Identifying exogenous and endogenous activity in social media

Kazuki Fujita,[1,*] Alexey Medvedev,[2,3,†] Shinsuke Koyama,[4,‡] Renaud Lambiotte,[5,§] and Shigeru Shinomoto[1,∥]

[1]*Department of Physics, Kyoto University, Kyoto 606-8502, Japan*
[2]*NaXys, Universite de Namur, 5000 Namur, Belgium*
[3]*ICTEAM, Universite Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium*
[4]*The Institute of Statistical Mathematics, Tokyo 190-8562, Japan*
[5]*Mathematical Institute, University of Oxford, Oxford OX2 6GG, United Kingdom*

(Received 2 August 2018; published 13 November 2018)

The occurrence of new events in a system is typically driven by external causes and by previous events taking place inside the system. This is a general statement, applying to a range of situations including, more recently, to the activity of users in online social networks (OSNs). Here we develop a method for extracting from a series of posting times the relative contributions that are exogenous, e.g., news media, and endogenous, e.g., information cascade. The method is based on the fitting of a generalized linear model (GLM) equipped with a self-excitation mechanism. We test the method with synthetic data generated by a nonlinear Hawkes process, and apply it to a real time series of tweets with a given hashtag. In the empirical dataset, the estimated contributions of exogenous and endogenous volumes are close to the amounts of original tweets and retweets respectively. We conclude by discussing the possible applications of the method, for instance in online marketing.

DOI: 10.1103/PhysRevE.98.052304

I. INTRODUCTION

In online social networks (OSNs), users have the ability to produce, consume and validate information by posting their own content and by reading the content written by others and sharing it to their own social circle [1]. The growing popularity of OSNs and the complexity and size of their data require the development of new tools for a variety of applications, from online marketing and tracking the pulse of society [2] to sociological studies on the emergence of grassroots movement [3]. The dynamics of information in OSNs is particularly rich due to the strong heterogeneity among users, typically associated with a broad degree distribution in the social network [4]; the competition between different keywords of hashtags [5,6]; the coexistence of different types of users [7], e.g., genuine versus bots; and also the interplay between OSNs and more traditional mass media [8].

Several works have focused on the structure and dynamics of the resulting information cascades, from their characterization in empirical data to the design of machine learning algorithms and mathematical models to predict their behavior [9–17]. Mathematically, information cascades are often modeled by self-exciting point processes [18–21], as previous events may trigger new events, in a way that generalizes the standard Hawkes process [22]. In their simplest instance, Hawkes processes are linear self-reinforced processes, where the occurrence of an event increases the likelihood of future events. Hawkes processes have a direct connection to SI models in epidemiology [23] with, as an additional ingredient, a temporal kernel determining the stochastic time between an event and its response. This family of models has been successfully applied to model and predict, among other phenomena, seismic dynamics [24,25], scientometrics [26], finance [27–29], and neuronal firing [30,31].

The main purpose of this work is to design a method to identify the main forces driving the activity in an OSN, and to characterise the importance of endogenous activity, generated organically by interactions between users, and exogenous factors perturbing the internal dynamics. Distinguishing between exogenous and endogenous forces is critical for understanding the mechanisms that drive dynamics of OSNs and has important practical applications, such as the quantification of marketing or external factors that may manipulate the social system [32,33]. A possible solution to this challenging problem is to consider how the number of events decays after a burst of activity, as different types of relaxation are expected to emerge if the system endogenously built up its bubble of activity or if it was caused by an external shock [34]. However, this method suffers from practical limitations as it only allows for a *post hoc* analysis after a sufficiently important burst happened. Instead of analyzing gross activity, we propose to focus on the precise time series of event occurrences. Inspired by the parallels between spike train and social media time series [35], we model the system with a generalized linear model (GLM) equipped with a self-exciting mechanism. GLMs have emerged as an important statistical framework for modeling neuronal spiking activity in a single-neuron and multineuronal networks [36,37], and their nonlinearity presents desirable properties for information spreading on networks, as synchronised activity tends to reinforce the response to a signal. As we will show, the model naturally allows one to

[*] fujita.kazuki.37n@st.kyoto-u.ac.jp
[†] alexey.medvedev@unamur.be
[‡] koyama0526@gmail.com
[§] renaud.lambiotte@maths.ox.ac.uk
[∥] shinomoto@scphys.kyoto-u.ac.jp

disentangle endogenous and exogenous contributions in time series.

The rest of the paper is organized as follows. After introducing the model and the associated parameter inference, we validate the method on artificial data before testing it on empirical time series of appearance of tweets with a particular hashtag, where we successfully determine the contributions of endogenous and exogenous forces. We then provide a critical discussion about our work and conclude with possible future steps.

II. METHODS

At the core of our method, we assume that the activity time series in OSNs, for example postings of tweets with a specific keyword, is modeled by a GLM, where the underlying rate is given by

$$\lambda(t) = \exp\left(\gamma(t) + \alpha \sum_k h(t - t_k)\right) \tag{1}$$

or, equivalently,

$$\lambda(t) = \exp\left[\gamma(t)\right] \prod_k \exp\left[\alpha h(t - t_k)\right], \tag{2}$$

where $\gamma(t)$ and α represent the time-varying external environment and the degree of internal self-excitation, respectively. $h(t)$ is a kernel representing the time profile of internal excitation, and t_k is the occurrence time of kth event. Here we have chosen $h(t) = (1/\tau)\exp(-t/\tau)$ for $t > 0$ and $h(t) = 0$ otherwise.

In contrast with standard linear Hawkes models, where the underlying rate has the form

$$\lambda(t) = \mu(t) + \alpha \sum_k h(t - t_k), \tag{3}$$

the effect of previous events multiply each other, as seen in Eq. (2), which results in a nonlinear dynamical process. The nonlinearity of the model has interesting implications for the stochastic dynamics, as it favors configurations when events appear in short bursts instead of over a long period. The model thus intrinsically rests on the importance of reinforcement, and of multiple contacts over short times to promote diffusion, as observed in complex contagion [38], but previously modeled by means of threshold models [39] on temporal networks [40]. This effect is illustrated in Fig. 1, where we compare examples of intensities of the linear Hawkes process and the nonlinear Hawkes process of the GLM type. We observe that linear reinforcement adds a constant contribution into secondary events, while multiplicative reinforcement give a stronger push if subsequent events arrive closer in time.

Given the occurrence rate $\lambda(t)$, the probability that events occur at times $\{t_k\} \equiv \{t_1, t_2, \ldots, t_n\}$ in the period of $t \in [0, T]$ is obtained as [41,42]

$$p(\{t_k\} \mid \lambda(t)) = \left[\prod_{k=1}^{n} \lambda(t_k)\right] \exp\left(-\int_0^T \lambda(t)\, dt\right), \tag{4}$$

where the exponential term is the survivor function that represents the probability that no more events have occurred in the interval.

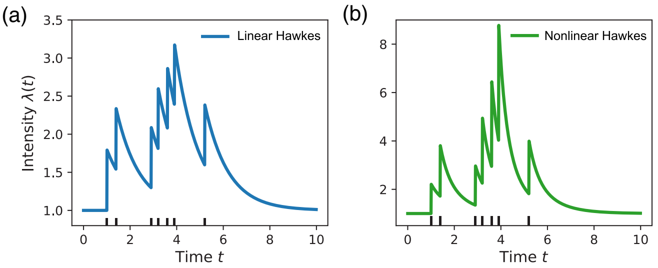

FIG. 1. Comparison of intensity self-reinforcement between (a) the linear Hawkes process and (b) the nonlinear Hawkes process. Both processes have background rates equal to 1 and exponential memory kernels. We consider the hypothetical situation where events get realized at the same times in each case and compare the resulting intensities. Linear reinforcement generates a constant number of secondary events, while multiplicative effect is stronger if subsequent events arrived closer in time, e.g., around $t = 4$.

When confronted with empirical time series, as is usual in practice, we invert the arguments of the conditional probability Eq. (4) with Bayes' rule so that the unknown underlying rate $\lambda(t)$ is inferred from the event series observed $\{t_k\}$:

$$p(\lambda(t) \mid \{t_k\}) = \frac{p(\{t_k\} \mid \lambda(t))\, p(\lambda(t))}{p(\{t_k\})}. \tag{5}$$

As a prior distribution of $\lambda(t)$, we assume that external modulation $\gamma(t)$ is slow. This is given by penalizing the large gradient, $|d\gamma(t)/dt|$:

$$p(\lambda(t)) \propto \exp\left[-\beta \int_0^T \left(\frac{d\gamma(t)}{dt}\right)^2 dt\right], \tag{6}$$

where β is a hyperparameter representing the slowness of the external fluctuations; the external stimulus is largely fluctuating if β is small, and we interpret that external stimulus as absent if $\beta = \infty$, because $\gamma(t)$ should be constant in time in this case.

We represent $p(\{t_k\} \mid \lambda(t))$ and $p(\lambda(t))$ respectively as $p_\alpha(\{t_k\} \mid \gamma(t))$ and $p_\beta(\gamma(t))$, by explicitly specifying the dependency on the external modulation $\gamma(t)$, internal excitation parameter α, and the stiffness parameter β. Then the probability of having event times $p(\{t_k\})$ is given as the marginal likelihood function or the evidence:

$$p_{\alpha,\beta}(\{t_k\}) = \int p_\alpha(\{t_k\} \mid \gamma(t))\, p_\beta(\gamma(t))\, D\{\gamma(t)\}, \tag{7}$$

where $\int D\{\gamma(t)\}$ represents a functional integration over all possible paths of external fluctuations $\gamma(t)$. The method of selecting the hyperparameters according to the principle of maximizing the marginal likelihood function is called the empirical Bayes method [43–46]. The marginalization path integral Eq. (7) for a given set of time series $\{t_k\}$ can be carried out by the expectation maximization (EM) method [47,48] or the Laplace approximation [49].

In this framework, the contributions of endogenous and exogenous origins that have influenced for the occurrence of events are judged by the hyperparameters $\{\hat{\alpha}, \hat{\beta}\}$ selected by maximizing the marginal likelihood $p_{\alpha,\beta}(\{t_k\})$ (Table I). Given the hyperparameters determined as $\{\hat{\alpha}, \hat{\beta}\}$, we can

TABLE I. Presence of endogenous and exogenous contributions can be inferred by the selected hyperparameters of the GLM.

	Self-excitation $\hat{\alpha}$	Stiffness $\hat{\beta}$
Endogenous	finite	∞
Exogenous	0	finite
Endo. + exo.	finite	finite

obtain the maximum *a posteriori* (MAP) estimate of the external circumstance $\hat{\gamma}(t)$, with which their posterior distribution,

$$p_{\hat{\alpha},\hat{\beta}}(\gamma(t) \mid \{t_k\}) \propto p_{\hat{\alpha}}(\{t_k\} \mid \gamma(t))\, p_{\hat{\beta}}(\gamma(t)), \tag{8}$$

is maximized. With the estimated $\hat{\gamma}(t)$ and the given series of event times $\{t_k\}$, we obtain the rate $\lambda_{\mathrm{GLM}}(t)$ as

$$\lambda_{\mathrm{GLM}}(t) = \exp\left(\hat{\gamma}(t) + \hat{\alpha}\sum_k h(t-t_k)\right). \tag{9}$$

III. RESULTS

A. Application to synthetic data

Here we test the efficiency of the method by fitting it to series of occurrence times derived from the following rate processes:

(a) *Exogenous modulation.* First we considered an inhomogeneous Poisson process in which events are drawn from a time varying rate:

$$\lambda(t) = \exp\left[\gamma_0 + b_0\sin(t/T)\right]. \tag{10}$$

We interpret this mode as purely exogenous because the rate variation is independent of past events. We fit our GLM to a series of occurrence times derived from this rate process. The left panel of Fig. 2(a) shows a contour plot of the log-likelihood [Eq. (7)], indicating that the self-excitation parameter $\hat{\alpha}$ is zero while the stiffness constant $\hat{\beta}$ is finite. Thus the method suggests that the rate modulation would have been exogenous. In the right panel of Fig. 2(a), the occurrence rate estimated with our GLM, $\exp[\hat{\gamma}(t)]$, is compared with a time histogram optimally fitted to the data [50], demonstrating that the GLM has succeeded in capturing the underlying rate properly.

(b) *Endogenous modulation with a small self-excitation.* We generated events with the nonlinear Hawkes process

$$\lambda(t) = \exp\left(\gamma_0 + \alpha_0\sum_k h_0(t-t_k)\right), \tag{11}$$

where we have taken the kernel $h_0(t) = (1/\tau_0)\exp(-t/\tau_0)$ for $t > 0$ and $= 0$ otherwise. Here, we have chosen the timescale of the GLM kernel (τ) to be identical to the timescale of this generative model (τ_0). By applying our GLM to the series of occurrence times, the self-excitation parameter $\hat{\alpha}$ is selected as nonzero, suggesting that the system had endogenous excitation [Fig. 2(b), left panel). Because the stiffness $\hat{\beta}$ is very large, the base rate $\exp[\hat{\gamma}(t)]$ is nearly constant, indicating that external circumstances were stationary. The total rate estimated by the GLM, Eq. (9), is very close to the original rate given in Eq. (11). For this data, the optimized bin size of the histogram diverges, indicating that the fluctuation in the rate was not detected. The estimated (constant) rate is above the baseline rate $\exp(\hat{\gamma})$, because the contribution of the self-excitation is included in the total rate [Fig. 2(b), right panel].

(c) *Endogenous modulation with a larger self-excitation.* We generated events with the nonlinear Hawkes process given by Eq. (11) with the self-excitation term α_0 greater than in the case (b), so that event occurrence exhibits large fluctuations. By applying the optimal histogram method, we obtained a fluctuating rate (i.e., the optimal bin size was finite), implying that the nonlinear Hawkes process may also exhibit the stationary-nonstationary (SN) transition that was found in the linear Hawkes process [20,21]: significant fluctuations appear even in the absence of external modulation. Although the rate estimation method suggested that the rate is fluctuating, our GLM was able to reveal that exogenous forcing was absent, and we conclude that the fluctuations appeared solely due to the self-excitation [Fig. 2(c), right panel].

(d) *Exogenous + endogenous modulation.* We derived events from the system receiving both external fluctuations and self-excitation:

$$\lambda(t) = \exp\left(\gamma_0 + b_0\sin(t/T) + \alpha_0\sum_k h_0(t-t_k)\right). \tag{12}$$

By applying our GLM to a series of occurrence times, we obtain both the self-excitation parameter $\hat{\alpha}$ and stiffness constant $\hat{\beta}$ as finite, as shown in the contour plot of the log-likelihood [Fig. 2(d), left panel], suggesting that the system would have been stimulated exogenously but there would have been endogenous self-excitation mechanisms as well.

In the above, we have seen that the GLM is able to decipher the original self-excitation mechanisms provided that the event generation process (the nonlinear Hawkes process in this case) is contained in a family of rate processes presumed for the GLM. In real applications, however, the precise underlying mechanisms of data generation are usually hidden, thus accordingly we have to assume that our GLM may not cover the original process. To examine whether or not the GLM may work even if the model does not contain the original process, we performed the following tests: We generated a series of events from a nonlinear Hawkes process with exponential self-excitation kernel and timescale $\tau_0 = 300$ seconds [Eq. (12), (α, β) both finite], and fitted GLMs whose self-excitation timescale τ is different from τ_0. We confirmed that the GLM suggests finite optimal $(\hat{\alpha}, \hat{\beta})$ for a rather wide range of timescales τ (between 10 and 600 seconds) [Figs. 3(a), 3(b), and 3(c)]. Furthermore, we tested the case in which the functional form of the self-excitation kernel in the GLM is different from that of the event generation process: We adopted the power law kernel $h(t) = c(t+a)^{-3}$ ($c = 2a^2$, $a = 300$ s) into the GLM and applied it to the same data of Fig. 2(d), which was generated using the exponential kernel. The contour plot of the log-likelihood function in Fig. 3(d) is very similar to the original one, Fig. 3(b). This observation suggests that the GLM may capture the presence of self-excitation and external fluctuation robustly even if the precise temporal profile of the self-excitation is not *a priori* known.

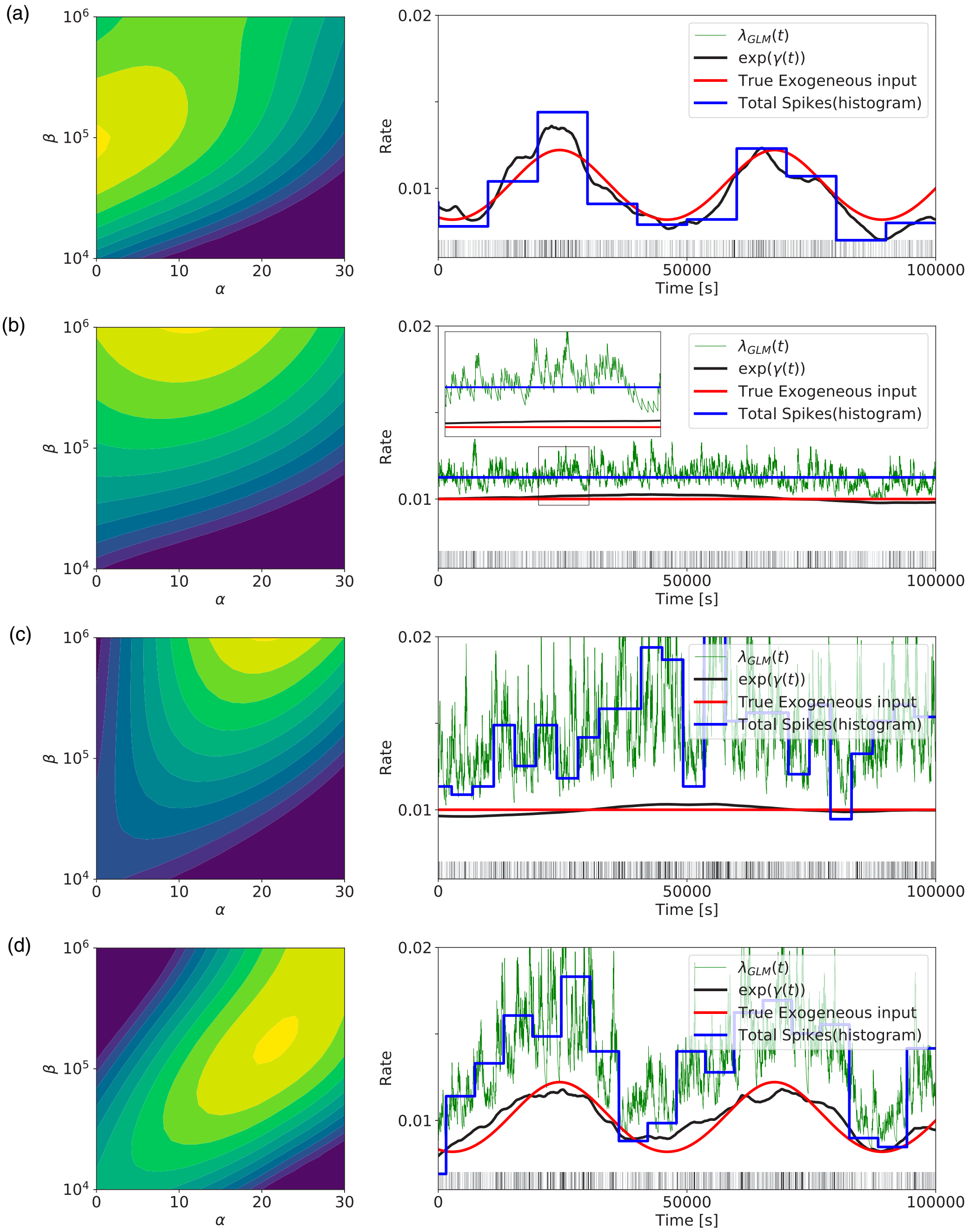

FIG. 2. Fitting the GLM to synthetic data. (a) Exogenously modulated rate process [Eq. (10), $\gamma_0 = \ln 0.01$; $b_0 = 0.2$; $T = 43200$]. (b) The nonlinear Hawkes process with small self-excitation [Eq. (11), $\gamma_0 = \ln 0.01$; $\alpha_0 = 10$; $\tau_0 = 300$). The inset magnifies the rate profile. (c) The nonlinear Hawkes process with the larger self-excitation [Eq. (11), $\gamma_0 = \ln 0.01$; $\alpha_0 = 25$; $\tau_0 = 300$]. (d) The system receiving external fluctuations and self-excitation [Eq. (12), $\gamma_0 = \ln 0.01$; $b_0 = 0.2$; $T = 43200$; $\alpha_0 = 20$; $\tau_0 = 300$]. Left panel: Contour plot of the log-likelihood [Eq. (7)]. Right panel: The solid blue line shows the optimal time histogram, the red line below shows the original exogenous activity $\exp[\gamma(t)]$, the black curve represents the inferred exogenous activity $\exp[\hat{\gamma}(t)]$, and green is the total rate given by the GLM, $\lambda_{\mathrm{GLM}}(t)$ [Eq. (9)].

B. Application to a series of tweet times

In an OSN, one may differentiate between the production of original content and the sharing of existing content over the network of peers. Content may be related to a topic or a real-world event, and its appearance in the digital space is modulated by its interest. When considering the total number of occurrences of a topic-related content, one may interpret the original posts as exogenous input, since the content arrives extrinsically into the social system, while following reshares or retweets may be considered as an endogenous self-exciting contribution into dynamics. These two processes are undoubtedly coupled together, thus it is hard to directly separate one type of activity from the other by observing only the global time series.

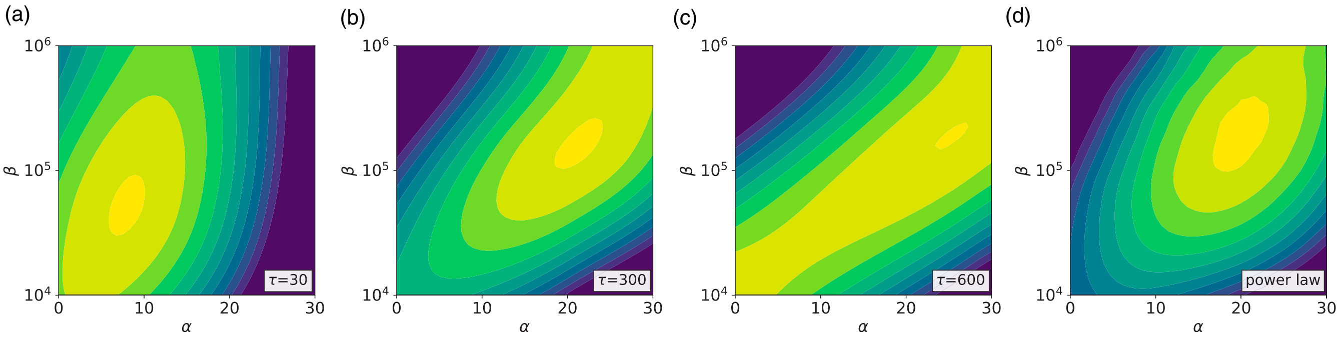

FIG. 3. Contour plots of the log-likelihood functions of various GLMs. Panels (a), (b), and (c) correspond to exponential kernels $h(t) = 1/\tau \exp(-t/\tau)$ of timescales $\tau = 30$, 300, and 600 seconds, respectively. Panel (d) corresponds to a power law kernel $h(t) = c(t + a)^{-3}$ ($c = 2a^2$, $a = 300$ s). The original data were derived from the nonlinear Hawkes process of the system receiving external fluctuations and self-excitation of the exponential kernel of timescale $\tau_0 = 300$ seconds as in Fig. 2(d).

We test our separation method on the data from Twitter, which is a perfect example of a content sharing social system. We consider the dataset of tweets, collected through the public API, posted between January and late August of 2017 that contain the hashtag #bitcoin. These tweets represent the topic of one cryptocurrency and public attention to it. The dataset contains 13 365 114 tweets, and for each tweet we have information about its creation time, its content, and whether the tweet is an original piece of content or a retweet. Note that no underlying network of followers was captured. From this information we infer two separate time series, one related to the *original* tweet postings with the hashtag and another represents the *total* hashtag appearance, including retweets. The average rate of appearance of these two types of tweets is drawn in Fig. 4. Both rates were approximated from daily bins for the sake of clarity of presentation. We observe an increase in appearance rate of retweeted content while the rate of original tweets remains practically stable. Since the tweets are related to the topic of cryptocurrencies, this may be explained by a growing attention to bitcoin related to its recent growth in volume and market capitalization [51].

We apply our GLM in order to separate the original tweeting rate from retweeting. Due to the large size of the observation window we select three one-week samples from the dataset and present our analysis of these samples (Fig. 4). We applied the model to other samples from the dataset and the results were comparable and are not shown here due to space limitation. Following the rapid nature of retweeting activity [18,19], we use the exponential kernel with timescale parameter τ *a priori* set to 60 seconds. Regarding the data examined, time stamps were recorded in seconds, and data contain a nonzero fraction of multiple time stamps falling into the same second. We confirm that randomization of these multiple events in a half-second radius around the given second times tamp performed worse than simply disregarding them. Therefore, in our experiment we stick to the latter option.

The results of the exogenous activity separation are shown in Fig. 5. For the sake of clarity of presentation, the original and total tweeting rates are shown using the 20 min binning of time stamps, and estimated tweeting rates are shown using 1 min binning. We first observe that large peaks in the total tweeting activity are not accompanied by peaks in the rate of original tweet arrivals, therefore those are clearly due to retweets. The GLM method succeeded in filtering out these bursts of activity, and the estimated exogenous rate $\exp[\hat{\gamma}(t)]$ is close to the rate of original tweets. The total estimated rate $\lambda_{\mathrm{GLM}}(t)$ is shown to precisely follow the total tweeting activity, which is expected, since the algorithm optimizes the difference between total tweeting rate and $\lambda_{\mathrm{GLM}}(t)$. However, there appears to be a slight discrepancy in Fig. 5(c), which may be explained by the growth of attention in combination with the drawback of one-second resolution. The contour plots for the time series (a) show clear finite optimal $(\widehat{\alpha}, \widehat{\beta})$ for various values of timescale parameter τ (Fig. 6).

IV. DISCUSSION

We have developed a GLM-based method to estimate the influence of exogenous and endogenous forces on observed temporal events. Using synthetic data generated by nonlinear Hawkes processes, we confirmed that the method is capable

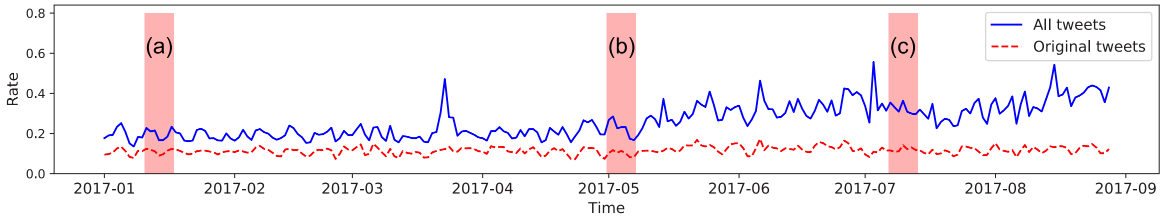

FIG. 4. Tweeting rate in the whole dataset from January to September 2017. The solid blue line shows the tweeting rate of the tweets that contain a hashtag related to "bitcoin." The dashed red line shows the rate of appearance of original tweets with the same hashtags. Both rates were approximated from daily bins.

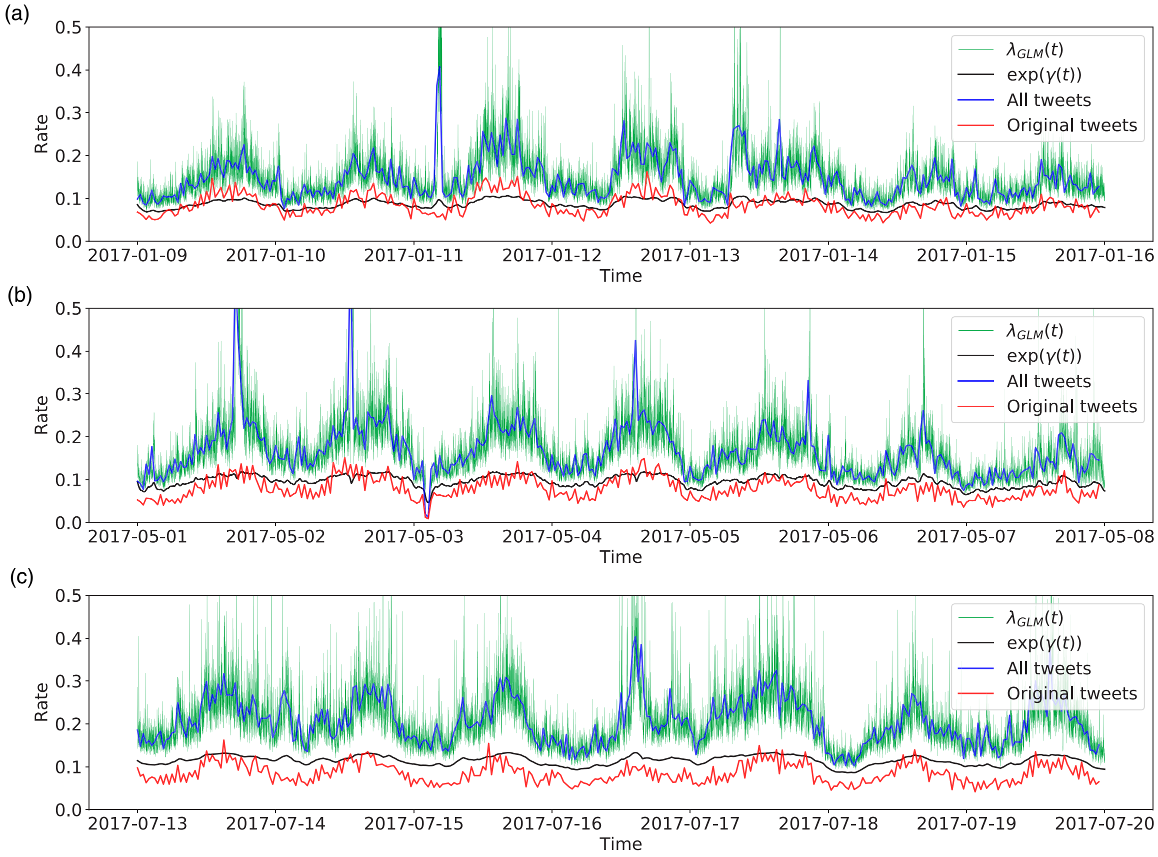

FIG. 5. Original tweeting rate estimation for one-week samples of tweets: (a) between January 9 and January 16, 2017, (b) between May 1 and May 8, 2017, and (c) between July 13 and July 20, 2017. The solid blue line shows the tweeting rate of all tweets, the red line below shows the rate of original tweets, the black curve represents the inferred exogenous activity $\exp[\hat{\gamma}(t)]$, and green is the total rate given by the GLM. The total and original tweeting rates were approximated using 20 min bins and the estimated rates are given using 1 min bins.

of estimating the respective contributions. Then we applied the method to time series of tweets with a given hashtag, and found that the estimated contributions of external and internal origins are close to the original tweets and retweets, respectively.

The concept of dividing the world into exogenous and endogenous categories is a controversial philosophical problem, and it might be considered as a subjective decision. However, the estimation of the exogenous component from a time series has important implications to design efficient models to predict its future behavior and, for instance, to infer the impact of a marketing campaign on the activity of a social network, judging whether items require extensive advertisement or word-of-mouth product mentions have already gone viral.

Note that another method has been designed for a similar purpose, based on the fitting of the linear Hawkes process using the EM method, and validated on a data set of violent civilian deaths occurring in the Iraqi conflict [52]. An advantage of this method is the linearity of the model, which avoids possible catastrophic divergences in the number of events [37]. However, our GLM-based approach has the advantage of determining the timescale of exogenous fluctuation β semiautomatically, according to the empirical Bayes method, while this timescale needs to be given manually in

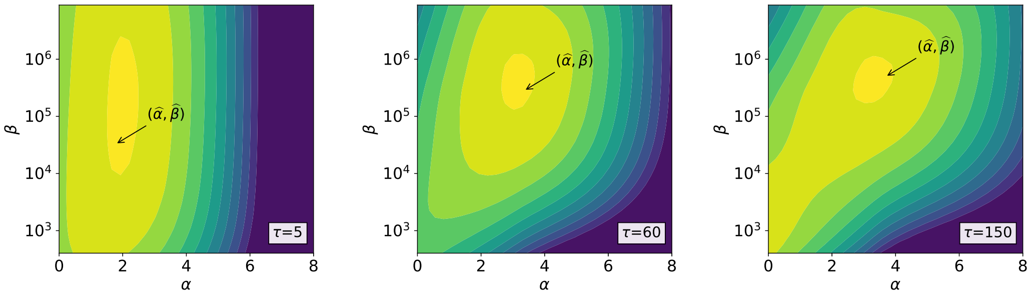

FIG. 6. Contour plots of the log-likelihood function for timescales $\tau = 5$, 60, and 150 seconds. The optimal value found by GLM method is depicted as $(\widehat{\alpha}, \widehat{\beta})$.

the linear model. For these reasons, our method is expected to perform well in situations when the exogenous activity has a slow modulation. In spite of this advantage, our method is bounded by the condition that external stimulus be slow, which is imposed due to the prior distribution, Eq. (6). An interesting venue of future research would be to design methods lifting this condition on the stimuli.

The continuous nature of the GLM suggests the recorded data be continuous as well. However, in practice, high precision temporal data are rarely available, and the time is rounded up to a second. Multiple events may thus occur in cases when the collected time series come from a process with high frequency, which can be subdued by improving the data measurement frequency. In contrast, this finer frequency would also increase the computational burden of the algorithm, a classic precision/speed tradeoff. Another practical issue is selection of the self-excitation kernel and its timescale τ. The proposed method showed to succeed when the provided value of τ_0 lies in a certain interval around the true τ of the process. Narrowing this interval down to a correct τ can be done using extra available information, e.g., retweet time distribution of a test sample of tweets.

ACKNOWLEDGMENTS

The authors would like to offer a special thanks to Alexandre Bovet for sharing the Twitter data. This study was supported in part by Grants-in-Aid for Scientific Research to S.S. from JSPS KAKENHI Grants No. 26280007 and No. 17H06028 and JST CREST Grant Number JPMJCR1304. A.M., R.L., and S.S. were supported by the Bilateral Joint Research Project between JSPS, Japan and FRS-FNRS, Belgium. A.M. was supported by ARC (Federation Wallonia-Brussels) and by the Russian Foundation of Basic Research (16-01-00499).

[1] M. Salganik, *Bit by Bit: Social Research in the Digital Age* (Princeton University Press, Princeton, 2017).

[2] R. Bandari, S. Asur, and B. Huberman, in *ICWSM '12, Proceedings of the Sixth International Conference on Weblogs and Social Media,* Dublin, June 4–7, 2012 (AAAI, Palo Alto, 2012), pp. 26–33.

[3] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno, Sci. Rep. **1**, 197 (2011).

[4] H. Kwak, C. Lee, H. Park, and S. Moon, in *WWW '10, Proceedings of the 19th International Conference on the World Wide Web*, April 26–30, 2010, Raleigh, NC (ACM, New York, 2010), pp. 591–600.

[5] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, Sci. Rep. **2**, 335 (2012).

[6] L. Weng, F. Menczer, and Y.-Y. Ahn, Sci. Rep. **3**, 2522 (2013).

[7] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)* (AAAI, 2017), pp. 280–289.

[8] C. Tan, A. Friggeri, and L. A. Adamic, in *ICWSM '16, Proceedings of the Tenth International Conference on Web and Social Media*, Cologne, May 17–20, 2016 (AAAI, Palo Alto, 2016), p. 378.

[9] K. Lerman and R. Ghosh, in *ICWSM '10, Proceedings of the Fourth International Conference on Weblogs and Social Media*, Washington, D.C., May 23–26, 2010 (AAAI, Menlo Park, 2010).

[10] D. M. Romero, B. Meeder, and J. Kleinberg, in *WWW '11, Proceedings of the 20th International Conference on the World Wide Web*, Hyderabad, March 28 – April 1, 2011 (ACM, New York, 2011), pp. 695–704.

[11] T. Zaman, E. B. Fox, E. T. Bradlow *et al.*, Ann. Appl. Stat. **8**, 1583 (2014).

[12] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, in *WWW '14, Proceedings of the 23rd International Conference on the World Wide Web*, Seoul, April 7–11, 2014 (ACM, New York, 2014), pp. 925–936.

[13] P. A. Dow, L. A. Adamic, and A. Friggeri, in *ICWSM '13, Proceedings of the Seventh International Conference on Weblogs and Social Media*, Cambridge, MA, July 8–11, 2013 (AAAI, Palo Alto, 2013), pp. 145–154.

[14] S. Petrovic, M. Osborne, and V. Lavrenko, in *ICWSM '11, Proceedings of the Fifth International Conference on Weblogs and Social Media*, Barcelona, July 17–21, 2011 (AAAI, Menlo Park, 2011), pp. 586–589.

[15] T. Aoki, T. Takaguchi, R. Kobayashi, and R. Lambiotte, Phys. Rev. E **94**, 042313 (2016).

[16] C. Cattuto, V. Loreto, and L. Pietronero, Proc. Natl. Acad. Sci. U.S.A. **104**, 1461 (2007).

[17] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse, Proc. Natl. Acad. Sci. U.S.A. **106**, 12640 (2009).

[18] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, in *KDD '15, Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, August 10–13, 2015 (ACM, New York, 2015), pp. 1513–1522.

[19] R. Kobayashi and R. Lambiotte, in *ICWSM '16, Proceedings of the Tenth International Conference on Web and Social Media*, Cologne, May 17–20, 2016 (AAAI, Palo Alto, 2016), pp. 191–200.

[20] T. Onaga and S. Shinomoto, Phys. Rev. E **89**, 042817 (2014).

[21] T. Onaga and S. Shinomoto, Sci. Rep. **6**, 33321 (2016).

[22] A. G. Hawkes, Biometrika **58**, 83 (1971).

[23] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).

[24] Y. Ogata, J. Am. Stat. Assoc. **83**, 9 (1988).

[25] A. Helmstetter and D. Sornette, J. Geophys. Res.: Solid Earth **108**, 2482 (2003).

[26] M. Golosovsky and S. Solomon, Phys. Rev. Lett. **109**, 098701 (2012).

[27] Y. Aït-Sahalia, J. Cacho-Diaz, and R. J. A. Laeven, J. Financial Econom. **117**, 585 (2015).

[28] E. Bacry, I. Mastromatteo, and J.-F. Muzy, Market Microstructure and Liquidity **1**, 1550005 (2015).

[29] S. J. Hardiman, N. Bercot, and J.-P. Bouchaud, Eur. Phys. J. B **86**, 442 (2013).

[30] V. Pernice, B. Staude, S. Cardanobile, and S. Rotter, PLoS Comput. Biol. **7**, e1002059 (2011).

[31] P. Reynaud-Bouret, V. Rivoirard, F. Grammont, and C. Tuleau-Malot, J. Math. Neurosci. **4**, 3 (2014).

[32] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, Science **359**, 1094 (2018).

[33] T. Omi, Y. Hirata, and K. Aihara, Phys. Rev. E **96**, 012303 (2017).

[34] R. Crane and D. Sornette, Proc. Natl. Acad. Sci. U.S.A. **105**, 15649 (2008).

[35] C. Sanlı and R. Lambiotte, PloS One **10**, e0131704 (2015).

[36] R. E. Kass, S.-I. Amari, K. Arai, E. N. Brown, C. O. Diekman, M. Diesmann, B. Doiron, U. T. Eden, A. L. Fairhall, G. M. Fiddyment, T. Fukai, S. Grün, M. T. Harrison, M. Helias, H. Nakahara, J.-n. Teramae, P. J. Thomas, M. Reimers, J. Rodu, H. G. Rotstein, E. Shea-Brown, H. Shimazaki, S. Shinomoto, B. M. Yu, and M. A. Kramer, Annu. Rev. Stat. Appl. **5**, 183 (2018).

[37] F. Gerhard, M. Deger, and W. Truccolo, PLoS Comput. Biol. **13**, e1005390 (2017).

[38] D. Centola, Science **329**, 1194 (2010).

[39] P. S. Dodds and D. J. Watts, Phys. Rev. Lett. **92**, 218701 (2004).

[40] T. Takaguchi, N. Masuda, and P. Holme, PloS One **8**, e68629 (2013).

[41] D. Cox and P. Lewis, *The Statistical Analysis of Series of Events* (John Wiley and Sons, New York, 1966).

[42] D. Daley and D. Vere-Jones, *Probability and Its Applications*, Vol. 1 (Springer, Berlin, 2003).

[43] I. J. Good, *The Estimation of Probabilities* (MIT Press, Cambridge, 1965).

[44] H. Akaike, in *Selected Papers of Hirotugu Akaike* (Springer, Berlin, 1980), pp. 333–345.

[45] D. J. MacKay, Neural Comput. **4**, 415 (1992).

[46] B. P. Carlin and T. A. Louis, J. Am. Stat. Assoc. **95**, 1286 (2000).

[47] A. P. Dempster, N. M. Laird, and D. B. Rubin, J. R. Stat. Soc. B **39**, 1 (1977).

[48] A. C. Smith and E. N. Brown, Neural Comput. **15**, 965 (2003).

[49] S. Koyama and L. Paninski, J. Comput. Neurosci. **29**, 89 (2010).

[50] H. Shimazaki and S. Shinomoto, Neural Comput. **19**, 1503 (2007).

[51] Bitcoin price and market capitalisation in time, https://coinmarketcap.com/currencies/bitcoin/, accessed 2018-05-24.

[52] E. Lewis, G. Mohler, P. J. Brantingham, and A. L. Bertozzi, Security J. **25**, 244 (2012).