Quantifiable Interactivity of Malicious URLs and the Social Media Ecosystem

Chun-Ming Lai 1,*, Hung-Jr Shiu 1 and Jon Chapman 2

1 Department of Computer Science, Tunghai University, Xitun District, Taichung 407224, Taiwan; hjshiu@thu.edu.tw
2 Department of Computer Science, University of California, Davis, One Shields Ave., Davis, CA 95616, USA; jwchapman@ucdavis.edu
* Correspondence: cmlai@thu.edu.tw

Received: 19 October 2020; Accepted: 23 November 2020; Published: 30 November 2020

Abstract: Online social network (OSN) users are increasingly interacting with each other via articles, comments, and responses. When access control mechanisms are weak or absent, OSNs are perceived by attackers as rich environments for influencing public opinions via fake news posts or influencing commercial transactions via practices such as phishing. This has led to a body of research looking at potential ways to predict OSN user behavior using social science concepts such as conformity and the bandwagon effect. In this paper, we address the question of how social recommendation systems affect the occurrence of malicious URLs on Facebook, based on the assumption that there are no differences among recommendation systems in terms of delivering either legitimate or harmful information to users. Next, we use temporal features to build a prediction framework with >75% accuracy to predict increases in certain user group behaviors. Our effort involves the demarcation of URL classes, from malicious URLs viewed as causing significant damage to annoying spam messages and advertisements. We offer this analysis to better understand OSN user sensors reactions to various categories of malicious URLs in order to mitigate their effects.

Keywords: facebook; malicious URL; social influence

1. Introduction

The attack vectors that users of Online Social Networks (OSNs) face have been evolving as the various bad actors learn to manipulate this new aspect of the cyber landscape. One of these newly evolving attack vectors is the news creation and dissemination cycle. Traditionally the structure for media dissemination was a top-down arrangement; news was typically published by well-trained reporters and edited by a skilled team. In this fashion, these professionals acted as a gatekeeper of sorts, ensuring higher journalistic integrity and correctness of the news. In contrast, news is now being created for use and spread on OSNs by users. This opens the door to the news generation process, and the associated URLs, to be utilized as nascent attack vectors.

With the modern structuring of this information creation and consumption process, social recommendation systems play an increasingly critical role as they determine which users will see exactly what information based on the characteristics of the individual users as well as specific features from the articles. Unfortunately, inappropriate information will diffuse on user-generated content platforms much more readily than traditional media, attributable to two primary factors: (1) The connectivity of social media makes information diffusion deeper and wider. (2) Users may wittingly or unwittingly boost inappropriate dissemination cascades with their own comments on the articles.
The major research efforts in the area of OSN security are concerned with detecting malicious accounts rather than normal user accounts. However, attackers can exploit either newly created fake or existing latent compromised accounts to avoid state-of-the-art defense schemes since most are based on verified attacker behavior and trained by machine learning algorithms either by lexical features or by account characteristics. Relatively less work has been done to measure and consider the influence of the actual malicious content. Therefore, our motivation is based on two research questions: (RQ1) For discussion threads having clearly malicious content, do they have a larger cascade size when compared to other discussion threads that do not contain malicious content? (RQ2) In a discussion thread, is there a significant influence between the prior and the latter comments surrounding a malicious comment? Specifically, we are wondering whether audiences will change their behavior when they see a malicious comment which is being promoted by a social recommendation system.

In this paper, we design two experiments trying to answer the above two questions. For cascade size between target and non-target post threads, we evaluate with the bandwagon effect experiment. The findings indicate that they both basically follow the same cascade model. However, their final cascade size are extremely different for at least two reasons: (1) users’ reactions; and (2) social recommendation design. Afterwards, we turn our attention to the fine-grained influence of a user-generated comment. We define an Influence Ratio (IR) for every comment to evaluate its influence based on the ratio over its upcoming activities and its preceding activities. Our framework achieves more than 75% accuracy on both critical and light damage URLs in predicting the upcoming activity increase or decrease.

Our results also indicate that the relative position, in the context of the chronological positioning, of a comment plays a critical role in contributing to the influence that the comment yields. For example, Figure 1 (post_id = 10152998395961509) shows a labeled advertisement URL occurred in the very late stage (2802th out of 2844 total comments) while Figure 2 (post_id = 10153908916401509) shows a labeled pornography URL occurred near the middle of the comment chronology (470th comment of 942 total). We found that, with regards to position, critical and those lesser threat level URLs are presenting at different times in the discussion—the light threat URLs tend to be posted later chronologically while critical threat URLs tend to be in the middle of a post’s timeline. This phenomenon is borne out very obviously in the CNN public page, however this presentation is not as dramatic on FOX News. There are at least two reasons for the chronological disparity between the different threat level URLs: (1) Users tend to leave the discussion when they feel there was an obvious ad posted, as in Figure 1. (2) Compared to the lower level threats, attackers who spread critical malicious URLs act in a more strategic manner—they choose the most opportune timing to achieve the greatest amount of influence.

The rest of this paper is organized as follows. Section 2 illustrates how Facebook and social recommendation systems work. In Section 3, we define necessary terms and provide a detailed description for our dataset. The bandwagon effect cross-validation is described in Section 4. We define and predict the influence ratio in Section 5. Related work and our conclusion are given in Sections 6 and 7, respectively.
2. Facebook and Social Recommendation System

In this section, we introduce one of the most popular online services of social media—Facebook— from its humble beginnings as a sort of simple digitized social yearbook limited to
only certain universities to a worldwide, incredibly complex, and multi-functional platform. We also describe the information consumption process between Facebook and OSNs Users.

2.1. Facebook Public Pages

Facebook was launched in 2004, initially providing a platform for students to search for people at the same school and look up friends of friends. Users updated personal information, likes, dislikes, and current activities. While doing this, they also kept track of what others were doing, and used Facebook to check the relationship availability of anyone they might meet and were interested in romantically (https://www.eonline.com/news/736769/this-is-how-facebook-has-changed-over-the-past-12-years). As Facebook grew quickly, users were not satisfied with merely following the personal status of their close friends on the network. Furthermore, users demonstrated an interest in public affairs and news. For this reason, public pages on Facebook were created and have become places where users receive news and information selected and promoted by news feeds, which are constantly updating lists of stories in the middle of one’s homepage, including stories regarding: (1) friends’ activities on Facebook; (2) articles from pages where an user is interested (Liked or Followed); (3) articles that your friends like or comment on from people you are not friends with; and (4) advertisements from sponsoring companies and organizations (https://www.facebook.com/help/327131014036297/). With these new media publication venues on Facebook, users interact with strangers on various public pages—discussing news published by commercial media companies, announcements by public figures, sharing movie reviews, gossiping about an actor, or criticizing the poor performance of a particular sports teams. According to Hong et al. [1], there are more than 38,831,367 public pages covering multiple topics including Brands and Products, Local Business and Places, and Companies and Organizations (https://www.facebook.com/pages/create).

2.2. Social Recommendation System

Most highly trafficked online social media sites contain some variation of a dynamic social recommendation system [2]. It is a continuous process cycle, which includes two entities, social computing platform and active users, and the four processes shown in Figure 3. Here, we explain the processes in more detail.

![Figure 3. The framework of users activities regarding online social media.](image)

1. Deliver: Large-scale and user-generated data have been disseminated on OSNs. However, only appropriate information is delivered to corresponding audiences.
2. Digest: When users see the news, they have the chance to join the discussion by actively typing their opinion, less actively clicking reactions, or passively doing nothing.
3. Derive and Evaluate: Recommendation systems will collect a large volume of user interaction data and modify the algorithm to better attract attention from users (mostly because of attention economy [3]). The evaluation step gives a chance for Facebook to modify the social algorithms to deliver more appropriate (Step 1) information to users. The primary concern for Facebook is to
maximize clicks on advertisements. This is primarily accomplished by maximizing time spent on Facebook by the user.

Attackers logically attempt to maximize the influence they wield for every malicious campaign—in effect, having more people see their malicious content, click it, interact with it, or trust it. With the application of behavioral targeting, we believe the bad actors spread URLs that will be more relevant to audiences, whose patterns could be collected from data mining or speculation. For example, the collection of bad actors involved with spreading fake news tend to chronologically target the planting of their fake news as well as topographically targeting the best locations to plant fake news (e.g., politics-related Facebook pages or articles). Other bad actors run accounts that are hired by commercial enterprises that have a more limited scope and primarily care about their business. The common thread is they all make use of the social recommendation system. We have seen that social recommendation system design actually increases the damage of malicious URLs since it offers a way for attackers to spread harmful content at the right place and at the right time. Vosoughi et al. indicated that false news is more novel than true news and humans are more likely to share novel information online [4]. Therefore, the social recommendation system will boost the “rich get richer” effect.

3. Data Description and Labeling

In this section, we define the necessary terminology used in this paper. After providing a high level overview of the discussion groups dataset, we show how we label filtered URLs into different categories.

3.1. Terminology

We use the following terminology to describe the concepts in our work more exactly:

- Page: A public discussion group. In this study, we only consider two main media pages: CNN (page_id = 5550296508) and FOX News (page_id = 15704546335).
- Original Post: An article on a Facebook discussion group.
- Comments: Text written in response to an original post.
- Reaction: Emoji to comments or original posts, including “Like”, “Love”, “Haha”, “Love”, “Sad”, and “Angry”.
- Post thread: The original post and all corresponding user activities (Comments and corresponding Reactions), ordered by their timestamps.
- Target post thread: Post threads which have at least one comment embedded with malicious URL(s).
- Non-Target Posts: Post threads which have no embedded malicious URLs.
- Time Series (TS): \( TS_{\text{created}} \) indicates a time period \( j \) following the time of the original post, measured in minutes. \( TS_{\text{final}} \) refers to the precise time 1 h after the original post was created (i.e., \( \text{final} = 60 \)).
- Number of comments \( (N_{\text{comment}}(post, TS_i)) \): The number of post comments collected at \( TS_i \)
- Accumulated number of participants \( (AccN_{\text{comment}}(post, TS_i)) \): The number of post comments between \( TS_i \) and \( TS_{i-1} \)

3.2. Crawled Dataset

To this end, our data were cataloged with the use of an open source social crawler called SINCERE (https://github.com/dsfaithdev/SocialCrawler) (Social Interactive Networking and Conversation Entropy Ranking Engine), which has been created and refined by our research group over several years. We employed it from 2014 to 2016 to collect post threads on both CNN and FOX News public pages to see the difference between left-wing and right-wing discussion. Detail information was stored, including timestamps of each comment, Facebook account identification numbers, and the raw text of
comments and articles. In total, we have 48,087 posts, 88,834,886 comments and 189,460,056 reactions for both pages. We describe the full dataset in Table 1.

<table>
<thead>
<tr>
<th>Page Name</th>
<th>Total Posts</th>
<th>Total Comments</th>
<th>Comments with URLs</th>
<th>Total Reactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>20,922</td>
<td>11,882,590</td>
<td>412,001 (3.47%)</td>
<td>24,174,160</td>
</tr>
<tr>
<td>FOX News</td>
<td>27,165</td>
<td>76,952,296</td>
<td>1,026,525 (1.33%)</td>
<td>165,285,896</td>
</tr>
</tbody>
</table>

3.3. Labeling URLs

Typically, a URL contains three parts: (1) a protocol; (2) a host name; and (3) a file name. In this paper, we focus on URLs which use HTTP and HTTPS protocols. Moreover, we focus on the host name itself. We first use a well-known Whitelist 'facebook.com', 'youtube.com', 'twitter', 'on.fb.me', 'en.wikipedia', 'huffingtonpost.com', 'foxnews.com', 'cnn.com', 'google.com', 'bbc.co.uk', 'nytimes.com', 'washingtonpost.com' to do a first-step filter. There is no doubt that there is lots of inappropriate content on our whitelists, such as 'facebook.com' and 'youtube.com'. However, in this work, the scope is on normal blacklists adopted by third-party cyber-security engines. Accordingly, we then employ the daily-updated Shalla Blacklist service [5], which is a collection of URL lists grouped into several categories intended for the usage with URL filters, to label and trace the behavior of URL influence. Note that we do not assume all URLs filtered by Shalla are completely malicious. Among the 74 categories listed, we manually divided targeted URLs into two classes: Light and Critical. The explanation of each category is as follows:

- **Light**:
  - Advertising: Includes sites offering banners and advertising companies.
  - Shopping: Sites offering online shopping and price comparisons.
  - Gamble: Poker, casino, bingo, and other chance games as well as betting sites.
  - Porn: Sites with sexual content.

- **Critical**:
  - Download: This covers mostly file-sharing, p2p, torrent sites, and drive-by-downloads.
  - Hacking: Sites with information and discussions about security weaknesses and how to exploit them.
  - Spyware: Sites that try to actively install software (or lure the user to do so) in order to spy on user behavior. This also includes trojan and phishing sites.
  - Aggressive: Sites of obvious aggressive content, including hate speech and racism.
  - Drugs: Sites offering drugs or explaining how to make drugs (legal and illegal).
  - Weapons: Sites offering weapons or accessories for weapons.
  - Violence: Sites about killing or harming people or animals.

We classify others as Benign if they are not in the Whitelist, Light, or Critical classes. The detailed number of each category is listed in Table 2.

<table>
<thead>
<tr>
<th>Page Name</th>
<th>URL in WhiteList</th>
<th>URL in Light</th>
<th>URL in Critical</th>
<th>URL in Benign</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>194,372</td>
<td>3762</td>
<td>636</td>
<td>213,231</td>
</tr>
<tr>
<td>FOX News</td>
<td>503,480</td>
<td>8125</td>
<td>1571</td>
<td>513,349</td>
</tr>
</tbody>
</table>

4. Post-Level Influence

Heterogeneous posts are updated and refreshed at tremendous speed and include videos, photos with attractive headlines, and assorted topics such as international affairs, elections or entertainment.
We are interested in why malicious URLs often occur on only some post threads. We applied the model proposed by Wang et al. [6] to gain insight into how those malicious URLs may influence the growth of the conversation.

4.1. Bandwagon Effect and Attacker Cost

A phenomenon known as the bandwagon effect explains how individuals will agree with a larger group of people who they may not normally agree with, but do so to feel a part of a group—individuals are more likely to be with those sub-groups that share similar thoughts but feel uncomfortable in the presence of minority groups that have different ideas [7]. Many voting behaviors are related to this effect; voters may or may not follow their own conscience to make a voting decision but may just follow the majority opinion [8].

From the result obtained by Lai et al. [9], when considering the posts targeted with a comment that includes a malicious URL, we see the most commonly attacked articles tend to generate large amounts of discussion. Moreover, targets may be those suggested by the Facebook social recommendation system to the attacker.

The following example indicates how a large number of majority opinions might be identified from a Facebook discussion. Assume three posts—post_A, post_B, and post_C—have been posted on a public page at around the same time, the original posters’ identities are irrelevant. In addition, assume there exist three different users—user_A visits the page and browses all three posts, has no signals from others for making a decision to engage with the post, and the user subsequently decides to only comment on post_B because it was subjectively the most interesting one to them. Five minutes later, user_B visits the same page and sees that only post_B has a comment. This user then checks post_B first and decides to add their own reply, either in response to the original post or user_A. Note that until now there are no comments on either post_A or post_C. Some short time later, user_C checks the page and finds that post_B has more than 10 comments, while post_A and post_C still have 0 comments. He then decides to add a comment to post_B since post_B is the first post that is pushed to the user by Facebook because, at that moment, post_B has a relatively larger share of public attention compared to post_A or post_C. This is an example of the information cascade phenomenon first proposed by Bikhchandani, Hirshleifer, and Welch [10], and most social media recommendation systems intensify this phenomenon—information and user activities are automatically selected by an algorithm, although most users do not realize this when participating in OSN discussion groups.

4.2. Prediction Model and Evaluation Method

To differentiate the information cascade model between target and non-target post threads, we describe a system designed to use the time series and number of current comments to predict how many new users are likely to participate in each respective thread. The Discussion Atmosphere Vector (DAV) [9] defined in our previous work used the definition of accumulated number of participants given in Section 3, using 5 min for the i value and 2 h for the t_{final} value.

\[
DAV(Post)_{t_n} = [AccNcomment(Post,t_1),
\quad AccNcomment(Post,t_2),...,
\quad AccNcomment(Post,t_n)]
\]  

(1)

In the bandwagon effect model proposed by Wang et al. [6], the numbers of comments with respect to each time window after a post has been created can be used to build matrices for each public page G to predict the final number of comments by machine learning and statistical methods. In other words, two post threads post_A and post_B are likely to have the same scale of cascades if in each timestamp i such that:

\[
DAV(Post_A)_{t_i} \approx DAV(Post_B)_{t_i}
\]  

(2)
We then defined a distribution matrix $D$, with each element $D_{ij}$ representing a set of posts $\text{post} \in G$, including the final number of comments we crawled $N_{\text{comment}}(\text{post}, TS_{\text{final}})$ and aggregate number of comments $j = N_{\text{comment}}(\text{post}, TS_{i})$ at time $i$.

$$D_{ij}(G) = \{N_{\text{comment}}(\text{post}, TS_{\text{final}}(\text{post}))\}$$

(3)

Based on the distribution matrix $D$, we used a bootstrapping method [11] to construct prediction matrix $M$.

$$M_{ij} = \text{Bootstrapping}(D_{ij})$$

(4)

The matrix $M$ is used to create a prediction function $F_{\text{predict}}$ that collects two inputs from any new post thread: observed time series $TS_{\text{ob}}(\text{post})$ and corresponding feature $N_{\text{comment}}(\text{post}, TS_{\text{ob}})$. According to $M$, we obtain the result using the following equation:

$$F_{\text{predict}}(TS_{\text{ob}}(\text{post}), N_{\text{comment}}(\text{post}, TS_{\text{ob}})) = M_{TS_{\text{ob}}(\text{post}, N_{\text{comment}}(\text{post}, TS_{\text{ob}}))}$$

(5)

4.3. Result and Discussion

Our bandwagon cross-validation between target and non-target post threads are offered in Tables 3 and 4. Note that the prediction function can sometimes fail for either of two reasons: insufficient features for a post to match $M$ in the testing data or insufficient existing posts within the training data. In our experiment, we have enough training data, so unpredictable posts $\approx 1\%$ for both pages are posts which do not have enough activities for $M$ to predict the final size of cascade.

<table>
<thead>
<tr>
<th>Page Name</th>
<th>Precision/Predictable (%)</th>
<th>Predictable/All (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>15,633/15,866 (97%)</td>
<td>15,866/15,869 (99%)</td>
</tr>
<tr>
<td>FOX News</td>
<td>17,338/17,448 (99%)</td>
<td>17,448/17,453 (99%)</td>
</tr>
</tbody>
</table>

Table 3. Bandwagon effect cross validation—target to non-target, observed time = 120 min.

<table>
<thead>
<tr>
<th>Page Name</th>
<th>Precision/Predictable (%)</th>
<th>Predictable/All (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>5013/5014 (99%)</td>
<td>5014/5053 (99%)</td>
</tr>
<tr>
<td>FOX News</td>
<td>9,706/9,712 (99%)</td>
<td>9,712/9,712 (100%)</td>
</tr>
</tbody>
</table>

Table 4. Bandwagon effect cross validation—non-target to target, observed time = 120 min.

Basically, there are no obvious differences regarding the first 2 h of activities with respect to final number of comments between target and non-target post threads. This suggests that malicious URLs did not affect the life cycle of post threads—people still engaged the target post threads, under the threat of malicious URLs. On the other hand, our result also indicate that the Facebook social recommendation system continued to deliver post threads which have malicious URLs to audiences, similar to the way it treats normal post threads.

Recall that the bootstapping method in this experiment is only providing the lower bound. For example, if $M_{5,5} = 100$, this means that any post that satisfies five comments in the first five minutes would have 100 comments or more. However, 200 and 2000 are both greater than 100, but the scale are not the same. To consider the final cascade of comments between targets and non-targets, we also conduct two-sample Kolmogorov–Smirnov (KS) tests to compare the distributions of the final number of comments between those two sets. The results are shown in Tables 5 and 6. In general, for both the FOX News and CNN pages, the final number of comments of target post threads is obviously greater than non-target ones. This can be attributed that attackers are led by the Facebook social recommendation systems. In other words, their targets are not chosen by themselves but mostly
by social algorithms. Moreover, we also noticed that FOX News attracts more people to join the discussion, rather than CNN (about four times the mean number of comments per post threads).

Table 5. CNN statistics on the cascade size. KS-test for target and non-targets: \( D = 0.068, p \approx 0.0. \)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SE</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>5053</td>
<td>1010</td>
<td>1558</td>
<td>14</td>
<td>39,929</td>
</tr>
<tr>
<td>Non-Target</td>
<td>15,869</td>
<td>427</td>
<td>698</td>
<td>1</td>
<td>35,591</td>
</tr>
</tbody>
</table>

Table 6. FOX statistics on the cascade size. KS-test for target and non-targets: \( D = 0.066, p \approx 0.0. \)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SE</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>9712</td>
<td>4712</td>
<td>11,740</td>
<td>24</td>
<td>412,621</td>
</tr>
<tr>
<td>Non-Target</td>
<td>17,453</td>
<td>1786</td>
<td>3436</td>
<td>1</td>
<td>115,669</td>
</tr>
</tbody>
</table>

5. Influence Ratio of a Comment

Post thread is a basic unit to consider the interaction of users. In the previous section, we show that the final size of cascade (number of comments) and the first 2 h of activities between target posts and non-targets are very similar, however the scales are extremely different from KS test. In this section, we turn our attention to the temporal neighbor—users interact with a time period on the same post threads, even though they are not mutual friends on Facebook.

5.1. Preceding and Upcoming Activities

Consider an original post released by a news media outlet on its public page. This post can be a video, a photo, or even just a short paragraph of text. There can be many users activities toward this particular post. Consider an original article (post). The corresponding comments are \( C_1, C_2, ..., C_n \), ordered by their created timestamps \( (Time(C_1), Time(C_2), ..., Time(C_n)) \). To evaluate the influence of a comment \( C_i \) with its created time \( Time(C_i) \), given a time window \( \Delta T \), we define Influence Ratio (IR) as the log ratio between all activities which occurred in the previous time window \( Time(C_i) - \Delta T \) and the upcoming time window \( Time(C_i) + \Delta T \).

\[
\text{InfluenceRatio}(C_i, \Delta T) = \log \left( \frac{\text{count(activities)} \in (Time(C_i) + \Delta T)}{\text{count(activities)} \in (Time(C_i) - \Delta T)} \right)
\]

we classify the comment itself in the time period \( (Time(C_i) - \Delta T) \) to avoid the denominator becoming zero. Activities include all comments, likes, and reactions. If IR is greater than 0, it means people will be more interested in this post and this comment in the next time slot \( Time(C_i) + \Delta T \).

5.2. Predict and Evaluate Influence Ratio

The time differences between two consecutive comments \( C_n - C_{n-1} \) vary greatly. For example, several studies have shown that post threads would have a rich get richer \[12\] and bandwagon effect \[13\], which indicates the nearby comments and reactions are critical to interact and influence with each other—everyone is a potential amplifier. Consider two users User(\( C_i \)) and User(\( C_j \)) who contribute \( C_i \) and \( C_j \). If \(| i - j | \) is quite close to 1, they would have a higher chance to interact with each other since: (1) the social recommendation system delivered this post to both users because of their past activities and browsing footprints; and (2) they remain online on social media at around the same time (this is not always true since we also need to consider the time difference between \( Time(C_i) \) and \( Time(C_j) \)). However, they may not be friends with each other but just have overlapping active time on Facebook. To consider the volume of specific time period, we define a function CountActivity(\( post, [Time(\text{Begin}), Time(\text{After})] \)) which refers to all activities (comments,
likes, reactions, and replies) for post within $Time(Begin)$ to $Time(After)$. To consider the influence and role of a comment in a post thread, given an influenced threshold $\delta T$ and preceding audience number $N_{prev}$, we define Preceding Influenced Vector (PIV) of a comment $C_k$ in a post thread $Post$ as follows:

$$PIV_i(Post, C_k) = CountActivity(Post, \left[Time(C_k) - i*\delta T, Time(C_k) + i*\delta T\right])$$ (7)

Our goal is to predict IR, the volume of the upcoming time windows. In other words, for any comment $C_k$ in an article post, the influence ratio problem predicts whether the upcoming audiences will be greater than the preceding audiences via a classifier ($c \rightarrow target, nontarget$) based on one set of $PIV(Post, C_k)$. Hence, for two arbitrary comments $C_m$ of $Post_m$ and $C_n$ of $Post_n$, they will be more likely to have the same trend for the upcoming number of activities such that:

$$PIV(Post_m, C_m) \approx PIV(Post_n, C_n)$$

We use comments with benign URLs as a training data, trying to predict the IR trends for both light URLs and critical URLs. We set time window $\delta T$ as 1 min and the number of components of PIV as 60, which means that for each comment, we assume the time period up to 1 h will influence the IR. We then normalized the input PIV to prevent overfitting. As output, we labeled the positive value of IR as increase while negative value of IR as decrease. We applied two popular machine learning classifiers (1) Adaboost; and (2) Gaussian Naive Baynes from scikit-learn [14]. For Adaboost, we set the number of estimators = 50 and learning rate = 1. The detailed results are shown in Tables 7 and 8. Overall, we can achieve greater than 75% F1-score on predicting the Influence ratio for both Light and Critical URLs, and the result for Light category is better than Critical category. Moreover, we summarize our findings as follows:

- **CNN vs. FOX News**: There are no obvious difference between CNN and FOX News with regards to predicting IR for the more critical threats versus the lower threat malicious campaigns. We think the reason may be both CNN and FOX News Feeds were controlled by the same social recommendation system. Hence, user activities with respect to temporal features can be predicted with the same amount of ease on either feed.

- **Increase vs. Decrease**: For most cases, the F1-score on predicting increase is better than decrease (CNN Benign to Critical and FOX News both cases). We think it is related to our previous experiment regarding bandwagon effect. We also notice that, on the CNN page, the IR of light malicious campaigns tend to decrease, which can either be audiences leaving the discussion because of the URL or the attacker strategies being inefficient to cause popularity.

- **Classifiers**: Better results were obtained by Adaboost. For social media data, since PIVs are not independent with respect to one another, Naive Baynes does not work well given the dependent variable constant. In other words, when considering group behaviors on OSNs, we believe that reinforcement learning classifier is better than Naive Baynes, which is based on probabilistic classifiers.
Table 7. Influence ratio prediction—Benign to Light, observed time = 60 min, $\delta T = 1$ min.

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>Naive Baynes</th>
<th>Adaboost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precise</td>
<td>Recall</td>
<td>F1-Score</td>
</tr>
<tr>
<td>CNN</td>
<td>0.86</td>
<td>0.49</td>
<td>0.62</td>
</tr>
<tr>
<td>Naive Baynes</td>
<td>0.53</td>
<td>0.88</td>
<td>0.66</td>
</tr>
<tr>
<td>Decrease</td>
<td>0.73</td>
<td>0.64</td>
<td>0.64</td>
</tr>
<tr>
<td>Increase</td>
<td>0.84</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>avg/total</td>
<td>0.72</td>
<td>0.77</td>
<td>0.74</td>
</tr>
<tr>
<td>FOX News</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>Naive Baynes</td>
<td>0.69</td>
<td>0.42</td>
<td>0.52</td>
</tr>
<tr>
<td>Increase</td>
<td>0.65</td>
<td>0.85</td>
<td>0.74</td>
</tr>
<tr>
<td>avg/total</td>
<td>0.67</td>
<td>0.66</td>
<td>0.64</td>
</tr>
<tr>
<td>Adaboost</td>
<td>0.81</td>
<td>0.69</td>
<td>0.74</td>
</tr>
<tr>
<td>Increase</td>
<td>0.78</td>
<td>0.87</td>
<td>0.82</td>
</tr>
<tr>
<td>avg/total</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 8. Influence ratio prediction—Benign to Critical, observed time = 60 min, $\delta T = 1$ min.

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>Naive Baynes</th>
<th>Adaboost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precise</td>
<td>Recall</td>
<td>F1-Score</td>
</tr>
<tr>
<td>CNN</td>
<td>0.88</td>
<td>0.42</td>
<td>0.57</td>
</tr>
<tr>
<td>Naive Baynes</td>
<td>0.62</td>
<td>0.94</td>
<td>0.75</td>
</tr>
<tr>
<td>Decrease</td>
<td>0.75</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td>Increase</td>
<td>0.83</td>
<td>0.71</td>
<td>0.77</td>
</tr>
<tr>
<td>avg/total</td>
<td>0.75</td>
<td>0.86</td>
<td>0.80</td>
</tr>
<tr>
<td>FOX News</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td>Naive Baynes</td>
<td>0.59</td>
<td>0.48</td>
<td>0.53</td>
</tr>
<tr>
<td>Increase</td>
<td>0.73</td>
<td>0.81</td>
<td>0.77</td>
</tr>
<tr>
<td>avg/total</td>
<td>0.68</td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td>Adaboost</td>
<td>0.75</td>
<td>0.56</td>
<td>0.64</td>
</tr>
<tr>
<td>Increase</td>
<td>0.78</td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td>avg/total</td>
<td>0.77</td>
<td>0.77</td>
<td>0.77</td>
</tr>
</tbody>
</table>

5.3. Life Cycle Stage

We noticed that the temporal ordering of activities on post threads are quite interesting. The audience generally rapidly increases to the peak, and then growth of the audience decays more slowly as time goes on. Figure 4 visualizes the relationship among ratio, IR, and elapsed time from the last comment. We model the life cycle of post threads into three stages:

1. Rapid growth: On the facebook page for FOX News, there is an obvious watershed at 50%: From the first comment to about the midway point in post threads, many users join the discussion; usually in the next time window each comment will have five times the activities as the previous time window, and the time difference with the last comment is usually smaller than 1 min. However, on the CNN page, we observe that IR experiences several huge discussions; the reason may be people provide lots of reactions for some interesting comments.
2. Slow Decay: For both CNN and FOX News, we observed that, from 50% to about 85%, the volume of comments faces an obvious decline. At the same time, IR lightly decays and elapsed time becomes larger.

3. Dormancy: At this stage, the thread has basically passed its shelf life. The elapsed time goes up to more than 10 min while the IR has fallen to almost 0.

We also noticed that attackers are more likely to spread malicious URLs at either the Slowly Decay or Dormancy stages on CNN, while, on FOX, the ratio seems to be a uniform distribution, as shown in Figure 5.

![Life cycle of two pages regarding IR and elapsed time.](a) CNN  (b) FOX News

**Figure 4.** Life cycle of two pages regarding IR and elapsed time.

![CDFs of different categories occurrences plotted against life stages of targeted threads.](a) CNN  (b) FOX News

**Figure 5.** CDFs of different categories occurrences plotted against life stages of targeted threads.

5.4. Attackers’ Footprint

In addition to the post thread life cycle described above, we are also interested in the user’s other activities. Are they actively expressing their opinions, or just acting as a one-time inappropriate information generator? We consider user activities on more than 40,000 public pages around the world from 2011 to 2016 on Facebook. Figure 6 shows that, no matter the numbers of comments, number of likes, or number of reactions, those users who spread Critical-type malicious URLs occur more often than Benign- and Light-type users. Considering their purpose, we noticed that Light-type users tend to lure users to commercial websites. However, Critical-type users comments usually advocate relatively personal belief and values—which makes them heavier Facebook users and tend to influence others. However, on FOX News, only Light-type users have less activities than the other two, so there is no obvious difference between Critical-type users and Benign users.
6. Related Work

Security issues surrounding OSN platforms have been growing in importance and profile due to the increasing number of users (and subsequent potential targets) of social media applications. Our related work mainly falls into two categories: (1) popularity on social media; and (2) cyber attack techniques on social media.

6.1. Information Diffusion and Influence

Castillo et al. showed different categories of news (news vs. in-depth) will have different life cycles regarding social media reactions [15]. In fact, there are many works aiming at observing and predicting the final cascade size of a given post or topic. According to Cheng et al., these factors may include content features, authors features, resharer features, and temporal features which make a cascade size more predictable [16]. As to the last issue, several papers exploit the reaction for a given fixed time frame to predict whether a post thread will be popular or not [17–19].

Cascade can also be interpreted from the audience’s perspective, i.e., why people spend lots of time on social media to share her own opinions to public. Marwick et al. showed a many-to-many communication model through which individuals conceptualize an imagined audience evoked through their content [20]. Hall et al. [21] demonstrated the impact of the individual on an information cascade. In the interactive communication model [22], in order to participate in the so-called attention economy, people want to attract “eyeballs” in a media-saturated, information-rich world and to influence audiences to like their comments/photos [23–25]. Hence, users strategically formulate their profile and participate in many discussion groups to increase attention.

6.2. Cyber attack Analysis on Social Media

Although increasing importance has been attached to security issues on social media, most works focus on pursuing a perfect classifier to detect malicious accounts or users with commonly used profile characteristics such as age, number of followers, geo-location, and total number of activities [26,27]. With the development of new security threats on social media such as cyberbullying or fake news,
recent research uses social science to understand collective human behavior. Cheng et al. studied the activity differences between organized groups and individuals [16]. Charzakou et al. noted people who spread hate are more engaged than typical users [28]. Vosoughi et al. found that false news was more novel than true news mainly because of humans, not robots [4].

7. Conclusions

In this paper, we describe our work regarding attacker intention and influence from large-scale malicious URLs campaigns using the public Facebook Discussion Groups dataset. Specifically, we focus on examining the differing characteristics between CNN and FOX News discussion threads ranging from 2014 to 2016.

We describe how social recommendation systems work for both target and non-target threads. Moreover, we define an Influence Ratio (IR) for every visible comment on Facebook based on the ratio between the upcoming activities and the preceding activities. We also propose a context-free prediction system to predict whether the trends will decrease or increase with a F1-score over 75%. From these results, we perform an in-depth analysis on different categories of malicious campaigns. Compared to those comments embedded with more critical level threats such as malicious URLs, some lower level threats, such as advertising or commercial shopping URLs, appeared at the very end of the discussion thread. The IR for those commercial sites for at least two reasons. (1) People just ignored those since they already know it only hinders the readability. (2) People do not want to check those posts anymore. However, the program bot did not update to the new-coming information.

The initial results we obtained provide us new insight regarding how malicious URLs influence both post thread life cycle and audience activities with the Facebook social recommendation algorithm. Our current observations enables us to reconsider new response strategies in handling inappropriate information on social media.

Author Contributions: Conceptualization: C.-M.L. and H.-J.S.; Writing—original draft: C.-M.L. and J.C.; and Writing—review and editing C.-M.L. and J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References


18. Ma, Z.; Sun, A.; Cong, G. On predicting the popularity of newly emerging hashtags in twitter. *J. Assoc. Inf. Sci. Technol.* 2013, 64, 1399–1410. [CrossRef]


**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.