Hawkes Process for Understanding the Influence of Pathogenic Social Media Accounts

Hamidreza Alvari, Paulo Shakarian

Arizona State University
Tempe, USA
{halvari, shak}@asu.edu

Abstract—Over the past years, political events and public opinion on the Web have been allegedly manipulated by accounts dedicated to spreading disinformation and performing malicious activities on social media. These accounts hereafter referred to as “Pathogenic Social Media (PSM)” accounts, are often controlled by terrorist supporters, water armies or fake news writers and hence can pose threats to social media and general public. Understanding and analyzing PSMs could help social media firms devise sophisticated and automated techniques that could be deployed to stop them from reaching their audience and consequently reduce their threat. In this paper, we leverage the well-known statistical technique “Hawkes Process” to quantify the influence of PSM accounts on the dissemination of malicious information on social media platforms. Our findings on a real-world ISIS-related dataset from Twitter indicate that PSMs are significantly different from regular users in making a message viral. Specifically, we observed that PSMs do not usually post URLs from mainstream news sources. Instead, their tweets usually receive large impact on audience, if contained URLs from Facebook and alternative news outlets. In contrary, tweets posted by regular users receive nearly equal impression regardless of the posted URLs and their sources. Our findings can further shed light on understanding and detecting PSM accounts.

Index Terms—Pathogenic Social Media Accounts, Malicious Activity, Misinformation, Hawkes Processes

I. INTRODUCTION

Online social media play major role in dissemination of information. However, recent years have witnessed evidence of spreading huge amount of harmful disinformation on social media and manipulating public opinion on the Web, attributed to accounts dedicated to spreading malicious information. These accounts are referred to “Pathogenic Social Media” (PSM) accounts and can pose threats to social media firms and general public. PSMs are usually controlled by terrorist supporters, water armies or fake news writers and they are owned by either real users or bots who seek to promote or degrade certain ideas by utilizing large online communities of supporters to reach their goals. Identifying PSM accounts could have immediate applications, including countering terrorism [1], [2], fake news detection [3], [4] and water armies detection [5], [6].

To better understand the behavior and impact of PSM accounts on the Web and normal users, and be able to counter their malicious activity, social media authorities need to deploy certain capabilities which could ultimately lead to reducing their threats. To make this happen, social media platforms are required to design sophisticated techniques that could automatically detect and suspend these accounts as quickly as possible, before they can reach their vast audience and spread malicious content. However, for the most part, the social media firms usually rely on reports they receive from their normal users or even their assigned teams to manually shut down these accounts. First of all, this mechanism is not always feasible due to the limited manpower and since not many real users are willing to put aside time and report the malicious activities. Also, it cannot be done in a timely manner since it takes time for these firms to review the reports and decide whether they are legit or not. On the other hand, the fact that these accounts simply return to social media using different accounts or even migrate to other social media, makes all these efforts almost useless. Therefore, the burden falls to automatic approaches that can identify these malicious actors on social media.

Present work. In this work, to address the above mentioned challenges, we aim to understand PSM accounts by (1) analyzing their behavior in terms of their posted URLs, and (2) estimate their influence by conducting experiments on a real-world dataset from Twitter. We deploy a mathematical technique known as “Hawkes process” [7] to quantify the impact of PSMs on normal users and the greater Web, by looking at their posted URLs on Twitter. Hawkes processes are special forms of point processes and have shown promising results in many problems that require modeling complicated event sequences where historical events have impact on future ones, including financial analysis [8], seismic analysis [9] and social network modeling [10] to name a few. This study uses an ISIS-related dataset from Twitter [11]. The dataset contains an action log of users in the form of cascades of retweets. In this work, we consider URLs posted by two groups of users: (1) PSM accounts and (2) normal users. The URLs can belong to any platform including the major social media (e.g., facebook.com), mainstream news (e.g., nytimes.com) and alternative news outlets (e.g., rt.com). For each group of users, we fit a multi-dimensional Hawkes processes model wherein each process correspond to a platform referenced in at least one tweet. Furthermore, every process can influence all the others including itself, which allows estimating the strength of connections between each of the social media platforms and news sources, in terms of how likely an event (i.e., the posted URL) can cause subsequent events in each of the groups. In other words, in this study we are interested to investigate if a given URL $u_1$ has influence on another URL $u_2$ (i.e.,
Main Findings. This paper makes the following main observations:

- Among all platforms studied here, URLs shared from facebook.com and alternative news media contributed the most to the dissemination of malicious information from PSM accounts. Simply put, they had the largest impact on making a message viral and causing the subsequent events.
- Posts that were tweeted by the PSM accounts and contained URLs from facebook.com, demonstrated more influence on the subsequent retweets containing URLs from youtube.com, in contrary to the other way around. This means that ultimately tweets with URLs from Facebook will high likely end up inducing more external impulse on YouTube than YouTube might have on Facebook.
- URLs posted by the normal users have nearly the same impact on the subsequent events regardless of the social media or news outlet used. This basically means that normal users do not often prefer specific social media or news sources over the others.

II. DATA ANALYSIS

In this section, we first explain the dataset and provide the list of the main social media platforms and news sites used in this study. Finally, we present our data analysis to demonstrate the differences between PSM accounts and normal users.

A. Dataset

We collect a dataset of 2.8M ISIS related tweets/retweets in Arabic between February 22, 2016 and May 27, 2016. The dataset contains different fields including user ID, retweet ID, hashtags, content, posting time. The dataset also contains user profile information including name, number of followers/followees, description, location, etc. The tweets were collected using different ISIS-related hashtags such as #stateoftheislamiccaliphate. In this dataset, about 600K tweets have at least one URL (i.e., event) referencing one of the social media platforms or news outlets. There are about 1.4M of paired URLs which we denote by u₁ → u₂ and indicates a retweet (with the URL u₂) of the original tweet (with the URL u₁).

In this study, we are interested in investigating the impact of the URL u₁ on u₂. Accordingly, the dataset contains 35K cascades (i.e., sequences of events) of different sizes and durations, some of which contain paired URLs in the aforementioned form. After pre-processing and removing duplicate users from cascades (those who retweet themselves multiple times), cascades sizes (i.e. number of associated postings) vary between 20 to 9,571 and take from 10 seconds to 95 days to finish. The log-log distribution of cascades vs. cascade size and the cumulative distribution of duration of cascades are depicted in Figure 1.

The statistics of the dataset are presented in Table I. For labeling, we check through Twitter API to examine whether the users have been suspended (labeled as PSM) or they are still active (labeled as normal). According to Table I, 11% of the users in our dataset are PSMs and others are normal. We also depict the total number of PSM accounts that have been suspended by Twitter in each cascade, in Figure 2. Finally, to reiterate, we note that these accounts mostly get suspended manually by Twitter based on reports the platform receives from its own users.

\[^1\text{https://blog.twitter.com/official/en_us/a/2016/an-update-on-our-efforts-to-combat-violent-extremism.html}\]

| Name                  | Value         |
|-----------------------|---------------|
| # of Cascades         | 35K           |
| # of Tweets/Retweets  | 2.8M          |
| # of Users            | 64,484       | 536,609    |
| # of Single URLs      | 104,948       | 536,046    |
| # of Paired URLs      | 200,892       | 1,123,434  |

**TABLE I**

DESCRIPTION OF THE DATASET.

![Image of graphs showing log-log distribution of cascades vs. cascade size and cumulative distribution of duration of cascades.]

![Image of graph showing total number of inactive users in each cascade.]

---

[1] The statistics of the dataset are presented in Table I. For labeling, we check through Twitter API to examine whether the users have been suspended (labeled as PSM) or they are...
TABLE II
SOCIAL MEDIA PLATFORM’S TOTAL NUMBER OF PAIRED URLs OF THE FORM u₁ → u₂ WITH AT LEAST ONE EVENT IN THE DATASET FOR THE PSM AND NORMAL USERS.

| Platform  | PSM     | Normal  |
|-----------|---------|---------|
| Twitter   | 139,940 | 918,803 |
| Facebook  | 4,017   | 878     |
| Instagram | 2,857   | 0       |
| Google    | 163     | 132     |
| Youtube   | 24,724  | 72,890  |

TABLE III
NEWS SOURCES’ TOTAL PAIRED URLs (u₁ → u₂) WITH AT LEAST ONE EVENT IN THE DATASET FOR THE PSM AND NORMAL USERS.

| News Source | PSM | Normal |
|-------------|-----|--------|
| Mainstream  | 0   | 286    |
| Alternatives| 35,187 | 124,449 |

B. Social Media Platforms and News Outlets

Twitter deploys a URL shortener technique to leave more space for content and protect users from malicious sites. To obtain the original URLs, we use a URL unshortening tool to obtain the original links contained in the tweets in our dataset.

We consider a number of major and well-known social media platforms including Twitter, Facebook, Instagram, Google and Youtube. About the dichotomy of mainstream and alternative media, it is notable to mention that most criteria for determining whether a news source counts as either of them, are based on a number of factors including but not limited to the content and whether or not it is corporate owned. However, a key difference between these two sources of media comes from the fact that all of mainstream media is profit-oriented, in contrast to the alternative media. We further note that for the most part, mainstream media is considered as a more credible source than alternative media, although the reputation has been recently tainted by the fake news.

In this work, following the commonsense, we consider popular news outlets such as The New York Times, and The Wall Street Journal as mainstream and less popular ones as alternatives. In Table II, we summarize the total number of paired URLs (i.e., u₁ → u₂) in which the original URL (i.e., u₁) corresponds to each social media platform with at least one event in our dataset. We also summarize in Table III the total number of paired URLs whose original URL belongs to the mainstream and alternative news sources. In Table IV, we see the break down of number of paired URLs for the PSM and normal users. We further demonstrate in Table V some examples of the mainstream and alternative news URLs occurrence used in this work.

C. Temporal Analysis

Here, we present the differences between the PSM accounts in our dataset with their counterparts, normal users through

A. Hawkes Processes

In many scenarios, one needs to deal with timestamped events such as the activity of users on a social network recorded in continuous time. An important task then is to estimate the influence of the nodes based on their timestamp patterns. Point process is a principled framework for modeling such event data, where the dynamic of the point process can be captured by its conditional intensity function as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}(N(t + \Delta t) - N(t)|\mathcal{H}_t)}{\Delta t} = \frac{\mathbb{E}(dN(t)|\mathcal{H}_t)}{dt}$$ (1)

where \(\mathbb{E}(dN(t)|\mathcal{H}_t)\) is the expectation of the number of events happen in the interval \((t, t + dt)\) given the historical observations \(\mathcal{H}_t\) and \(N(t)\) records the number of events before time \(t\). Point process can be equivalently represented as a counting process \(N = \{N(t)|t \in [0, T]\}\) over the time interval \([0, T]\).
The Hawkes process framework [7] has been used in many problems that require modeling complicated event sequences where historical events have impact on future ones. Examples include but are not limited to financial analysis [8], seismic analysis [9] and social network modeling [10]. One-dimensional Hawkes process is a point process \( N_t \) with the following particular form of intensity function:

\[
\lambda(t) = \mu + a \int_{-\infty}^{t} g(t-s) dN_s = \mu + a \sum_{i: t_i < t} g(t-t_i) \tag{2}
\]

where \( \mu > 0 \) is the exogenous base intensity (i.e., background rate of events), \( t_i \) are the time of events in the point process before time \( t \), and \( g(t) \) is the decay kernel.

In this paper, we use exponential kernel of the form \( g(t) = e^{-\omega t} \), but adapting to the other positive forms is straightforward. The second part of the above formulation captures the self-exciting nature of the point processes—the occurrence of events in the past has a positive impact on the future ones. Given a sequence of events \( \{t_i\}_{i=1}^{n} \) observed in \([0,T]\) and generated from the above intensity function, the log-likelihood function can be obtained as follows [10]:

\[
\mathcal{L} = \log \frac{\prod_{i=1}^{n} \lambda(t_i)}{\exp \int_{0}^{T} \lambda(t) dt} = \sum_{i=1}^{n} \log \lambda(t_i) - \int_{0}^{T} \lambda(t) dt \tag{3}
\]

In this paper, we focus on multi-dimensional Hawkes processes which is defined by a \( U \)-dimensional point process \( N^u_t, u = 1, \ldots, U \). In other words, we have \( U \) Hawkes processes coupled with each other—each Hawkes process correspond to one of the platforms and the influence between them is modeled using the mutually-exciting property of the multi-dimensional Hawkes processes. We formally define the following formulation to model the influence of different events on each other:

\[
\lambda_u(t) = \mu_u + \sum_{i: t_i < t} a_{uu_i} g(t-t_i) \tag{4}
\]

where \( \mu_u \geq 0 \) is the base intensity for the \( u \)-th Hawkes process. The coefficient \( a_{uu_i} \geq 0 \) captures the mutually-exciting property between the \( u \)-th and \( u_i \)-th processes. Larger value of \( a_{uu_i} \) shows that events in the \( u_i \)-th dimension are more likely to trigger an event in \( u \)-th dimension in future. More intuitively, an event on one point process can cause an impulse response on other processes, which increases the probability of an event occurring above the processes’ background rates. We reiterate that in this study each URL is attributed to an event, i.e., if the URL \( u_1 \) triggers the URL \( u_2 \) (i.e., \( u_1 \rightarrow u_2 \)), then \( a_{uu_{1u_2}} \geq 0 \).

In Figure 4 we depict a multivariate example of three different streams of events, \( e_0, e_1 \) and \( e_2 \). As illustrated, \( e_0 \) is caused by the background rate \( \lambda(0) \) and has an influence on itself and \( e_1 \). On the other hand, \( e_1 \) is caused by \( \lambda(t_1) \) and has an influence on \( e_2 \). Simply put, a background event in \( e_0 \) induces impulse on responses on processes \( e_1 \) and \( e_2 \). Accordingly, the caused child event in \( e_1 \) leads to another child event in \( e_2 \).

Table IV: Total number of paired URLs of the form \( u_1 \rightarrow u_2 \) with at least one event for PSM/normal users and for all platforms.

| URL A | URL B | Twitter → | Facebook → | Instagram → | Google → | Youtube → | Mainstream → | Alternatives → |
|-------|-------|-----------|------------|-------------|----------|-----------|--------------|----------------|
|       |       | 109,354/676,617 | 598,343/229,246 | 120,382/611,992/59,889 | 90/688 | 17,557/84,923 | 0/25 | 0/189 |
|       |       | 655/3,108 | 4/11 | 3/0 | 2/1 | 87/281 | 0/1 | 127/576 |
|       |       | 0/2,362 | 0/11 | 0/25 | 0/2 | 0/161 | 0/2 | 0/294 |
|       |       | 134/74 | 0/0 | 0/0 | 0/0 | 12/53 | 0/0 | 17/4 |
|       |       | 14,004/56,545 | 132/312 | 23/211 | 22/32 | 23/52 | 13/48 | 3/31/8,213 |
|       |       | 0/189 | 0/1 | 0/1 | 0/0 | 0/0 | 26/122 | 0/1 |
|       |       | 21,047/95,641 | 145/767 | 45/318 | 59/64 | 3,862/91,199 | 2/1 | 10,003/18,338 |

Fig. 4. Illustration of the Hawkes Process. Events induce impulse on other processes and cause child events. Background event in \( e_0 \) induces impulse on responses on processes \( e_1 \) and \( e_2 \).
B. Methodology

We aim to assess the influence of the PSM accounts in our dataset via their posted URLs. We consider the URLs posted by two groups of users: (1) PSM accounts and (2) normal users. For both groups, we fit a Hawkes model with $K = 7$ point processes each for the seven categories of social media and news outlets discussed earlier. In each of the Hawkes models, every process is able to influence all the others including itself, which allows us to estimate the strength of connections between each of the seven categories for both groups of users, in terms of how likely an event (i.e., the posted URL) can cause subsequent events in each of the groups.

We use the ADM4 algorithm presented by [10] and follow the methodology presented by [14] for fitting the Hawkes processes for both PSM and normal users. ADM4 [10] is an efficient optimization that estimates the parameters $\mathbf{A}$ and $\boldsymbol{\mu}$ by maximizing the regularized log-likelihood $L(A, \mu)$:

$$
\min_{\mathbf{A} \geq 0, \mu \geq 0} -L(\mathbf{A}, \mu) + \lambda_1 ||\mathbf{A}||_*, + \lambda_2 ||\mathbf{A}||_1 \tag{5}
$$

where $L(\mathbf{A}, \mu)$ can be obtained by substituting $\lambda_u(t)$ from Equation 4 into Equation 3. Also, $||\mathbf{A}||_*$ is the nuclear norm of matrix $\mathbf{A}$, and is defined as the sum of its singular value.

We consider two different sets of URLs posted by the PSM accounts and normal users by selecting URLs that have at least one event in Twitter (i.e., posted by a user). For each group, we construct a matrix $\mathbf{W} \in \mathbb{N}^{T \times U}$ with $U = 7$, whose entries are sequences of events (i.e., posted URLs) observed during a time period $T$. We note that each sequence of events is of the form $S = \{(t_i, u_i)\}_{i=1}^{n_i}$ where $n_i$ is the number of the events occurring at the $u_i$-th dimension (i.e., URLs posted containing one of the 7 platforms).

IV. Experimental Results

Here, we conduct experiments to gauge the effectiveness of Hawkes process for modeling influence of PSMs.

A. Settings

In this work, we adopt the ADM4 algorithm [10] which implements parametric inference for Hawkes processes with an exponential kernel and a mix of Lasso and nuclear regularization. We initialize infectivity matrix $\mathbf{A}$, base intensities $\boldsymbol{\mu}$ and decays $\beta \in \mathbb{R}$ randomly.

We further set the number of nodes $U = 7$ to reflect the 7 platforms used in this study. Level of penalization is set to $C = 1000$, and the ratio of Lasso-Nuclear regularization mixing parameter is set to 0.5. Finally, maximum number of iterations for solving the optimization is set to 50 and the tolerance of solving algorithm is set to $10^{-5}$.

B. Results

We estimate infectivity matrix for both PSM and normal users by fitting the Hawkes model described earlier. In our study, this matrix characterizes the strength of the connections between the platforms and news sources. More specifically, each weight value represents the connection strength from one platform to another. In other words, each entry in this matrix can be interpreted as the expected number of subsequent events that will occur on the second group after each event on the first $[14]$. In Figure 5 we depict the estimated weights for all paired URLs for both PSM and normal users. Looking at the weights in both of the plots, we realize that greater weights belong to processes that have impact on Twitter, i.e. “→ Twitter”. This implies that both of the groups in our Twitter dataset often post URLs that ultimately have greater impact on Twitter.

Overall, we observe the followings:

- URLs referencing all platforms and posted by the PSMs and regular users, mostly trigger URLs that contain the Twitter domain.
- Among all platforms studied here, URLs shared from facebook.com and alternative news media contributed the most to the dissemination of malicious information from PSM accounts. In other words, they had largest impact on making a message viral and causing the subsequent events.
- Posts that were tweeted by the PSM accounts and contained URLs from facebook.com, demonstrated more influence on the subsequent retweets containing URLs from youtube.com, in contrary to the other way around. This means that ultimately tweets with URLs from facebook will likely end up inducing external impulse on youtube.com. In contrast, URLs posted by the normal users have nearly the same impact on the subsequent events regardless of the social media or news outlet used.

The above mentioned observations demonstrate the effectiveness of leveraging Hawkes process to quantify the impact of URLs posted by PSMs and regular users on the dissemination of content on Twitter. The observations we make here show that PSM accounts and regular users behave differently in terms of the URLs they post on Twitter, in that they have different tastes while disseminating URL links. Accordingly their impact on the subsequent events significantly differ from each other.

V. Related Work

The explosive growth of the Web has raised numerous challenges and attracted several researchers with different background [11], [15]–[29]. Below, we will review some of the closest research directions to our research.

Point Process. When dealing with timestamped events in continuous time such as the activity of users on social media, point process could be leveraged for modeling such events. Point processes have been extensively used to model activities in networks [30]. Hawkes process is a special form of point processes which models complicated event sequences with historical events influencing future ones. Hawkes processes have been applied to a variety of problems including financial analysis [8], seismic analysis [9] and social network modeling [10], community detection [31], and causal inference [32]. Social Spam/Bot Detection. Social bot is a computer program that automatically generate content and interacts with real
people on social media, trying to emulate and possibly alter their behavior [15]. Recently, DARPA organized a Twitter bot challenge to detect “influence bots” [33], where supervised and semi-supervised approaches were proposed using different features. The work of [34] for example, use similarity to cluster accounts and uncover groups of malicious users. The work of [35] presents a supervised framework for bot detection which uses more than thousands features. In a different attempt, the work of [36] studied the problem of spam detection in Wikipedia using different spammers behavioral features. For a comprehensive survey on the ongoing efforts to fight social bots, we direct the reader to [15].

Fake News Identification. Fake news detection has recently attracted a growing interest of general public and researchers, as the spread of misinformation on social media and the Web increases on a daily basis. A growing body of work has been devoted to addressing the impact of bots in manipulating political discussion and spreading fake news, including the 2016 U.S. presidential election [19], [37], [38] and the 2017 French election [39]. For example, [38] analyzes tweets following recent U.S. presidential election and found evidences that bots played key roles in spreading fake news.

Identifying Instigators. Given a snapshot of the diffusion process at a given time, these works aim to detect the source of the diffusion. For instance, [40] designed an approach for information source detection and in particular initiator of a cascade. In contrast, we are focused on a set of users who might or might not be initiators. Other similar works on finding most influential spreaders of information such as [41], [42] and outbreak prediction such as [43] also exist in the literature. For example, the work of [44] performed classification to detect users who adopt popular items.

Extremism Detection. Several studies have focused on understanding extremism in social networks [2], [22], [28], [45]–[48]. The work of [2], uses Twitter and proposes an approach to predict new extremists, determine if the newly created account belongs to a suspended extremist, and predict the ego-network of the suspended extremist upon creating her new account. Authors in [22], [45] performed iterative vertex clustering and classification to identify Islamic Jihadists on Twitter.

Detection of Internet Water Armies. The term “Internet water armies” refers to a special group of online users who get paid for posting comments for some hidden purposes such as influencing other users towards social events or business markets. Therefore, they are also called “hidden paid posters”. The works of [5], [6], [25] use user behavioral and domain-specific attributes and designed approaches to detect Internet water armies.

VI. CONCLUSION AND FUTURE WORK

In this study, we presented an analysis on a real-world ISIS-related dataset from Twitter to demonstrate how Pathogenic Social Media (PSM) and normal users usually post on Twitter in terms of the URLs they post. More specifically, we leveraged a statistical technique known as Hawkes Process for modeling the influence of PSM accounts on dissemination of malicious content on the Web. In this work, we used URLs posted by two groups of users, PSMs and normal users, on major social media and mainstream and alternative news outlets. Overall, our findings indicate that the URLs posted by the PSM accounts have the largest impact if contained either facebook.com or alternative news media. In contrast, their counterparts, i.e., normal users, often post URLs that have nearly the same impact on the Web, no matter what social media or news outlet they use. There are potential avenues for future work. First, we would like to extend the study by proposing a prediction mechanism for distinguishing PSMs from normal users, based on their different impact on the greater Web. Another research direction would also be learning causal inference for Hawkes process and investigating the relation between the two concepts, while considering the problem of identification of PSM accounts.

ACKNOWLEDGMENT

This work was supported through DoD Minerva program.

REFERENCES

[1] M. Khader, Combating Violent Extremism and Radicalization in the Digital Era, ser. Advances in Religious and Cultural Studies. IGI Global, 2016.

[2] J. Klausen, C. Marks, and T. Zaman, “Finding online extremists in social networks,” CoRR, vol. abs/1610.06242, 2016.
