RESEARCH ARTICLE

Modeling the spread of fake news on Twitter

Taichi Murayama[1], **Shoko Wakamiya**[1], **Eiji Aramaki**[1], **Ryota Kobayashi**[2,3]*

1 Nara Institute of Science and Technology (NAIST), Ikoma, Japan, **2** The University of Tokyo, Tokyo, Japan, **3** JST PRESTO, Kawaguchi, Japan

* r-koba@k.u-tokyo.ac.jp

Abstract

Fake news can have a significant negative impact on society because of the growing use of mobile devices and the worldwide increase in Internet access. It is therefore essential to develop a simple mathematical model to understand the online dissemination of fake news. In this study, we propose a point process model of the spread of fake news on Twitter. The proposed model describes the spread of a fake news item as a two-stage process: initially, fake news spreads as a piece of ordinary news; then, when most users start recognizing the falsity of the news item, that itself spreads as another news story. We validate this model using two datasets of fake news items spread on Twitter. We show that the proposed model is superior to the current state-of-the-art methods in accurately predicting the evolution of the spread of a fake news item. Moreover, a text analysis suggests that our model appropriately infers the correction time, i.e., the moment when Twitter users start realizing the falsity of the news item. The proposed model contributes to understanding the dynamics of the spread of fake news on social media. Its ability to extract a compact representation of the spreading pattern could be useful in the detection and mitigation of fake news.

Introduction

As smartphones become widespread, people are increasingly seeking and consuming news from social media rather than from the traditional media (e.g., newspapers and TV). Social media has enabled us to share various types of information and to discuss it with other readers. However, it also seems to have become a hotbed of fake news with potentially negative influences on society. For example, Carvalho et al. [1] found that a false report of United Airlines parent company's bankruptcy in 2008 caused the company's stock price to drop by 76% in a few minutes; it closed at 11% below the previous day's close, with a negative effect persisting for more than six days. In the field of politics, Bovet and Makse [2] found that 25% of the news outlets linked from tweets before the 2016 U.S. presidential election were either fake or extremely biased, and their causal analysis suggests that the activities of Trump's supporters influenced the activities of the top fake news spreaders. In addition to stock markets and elections, fake news has emerged for other events, including natural disasters such as the East Japan Great Earthquake in 2011 [3, 4], often facilitating widespread panic or criminal activities [5].

Competing interests: The authors have declared that no competing interests exist.

In this study, we investigate the question of how fake news spreads on Twitter. This question is relevant to an important research question in social science: how does unreliable information or a rumor diffuses in society? It also has practical implications for fake news detection and mitigation [6, 7]. Previous studies mainly focused on the path taken by fake news items as they spread on social networks [8, 9], which clarified the structural aspects of the spread. However, little is known about the temporal or dynamic aspects of how fake news spreads online.

Here we focus on Twitter and assume that fake news spreads as a two-stage process. In the first stage, a fake news item spreads as an ordinary news story. The second stage occurs after a correction time when most users realize the falsity of the news story. Then, the information regarding that falsehood spreads as another news story. We formulate this assumption by extending the Time-Dependent Hawkes process (TiDeH) [10], a state-of-the-art model for predicting re-sharing dynamics on Twitter. To validate the proposed model, we compiled two datasets of fake news items from Twitter.

The contribution of this study is summarized as follows:

- We propose a simple point process model based on the assumption that fake news spreads as a two-stage process.

- We evaluate the predictive performance of the proposed model, which demonstrates the effectiveness of the model.

- We conduct a text mining analysis to validate the assumption of the proposed model.

Related work

Predicting future popularity of online content has been studied extensively [11, 12]. A standard approach for predicting popularity is to apply a machine learning framework, such that the prediction problem can be formulated as a classification [13, 14] or regression [15] task. Another approach to the prediction problem is to develop a temporal model and fit the model parameters using a training dataset. This approach consists of two types of models: time series and point process models. A time series model describes the number of posts in a fixed window. For example, Matsubara et al. [16] proposed SpikeM to reproduce temporal activities on blogs, Google Trends, and Twitter. In addition, Proskurnia et al. [17] proposed a time series model that considers a promotion effect (e.g., promotion through social media and the front page of the petition site) to predict the popularity dynamics of an online petition. A point process model describes the posted times in a probabilistic way by incorporating the self-exciting nature of information spreading [18, 19]. Point process models have also motivated theoretical studies about the effect of a network structure and event times on the diffusion dynamics [20]. Various point process models have been proposed for predicting the final number of re-shares [19, 21] and their temporal pattern [10] on social media. Furthermore, these models have been applied to interpret the endogenous and exogenous shocks to the activity on YouTube [22] and Twitter [23]. To the best of our knowledge, the proposed model is the first model incorporating a two-stage process that is an essential characteristic of the spread of fake news. Although some studies [24] proposed a model for the spread of fake news, they focused on modeling the qualitative aspects and did not evaluate prediction performances using a real data set.

Our contribution is related to the study of fake news detection. There have been numerous attempts to detect fake news and rumors automatically [6, 7]. Typically, fake news is detected based on the textual content. For instance, Hassan et al. [25] extracted multiple categories of features from the sentences and applied a support vector machine classifier to detect fake
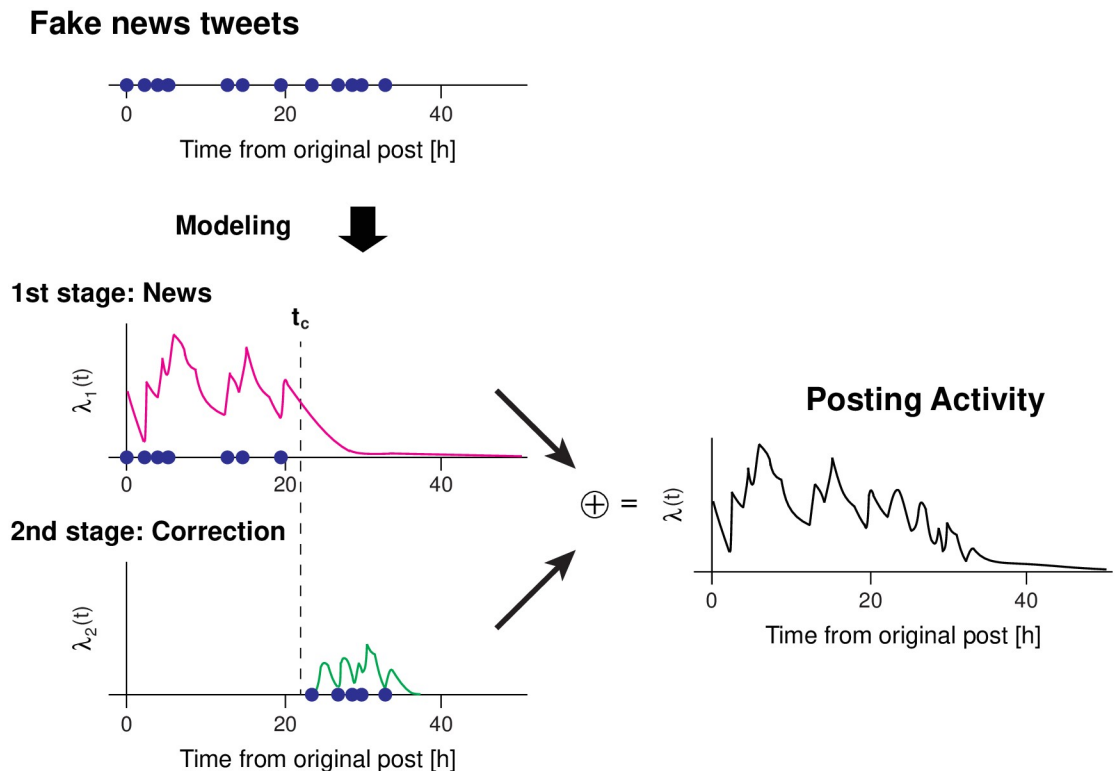
Fake news tweets

Fig 1. Schematic of the proposed model. We propose a model that describes how posts or re-shares that are related to a fake news item spread on social media (Fake news tweets). Blue circles represent the time stamp of the tweets. The proposed model assumes that the information spread is described as a two-stage process. Initially, a fake news item spreads as a novel news story (1st stage). After a correction time t_c, Twitter users recognize the falsity of the news item. Then, the information that the original news item is false spreads as another news story (2nd stage). The posting activity related to the fake news $\lambda(t)$ (right: black) is given by the summation of the activity of the two stages (left: magenta and green).

https://doi.org/10.1371/journal.pone.0250419.g001

news. Rashkin et al. [26] developed a long short-term memory (LSTM) neural network model for the fact-checking of news. The temporal information of a cascade, e.g., timings of posts and re-shares triggered by a news story, might improve fake news detection performance. Kwon et al. [27] showed that temporal information improves rumor classification performance. It has also been shown that temporal information improves the fake news detection performance [28], rumor stance classification [29], source identification of misinformation [30], and detection of fake retweeting accounts [31]. A deep neural network model [28] can also incorporate temporal information to improve the fake news detection performance. However, a limitation of the neural network model is that it can utilize only a part of the temporal information and cannot handle cascades with many user responses. The proposed model parameters can be used as a compact representation of temporal information, which helps us overcome this limitation.

Modeling the information spread of fake news

We develop a point process model for describing the dynamics of the spread of a fake news item. A schematic of the proposed model is shown in Fig 1. The proposed model is based on the following two assumptions.

- Users do not know the falsity of a news item in the early stage. The fake news spreads as an ordinary news story (Fig 1: 1st stage).

- Users recognize the falsity of the news item around a correction time t_c. The information that the original news is fake spreads as another news story (Fig 1: 2nd stage).

In other words, the proposed model assumes that the spread of a fake news item consists of two cascades: 1) the cascade of the original news story and 2) the cascade asserting the falsity of the news story. In this study, we use the term *cascade* meaning tweets or retweets triggered by a piece of information. To describe each cascade, we use the Time-Dependent Hawkes process model, which properly considers the circadian nature of the users and the aging of information.

Time-Dependent Hawkes process (TiDeH): Model of a single cascade

We describe a point process model of a single cascade: the information spreading triggered by a news story. In point process models [32], the probability of obtaining a post or reshare in a small time interval $[t, t + \Delta t]$ is written as $\lambda(t)\Delta t$, where $\lambda(t)$ is the instantaneous rate of the cascade, that is, the intensity function. The intensity function of the TiDeH model [10] depends on the previous posts in the following manner:

$$\lambda_{\text{TiDeH}}(t) = p(t)h(t), \tag{1}$$

and the memory function $h(t)$ is defined as follows:

$$h(t) = \sum_{i:t_i<t} d_i \phi(t - t_i), \tag{2}$$

where $p(t)$ is the infection rate, t_i is the time of the i-th post, and d_i is the number of followers of the i-th post. The infection rate $p(t)$ incorporates two main properties in the cascade: the circadian rhythm and decay owing to the aging of information

$$p(t) = a\left\{1 - r\,\sin\left(\frac{2\pi}{T_m}(t + \theta_0)\right)\right\}e^{-(t-t_0)/\tau},$$

where the time of the original post is assumed to be $t_0 = 0$ and $T_m = 24$ hours is the period of oscillation. The parameters, a, r, θ_0, and τ, correspond to the intensity, the relative amplitude, the phase of the oscillation, and the time constant of decay, respectively. The memory kernel $\phi(t)$ represents the probability distribution for the reaction time of a follower. A heavy-tailed distribution was adopted for the memory kernel [10, 19]

$$\phi(s) = \begin{cases} c_0 & (0 \leqq s \leqq s_0) \\ c_0(s/s_0)^{-(1+\gamma)} & (\text{Otherwise}) \end{cases}$$

The parameters were set to $c_0 = 6.94 \times 10^{-4}$ (/seconds), $s_0 = 300$ seconds, and $\gamma = 0.242$.

Proposed model of the spread of fake news

We formulate a point process model for the spread of a fake new item. Let us assumes that the spread consists of two cascades, namely, the one owing to the original news item and the other owing to the correction of the news item. The activity of the fake news cascade can be written as the sum of two cascades using TiDeH

$$\lambda_{\text{prop}}(t) = p_1(t)h_1(t) + p_2(t)h_2(t). \tag{3}$$

The first term $p_1(t)h_1(t)$ represents the rate of the cascade caused by the original news item.

$$p_1(t) = a_1 \left\{ 1 + r \, \sin\left(\frac{2\pi}{T_m}(t + \theta_0)\right) \right\} e^{-t/\tau_1}, \quad h_1(t) = \sum_{i:t_i < \min(t,t_c)} d_i \phi(t - t_i), \qquad (4)$$

where a_1 represents the impact of the original news item on the spreading, τ_1 is the decay time constant, $\min(t, t_c)$ represents the smaller of the two values (t or t_c), and t_c is the correction time of the fake news item. The second term $p_2(t)h_2(t)$ represents the cascade induced by the correction.

$$p_2(t) = a_2 \left\{ 1 + r \, \sin\left(\frac{2\pi}{T_m}(t + \theta_0)\right) \right\} e^{-(t-t_c)/\tau_2}, \quad h_2(t) = \sum_{i:t_c < t_i < t} d_i \phi(t - t_i), \qquad (5)$$

where a_2 represents the impact of the falsity of the news on the spreading, and τ_2 is the decay time constant. It is assumed that the circadian parameters of $p_2(t)$ are the same as those of $p_1(t)$. Mathematically, the proposed model includes TiDeH as a special case. Let us consider the proposed model that satisfies the following conditions

$$\tilde{a} = a_1 = a_2 e^{-t_c/\tilde{\tau}}, \quad \tilde{\tau} = \tau_1 = \tau_2. \qquad (6)$$

We can see that the proposed model is equivalent to TiDeH (with parameters $a = \tilde{a}$ and $\tau = \tilde{\tau}$) by substituting Eq (6) into Eqs (3), (4) and (5).

Parameter fitting

Here, we describe the procedure for fitting the parameters from the event time series (e.g., the tweeted times). Seven parameters $\{a_1, \tau_1; a_2, \tau_2; r, \theta_0; t_c\}$ were determined by maximizing the log-likelihood function

$$l = \sum_i \log \lambda(t_i) - \int_0^{T_{obs}} \lambda(s)ds, \qquad (7)$$

where t_i is the i-th tweeted time, $\lambda(t)$ is the intensity given by Eq (3), and T_{obs} is the observation time. We first fix the correction time t_c and the other parameters are optimized using the Newton method [33], provided by Scipy [34], within a range of $12 < \tau_1, \tau_2 < 2T_{obs}$ (hours). The correction time is separately optimized using Brent's method [35] within a range of $0.1T_{obs} < t_c < 0.9T_{obs}$. The code for fitting parameters from the tweeted times is available in Github [36].

We validate the fitting procedure by applying synthetic data generated by the proposed model (Eq 3). Fig 2 shows the dependence of the estimation accuracy on the observation time T_{obs}. To evaluate the accuracy, we calculated the median and interquartile ranges of the estimates from 100 trials. The estimation error decreases as the observation time increases. The result suggests that this fitting procedure can reliably estimate the parameters for sufficiently long observations ($\geq$36 hours). The medians of the absolute relative errors obtained from 36 hours of synthetic data are 18%, 11%, 38%, 38%, and 10% for a_1, τ_1, a_2, τ_2, and t_c, respectively. The estimation accuracy of the second cascade parameters (a_2, τ_2) is worse than that of the first cascade parameters (a_1, τ_1). This seems to be caused by the insufficiency of the observed data. While the first cascade parameters are estimated from the entire data, the second cascade parameters are estimated from the observation data after the correction time t_c. Moreover, the model parameters are not identifiable [37, 38] in the case of $a_1 = a_2 e^{-t_c/\tau_2}$ and $\tau_1 = \tau_2$. Because the proposed model is equivalent to TiDeH ($a_2 = 0$, $t_c \geq T_{obs}$) in this case, other parameter sets can also reproduce the observed data. Fig 3 shows that the fitting procedure can estimate the parameters accurately except for the non-identifiable domain.
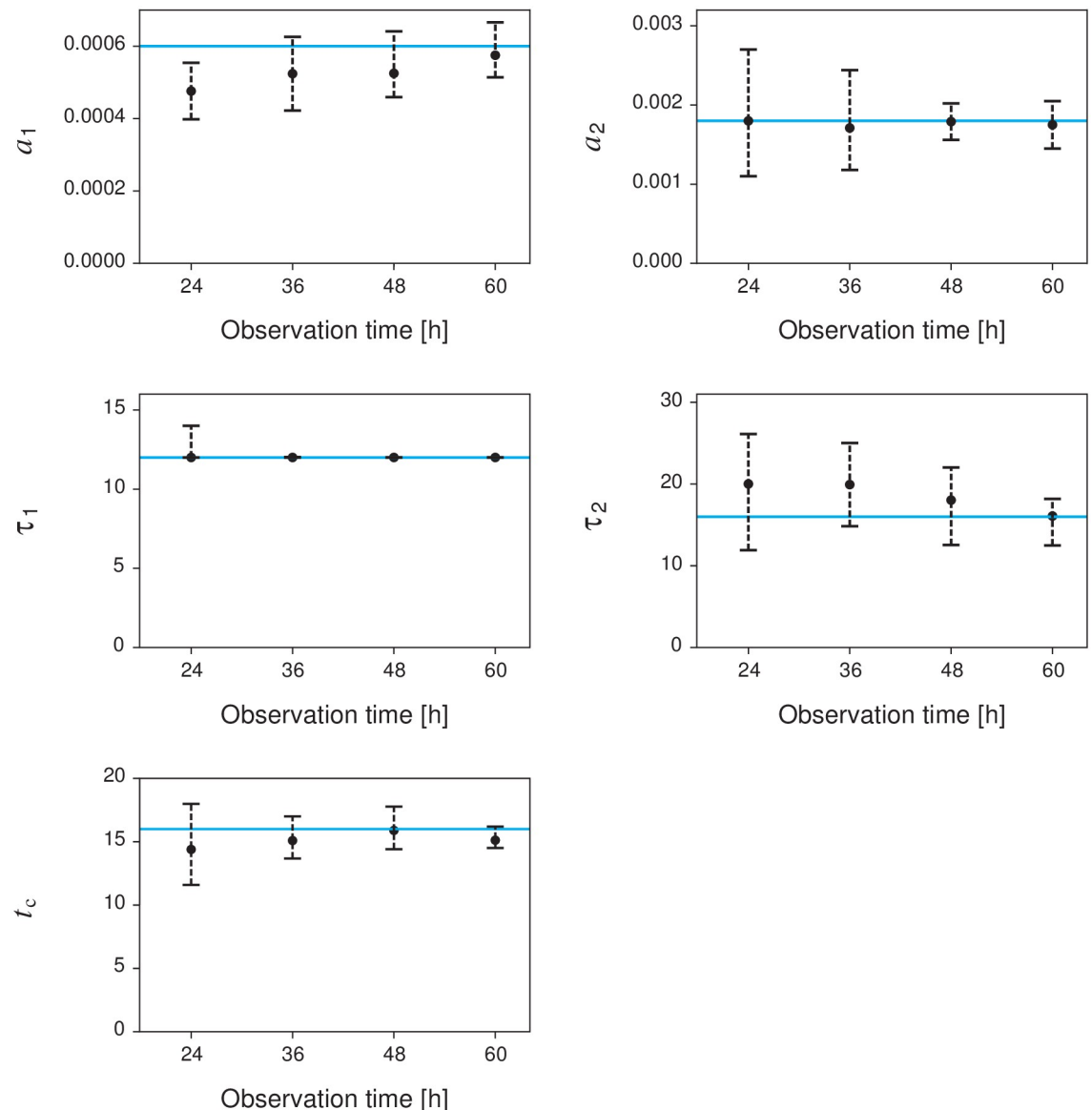
Fig 2. Dependence of the estimation accuracy of parameters $\{a_1, \tau_1; a_2, \tau_2; t_c\}$ on the observation time. Black circles and error bars represent the median and interquartile ranges of the estimates obtained from 100 synthetic data. Cyan lines indicate the true value: $a_1 = 0.0006$, $a_2 = 0.0018$, $\tau_1 = 12$, $\tau_2 = 16$, and $t_c = 16$.

https://doi.org/10.1371/journal.pone.0250419.g002

Dataset

We evaluate the proposed model and examine the correction time of fake news based on two datasets of the spread of fake news items. Datasets of the spread of fake news based on retweets of the original news post [39, 40] are publicly available. However, rather than a simple retweet, the information sharing of fake news can be complex. To cover the information spread in detail, we manually compiled two datasets of fake news items spread on Twitter. In our dataset, 61% and 20% of the tweets are retweets of original posts in the Recent Fake News dataset and the 2011 Tohoku Earthquake and Tsunami dataset, respectively.
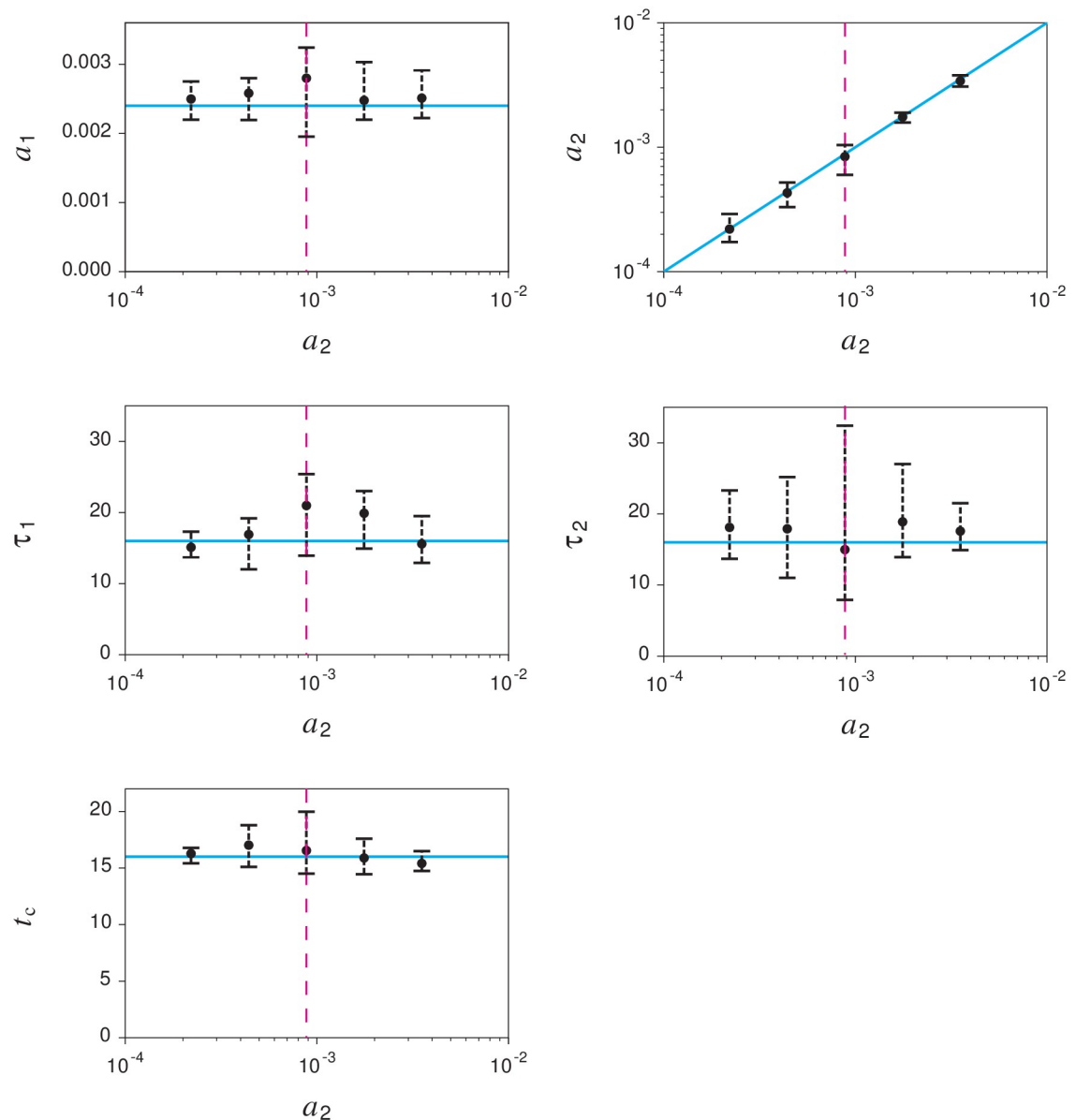
Fig 3. Estimation accuracy of parameters around the non-identifiable domain. Black circles and error bars represent the median and interquartile ranges of the estimates obtained from 100 synthetic data. Dashed magenta lines represent the non-identifiable domain satisfying $a_2 = a_1 e^{-t_c/\tau_2}$. Cyan lines indicate the true value: $a_1 = 0.0024$, $\tau_1 = \tau_2 = 16$, and $t_c = 16$, and a_2 is changed between 2.2×10^{-4} and 3.5×10^{-3}.

Recent Fake News (RFN)

We collected the spread of 10 fake news items from two fact-checking sites, Politifact.com [41] and Snopes.com [42] between March and May, in 2019. PolitiFact is an independent, non-partisan site for online fact-checking, mainly for U.S. political news and politicians' statements. Snopes.com, one of the first online fact-checking websites, handles political and other social and topical issues. Using the Twitter API, tweets highly relevant to the fake news stories were crawled based on the keywords and the URLs. We selected six fake news stories based on two conditions: 1) the number of posts must be greater than 300 and 2) the observation period

Table 1. Recent Fake News (RFN): Details of 6 U.S. fake news items.

News No.	Title	Date	No. Posts	T_{max}
a. Abolish	America came along as the first country to end (slavery) within 150 years.[a]	2019-03-21	1159	36
b. Notredame	A video clip from the Notre Dame cathedral fire shows a man walking alone in a tower of the church "dressed in Muslim garb."	2019-04-16	1641	132
c. Islamic	Did Ilhan Omar hold 'Secret Fundraisers with 'Islamic Groups Tied to Terror'?	2019-03-27	10811	130
d. Lionhunter	Was a trophy hunter eaten alive by lions after he killed 3 baboon families?	2019-03-25	25071	88
e. Newzealand	Did New Zealand take Fox News or Sky News off the air in response to mosque shooting coverage?	2019-03-25	11711	88
f. Sonictrans	Will the animated character of Sonic the Hedgehog be transgender in a new film?	2019-05-06	2319	132

[a]Verbatim quote from Katie Pavlich on Politifact.com, March 19, 2019.

https://doi.org/10.1371/journal.pone.0250419.t001

must be longer than 36 hours (as indicated by the experiments conducted on synthetic data, Fig 2). A summary of the collected fake news stories is presented in Table 1.

Fake news on the 2011 Tohoku earthquake and tsunami (Tohoku)

Numerous fake news stories emerged after the 2011 earthquake off the Pacific coast of Tohoku [3, 4]. We collected tweets posted in Japanese from March 12 to March 24, 2011, by using sample streams from the Twitter API. There were a total of 17,079,963 tweets. We first identified 80 fake news items based on a fake news verification article [43] and obtained the keywords and related URLs of the news items. Then, we extracted the tweets highly relevant to the fake news. Finally, we selected 19 fake news stories using the same conditions as in the RFN dataset. A summary of the collected fake news items is presented in Table 2.

Table 2. 2011 Tohoku earthquake and tsunami (Tohoku): Details of 19 Japanese fake news items.

News No.	Title	Date	No. Posts	T_{max}
a. Saveenergy	Large-scale power saving required even in the Kansai region.	2011-03-12	2846	174
b. EscapeTokyo	The bureaucracy in the Ministry of Defense says "You should escape from Tokyo"	2011-03-18	1056	92
c. Isodin	Isodin is effective against radiation.	2011-03-12	2421	118
d. Seaweed	Seaweed is effective against radiation.	2011-03-12	1798	118
e. Blog	The blog "I want you to know what a nuclear plant is."	2011-03-13	501	170
f. Hutaba	Officials in Hutaba hospital left patients behind and fled.	2011-03-17	1525	118
g. Remark1	Former chief cabinet secretary Sengoku's remark in Tokushima was inappropriate.	2011-03-13	638	170
h. Remark2	Former prime minister Hatoyama remarked "We cannot live within a 200-kilometer radius of the nuclear power plant."	2011-03-16	955	120
i. Visit	Chief Cabinet Secretary Edano visits Korea a few days after the earthquake.	2011-03-15	1973	168
j. Regulation	Ms. Renho proposes to regulate convenience stores to save energy.	2011-03-12	7561	156
k. Rescue	Ms. Tsujimoto protests U.S. military's rescue activities.	2011-03-16	1887	144
l. Taiwan	Taiwan's aid is rejected by the Japanese government.	2011-03-12	2736	156
m. School seismic	Budget for school seismic retrofitting was cut by the project screening.	2011-03-12	1044	174
n. Debt	South Korea asks Japan to borrow money. Moreover, Japan agrees to this.	2011-03-16	399	174
o. Sanjyo	Sanjo Junior High School stopped functioning due to international students.	2011-03-17	379	162
p. Fujitv	Japanese TV company Fuji donated to UNICEF Japan.	2011-03-16	885	124
q. Cartoonist	Japanese cartoonist Mr.Oda donated 1.5 billion yen.	2011-03-12	2546	171
r. Starvation	An infant in Ibaraki died of starvation.	2011-03-16	2025	144
s. Turkey	Turkey donates 10 billion yen for Japan.	2011-03-12	2380	158

https://doi.org/10.1371/journal.pone.0250419.t002

Experimental evaluation

To evaluate the proposed model, we consider the following prediction task: For the spread of a fake news item, we observe a tweet sequence $\{t_i, d_i\}$ up to time T_{obs} from the original post ($t_0 = 0$), where t_i is the i-th tweeted time, d_i is the number of followers of the i-th tweeting person, and T_{obs} represents the duration of the observation. Then, we seek to predict the time series of the cumulative number of posts related to the fake news item during the test period $[T_{obs}, T_{max}]$, where T_{max} is the end of the period. In this section, we describe the experimental setup and the proposed prediction procedure, and compare the performance of the proposed method with state-of-the-art approaches.

Setup

The total time interval $[0, T_{max}]$ was divided into the training and test periods. The training period was set to the first half of the total period $[0, 0.5T_{max}]$ and the test period was the remaining period $[0.5T_{max}, T_{max}]$. The prediction performance was evaluated by the mean and median absolute error between the actual time series and its predictions:

$$\text{Mean Absolute Error} = \frac{1}{n_b} \sum_{k=1}^{n_b} |\hat{N}_k - N_k|,$$

$$\text{Median Absolute Error} = \text{Median}(|\hat{N}_k - N_k|) \quad (k = 1, 2, \cdots n_b),$$

where $\hat{N}_k$ and N_k are the predicted and actual cumulative numbers of tweets in a k-th bin $[(k-1)\Delta + T_{obs}, k\Delta + T_{obs}]$, respectively, n_b is the number of bins, and $\Delta = 1$ hour is the bin width.

Prediction procedure based on the proposed model

First, we fit the model parameters using the maximum likelihood method from the observation data (see Section 4). Second, we calculate the intensity function $\hat{\lambda}(t)$ during the prediction period $t \in [T_{obs}, T_{max}]$

$$\hat{\lambda}_{prop}(t) = \hat{\lambda}_1(t) + \hat{\lambda}_2(t) \tag{8}$$

with

$$\hat{\lambda}_1(t) = p_1(t) \sum_{i:t_i < t_c} d_i \phi(t - t_i), \tag{9}$$

where $\hat{\lambda}_1(t)$ and $\hat{\lambda}_2(t)$ are the intensities of the first and second cascades, respectively. The intensity due to the original news item $\hat{\lambda}_1(t)$ is calculated using the fitted parameters $\{a_1, \tau_1; r, \theta_0\}$ and the observations $\{t_i, d_i\}$ before the inferred correction time t_c. The number of followers was fixed as 1 ($d_i = 1$) for the Tohoku dataset, because the follower information was not available in the data. The intensity due to the correction $\hat{\lambda}_2(t)$ is given by the solution of the integral equation:

$$\hat{\lambda}_2(t) = f(t) + d_p p_2(t) \int_{T_{obs}}^{t} \hat{\lambda}_2(s) \phi(t - s) ds, \tag{10}$$

where

$$f(t) = p_2(t) \sum_{i:t_c < t_i < T_{\text{obs}}} d_i \phi(t - t_i),$$

and d_p is the average number of followers during the observation period.

Prediction results

We evaluated the prediction performance of the proposed model and compared it with three baseline methods: linear regression (LR) [15], reinforced Poisson process (RPP) [44] and TiDeH [10]. We used the Python code in Github [45] to implement TiDeH. Details of the LR and RPP methods are summarized in the S1 Appendix. Fig 4 shows three examples of the time series of the cumulative number of posts related to fake news items and their prediction results. The proposed method (Fig 4: magenta) follows the actual time series more accurately than the baselines. While the proposed method reproduces the slowing-down effect in the posting activity, the baseline models tend to over-estimate the number of posts.
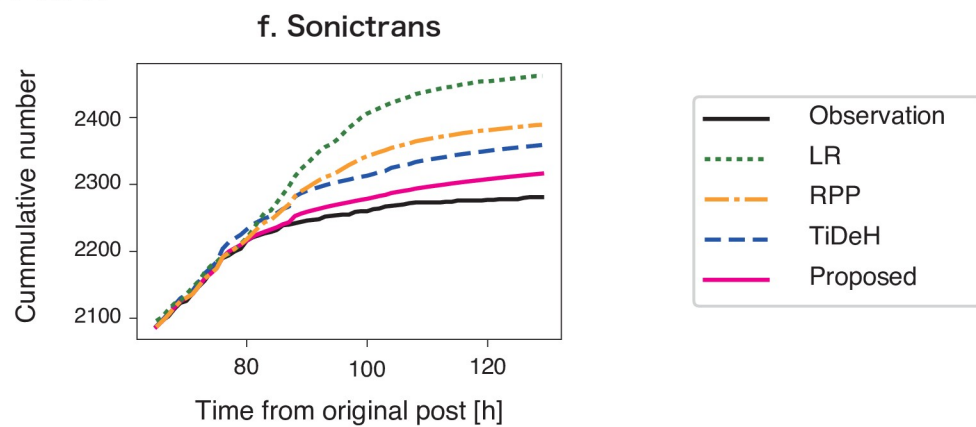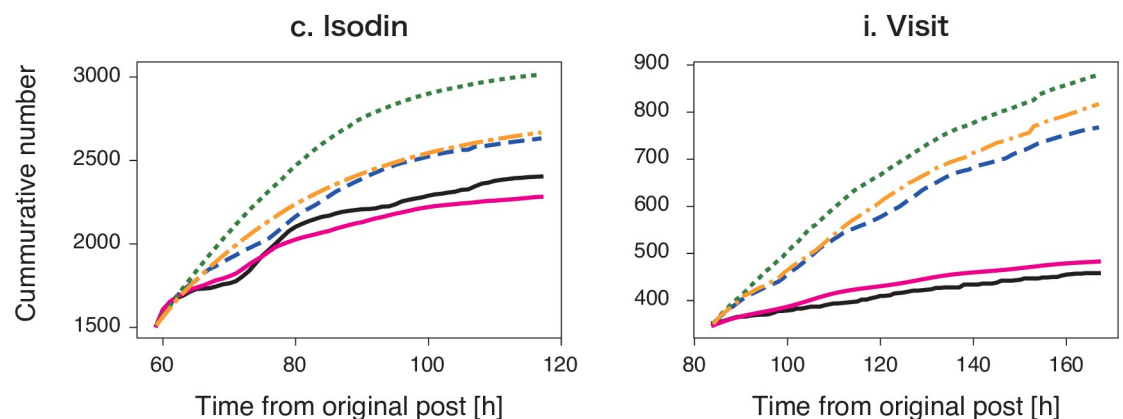
Fig 4. Predicting time series of the cumulative number of posts related to a fake news item. Prediction results from (A) RFN and (B) Tohoku datasets are shown. Green, orange, and blue dashed lines represent the prediction results of the baselines (LR, RPP, and TiDeH, respectively). The black and magenta lines represent the observations and their prediction results of the proposed model.

https://doi.org/10.1371/journal.pone.0250419.g004

Table 3. Prediction performance on the two datasets: Mean and median absolute errors per hour. The best results are shown in bold for each case.

Datasets	RFN		Tohoku	
Metric	Mean	Median	Mean	Median
LR	88.3	5.08	13.9	4.51
RPP	61.8	3.12	8.23	2.30
TiDeH	54.2	1.89	4.12	1.99
Proposed	**36.9**	**1.37**	**2.40**	**1.80**

Next we examine the distribution of the proposed model's parameters. The spreading effect of the falsity of the news item a_2 is weaker than that of the news story itself a_1 for most fake news items (67% and 79% in the RFN and Tohoku datasets, respectively). The result can be attributed to the fact that the news story itself is more surprising for the users than the falsity of the news. The decay time constant of the first cascade τ_1 is approximately 40 (hours) in both datasets: the median (interquartile range) was 35 (22−92) hours and 40 (19−54) hours for the RFN and Tohoku datasets, respectively. The time constant of the second cascade τ_2 is widely distributed in both datasets, which is consistent with the result observed in the synthetic data (Fig 2). The correction time t_c tends to be around 30−40 hours after the original post: 32 (21−54) hours and 37 (31−61) hours for the RFN and Tohoku datasets, respectively. A previous study [46] reported that the fact-checking sites detect the fake news in 10−20 hours after the original post. The result implies that Twitter users recognize the falsity of a fake news item after 10−20 hours from the initial report by the fact-checking sites.

Finally, we evaluated the prediction performance using the two fake news datasets (Table 3). Table 3 demonstrates that the proposed method outperforms the baseline methods in both datasets and metrics. Comparison of the mean error for the proposed model and TiDeH suggests that the two-stage spreading mechanism reduces the mean error by 32% and 42% in the RFN and Tohoku datasets, respectively. Consistent with previous studies [10, 19], the methods based on the point process model (the proposed method, TiDeH, and RPP) perform better than the linear regression (LR) method. Indeed, the proposed model performs best for most fake news items (100% and 89% in the RFN and Tohoku datasets, respectively). While TiDeH performs better than the proposed model for the other dataset (8%), the proposed model still performs much better than the other baselines (RPP and LR). Furthermore, we evaluated the goodness-of-fit of the model using Akaike's information criterion (AIC) [47]. Comparison of AIC values implies that the proposed model achieves a better fit than TiDeH for most fake news items (100% and 89% in the RFN and Tohoku datasets, respectively). These results suggest that the fake news occasionally spreads in a single cascade rather than in two cascades. This might happen when the users already know the falsity of the news in advance (e.g., April Fool's Day) or they are not interested in the falsity of the news at all. Overall, these results show that the proposed method is effective for predicting the spread of fake news posts on Twitter.

Inferring the correction time

We have demonstrated that the proposed method outperforms the existing methods for predicting the evolution of the spread of a fake news item. The proposed model assumes that Twitter users realize the falsity of the news around the correction time t_c. In this section, we examine the validity of this assumption through text mining.

First, we compared the frequency of fake words with inferred correction time t_c (Fig 5). The fake word frequency is regarded as the number of the tweets having fake words (e.g., false

A RFN

b. Notredame

f. Sonictrans

B Tohoku

a. Saveenergy

l. Taiwan
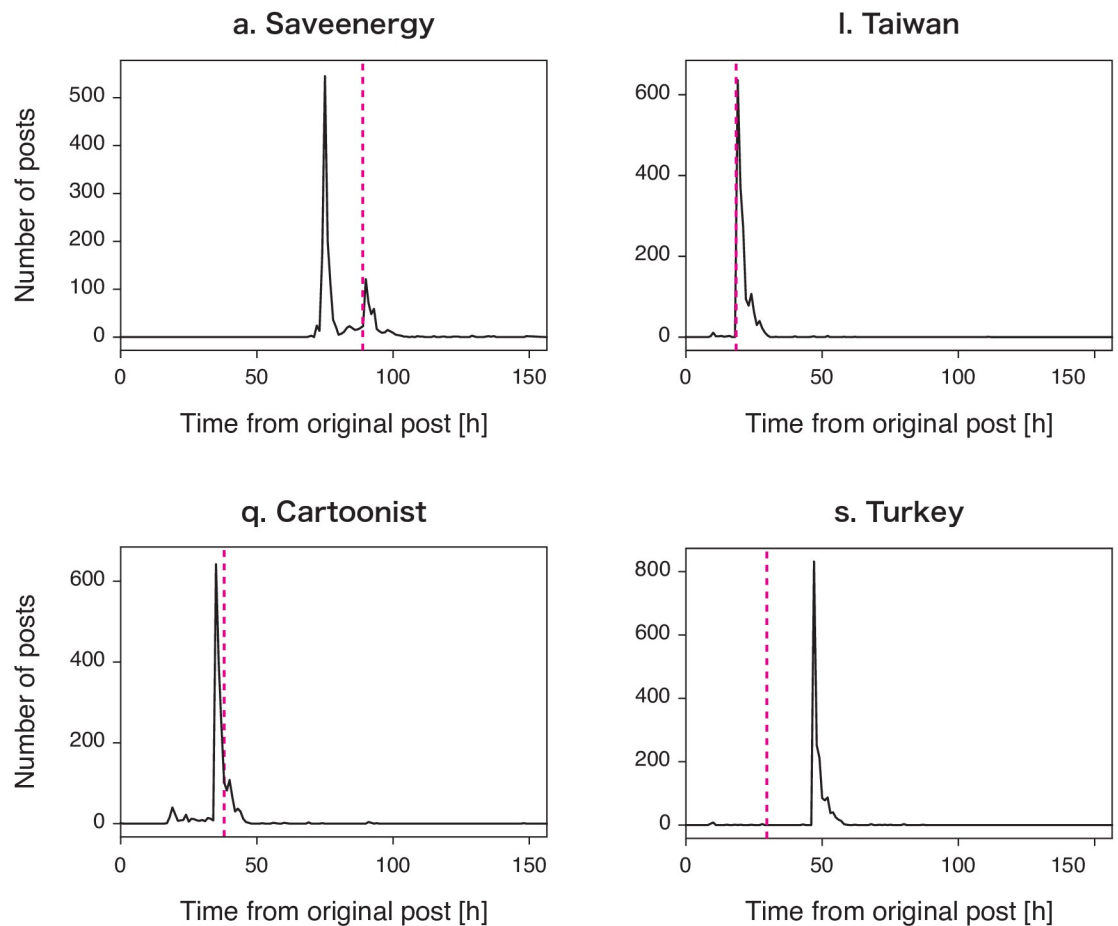
q. Cartoonist

s. Turkey
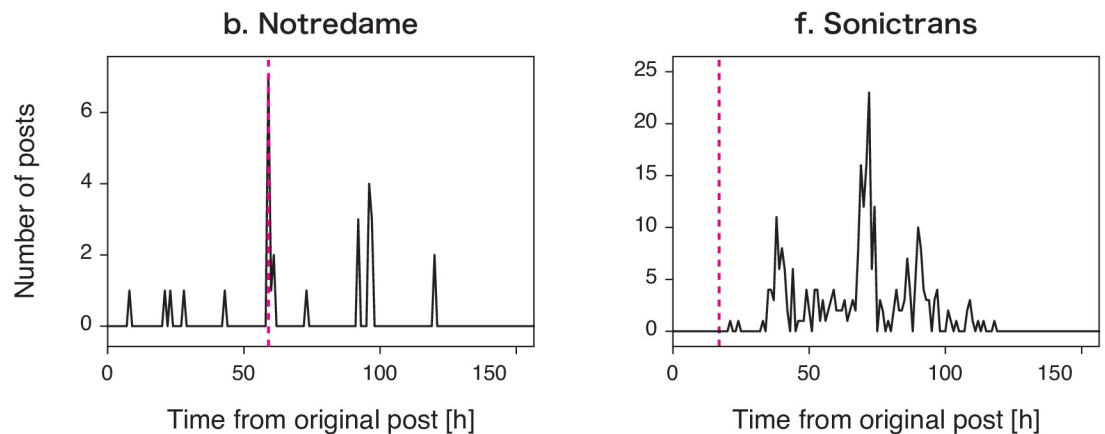
Fig 5. Time series of the fake word frequency for fake news items: (A) RFN and (B) Tohoku datasets. In each panel, the black line represents the time series of the "fake" word count per hour for the tweets related to the fake news item and the magenta vertical lines represent the correction time t_c.

https://doi.org/10.1371/journal.pone.0250419.g005

Before the correction time   **After the correction time**

Fig 6. Example of word cloud before (left) and after (right) the correction time t_c. Each cloud shows the top 10 most frequent words in the fake news story (Turkey in the Tohoku dataset).

rumors, fake, not true, and not real) in each hour. The spread of fake news items in the RFN dataset contained fewer "fake" words than those in the Tohoku dataset: 29 and 277 fake words in the tweets of b. Notredome and f. Sonictrans in the RFN dataset, and 1,752, 1,616, 1,723, and 1,930 fake words in the tweets of a. Saveenergy, l. Taiwan, q. Cartoonist, and s. Turkey in the Tohoku dataset during the observation period (150 hours), respectively. This is because most of the tweets in the RFN dataset are retweets of the original post. We observed that the fake words were posted around the correction time. The peak of the fake word frequency is close to the correction time for Taiwan and Cartoonist in the Tohoku dataset (Fig 5).

Next, we compared the word cloud before and after the correction time t_c. Fig 6 demonstrates an example of a fake news item spreading "Turkey" in the Tohoku dataset. The fake news story is about the huge financial support (10 billion yen) from Turkey to Japan. The word cloud before the correction time implies that this fake news item spread due to the fact that Turkey is considered as a pro-Japanese country. The term "False rumor" starts to appear frequently after the correction time. The word "Taiwan" also appears after the correction time, which is related to another fake news story about Taiwan. These results suggest that Twitter users realize the falsity of the news after the correction time, which supports the key assumption of the proposed model.

Conclusion

We have proposed a point process model for predicting the future evolution of the spreading of fake news on Twitter (i.e., tweets and re-tweets related to a fake news story). The proposed model describes the fake news spread as a two-stage process. First, a fake news item spreads as an ordinary news story. Then, the users recognize the falsity of the news story and spread it as another news story. We have validated this model by compiling two datasets of fake news items spread on Twitter. We have shown that the proposed model outperforms the state-of-the-art methods for accurately predicting the spread of fake news items. Moreover, the proposed model was able to infer the correction time of the news story. Our results based on text mining indicate that Twitter users realize the falsity of the news story around the inferred correction time.

There are several interesting directions for future works. The first direction is to investigate cascades exhibiting multiple bursts. While most fake news cascades exhibit the two-stage spreading pattern, this pattern can also be observed associated with cascades in general. A previous study [48] found that the cascades of image memes in Facebook consists of multiple popularity bursts and argued that the content virality is the primary driver of cascade recurrence.

Our work implies that the change in the perception of the content can be another driver. Additional research is needed to determine whether this hypothesis explains the cascade recurrence better than the content virality or not. A second direction would be to extend the proposed model. While we simply assumed the two-stage process for the spread of a fake news item, this could be extended to describe the spread of fake news in more detail. For example, we can consider multiple types of tweets or a hidden variable to incorporate a soft switch to the second stage from the first one. Another direction would be to apply the proposed model to the practical problems such as fake news detection and mitigation. We believe that the proposed model provides an important contribution to the modeling of the spread of fake news, and it is also beneficial for the extraction of a compact representation of the temporal information related to the spread of a fake news item.

Supporting information

S1 Appendix. Baseline methods. We summarize the baseline methods for predicting the evolution of the spread of a fake news item: linear regression (LR) and reinforced Poisson process (RPP).
(PDF)

Acknowledgments

We thank Takeaki Uno for stimulating discussions and JST ACT-I for providing us the opportunity for this collaboration.

Author Contributions

Conceptualization: Taichi Murayama, Shoko Wakamiya, Eiji Aramaki, Ryota Kobayashi.

Data curation: Taichi Murayama.

Formal analysis: Taichi Murayama, Ryota Kobayashi.

Funding acquisition: Shoko Wakamiya, Eiji Aramaki, Ryota Kobayashi.

Investigation: Taichi Murayama, Ryota Kobayashi.

Methodology: Taichi Murayama, Ryota Kobayashi.

Project administration: Shoko Wakamiya, Eiji Aramaki, Ryota Kobayashi.

Resources: Taichi Murayama, Ryota Kobayashi.

Software: Taichi Murayama.

Supervision: Ryota Kobayashi.

Validation: Taichi Murayama.

Visualization: Taichi Murayama, Ryota Kobayashi.

Writing – original draft: Ryota Kobayashi.

Writing – review & editing: Taichi Murayama, Shoko Wakamiya, Eiji Aramaki, Ryota Kobayashi.

References

1. Carvalho C, Klagge N, Moench E. (2011) The persistent effects of a false news shock. Journal of Empirical Finance 18: 597–615.

2. Bovet A, Makse HA. (2019) Influence of fake news in Twitter during the 2016 US presidential election. Nature communications 10: 1–14. https://doi.org/10.1038/s41467-018-07761-2 PMID: 30602729

3. Takayasu M, Sato K, Sano Y, Yamada K, Miura W, Takayasu H. (2015) Rumor diffusion and convergence during the 3.11 earthquake: a Twitter case study. PLoS one 10: e0121443. https://doi.org/10.1371/journal.pone.0121443 PMID: 25831122

4. Hashimoto T, Shepard DL, Kuboyama T, Shin K, Kobayashi R, Uno T. (2021) Analyzing temporal patterns of topic diversity using graph clustering. The Journal of Supercomputing 77: 4375–4388.

5. Marc F, Cox JW, Hermann P. (2016) Pizzagate: From rumor, to hashtag, to gunfire in dc. Washington Post.

6. Shu K, Sliva A, Wang S, Tang J, Liu H. (2017) Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter 19: 22–36.

7. Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y. (2019) Combating fake news: A survey on identification and mitigation techniques. ACM Transactions on Intelligent Systems and Technology 10: 1–42.

8. Vosoughi S, Roy D, Aral S. (2018) The spread of true and false news online. Science 359: 1146–51.

9. Zhao Z, Zhao J, Sano Y, Levy O, Takayasu H, Takayasu M, et al. (2020) Fake news propagates differently from real news even at early stages of spreading. EPJ Data Science 9:7. https://doi.org/10.1140/epjds/s13688-020-00224-z

10. Kobayashi R, Lambiotte R. (2016) TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. Proceedings of 10th International Conference on Web and Social Media, ICWSM 2016. p. 191-200.

11. Kursuncu U, Gaur M, Lokala U, Thirunarayan K, Sheth A, Arpinar IB. (2019) Predictive analysis on Twitter: Techniques and applications. Emerging research challenges and opportunities in computational social network analysis and mining, p. 67–104. Springer, Cham.

12. Tatar A, De Amorim MD, Fdida S, Antoniadis P. (2014) A survey on predicting the popularity of web content. Journal of Internet Services and Applications 5:8. https://doi.org/10.1186/s13174-014-0008-y

13. Cheng J, Adamic LA, Dow PA, Kleinberg JM, Leskovec J. (2014) Can cascades be predicted? Proceedings of the 23rd international conference on world wide web, WWW 2014, p. 925-936.

14. Petrovic S, Osborne M, Lavrenko V. (2011) Rt to win! predicting message propagation in twitter. International Conference on Web and Social Media, ICWSM 2011, p.586-589.

15. Szabo G, Huberman BA. (2010). Predicting the popularity of online content. Communications of the ACM, 53.8: 80–88. https://doi.org/10.1145/1787234.1787254

16. Matsubara Y, Sakurai Y, Prakash BA, Li L, Faloutsos C. (2012) Rise and fall patterns of information diffusion: model and implications. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD 2012, p.6-14

17. Proskurnia J, Grabowicz P, Kobayashi R, Castillo C, Cudré-Mauroux P, Aberer K. (2017) Predicting the success of online petitions leveraging multidimensional time-series. Proceedings of the 26th International Conference on World Wide Web, WWW 2017, p. 755-764.

18. Masuda N, Takaguchi T, Sato N, Yano K. (2013) Self-exciting point process modeling of conversation event sequences. Temporal networks, pp. 245–264. Springer, Berlin, Heidelberg.

19. Zhao Q, Erdogdu MA, He HY, Rajaraman A, Leskovec J. (2015) Seismic: A self-exciting point process model for predicting tweet popularity. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, KDD 2015, p.1513-1522.

20. Delvenne JC, Lambiotte R, Rocha LE. (2015) Diffusion on networked systems is a question of time or structure. Nature communications. 6.1:1–10. https://doi.org/10.1038/ncomms8366 PMID: 26054307

21. Medvedev AN, Delvenne JC, Lambiotte R. (2019) Modelling structure and predicting dynamics of discussion threads in online boards. Journal of Complex Networks. 7.1:67–82. https://doi.org/10.1093/comnet/cny010

22. Rizoiu MA, Xie L, Sanner S, Cebrian M, Yu H, Van Hentenryck P. (2017) Expecting to be HIP: Hawkes intensity processes for social media popularity. Proceedings of the 26th International Conference on World Wide Web, WWW 2017 p.735-744.

23. Fujita K, Medvedev A, Koyama S, Lambiotte R, Shinomoto S. (2018) Identifying exogenous and endogenous activity in social media. Physical Review E. 98.5:052304. https://doi.org/10.1103/PhysRevE.98.052304

24. Törnberg P. (2018) Echo chambers and viral misinformation: Modeling fake news as complex contagion. PLoS one. 13.9:e0203958. https://doi.org/10.1371/journal.pone.0203958 PMID: 30235239

25. Hassan N, Arslan F, Li C, Tremayne M. (2017) Toward automated fact-checking: Detecting check-worthy factual claims by ClaimBuster. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2017, p.1803-1812.

26. Rashkin H, Choi E, Jang JY, Volkova S, Choi Y. (2017) Truth of varying shades: Analyzing language in fake news and political fact-checking. Proceedings of the 2017 conference on Empirical Methods in Natural Language Processing, EMNLP 2017, p.2931-2937.

27. Kwon S, Cha M, Jung K, Chen W, Wang Y. (2013) Prominent features of rumor propagation in online social media. Proceedings of the 2013 IEEE 13th International Conference on Data Mining, ICDM 2013, p.1103-1108.

28. Ruchansky N, Seo S, Liu Y. (2017) Csi: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, p.797-806.

29. Lukasik M, Srijith PK, Vu D, Bontcheva K, Zubiaga A, Cohn T. (2016) Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, p.393-398.

30. Farajtabar M, Rodriguez MG, Zamani M, Du N, Zha H, Song L. (2015) Back to the past: Source identification in diffusion networks from partially observed cascades. Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, AISTATS 2015, p.232-240.

31. Dutta HS, Dutta VR, Adhikary A, Chakraborty T. (2020) HawkesEye: Detecting fake retweeters using Hawkes process and topic modeling. IEEE Transactions on Information Forensics and Security. 15: p. 2667–2678.

32. Daley DJ, Vere-Jones D. (2003) An introduction to the theory of point processes, volume 1: Elementary theory and methods. Verlag New York Berlin Heidelberg: Springer.

33. Nash SG. (1984) Newton-type minimization via the Lanczos method. SIAM Journal on Numerical Analysis 21: p. 770–788. https://doi.org/10.1137/0721052

34. Scipy.org, https://docs.scipy.org. Last accessed 19 Oct 2020

35. Brent RP. (2013) Algorithms for minimization without derivatives. Dover Publications.

36. https://github.com/hkefka385/extended_tideh. Last accessed 22 Feb 2021

37. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics 25: p. 1923–1929. https://doi.org/10.1093/bioinformatics/btp358 PMID: 19505944

38. Gontier C, Pfister JP. (2020) Identifiability of a Binomial Synapse. Frontiers in computational neuroscience 14:86. https://doi.org/10.3389/fncom.2020.558477 PMID: 33117139

39. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. (2020) FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. Big Data 8: p. 171–188.

40. Ma J, Gao W, Mitra P, Kwon S, Jansen BJ, Wong KF, Cha M. (2016). Detecting rumors from microblogs with recurrent neural networks. Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI 2016, p. 3818-3824.

41. Politifact, https://www.politifact.com/. Last accessed 19 Oct 2020

42. Snopes, https://www.snopes.com/. Last accessed 19 Oct 2020

43. The Social Psychology of Panic Revealed by Categorizing 80 Post-Disaster Hoaxes, https://blogos.com/article/2530/. Last accessed 19 Oct 2020

44. Gao S, Ma J, Chen Z. (2015) Modeling and predicting retweeting dynamics on microblogging platforms. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, p. 107-116.

45. https://github.com/NII-Kobayashi/TiDeH. Last accessed 19 Oct 2020

46. Shao C, Ciampaglia GL, Flammini A, Menczer F. (2016) Hoaxy: A platform for tracking online misinformation. Proceedings of the 25th international conference companion on world wide web, WWW 2016, p. 745-750.

47. Akaike H, (1974) A new look at the statistical model identification. IEEE transactions on automatic control 19: 716–723. https://doi.org/10.1109/TAC.1974.1100705

48. Cheng J, Adamic LA, Kleinberg JM, Leskovec J. (2016) Do cascades recur? In Proceedings of the 25th international conference on world wide web, WWW 2016, p. 671-681.