impact of fake news should be limited because the strong ties of the fake news publishing account are unlikely to forward fake news due to relational concerns. As a result, the number of fake news forwards will be significantly reduced, ultimately leading to the faster and earlier disappearance of the post. Therefore, we propose the following hypothesis:

H2c: Fake news has a shorter survival time after the implementation of a forwarding restriction policy.

The next section describes the empirical context and data to test the proposed hypotheses.

Empirical Context and Data

Sina Weibo and Sina Community Management Center

Sina Weibo is one of China's largest and most popular microblogging websites. Launched by Sina Corporation on August 14, 2009, Sina Weibo has grown dramatically and had over 497 million monthly active users in the third quarter of 2019.¹⁰ Faced with the growing threat of fake news, Sina Weibo launched a Community Management Center¹¹ in May 2012 to take advantage of the collective intelligence of community users to control the spread of fake news. The center relies on a social reporting system, through which users can report a post if they believe it to be harmful information (e.g., a threat to national security, misleading advertising, or obscene information), a message related to personal attacks, or fake news ("misinformation").¹² This report is posted publicly, with details including the reporting user's ID, reasons for reporting, reported post, and processing stage (e.g., stage of proof, judgment, and publicity).

According to Sina Weibo Community Management Regulations,¹³ a reported post will be accepted for validation only if 1) the post has been forwarded more than 100 times or 2) the post has been reported by more than 10 users. We believe that this rule validates our study, as we can avoid issues such as malicious and indiscriminate reporting. We focus on posts that are reported and verified as fake news and exclude those being reported as harmful information, as the latter will be directly assessed and removed by the platform and will not be allowed to be freely distributed. We also exclude all posts related to personal attacks, as they do not necessarily contain fake content.

Content-level and Account-level Platform Interventions on Sina Weibo

Using the launch of this Community Management Center, we focus on two interventions implemented by Sina Weibo: a content-level intervention, i.e., fake news flag, and an account-level intervention, i.e., forwarding restriction policy.

Sina Weibo introduced the fake news flag on May 28, 2012, alongside the launch of its Community Management Center. Fake news is flagged with a message stating that “this post is identified as fake” after being reported and verified. Along with this warning message, users can follow the hyperlink, which directs them to the web page of the Community Management Center. This page provides users with various information about the fake news post, including the reporting time, the reporter, the reported proof, and the assessment process. [Figure 1](#) shows a screenshot of this assessment page.

On August 30, 2013, Sina Weibo implemented a forwarding restriction policy based on a credit scoring system. The credit scoring system punishes users with score deductions for misbehavior, such as publishing fake news, and the number of deductions increases with the number of fake news forwards. As the number of forwards increases, an account’s credit score will be continuously reduced until it reaches a certain low value and the account’s activities are restricted, e.g., posts can no longer be forwarded by others. Before August 30, 2013, no form of punishment prevented users from engaging with accounts with low credit scores. However, after the implementation of the forwarding restriction policy (August 30, 2013), accounts with a credit score of fewer than 60 points are restricted automatically by preventing their posts from being forwarded by others, regardless of the content. Accounts with a credit score of fewer than 40 points are further restricted by hiding all their posts from their followers. To emphasize again, the forwarding restriction policy intervention takes effect at the account level and naturally influences all posts from restricted accounts, even if the content is truthful. As this platform intervention is unexpected or unpredictable, we consider it an exogenous shock to platform users and examine its impact on the dissemination of both fake and truthful news. [Figure 2](#) shows how the intervention works to prevent platform users from forwarding fake news.

This represents the reporter.

The icon displayed after the “V” sign means that the account has a credit score more than 60 points. Therefore, its activities are not restricted.

This represents the account being reported.

The icon displayed after the user name means that the account has a credit score of fewer than 60 points. Therefore, its activities are restricted (see Site Judgment below).

(translation below)

Site Judgment

After investigation, the post saying “there was a case of burglary in a building of the North Taiping house, Haidian district, Beijing” is a fake news. @Haidian Public Security Bureau has dispelled this fake news. The reported person constituted “publishing fake news”, with serious consequences. In accordance of item 19 of the “Detailed Rules for the Operation of Weibo Complaints”, the reported person will be handled as follows: 20 credit score will be deducted, posting activities will be prohibited for 30 days, being followed by other users will be prohibited for 30 days. The above operations will take effect within 60 minutes after the announcement.

Figure 1. Fake News Assessment Page from Sina Weibo

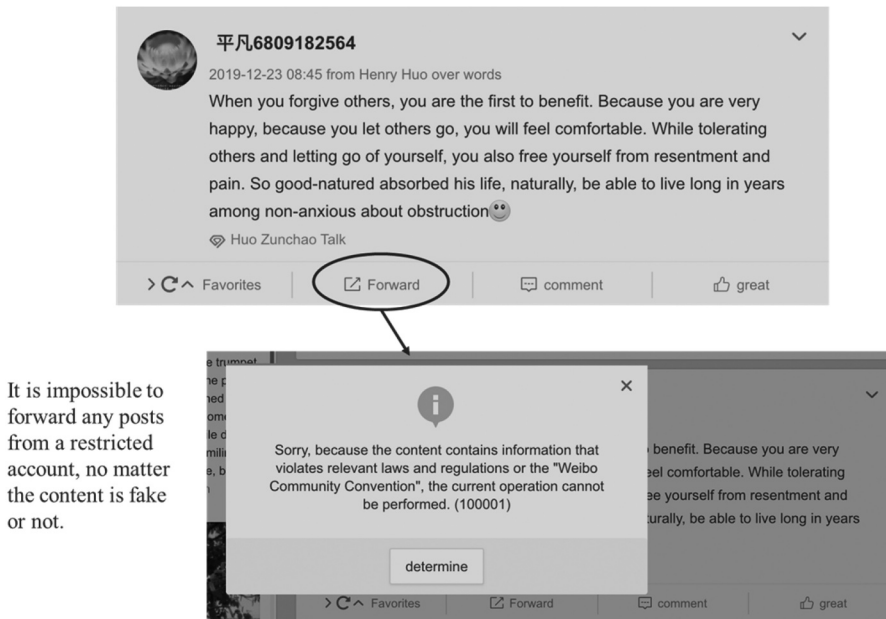


Figure 2. Forwarding Restriction Policy Intervention from Sina Weibo

Sina Weibo Datasets

To analyze the effects of the two platform interventions on fake news, we obtain our datasets from Sina Weibo through its open API.¹⁴ The Sina Weibo API provides a comprehensive interface to capture all relevant information of a post and its forwards. We restrict our sample to all posts published after June 2012, as the Community Management Center was officially launched on May 28, 2012. We then focus on a 2-year period, from June 2012 to May 2014. We identify a set of known fake news posts from the Community Management Center and only focus on posts reported as fake news by users. Notably, in addition to user reporting, Sina Weibo proactively identifies fake news either manually or by using machine learning algorithms. These fake news posts are likely to be identified quickly after publication due to sensitive keywords and images and then be deleted immediately. Therefore, they are unlikely to be disseminated and are not appropriate for our analysis of fake news dissemination pattern and survival. This proactive detection of fake news can also be seen as a form of censorship and is out of our scope. Therefore, we limit our scope to only fake news identified by the social reporting system rather than the platform. Based on our screening process, our dataset contains 1,514 fake news posts and all their forward/comment messages from 409,020 Weibo users. This sample of fake news posts covers various topics such as local news, international politics, and life-related news. In addition, we collect over 50,000 truthful news posts published during the same sample period.

For each post, we have information on the number of forwards, comments, likes received, and pictures included. We also know from which source the post is published (e.g., iPhone, website, or desktop app). In addition to post-specific information, we obtain

user-specific information, including gender, self-description, number of posts, number of followers, number of friends, verification status, location, and account age. For all fake news posts, we collect their reported date.

Operationalization of Variables

We study two platform interventions that are represented by two variables. For the fake news flag, we define *Marked* as a binary variable indicating the time before and after the flag, with a value of 1 when a post was flagged as fake news and 0 otherwise. For the forwarding restriction policy, we define *Restriction* as equal to 1 if a post was published after the implementation date of the restriction (August 30, 2013) and 0 otherwise. Of the 1,514 fake news posts, 137 were published after the implementation of the forwarding restriction policy.

For variables that capture the post dissemination patterns, two are directly adopted from the literature and one is adapted with minor modifications [40]. *Centrality* is measured by the standardized out-degree centrality score of a post in its dissemination network. *Influenceability* is measured by the eigenvector centrality score of a post in its dissemination network. *Dispersibility* is measured by taking the reciprocal of the standardized closeness centrality score of a post. As mentioned earlier, previous studies have used structural virality to capture the dispersibility of the spread of fake news [69]. This measure considers the average distance between all pairs of nodes in the dissemination network and is less accurate and reliable than our proposed measure that only considers the average distance between the focal node and all other nodes. When measuring the dispersibility of a post, we treat its dissemination network as an undirected network because forwarding is unidirectional, which causes a problem when calculating the shortest paths between nodes. Note that *Centrality* and *Dispersibility* are defined in a relative sense, as the two variables are based on standardized scores, although their rates of change are different. For example, centrality can decrease dramatically without a significant increase in dispersibility if indirect forwards occur within a few degrees of the focal account.

Following Tang and Ng [66], we operationalize fake news survival by two variables: *Discovery time* and *Stopping time*. *Discovery time* is measured (in minutes) by the time between the published time and the reported time. It captures how fast a post is identified and reported to the platform as fake news. *Stopping time* is measured (in minutes) by the time between the reported time and the time of the last reply (either a forward or comment), which serves as a proxy for fake news survival rather than an exact measure. As fake news has an effect when people read, comment, and forward it, an exact survival time should be measured by the time between the reported time and the time when no more users interact with that fake news post. However, it is impossible to identify whether any individual has read that fake news post or not. Thus, we consider the last observable user engagement as the length of time that fake news can survive after being verified and labeled.

Fake is a binary variable indicating whether a post is a fake news or not. Finally, to account for the heterogeneity of the post and user characteristics, we use a comprehensive set of control variables. The definitions of all variables used in this study are summarized in Appendix A. The summary statistics of the fake news variables are reported in Table 2.

Table 2. Summary Statistics of Variables for Fake News

Variable Name	Overall (n = 1,514)					Before Restriction (n = 1,377)					After Restriction (n = 137)				
	Mean	Std. Dev.	Max.	Min.	Mean	Std. Dev.	Max.	Min.	Mean	Std. Dev.	Max.	Min.			
Stopping Time (in min.)	115,731	216,957	1,714,810	0	124,395	223,064	1,714,810	0	28,653	108,750	1,006,050	0			
Discovery Time (in min.)	38,152	253,520	2,513,535	1	39,461	259,680	2,513,535	1	24,992	180,461	1,501,583	4			
Centrality	0.540	0.196	0.981	0.075	0.543	0.189	0.981	0.100	0.517	0.254	0.965	0.075			
Dispersibility	2.000	0.866	9.576	1.019	1.972	0.791	9.126	1.019	2.279	1.388	9.576	1.038			
Influencedability	0.004	0.002	0.014	0.001	0.004	0.002	0.014	0.001	0.004	0.003	0.012	0.001			
Forward	336	230	999	0	335	229	999	0	347	238	974	78			
Comment	92	127	2629	0	90	128	2629	0	115	116	577	0			
Like	11	41	605	0	5	16	605	0	68	111	605	0			
Picture	0.935	0.726	9	0	0.907	0.676	9	0	1.212	1.074	7	0			
Description	0.933	0.251	1	0	0.934	0.249	1	0	0.970	0.273	1	0			
Gender	0.631	0.483	1	0	0.630	0.483	1	0	0.635	0.483	1	0			
Message	17,664	30,933	358,663	0	17,837	31,792	358,663	0	15,925	20,389	107,017	0			
Follower	345,527	1,184,097	26,630,301	17	335,779	1,190,912	26,630,301	17	443,505	1,112,658	7,591,558	19			
Friend	1,044	893	4,981	0	1,043	886	4,981	0	1,055	965	4,745	17			
Account Age (in hr.)	15,945	7,808	40,329	1	15,344	7,383	34,854	1	21,984	9,291	40,329	912			
Length	110	44	193	3	111	44	193	3	94	48	172	9			
Number	0.606	0.489	1	0	0.614	0.487	1	0	0.518	0.502	1	0			

Empirical Methodology

We first use two matching strategies to alleviate potential selection bias and heterogeneity concerns to examine how the two platform interventions affect fake news dissemination.

Matching Strategies

Content-Based Matching: Latent Semantic Analysis

The first matching strategy is based on the idea of textual similarity, in which we match fake news and truthful news so that they are semantically similar. The basic logic is to quantify the news content in a numerical representation using natural language processing techniques and then apply similarity measures (e.g., cosine similarity and Euclidean distance) to infer semantic similarity between news posts. This approach has been implemented in various applications, such as collaborative filtering [29], incident risk factor identification [60], business proximity analysis [61], copycat detection [70], and customer agility measurement [74].

We start with truthful news posts. Due to the highly noisy dataset, a preliminary step is implemented to manually remove all posts that are (1) meaningless (with only emojis, numbers, or fewer than five words), (2) advertisements, and (3) forwarded posts. We remove all meaningless posts because they are not suitable for our content-based matching strategy. We ignore advertisements to avoid comparison with fake news posts, which are expected to be different from truthful news posts. Finally, we exclude forwarded posts because they are used to construct the post dissemination network. As a result, we obtain 23,535 truthful news posts, which are matched to our 1,514 fake news posts for further processing. We then tokenize all posts into a bag-of-words dictionary and remove all stop words and punctuation. Thus, each post is represented by a word vector, and each vector value indicates the frequency of a word occurring in the corresponding post. The term frequency-inverse document frequency (TF-IDF) technique is applied to all posts to normalize their corresponding word vectors. The result is a word-by-post matrix, with each row representing a post, each column representing a unique word, and each cell representing the TF-IDF value of the word in the post.

Next, we apply latent semantic analysis (LSA)¹⁵ to the word-by-post matrix to reduce the dimensionality and independency between words [39]. A dimensionality of 300 is chosen for LSA so that each word vector of a post is decomposed into a 300-dimensional feature vector. Finally, we match each fake news post to one or two truthful news posts based on the smallest angle calculated from the cosine similarity between their feature vectors. As a result, we obtain a matched sample of 1,586 truthful news posts and 1,514 fake news posts. The performance of the content-based matching method is reported in Appendix B1. Table 3 reports the summary statistics for truthful news posts after content-based matching.

Post-Based Matching: Propensity Score Matching

Based on the content-based matched sample, we implement a second matching strategy to control for post and user characteristics using the propensity score matching (PSM) approach [53]. This strategy helps us remove non-comparable fake and truthful news posts to minimize estimation bias arising from the post and user characteristics. One-to-

Table 3. Summary Statistics of Variables for Truthful News After Content-Based Matching

Variable Name	Overall (n = 1,586)					Before Restriction (n = 1,246)					After Restriction (n = 340)				
	Mean	Std. Dev.	Max.	Min.	0.028	Mean	Std. Dev.	Max.	Min.	0.028	Mean	Std. Dev.	Max.	Min.	
Centrality	0.667	0.233	1,000	0.028	0.674	0.674	0.229	1,000	0.028	0.644	0.246	1,000	0.030		
Dispersibility	1.583	0.818	16,746	1,000	1.561	0.651	8,727	8,727	1,000	1.665	1.250	16,746	1,000		
Influenceability	0.006	0.050	1,000	0.001	0.006	0.041	1,000	1,000	0.001	0.009	0.076	1,000	0.001		
Forward	463	250	999	13	449	251	999	999	13	514	242	982	13		
Comment	198	329	5,285	0	183	307	5,285	5,285	0	253	393	4,816	0		
Like	74	240	4,348	0	24	91	1,551	1,551	0	257	443	4,348	0		
Picture	1.067	1.187	9	0	0.885	0.551	9	9	0	1.732	2.216	9	0		
Description	0.970	0.170	1	0	0.967	0.178	1	1	0	0.982	0.132	1	0		
Gender	0.602	0.490	1	0	0.601	0.490	1	1	0	0.603	0.490	1	0		
Message	40,429	42,461	295,309	40	40,797	43,447	295,309	295,309	40	39,081	38,664	228,462	95		
Follower	5,814,669	10,140,082	50,910,636	0	5,544,103	10,021,889	50,910,636	50,910,636	132	6,806,215	10,517,658	48,251,168	0		
Friend	805	849	5,000	0	805	838	5,000	5,000	0	807	891	5,000	0		
Account Age (in hr.)	20,138	8,886	39,985	1	18,364	7,841	34,933	34,933	1	26,641	9,457	39,985	1,386		
Length	113	41	226	10	112	41	226	226	10	116	41	194	14		
Number	0.512	0.500	1	0	0.509	0.500	1	1	0	0.524	0.500	1	0		

one matching is implemented with a caliper size of 0.01. The matching procedures and performance assessment are reported in Appendix B2. With both matching strategies, we obtained a matched sample of 703 fake news posts and 703 truthful news posts for regression analysis.

Regression Analysis

We first use the fake news flag intervention implemented by Sina Weibo to establish the causal impact of the platform intervention on fake news dissemination. As the fake news flag is only applied to fake news but not truthful news, we can consider a quasi-experimental design based on a matched DiD sample to tease out unobservable characteristics of posts that may bias our estimation [10,37]. The matched sample of truthful news posts is then used as the quasi-control group to reveal the impact of flagging on fake news dissemination. The following model specification is estimated:

$$Dissemination_{it} = \beta_0 + \beta_1 Fake_i + \beta_2 Marked_t + \beta_3 Fake_i \times Marked_t + \varepsilon_{it}, \quad (1)$$

where $Dissemination_{it}$ is one of the three dissemination variables of post i measured up to time t , $Fake_i$ indicates whether post i is fake news, $Marked_t$ is coded as 0 if t is before the flagged time and 1 otherwise, and t represents the daily time range from three days before to three days after the flagged time. A major challenge is that fake news posts have different discovery and end times. For instance, one fake news post may be identified three days after its publication and stop spreading five days after, and another fake news post may be identified two months after its publication and stop spreading one month later. To overcome this challenge, we focus on the dissemination patterns three days before and three days after the flagged date of a fake news post, which corresponds to one week. To do so, we perform time matching between the matched fake news and truthful news pairs to align the unflagged period before the fake news post is identified. Figure 3 illustrates our identification strategy. Consider a pair of matched posts. The fake news post was published on April 12, 2013, and

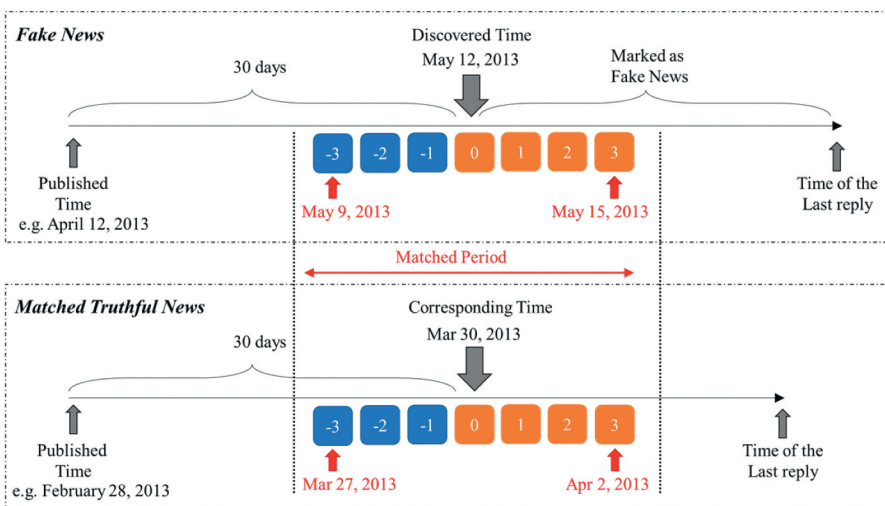


Figure 3. Illustration of the Identification Strategy for the Fake News Flag

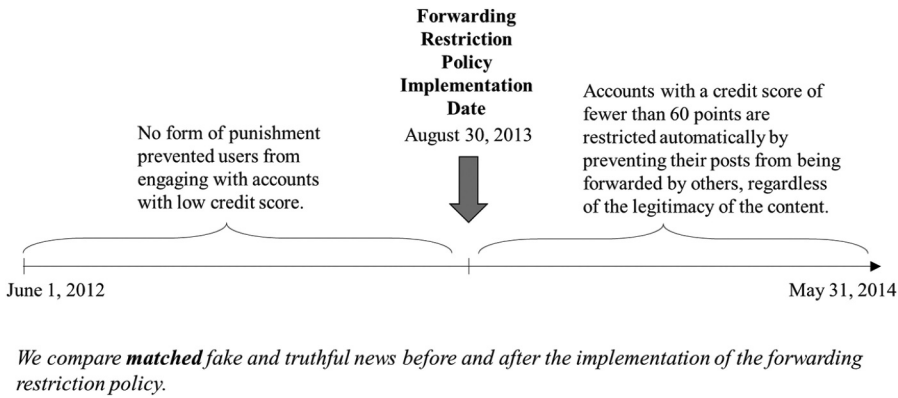


Figure 4. Illustration of the Identification Strategy for the Forwarding Restriction Policy

discovered on May 12, 2013, and the truthful news post was published on February 28, 2013. We compare the dissemination patterns of the fake news post three days before and after May 12, 2013. As it took 30 days for the fake news post to be discovered, we also use 30 days to identify a matched period of the matched truthful news post for comparison. From Figure 3, we subsample fake news posts between May 9, 2013, and May 15, 2013, and truthful news posts between March 27, 2013, and April 2, 2013, and specify a DiD strategy during this matched period. As some fake news posts only last for a very short time (e.g., less than a day), we exclude these and their matched truthful news posts, which leads to a matched sample of 1,014 pairs for our DiD analysis.

Our next model uses the forwarding restriction policy intervention implemented on August 30, 2013. As this intervention affects both fake and truthful news, we are interested in examining its different impacts on the dissemination of fake and truthful news. Our regression framework is specified below:

$$Dissemination_{it} = \beta_0 + \beta_1 Fake_i + \beta_2 Restriction_t + \beta_3 Fake_i \times Restriction_t + ControlVars_i + \gamma_t + \epsilon_{it}, \tag{2}$$

where $Dissemination_{it}$ is one of the three dissemination variables of post i published at time t , measured up to the post’s end time; $Restriction_t$ is coded as 0 if t is before the implementation date of the forwarding restriction policy and 1 otherwise; $ControlVars_i$ represents time-invariant post control variables; γ_t represents week fixed effects; and t represents the daily time range throughout our analysis period. The coefficient of the interaction term revealed by β_3 helps us to determine how fake news is disseminated after the forwarding restriction policy is implemented. We use our matched sample of 1,406 pairs obtained from the content-based and post-based matching approaches for this model specification. An illustration of this analysis framework is presented in Figure 4.

Finally, to examine the impact of the forwarding restriction policy on fake news survival, we specify two regression frameworks for a more comprehensive analysis. The first regression is an ordinary least squares (OLS) regression specified below:

$$Time_{it} = \beta_0 + \beta_1 Restriction_t + ControlVars_i + \gamma_t + \epsilon_{it}, \tag{3}$$

where $Time_{it}$ represents either the logarithm of *Discovery time* or *Stopping time* of fake news i measured at time t , and t represents the daily time range throughout our analysis period. The advantage of this framework is that we can incorporate time fixed effects γ_t to account for potential time-induced and trend effects. The second regression is a Cox proportional-hazards framework commonly used to investigate how multiple covariates are simultaneously related to survival time:

$$h(Time_{it}) = h_0(Time_{it}) \times \exp(\beta_1 Restriction_t + ControlVars_i), \tag{4}$$

where $h(\cdot)$ is the hazard function and $h_0(\cdot)$ the baseline hazard function with all variables set to 0. This framework is a natural choice because our variable of interest is the survival time of fake news. However, the two limitations of this framework are the lack of data censoring,

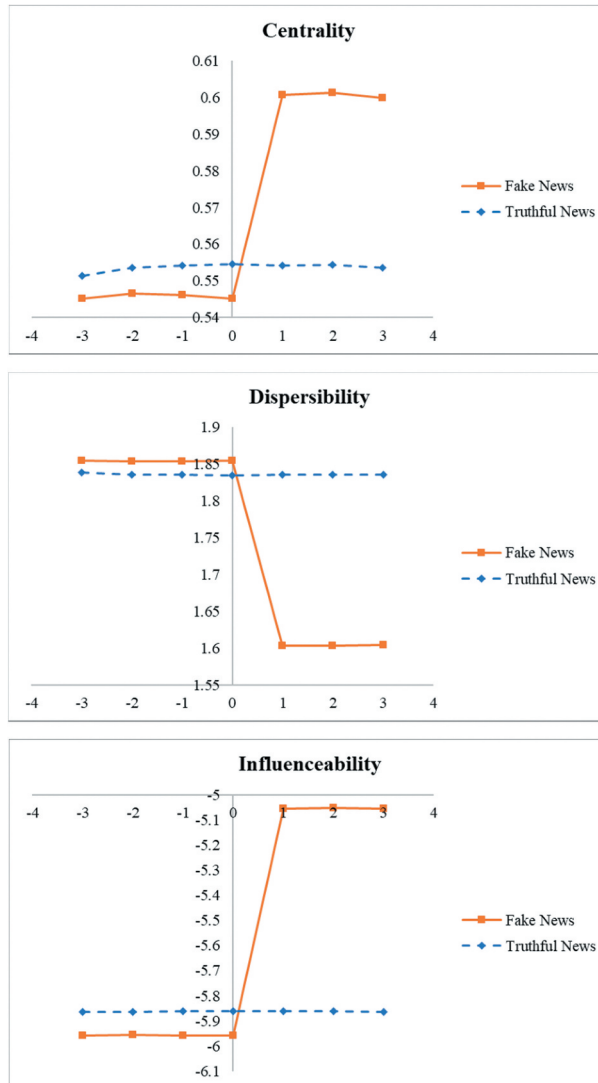


Figure 5. Model-Free Evidence of the Effect of the Fake News Flag on Fake News Dissemination

Table 4. Effect of the Fake News Flag on Fake News Dissemination

	Centrality	Dispersibility	Influenceability
<i>Fake</i>	-0.007 (0.005)	0.017 (0.030)	-0.093*** (0.030)
<i>Marked</i>	0.000 (0.007)	0.000 (0.032)	0.000 (0.033)
<i>Fake × Marked</i>	0.055*** (0.012)	-0.250*** (0.045)	0.903*** (0.012)
Log Likelihood	68	-9,639	-9,788
Observations	1,014	1,014	1,014

Notes. *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Standard errors are displayed in parentheses under the coefficient estimates.

as all fake news posts are eventually identified, and the proportional hazards assumption, which precludes the incorporation of time fixed effects. Therefore, we implement both frameworks to complement each other's limitations for a robust analysis of the impact of this platform intervention.

Results

Effect of the Fake News Flag on Fake News Dissemination

We present the results of our main analyses in this section. We first assess the parallel trend assumption in Figure 5. Figure 5 shows a sharp change in centrality, dispersibility, and influenceability after the intervention, which provides us with model-free evidence of the effect of the fake news flag. We also analyze pre-treatment trends to verify the parallel trend assumption and report the results in Appendix C. Table 4 reports the DiD regression results. We find that the intervention has a positive and significant impact on the centrality ($\beta = 0.055, p < 0.01$) of fake news, which means that after a fake news post is flagged as such, it is more likely to be disseminated through direct forwards (H1a is supported). We also find that the flag has a significant and negative effect on the dispersibility ($\beta = -0.250, p < 0.01$) of fake news, which means that fake news is less likely to be forwarded to distant others after the intervention (H1b is supported). Moreover, the intervention has a significant and positive impact on the influenceability ($\beta = 0.903, p < 0.01$) of fake news. In other words, the flag encourages influential users to forward fake news posts (H1c is supported).

To identify the underlying mechanism, we test whether the ambiguity of fake news decreases after being flagged as such. We perform a content analysis on all fake news forwarding comments. We use a popular psycholinguistic dictionary, the Linguistic Inquiry and Word Count (LIWC) constructed by Tausczik and Pennebaker [67], to infer the extent of ambiguity for each fake news forwarding comment [27, 28]. We consider the two most

Table 5. Ambiguity Analysis

Condition	Tentative Word Usage	Affective Word Usage
Before fake news flag	0.137 (0.401)	7.282 (5.595)
After fake news flag	0.100 (0.339)	5.036 (5.456)
Within-subject t-test	2.671***	13.048***
Observations	1,250	1,250

Notes. *** indicates statistical significance at the 1% level. Mean values are displayed with standard errors in parentheses.



Table 6. Examples of Fake News Forwarded by an Influencer

Original Fake News	Forwarding Comment from the Influencer
<p>In summer, watermelon is the top choice of refreshing food, but malicious merchants seek to make a profit from unripe watermelons. They inject the banned food additives, cyclamate and carmine! The injected watermelon is red and juicy but tasteless. The additives used contain toxic substances that destroy the function of the liver and kidneys and affect the intellectual development of children!</p> <p>China's Bone Marrow Bank, sponsored by the Red Cross Society of China, calls for bone marrow donations in the name of charity and creates a bank of samples. If there are matching patients, they will contact volunteers to donate bone marrow. How wonderful! However, the patient will be charged 500 yuan for each inquiry and at least 50,000 yuan for obtaining bone marrow! Therefore, Peking University students were outraged and set up a private bone marrow bank. From inquiry to transplantation, the process is free, and the bank is called "Sunshine Marrow Bank."</p> <p>Although I am not a celebrity, I notice that there are more and more bad people in this society, and I hope to do my best to get your attention. This child has just been abducted from RT-Mart Supermarket in Xiaolan. I hope more people can follow and forward this post.</p>	<p>To the blogger and the latest forwarder, couldn't you use your brain? It is so annoying that you post fake news every year. Please be conscientious when posting on Weibo to make money.</p> <p>China's Bone Marrow Bank search is free: 500 yuan is the HLA typing test fee for each blood sample in the bank, which is currently funded by the financial department, and the Bone Marrow Bank has never charged it. The number "50,000 yuan" is also wrong. Due to the international principle that donors should not meet patients, we collect 20,000 yuan from patients to be used for transportation, collection, and other donor expenses, and any overpayment will be refunded and any deficit will be repaid. The Bone Marrow Bank has never kept any money. China's Bone Marrow Bank will hold an opening day in the morning of May 8. Welcome on-site consultation.</p> <p>After due diligence, the information source is identified: user @storm0109 reported at around 8 a.m. that the daughter of his/her friend's coworker was abducted near RT-Mart Supermarket, Plaza South Road, Nanchang City, Jiangxi Province. Some certified users in Jiangxi Province forwarded this post. At around 12 p.m., the child was found. All information was provided by this single user, and its factuality could not be verified. However, Shinnosuke Nohara is to be blamed for posting this fake news.</p>

relevant word categories: *Tentative* and *Affect*. We expect the reduced ambiguity of fake news after being flagged to be reflected by the reduced use of tentative words, such as “maybe,” “perhaps,” and “guess,” in forwarding comments. We also expect a reduced use of affective words in forwarding comments, as ambiguity is strongly associated with various emotions such as anxiety and worry [9,16]. We only focus on fake news with forwarding comments containing at least one tentative or affective word, resulting in a sample of 1,250 fake news posts for analysis. We then perform a within-subjects *t*-test to statistically compare the proportions of tentative or affective words used in fake news forwarding comments before and after the implementation of the fake news flag. Table 5 shows a significant decrease in the use of tentative and affective words in forwarding comments after a post is flagged as fake news, which supports our conjecture that the fake news flag can reduce the ambiguity of fake news.

Table 7. Effect of the Forwarding Restriction Policy on Fake News Dissemination

	Centrality		Dispersibility		Influenceability	
<i>Fake</i>	0.042 (0.029)	0.108 (0.103)	-0.182 (0.119)	-0.450 (0.424)	-0.230*** (0.068)	0.051 (0.240)
<i>Restriction</i>	-0.058 (0.090)	-0.069 (0.101)	0.356 (0.366)	0.248 (0.413)	-0.298 (0.211)	-0.259 (0.233)
<i>Fake × Restriction</i>		-0.338* (0.179)		1.471** (0.736)		-0.102 (0.416)
<i>ln(Forward)</i>	-0.102*** (0.008)	-0.098*** (0.008)	0.274*** (0.034)	0.259*** (0.035)	-0.879*** (0.020)	-0.894*** (0.020)
<i>ln(Comment)</i>	0.068*** (0.006)	0.067*** (0.006)	-0.104*** (0.026)	-0.108*** (0.026)	-0.095*** (0.015)	-0.095*** (0.015)
<i>ln(Like)</i>	0.005 (0.006)	0.009 (0.006)	-0.084*** (0.022)	-0.075*** (0.023)	-0.041*** (0.013)	-0.021 (0.013)
<i>ln(Picture)</i>	-0.010 (0.023)	-0.012 (0.023)	-0.201** (0.094)	-0.143 (0.093)	0.220*** (0.054)	0.252*** (0.053)
<i>Description</i>	0.018 (0.025)	0.002 (0.025)	-0.206** (0.102)	-0.160 (0.101)	-0.055 (0.059)	-0.050 (0.057)
<i>Gender</i>	0.034*** (0.011)	0.033*** (0.010)	-0.079* (0.043)	-0.089** (0.042)	0.009 (0.025)	0.015 (0.024)
<i>ln(Message)</i>	-0.018*** (0.004)	-0.012*** (0.004)	0.115*** (0.017)	0.122*** (0.018)	-0.043*** (0.010)	-0.037*** (0.010)
<i>ln(Follower)</i>	0.039*** (0.003)	0.041*** (0.003)	-0.152*** (0.012)	-0.155*** (0.012)	0.042*** (0.007)	0.034*** (0.007)
<i>ln(Friend)</i>	-0.020*** (0.004)	-0.017*** (0.004)	0.036** (0.018)	0.030* (0.018)	-0.009 (0.010)	0.000 (0.010)
<i>ln(Account Age)</i>	0.001 (0.007)	-0.006 (0.007)	-0.041 (0.029)	-0.027 (0.030)	-0.011 (0.017)	-0.024 (0.017)
<i>ln(Length)</i>	0.002 (0.009)	0.009 (0.008)	-0.004 (0.035)	-0.026 (0.035)	0.015 (0.020)	0.008 (0.020)
<i>Number</i>	0.003 (0.011)	0.005 (0.011)	-0.068 (0.044)	-0.061 (0.044)	0.035 (0.026)	0.055** (0.025)
Verified Type	Yes	Yes	Yes	Yes	Yes	Yes
Location	Yes	Yes	Yes	Yes	Yes	Yes
Source	Yes	Yes	Yes	Yes	Yes	Yes
Week Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Log Likelihood	662	761	-1,305	-1,223	-529	-420
Observations	1,406	1,406	1,406	1,406	1,406	1,406

Notes. 1. *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Standard errors are displayed in parentheses under the coefficient estimates.

2. For fake news, the sample sizes before and after restriction are 623 and 80, respectively. For truthful news, the sample sizes before and after restriction are 586 and 117, respectively.

In summary, we find that the fake news flag leads the fake news dissemination network to be more centralized through direct forwards than dispersed through indirect forwards. These results confirm our theorization based on the basic law of rumor [35] that flagging fake news can reduce its ambiguity and prevent it from being disseminated farther and deeper. In terms of the effect of the fake news flag on fake news influenceability, influential users with a large number of followers are expected to behave more cautiously. Therefore, they tend to avoid disseminating an ambiguous post until it has been verified. The increased influenceability of fake news after the implementation of the fake news flag highlights the effort of influential users to help dispel fake news. Table 6 qualitatively supports this claim, as we find that influential users forward fake news posts with an intention to confirm their falsehood. In short, the fake news flag significantly alters the spread of fake news by generating more direct than indirect forwards, causing fake news posts to be disseminated more by influential users.

Effect of the Forwarding Restriction Policy on Fake News Dissemination

We now examine the impact of the forwarding restriction policy on post dissemination patterns. We first provide a descriptive analysis to understand how the forwarding restriction policy shapes the post dissemination network. We randomly select 24 fake news and 24 truthful news dissemination networks before and after the intervention for visual comparison, as shown in Appendix D. Before the intervention, the fake news dissemination networks were more centralized, and only a relatively small portion spread to distant accounts. A similar pattern can be observed for truthful news, with their dissemination network showing high centrality but low dispersibility. In general, both fake news and truthful news dissemination networks involved some degree of influenceability. However, after the implementation of the forwarding restriction policy, a clear distinction can be observed: the fake news dissemination networks become more dispersed with longer tails (i.e., more indirect forwards) and less centralized (i.e., fewer direct forwards). We observe no significant change in the truthful news dissemination networks from before to after the intervention. In short, the visualization of these networks is consistent with our expectation that the forwarding restriction policy influences fake news and truthful news differently.

We present the regression results in Table 7. The significant and negative coefficient of *Fake* \times *Restriction* on the centrality ($\beta = -0.338$, $p < 0.1$) of fake news indicates that compared with truthful news, fake news is disseminated via less direct forwards after the implementation of the forwarding restriction policy (H2a is supported). However, the significant and positive coefficient of *Fake* \times *Restriction* on the dispersibility ($\beta = 1.471$, $p < 0.05$) of fake news suggests that compared with truthful news, fake news is disseminated in a more dispersed manner after the intervention (H2b is supported). We find no significant effect ($\beta = -0.102$, $p > 0.1$) of the forwarding restriction policy on the influenceability of fake news. Taken together, the forwarding restriction policy leads to significantly less centralized and more dispersed dissemination of fake news as compared with truthful news.

The above results are consistent with our theorization based on the social tie theory. After the implementation of the forwarding restriction policy, the strong ties of fake news publishing accounts are prevented from forwarding fake news posts by relational concerns,

whereas their weak ties are not affected by relational concerns and continue to forward fake news posts. Thus, the dissemination of fake news becomes less centralized and more dispersed than truthful news.

Robustness Checks

We run two additional tests to check whether the impact of the forwarding restriction policy on fake news dissemination changes based on the topic and sentiment. The results are reported in Appendix E. We find that the forwarding restriction policy only affects life-related fake news posts but does not affect international and local fake news posts (Table E2). Relational concerns are not likely to occur among the strong ties of fake news publishing accounts that publish international and local fake news, as these types of posts have more effect on society than life-related fake news posts. This finding indirectly supports our proposed mechanism of relational concerns arising from strong ties. In addition, we find that fake news sentiment does not affect the dissemination patterns of fake news, which confirms that the identified effect is not driven by the emotions expressed in the content of fake news (Table E4).

Another major concern regarding the dissemination of fake news is whether social bot accounts spread fake news on social media on a large scale [59]. Previous research has shown that social bots play a significant role in promoting fake news that distorts various social events [6], such as the 2016 U.S. election. However, this is not a concern in our study due to the real-name policy implemented by Sina Weibo since 2011.¹⁶ The policy states that Weibo users must verify their accounts by using their real names for account registration. This policy was formally launched on March 3, 2012, before our analysis period. According to this real-name policy, accounts registered without real-name verification can only read posts on Weibo but cannot publish, comment, or forward posts. We believe that the real-name policy alleviates the concern that our findings might be contaminated by social bots.

In summary, this study reveals that the fake news flag and the forwarding restriction policy have different effects on fake news dissemination. In a broader sense, the fake news flag (content-level intervention) affects the spread of fake news by reducing its ambiguity, leading to more centralized and less dispersed dissemination of fake news and more forwards by influential users. The forwarding restriction policy (account-level intervention) creates relational concerns among the strong ties of fake news publishing accounts, leading to less centralized and more dispersed dissemination of fake news. These findings provide insights into the effectiveness of different types of platform interventions in fake news mitigation.

Effect of the Forwarding Restriction Policy on Fake News Survival

As little research focuses on the impact of account-level interventions (e.g., forwarding restriction policy), we conducted an additional analysis to examine how the forwarding restriction policy influences fake news survival, which is a more direct way to examine the effectiveness of this type of intervention. We report the results in Table 8. As the lifespans of some fake news posts may overlap with the intervention implementation date, we remove these posts for a more robust analysis, which reduces our sample to 1,368 fake news posts. We obtain consistent results across the two regression models. For the linear regression model, the significant and negative coefficient of *Restriction* on *Stopping time* ($\beta = -8.779$,

Table 8. Effect of the Forwarding Restriction Policy on Fake News Survival

	OLS		Cox Proportional-Hazards Model	
	<i>log</i> (Stopping Time)	<i>log</i> (Discovery Time)	Stopping Hazard	Discovery Hazard
<i>Restriction</i>	-8.779*** (2.848)	-0.673 (2.341)	0.899*** (0.154)	0.013 (0.122)
<i>log</i> (Forward)	0.817*** (0.142)	-0.010 (0.116)	-0.427*** (0.052)	-0.098* (0.052)
<i>log</i> (Comment)	-0.175* (0.099)	-0.106 (0.081)	0.037 (0.039)	0.051 (0.039)
<i>log</i> (Like)	0.148 (0.094)	0.013 (0.077)	0.082** (0.037)	0.081** (0.032)
<i>log</i> (Picture)	0.042 (0.436)	0.267 (0.359)	0.651*** (0.170)	-0.216 (0.146)
<i>Description</i>	-0.359 (0.339)	0.355 (0.278)	-0.108 (0.142)	-0.085 (0.120)
<i>Gender</i>	-0.311* (0.167)	-0.011 (0.138)	0.130* (0.072)	-0.002 (0.063)
<i>log</i> (Message)	-0.147*** (0.056)	-0.001 (0.046)	0.100*** (0.024)	-0.009 (0.020)
<i>log</i> (Follower)	-0.058 (0.050)	0.033 (0.041)	-0.061*** (0.021)	0.001 (0.017)
<i>log</i> (Friend)	0.068 (0.070)	-0.083 (0.057)	-0.019 (0.031)	0.023 (0.025)
<i>log</i> (Account Age)	0.157 (0.112)	-0.166* (0.092)	0.089* (0.048)	0.050 (0.039)
<i>log</i> (Length)	-0.073 (0.128)	0.279*** (0.106)	-0.077 (0.052)	-0.106** (0.045)
<i>Number</i>	0.005 (0.179)	-0.275* (0.147)	0.137* (0.077)	0.053 (0.066)
Verified Type	Yes	Yes	Yes	Yes
Location	Yes	Yes	Yes	Yes
Source	Yes	Yes	Yes	Yes
Week Fixed Effect	Yes	Yes	No	No
Log Likelihood	-3,111	-2,843	-8,244	-8,535
Observations	1,368	1,368	1,368	1,368

Notes. *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. Standard errors are displayed in parentheses under the coefficient estimates.

$p < 0.01$) indicates that fake news has a shorter average survival time after the intervention. In the Cox proportional-hazards model analysis, the positive and significant coefficient of *Restriction* on *Stopping time* ($\beta = 0.899$, $p < 0.01$) suggests a higher hazard rate of fake news after the intervention. Specifically, the expected hazard of fake news increases by 146% after the intervention, compared with that before the intervention. We find no evidence of the impact of the forwarding restriction policy on the discovery time of fake news. In all cases, we examine the Kaplan-Meier curves of the Cox proportional-hazards models and find no violation of the proportional hazard assumption [26].

We find that the forwarding restriction policy effectively combats the spread of fake news by shortening its lifespan because relational concerns of strong ties prevent the further spread of fake news. However, the forwarding restriction policy has no effect on the discovery time of fake news, likely because there are no relational concerns before the verification of fake news. We also conduct a placebo test by repeating the same regression analyses on truthful news as a robustness check and find that the forwarding restriction policy has no effect on the lifespan of truthful news. The placebo test is presented in Appendix F.

Overall, after the implementation of the forwarding restriction policy, 1) fake news is disseminated in a less centralized and more dispersed manner, compared with truthful news, and 2) fake news survives for a much shorter time but is not discovered sooner. These findings have several implications for social media platforms, as they show how the account-level intervention affects user engagement with fake news and suggest that this type of platform intervention can effectively mitigate fake news by shortening its survival time.

Discussion

Contributions

The study makes two main contributions to the literature. First, we extend the fake news literature by establishing a causal relationship between the fake news flag and fake news dissemination using empirical field data. To the best of our knowledge, previous studies have mainly focused on the cognitive and psychological impacts of the fake news flag by using experimental laboratory data, and few have addressed the actual behavior change caused by this flag. Thus, we complement the literature by using field data and providing causal empirical evidence to better understand the effectiveness of the fake news flag.

Second, we extend the literature on platform interventions that target fake news by examining the impact of the forwarding restriction policy on fake news dissemination and survival time. Previous studies have mainly focused on the content-level intervention (i.e., fake news flag) from a cognitive perspective to reduce people's willingness to engage with fake news [25]. We find that the forwarding restriction policy, a type of stricter platform intervention applied at the account level, can shorten the survival time of fake news, with no detrimental effect on the dissemination of truthful news. Furthermore, we highlight the different impacts of the two interventions on fake news dissemination, thereby solving a conundrum that has never been adequately addressed. This missing piece of the puzzle advances our understanding of the effectiveness of platform interventions by providing a holistic view of how content-level and account-level interventions work differently to combat the spread of fake news.

From a practical perspective, our findings provide important insights for online platform owners and policymakers. Today, the Internet and social media have become the primary sources of information consumption for most people. Online platforms are important mediators that ensure the quality of information to prevent consumers from exposure to misleading and manipulative news. However, the effectiveness and efficacy of platform interventions are questionable [41], and more effort should be devoted to better understanding how different policies work. Our study responds to this call by empirically examining the effects of two platform interventions in a rigorous framework. Our analysis of the fake news flag provides a result consistent with that of Moravec et al. [45], indicating that people continue to forward fake news posts even after they are identified as fake news. Our proposed mechanism is supported, showing that the fake news flag reduces the ambiguity of fake news posts and, therefore, helps control the spread of fake news within a smaller network, i.e., direct forwards within one degree of separation. However, our findings also suggest a possible consequence of the increased echo chamber, as the flagged fake news posts are likely to be forwarded more by like-minded people [38], adversely

reinforcing polarization and fueling extreme emotions within the online user community. In brief, the fake news flag may have both positive and potential negative effects on controlling the spread of fake news. Contrary to the concern that the account-level intervention may interfere with the spread of truthful news due to its strict “one-size-fits-all” approach, our results reveal different effects for fake and truthful news. Although we find that fake news spreads much farther and more broadly after the implementation of the forwarding restriction policy, it stops the spread of fake news much more quickly. Altogether, these findings have important implications for online platforms in designing interventions to mitigate the spread of fake news.

For policymakers, this study suggests that platform interventions on social media are an effective way to combat the spread of fake news. Specifically, the two platform interventions have different objectives. The fake news flag is effective because it can prevent fake news from spreading farther and deeper within the dissemination network. In addition, as the fake news flag encourages influential users to spread fake news posts, we recommend that the platform develop an effective policy to motivate influential users to dispel fake news. This study reveals that the account-level intervention can effectively mitigate fake news by shortening its survival time. As this intervention only affects the dissemination of fake news but not truthful news, platform owners may consider implementing this account-level intervention to stop the spread of fake news quickly.

Limitations and Future Research

We acknowledge that this paper has several limitations. Measuring the fake news survival time may be of concern, as users may remember a post after several years and forward or comment on it, which may potentially compromise our survival time measure. We address this concern with two points. First, we ensure that all forwards and comments collected from the posts cover the period up to 2018, so any forward or comment after a sufficiently long period is unlikely to happen. Second, the mean and standard deviation of the number of forwards are very similar before and after the implementation of the forwarding restriction policy (Before: $M = 335$, $SD = 229$; After: $M = 347$, $SD = 238$).¹⁷ Thus, we believe that the effect of the forwarding restriction policy on the survival time of fake news is unlikely to be affected by forwards or comments omitted during data collection.

Another limitation is that this study only investigates one social media platform in China. Although Sina Weibo offers the opportunity to investigate the effectiveness of both platform interventions, the findings of this study may not be widely generalizable to other cultures, e.g., the United States. Referring to the influential theory of cultural dimensions [30, 31], we suggest that two dimensions should be taken into consideration when applying our findings to other cultures. One dimension is power distance, which is defined as “the extent to which the members of a society accept that power in institutions and organizations is distributed unequally” [30]. Brockner et al. [8] showed that people tend to react unfavorably when they have little voice in a decision-making process, but this tendency is weaker for people in high power distance cultures (e.g., China) than in low power distance cultures (e.g., the United States). Therefore, Chinese people might be more willing to accept a fake news flag verified by a reliable or mainstream information source. In contrast, people in low power distance cultures may not react favorably to the fake news flag, as they have little say in investigating and claiming the authenticity of a post.

The second cultural dimension is individualism/collectivism, which is defined as the extent to which members of a society emphasize their own needs over those of the group and tend to act as individuals rather than as members of a group [30]. It has been argued that members of a collectivist culture like China value harmony and group consensus more than freedom of expression, compared with members of an individualistic culture like the United States [36]. Hence, social media users in individualistic cultures may be more tolerant of extreme, irrational, and harmful posts and may be less likely to spontaneously report such posts.

Our results show that the fake news flag leads to more centralized and less dispersed dissemination of fake news. We argue that this observation can be explained by the reduced ambiguity of fake news content and provide empirical evidence to support this proposed mechanism. However, we note that the more centralized fake news dissemination network can also be explained by alternative theories such as the echo chamber [38] or social bot sharing [7]. Social bot sharing is less likely to be a concern due to the above-mentioned real-name policy implemented by Sina Weibo. Regarding the echo chamber, like-minded individuals may be more likely to forward fake news posts after flagging. Multiple mechanisms may occur simultaneously to explain why individuals continue to engage with fake news after it is flagged as such, and our study proposes and empirically tests one mechanism, reduced ambiguity. Therefore, future research could conduct a more in-depth investigation to determine how different mechanisms interact to strengthen our study results.

Our study offers several research opportunities for future studies. First, our sample of fake news posts relies on a social reporting system operated through the collective intelligence of the crowd, but social media platforms are increasingly using machine learning-based fake news detection systems. Future research could compare the effectiveness of these two types of systems in terms of fake news mitigation. Second, our study only focuses on a simple type of fake news flag. More empirical studies should be conducted to examine a fake news flag with various manipulations, such as a strong warning or a high level of severity, which is expected to have a more salient deterrent effect on fake news dissemination. Third, as the forwarding restriction policy depends on a credit scoring system, future research could explore the impact of the credit score on users' fake news posting and forwarding behavior and study its interaction with the platform intervention. Fourth, future research could investigate the dynamic change in the survival time of fake news. An interesting research question would be to study how two fake news posts with the same survival time differ if one is disseminated with an initial surge followed by a decline and the other is disseminated with a slow start followed by a huge surge.

Conclusions

This study exploits fake news data and two types of platform interventions in Sina Weibo to study the impact of platform interventions on fake news dissemination and survival. First, we exploit a natural experiment using a DiD approach to identify the causal relationship between the fake news flag (content-level intervention) and three fake news dissemination patterns. We show that after a post is flagged as fake news, its dissemination network instantly becomes more centralized and less dispersed. Furthermore, fake news is more likely to be spread by influential users. We attribute these findings to the reduced ambiguity of fake news [35]. Second, we investigate how the forwarding restriction policy (account-

level intervention) influences the spread of fake and truthful news. We show that the implementation of the forwarding restriction policy leads to less direct and more indirect forwards of fake news, compared with truthful news. This phenomenon can be explained by social tie theory [20], as relational concerns prevent the strong ties of fake news publishing accounts from spreading fake news but have no effect on the weak ties. We also show that the forwarding restriction policy shortens the survival time of fake news. This study is among the first to provide empirical evidence of the effectiveness of platform interventions in combating the spread of fake news. Thus, our study constitutes an early effort, and we hope that our work will shed light on subsequent understandings of platform interventions and fake news dissemination.

Notes

1. <https://edition.cnn.com/2020/03/23/health/arizona-coronavirus-chloroquine-death/index.html>
2. <https://news.un.org/en/story/2020/04/1061682>
3. <https://www.theguardian.com/technology/2020/apr/07/whatsapp-to-impose-new-limit-on-forwarding-to-fight-fake-news>
4. <https://chinacopyrightandmedia.wordpress.com/2012/05/08/sina-weibo-community-management-regulations-trial/>
5. <https://www.facebook.com/journalismproject/programs/third-party-fact-checking>
6. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html; <https://economictimes.indiatimes.com/magazines/panache/twitter-tightens-rules-will-label-tweets-that-spread-fake-news-to-ensure-a-fair-us-election/articleshow/78617901.cms>
7. <https://www.bbc.com/news/blogs-trending-37846860>
8. <https://theconversation.com/governments-are-making-fake-news-a-crime-but-it-could-stifle-free-speech-117654>
9. <https://www.forbes.com/sites/nicolemartin1/2018/11/30/how-social-media-has-changed-how-we-consume-news/?sh=18400d243c3c>
10. <https://www.chinainternetwatch.com/statistics/weibo-mau/>
11. <http://service.account.weibo.com/?type=0&status=4>
12. Full details of the difference between harmful information and fake news can be found in this policy document: <https://chinacopyrightandmedia.wordpress.com/2012/05/08/sina-weibo-community-management-regulations-trial/>
13. <https://chinacopyrightandmedia.wordpress.com/2012/05/08/sina-weibo-community-management-regulations-trial/>
14. <http://www.open.weibo.com/>
15. <https://radimrehurek.com/gensim/models/lmodel.html>
16. <https://baike.baidu.com/item/%E5%BE%AE%E5%8D%9A%E5%AE%9E%E5%90%8D%E5%88%B6>; <https://www.globaltimes.cn/content/700489.shtml>
17. The mean difference is not statistically significant based on independent two-samples *t*-tests (*t*-statistic = -0.564).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Ka Chung Ng (borisnkc@gmail.com) is an assistant professor in the Department of Management and Marketing, Faculty of Business, Hong Kong Polytechnic University. He received his Ph.D. in Information Systems from the Hong Kong University of Science and Technology. His research interests lie in fake news, business analytics, and fintech. His work has appeared in the *Journal of Management Information Systems* and *ACM Transactions on Management Information Systems*.

Jie Tang (tangjie@connect.hku.hk) is a Ph.D. student in Information Systems in HKU Business School at the University of Hong Kong. Her research interests include social media, information privacy, and human-computer interaction. Her research has appeared in the proceedings of several international conferences including International Conference on Information Systems and Pacific Asia Conference on Information Systems.

Dongwon Lee (dongwon@ust.hk; corresponding author) is an assistant professor in the Information Systems, Business Statistics, and Operations Management Department at the Hong Kong University of Science and Technology (HKUST). He received his Ph.D. in Information Systems from University of Maryland. His research interests focus on customer analytics, mobile commerce, digital nudging, digital transformation, and economics of information systems. Dr. Lee's work has been appeared in a number of premier journals as well as major conferences and workshops in Information Systems.

ORCID

Ka Chung Ng  <http://orcid.org/0000-0001-7875-8194>

Jie Tang  <http://orcid.org/0000-0002-9588-8756>

Dongwon Lee  <http://orcid.org/0000-0001-7450-4437>

References

- Allcott, H. and Gentzkow, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31, 2 (2017), 211–236.
- Allcott, H., Gentzkow, M., and Yu, C. Trends in the diffusion of misinformation on social media. *Research and Politics*, 6, 2 (2019), 1–8.
- Amoruso, M., Anello, D., Auletta, V., Cerulli, R., Ferraioli, D., and Raiconi, A. Contrasting the spread of misinformation in online social networks. In K. Larson, M. Winikoff, S. Das and E. Durfee (eds.), *Journal of Artificial Intelligence Research*. International Foundation for Autonomous Agents and Multiagent Systems, São Paulo, Brazil, 2020, pp. 847–879.
- Audrezet, A., de Kerviler, G., and Guidry Moulard, J. Authenticity under threat: When social media influencers need to go beyond self-presentation. *Journal of Business Research*, 117, (2020), 557–569.
- Balkin, J.M. Free speech in the algorithmic society: Big data, private governance, and new school speech regulation. *SSRN Electronic Journal*, 51, (2017), 1149–1210.
- Bessi, A. and Ferrara, E. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, 21, 11 (2016).
- Boichak, O., Jackson, S., Hemsley, J., and Tanupabrungrsun, S. Automated diffusion? Bots and their influence during the 2016 U.S. Presidential election. In G. Chowdhury, J. McLeod, V. Gillet and P. Willett (eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Cham, 2018, pp. 17–26.
- Brockner, J., Ackerman, G., Greenberg, J., et al. Culture and procedural justice: The influence of power distance on reactions to voice. *Journal of Experimental Social Psychology*, 37, 4 (2001), 300–315.
- Buhr, K. and Dugas, M.J. The intolerance of uncertainty scale: Psychometric properties of the English version. *Behaviour Research and Therapy*, 40, 8 (2002), 931–945.

10. Chan, J. and Wang, J. Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science*, 64, 7 (2018), 2973–2994.
11. Chen, D., Lü, L., Shang, M.S., Zhang, Y.C., and Zhou, T. Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and its Applications*, 391, 4 (2012), 1777–1787.
12. Craig, S. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed News*, 2016, 1–7. <https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook> .
13. Enke, N. and Borchers, N.S. Social media influencers in strategic communication: A conceptual framework for strategic social media influencer communication. *International Journal of Strategic Communication*, 13, 4 (2019), 261–277.
14. Figl, K., Rank, C., Kießling, S., and Vakulenko, S. Fake news flags, cognitive dissonance, and the believability of social media posts. In H. Krčmar, J. Fedorowicz, W.F. Boh, J.M. Leimeister, and S. Wattal (eds.), *International Conference on Information Systems, Munich, Germany, 2019*.
15. Friedkin, N.E. Theoretical Foundations for Centrality Measures. *American Journal of Sociology*, 96, 6 (1991), 1478–1504.
16. Gao, G. and Gudykunst, W.B. Uncertainty, anxiety, and adaptation. *International Journal of Intercultural Relations*, 14, (1990), 301–317.
17. Garrett, R.K. and Poulsen, S. Flagging Facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings. *Journal of Computer-Mediated Communication*, 24, 5 (2019), 240–258.
18. Gimpel, H., Heger, S., Olenberger, C., and Utz, L. The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems*, 38, 1 (2021), 196–221.
19. Goel, S., Anderson, A., Hofman, J., and Watts, D.J. The structural virality of online diffusion. *Management Science*, 62, 1 (2016), 180–196.
20. Granovetter, M. The strength of weak ties: A network theory revisited. *Sociological Theory*, 1, (1983), 201.
21. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. Political science: Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363, 6425, (2019) 374–378.
22. Gu, B., Konana, P., Raghunathan, R., and Chen, H.M. The allure of homophily in social media: Evidence from investor responses on virtual communities. *Information Systems Research*, 25, 3 (2014), 604–617.
23. Guess, A., Nyhan, B., and Reifler, J. *Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign*. *European Research Council* 9.3 (2018): 4.
24. Harmon-Jones, E. and Mills, J. *An introduction to cognitive dissonance theory and an overview of current perspectives on the theory*. Stanford University Press, 2004.
25. Hartley, K. and Vu, M.K. Fighting fake news in the COVID-19 era: policy insights from an equilibrium model. *Policy Sciences*, 53, 4 (2020), 735–758.
26. Hess, K.R. Graphical methods for assessing violations of the proportional hazards assumption in cox regression. *Statistics in Medicine*, 14, 15, (1995), 1707–1723.
27. Hinz, O. and Spann, M. The impact of information diffusion on bidding behavior in secret reserve price auctions. *Information Systems Research*, 19, 3 (2008), 351–368.
28. Ho, S.M., Hancock, J.T., Booth, C., and Liu, X. Computer-mediated deception: Strategies revealed by language-action cues in spontaneous communication. *Journal of Management Information Systems*, 33, 2 (2016), 393–420.
29. Hofmann, T. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22, 1 (2004), 89–115.
30. Hofstede, G. The interaction between national and organizational value systems[1]. *Journal of Management Studies*, 22, 4 (1985), 347–357.
31. Jalili, M. and Perc, M. Information cascades in complex networks. *Journal of Complex Networks*, 5, 5 (2017), 665–693.

32. Katz, E. and Shibutani, T. *Improvised News: A Sociological Study of Rumor*. The Bobbs-Merrill Company Inc., 1969.
33. Kim, A. and Dennis, A.R. Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, 43, 3 (2019), 1025–1039.
34. Kim, A., Moravec, P.L., and Dennis, A.R. Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, 36, 3 (2019), 931–968.
35. Knapp, R.H. A psychology of rumor. *Public Opinion Quarterly*, 8, 1 (1944), 22–37.
36. Koh, N.S., Hu, N., and Clemons, E.K. Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications*, 9, 5 (2010), 374–385.
37. Kuang, L., Huang, N., Hong, Y., and Yan, Z. Spillover effects of financial incentives on non-incentivized user engagement: Evidence from an online knowledge exchange platform. *Journal of Management Information Systems*, 36, 1 (2019), 289–320.
38. Kwon, H.E., Oh, W., and Kim, T. Platform structures, homing preferences, and homophilous propensities in online social networks. *Journal of Management Information Systems*, 34, 3 (2017), 768–802.
39. Landauer, T.K., Foltz, P.W., and Laham, D. An introduction to latent semantic analysis. *Discourse Processes*, 25, 2–3 (1998), 259–284.
40. Landherr, A., Friedl, B., and Heidemann, J. A critical review of centrality measures in social networks. *Business & Information Systems Engineering*, 2, 6 (2010), 371–385.
41. Lazer, D.M.J., Baum, M.A., Benkler, Y., et al. The science of fake news. *Science*, 359, 6380 (2018), 1094–1096.
42. Maasberg, M., Ayaburi, E., Liu, C., and Au, Y. Exploring the propagation of fake cyber news: An experimental approach. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. Curran Associates Inc., Hawaii, USA, 2018.
43. Marett, K. and Joshi, K.D. The decision to share information and rumors: Examining the role of motivation in an online discussion forum. *Communications of the Association for Information Systems*, 24, 1 (2009), 47–68.
44. Moravec, P.L., Kim, A., and Dennis, A.R. Flagging fake news: System 1 vs. System 2. In J.P. Heje, S. Ram, and M. Rosemann (eds.), *International Conference on Information Systems*, San Francisco, USA, 2018.
45. Moravec, P.L., Minas, R.K., and Dennis, A.R. Fake news on social media: People believe what they want to believe when it makes no sense at All. *MIS Quarterly*, 43, 4 (2019), 1343–1360.
46. Oh, O., Agrawal, M., and Rao, H.R. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly*, 37, 2 (2013), 407–426.
47. Papanastasiou, Y. Fake news propagation and detection: A sequential model. *Management Science*, 66, 5, (2020), 1826–1846.
48. Park, J.H., Konana, P., Gu, B., Kumar, A., and Raghunathan, R. Information valuation and confirmation bias in virtual communities: Evidence from stock message boards. *Information Systems Research*, 24, 4 (2013), 1050–1067.
49. Pennycook, G., Bear, A., Collins, E.T., and Rand, D.G. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66, 11 (2020), 4944–4957.
50. Pennycook, G., Cannon, T.D., and Rand, D.G. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147, 12, (2018), 1865–1880.
51. Pennycook, G. and Rand, D.G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, (2019), 39–50.
52. Pickett, J.T., Roche, S.P., and Pogarsky, G. Toward a Bifurcated Theory of Emotional Deterrence. *Criminology*, 56, 1 (2018), 27–58.
53. Rosenbaum, P.R. and Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 1 (1983), 41–55.
54. Rosnow, R.L. Inside rumor: A personal journey. *American Psychologist*, 46, 5 (1991), 484–496.

55. Ross, B., Heisel, J., Jung, A.K., and Stieglitz, S. Fake news on social media: The (in)effectiveness of warning messages. In J.P. Heje, S. Ram, and M. Rosemann (eds.), *International Conference on Information Systems*, San Francisco, USA, 2018.
56. Sarker, S., Ahuja, M., Sarker, S., and Kirkeby, S. The role of communication and trust in global virtual teams: A social network perspective. *Journal of Management Information Systems*, 28, 1 (2011), 273–310.
57. Schulze, E. EU tells Facebook, Google and Twitter to take more action on fake news. In *CNBC*. 2019.
58. Shao, C., Hui, P.M., Cui, P., Jiang, X., and Peng, Y. Tracking and characterizing the competition of fact checking and misinformation: Case Studies. *IEEE Access*, 6, (2018), 75327–75341.
59. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, 10, 3 (2019), 1–41.
60. Shi, D., Guan, J., Zurada, J., and Manikas, A. A data-mining approach to identification of risk factors in safety management systems. *Journal of Management Information Systems*, 34, 4 (2017), 1054–1081.
61. Shi, Z.M., Lee, G.M., and Whinston, A.B. Toward a better measure of business proximity: Topic modeling for industry intelligence. *MIS Quarterly*, 40, 4 (2020), 1035–1056.
62. Stieglitz, S. and Dang-Xuan, L. Emotions and information diffusion in social media - Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29, 4 (2013), 217–248.
63. Suh, A., Shin, K.S., Ahuja, M., and Kim, M. The influence of virtuality on social networks within and across work groups: A multilevel approach. *Journal of Management Information Systems*, 28, 1 (2011), 351–386.
64. Sutton, J., Spiro, E.S., Fitzhugh, S., Johnson, B., Gibson, B., and Butts, C.T. Terse message amplification in the Boston bombing response. In S.R. Hiltz, M.S. Pfaff, L. Plotnick and P.C. Shih (eds.), *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*. University Park, USA, 2014, pp. 612–621.
65. Sutton, J., Spiro, E.S., Johnson, B., Fitzhugh, S., Gibson, B., and Butts, C.T. Warning tweets: serial transmission of messages during the warning phase of a disaster event. *Information Communication and Society*, 17, 6 (2014), 765–787.
66. Tang, J., and Ng, K.C. Reposts influencing the effectiveness of social reporting system: An empirical study from sina weibo. In H. Krcmar, J. Fedorowicz, W.F. Boh, J.M. Leimeister, and S. Wattal (eds.), *International Conference on Information Systems*, Munich, Germany, 2019.
67. Tausczik, Y.R. and Pennebaker, J.W. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29, 1 (2010), 24–54.
68. Torres, R., Gerhart, N., and Negahban, A. Combating fake news: An investigation of information verification behaviors on social networking sites. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. Curran Associates Inc., Hawaii, USA, 2018.
69. Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *Science*, 359, 6380 (2018), 1146–1151.
70. Wang, Q., Li, B., and Singh, P.V. Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis. *Information Systems Research*, 29, 2 (2018), 273–291.
71. Zhang, X. and Venkatesh, V. Explaining employee job performance: The role of online and offline workplace communication networks. *MIS Quarterly*, 37, 3 (2013), 695–722.
72. Zhang, X., Zhu, J., Wang, Q., and Zhao, H. Identifying influential nodes in complex networks with community structure. *Knowledge-Based Systems*, 42, (2013), 74–84.
73. Zhang, X.M. and Wang, C. Network positions and contributions to online public goods: The case of Chinese wikipedia. *Journal of Management Information Systems*, 29, 2 (2012), 11–40.
74. Zhou, S., Qiao, Z., Du, Q., Wang, G.A., Fan, W., and Yan, X. Measuring customer agility from online reviews using big data text analytics. *Journal of Management Information Systems*, 35, 2 (2018), 510–539.