The Annals of Applied Statistics
2020, Vol. 14, No. 3, 1285–1303
https://doi.org/10.1214/20-AOAS1338

DOES TERRORISM TRIGGER ONLINE HATE SPEECH? ON THE ASSOCIATION OF EVENTS AND TIME SERIES

BY ERIK SCHARWÄCHTER* AND EMMANUEL MÜLLER†

Data Science and Data Engineering, University of Bonn, *scharwaechter@bit.uni-bonn.de; †mueller@bit.uni-bonn.de

Hate speech is ubiquitous on the Web. Recently, the offline causes that contribute to online hate speech have received increasing attention. A recurring question is whether the occurrence of extreme events offline systematically triggers bursts of hate speech online, indicated by peaks in the volume of hateful social media posts. Formally, this question translates into measuring the association between a sparse event series and a time series. We propose a novel statistical methodology to measure, test and visualize the systematic association between rare events and peaks in a time series. In contrast to previous methods for causal inference or independence tests on time series, our approach focuses only on the *timing* of events and peaks and no other distributional characteristics. We follow the framework of event coincidence analysis (ECA) that was originally developed to correlate point processes. We formulate a discrete-time variant of ECA and derive all required distributions to enable analyses of peaks in time series with a special focus on serial dependencies and peaks over multiple thresholds. The analysis gives rise to a novel visualization of the association via quantile-trigger rate plots. We demonstrate the utility of our approach by analyzing whether Islamist terrorist attacks in Western Europe and North America systematically trigger bursts of hate speech and counter-hate speech on Twitter.

1. Introduction. The ubiquity of hate speech in online social media, while distressing, delivers insights into key emotive subjects within a society and globally. The terms and conditions of most online social media platforms prohibit hate speech. Providers ask users to report such contents in order to take further action, for example, by removing the contents, warning the involved users or suspending or deleting their profiles (Matias et al. (2015)). Deletion of harassing material and incitements to violence against individuals is important to protect the victims. However, bursts of group-based hate speech online, for example, anti-Muslim, antiimmigrant, anti-Black, antisemitic or homophobic, also help identifying triggers and mechanisms of hate and, thereby, inform policymakers and nongovernmental organizations. A recent publication by the United Nations Educational, Scientific and Cultural Organization (UNESCO) points out that the "character of hate speech online and its relation to offline speech and action are poorly understood" and that the "causes underlying the phenomenon and the dynamics through which certain types of content emerge, diffuse and lead—or not—to actual discrimination, hostility or violence" should be investigated more deeply (Gagliardone et al. (2015)).

We propose a novel statistical methodology that enables analyses of the systematic relation between rare offline events and online social media usage. Following a recent study (Olteanu et al. (2018)), we demonstrate the utility of our approach by analyzing whether Islamist terrorist attacks systematically trigger bursts of hate speech and counter-hate speech on Twitter. We operationalize these speech acts by tracking usage of the hashtags `#stopislam` (anti-Muslim hate speech) and `#notinmyname` (Muslim counter-hate speech) as well as the

Received February 2019; revised December 2019.

Key words and phrases. Time series, event series, statistical association, independence test, event coincidence analysis, event synchronization, causal inference, hate speech, social media, terrorism.

Arabic keyword `kafir` (jihadist hate speech against "nonbelievers") over a period of three years (2015–2017). We correlate usage of these terms with the occurrence of severe terrorist incidents in Western Europe and North America in the same time period. The key novelty of our approach is that we focus only on the *timing* of spikes in the resulting time series, not their *magnitude* or *duration* as in previous studies (Olteanu et al. (2018), Burnap et al. (2014)). If spikes in the time series coincide with events more often than expected under an independence assumption, there is evidence for a systematic statistical relationship between the two. Thus, we map the correlation problem into the framework of event coincidence analysis (ECA) (Donges et al. (2016)) that was recently proposed to measure coincidences for pairs of point processes.

We first provide a discrete-time formulation of ECA for pairs of event series that corresponds to the original continuous-time formulation for point processes. Building on this formulation, our methodological contributions are as follows. We replace the lagging event series with a thresholded time series and *derive the null distribution of the ECA statistic* for exceedances of a single threshold. The derivation is valid for a large class of strictly stationary time series with serial dependencies. Since a single threshold is often not sufficient to capture the association, we further *derive the joint distribution of the ECA statistic at multiple thresholds*. The derivations yield two hypothesis tests for the association at multiple thresholds. We further propose a novel visualization of trigger coincidences via *quantile-trigger rate plots* (QTR plots). With our method we are able to show that Islamist terrorist attacks in Western Europe and North America systematically trigger bursts of anti-Muslim hate speech on Twitter (`#stopislam`) which confirms case studies and explorative analyses from the literature. Source codes to reproduce our results are available at Scharwächter and Müller (2020).

2. Related work. Most research on the relation between offline actions/events and online hate speech so far is based on *case studies*. The UNESCO publication mentioned earlier describes a few qualitative case studies on the extreme right-wing online forum "Stormfront" (De Koster and Houtman (2008), Bowman-Grieve (2009), Meddaugh and Kay (2009)), and presents findings from nonacademic reports on online hate speech during elections in Kenya and against the Rohingya community in Myanmar (Gagliardone et al. (2015)). Burnap et al. (2014) perform a quantitative case study of the social media reaction after the Woolwich terrorist attack in the United Kingdom (May 23rd, 2013). They analyze the size and survival rate of posts that express *tension*, defined as antagonistic or accusatory content similar to hate speech. In follow-up works (Burnap and Williams (2014, 2016), Williams and Burnap (2016)), they exploit their findings to train hate-speech classifiers and predictive models for information flow following emotive offline events. Magdy, Darwish and Abokhodair (2015) perform a quantitative case study of Twitter usage after the Paris terrorist attacks (November 13th, 2015) and find 21.5% of the posts attacking Islam or Muslims, as opposed to 55.6% defending posts. In contrast to these case studies, we address the *systematic* relation between offline events and online hate speech in a longitudinal study with 17 relevant events over three years to uncover a potential causal link.

Müller and Schwarz (2019) empirically analyze the relation between online hate speech and hate crimes against refugees in Germany by performing fixed effects panel regression on data that covers the years 2015 to 2017. They exploit internet outages as sources of quasiexperimental exogenous variation to establish a causal link. The major difference to our work is that their events of interest are so numerous that they can be aggregated to a numerical value with weekly resolution and be analyzed with standard statistical methodology. In our setting events are very rare with respect to the daily resolution of the time series.

Most related to our work, Olteanu et al. (2018) analyze the impact of Islamist and Islamophobic terrorist attacks on anti-Muslim hate speech online in a longitudinal study with

13 relevant events over 19 months. They perform counterfactual analyses (Brodersen et al. (2015)) of a large number of time series representing the daily volumes of hundreds of anti-Muslim keywords, independently, for every event and report aggregated effects. However, the counterfactual approach is designed for singular, controlled interventions, not for observational studies with reoccurring, uncontrolled events. The reported aggregated effects are thus explorative and not corroborated by measures of statistical significance. We fill this gap by providing a novel statistical methodology to systematically analyze coincidences of rare events and peaks in a time series within the framework of ECA.

ECA was developed to measure the association between two types of reoccurring events. It was applied to assess whether floods systematically trigger epidemic outbreaks (Donges et al. (2016)) or whether natural disasters systematically trigger violent conflicts (Schleussner et al. (2016)). We give an introduction to ECA and discuss challenges when applied to the study of peaks in time series in Section 4. Event synchronization (Quiroga, Kreuz and Grassberger (2002)) is similar to ECA but allows the time tolerance for coincidences to vary. This increases model expressiveness at the cost of an analytical null distribution. The major difference between our ECA-based approach and other methods for causal inference in time series (Granger (1969), Box and Tiao (1975), Schreiber (2000), Bressler and Seth (2011), Brodersen et al. (2015)) or related independence tests (Besserve, Logothetis and Schölkopf (2013), Chwialkowski and Gretton (2014), Scharwächter and Müller (2020)) is that the only feature it uses is the *timing* of events and peaks, irrespective of other distributional characteristics. In particular, it does not assume an underlying predictive model that would have to explain the exact behavior of the time series after event occurrences. By focusing on peaks in the time series, it is closely related to measures and models for tail dependence of random variables (Frahm, Junker and Schmidt (2005), Yan, Wu and Zhang (2019)).

Causal inference techniques have been applied in social media studies before (Cunha, Weber and Pappa (2017), Chandrasekharan et al. (2017), Saha, Chandrasekharan and De Choudhury (2019)). These works differ from ours in that they do not focus on the association between time series and event series. Recent work on online hate speech has focused on the targets of hate (Silva et al. (2016), Mondal, Silva and Benevenuto (2017), ElSherief et al. (2018a)), characterizations of hateful users (ElSherief et al. (2018a), Ribeiro et al. (2018)) as well as geographic (Mondal, Silva and Benevenuto (2017)) and linguistic differences (ElSherief et al. (2018b)) in hate. Perhaps the largest body of research on online hate speech in the past decade has been on different approaches for its automatic identification (Warner and Hirschberg (2012), Kwok and Wang (2013), Burnap and Williams (2014), Djuric et al. (2015), Davidson et al. (2017)). An overview of the various approaches is given in a recent survey (Schmidt and Wiegand (2017)).

3. Data. For a quantitative analysis of social media usage in reaction to Islamist terrorist attacks we have to operationalize these terms.

3.1. *Islamist terrorist attacks.* We obtained a comprehensive list of global terrorist attacks from the publicly available Global Terrorism Database (GTD) (National Consortium for the Study of Terrorism (2018)). We filtered the GTD for attacks that occurred in Western Europe and North America between January 2015 and December 2017, left at least 10 people wounded and were conducted by the so-called *Islamic State of Iraq and the Levant* (ISIL), *Al-Qaida in the Arabian Peninsula* (AQAP), Jihadi-inspired extremists or Muslim extremists, according to the GTD. The resulting 17 severe Islamist terrorist attacks are shown in Table 1.

TABLE 1
Severe Islamist terrorist attacks in Western Europe and North America

Date	City	Date	City
2015-01-07	Paris, France	2016-12-19	Berlin, Germany
2015-11-13	Paris, France	2017-03-22	London, UK
2015-12-02	San Bernardino, USA	2017-04-07	Stockholm, Sweden
2016-03-22	Brussels, Belgium	2017-05-22	Manchester, UK
2016-06-12	Orlando, USA	2017-06-03	London, UK
2016-07-14	Nice, France	2017-08-17	Barcelona, Spain
2016-07-24	Ansbach, Germany	2017-09-15	London, UK
2016-09-17	New York City, USA	2017-10-31	New York City, USA
2016-11-28	Columbus, USA		

3.2. *Social media response.* To assess the online social media response to these events in terms of anti-Muslim hate speech, Muslim counter-hate speech and jihadist hate speech, we retrieved time series of the global Twitter volume in the same time period (2015–2017) for the three keywords `#stopislam`, `#notinmyname` and `kafir` ("nonbeliever") that represent the three speech acts. We retrieved daily time series of the global Twitter volume for our keywords. Details on keyword selection, data acquisition and preprocessing are given in Appendix A. The global daily volume for all queries after preprocessing is shown in Figure 1 along with all Islamist terrorist attacks from Table 1.

4. Methods. Our goal is to analyze the systematic relation between offline events and online social media usage. Formally, we model the occurrence of terrorist attacks by a (discrete-time) *event series* $\mathcal{E} = (E_t)_{t=1}^T$, where each E_t is a binary random variable with $E_t = 1$ if and only if there is a terrorist attack at time t and $E_t = 0$ otherwise. Social media usage is captured by a (discrete-time) *time series* $\mathcal{X} = (X_t)_{t=1}^T$, where each X_t is a continuous random variable that indicates the daily volume of posts. A *peak* in the time series is the exceedance of some large threshold $\tau \in \mathbb{R}$. The problem is to decide whether the number of events in $\mathcal{E}$ that trigger peaks in $\mathcal{X}$ is so high that the association should be considered statistically significant: in this case, there is a potential causal link between event occurrences and peaks in the time series. Observe that, for a time series $\mathcal{X}$ and threshold τ, the *threshold exceedance series*

$$\mathcal{A} = \big(\mathcal{I}(X_1 > \tau), \dots, \mathcal{I}(X_T > \tau)\big) \tag{1}$$

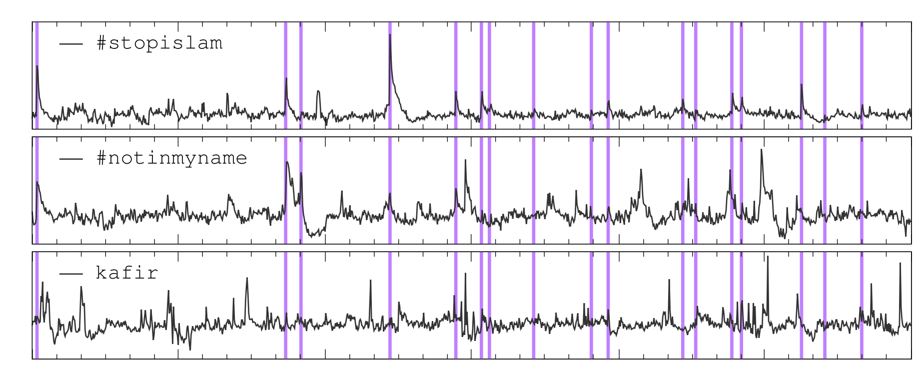

FIG. 1. *Daily Twitter volume of the keywords analyzed in this study. The vertical lines indicate dates of severe Islamist terrorist attacks in Western Europe and North America.*

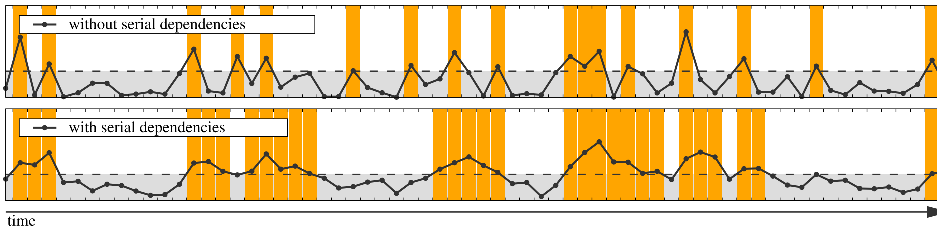

FIG. 2. *Threshold exceedance series* (*bars*) *for time series with and without serial dependencies* (*lines*).

is itself an event series. Here, $\mathcal{I}(C)$ is an indicator function that is 1 if and only if the condition C is true and 0 otherwise; see Figure 2 for an example. The threshold exceedance series retains only information on the timing of exceedances and disregards all other distributional characteristics. The problem of correlating event series with peaks in a time series can thus directly be mapped to the problem of correlating two event series, for example, using ECA.

There are two challenges that need to be addressed when applying measures designed for pairs of event series in this context, *serial dependencies* and *threshold selection*. Serial dependencies in the time series lead to clustering of events in the threshold exceedance series. This effect can be observed when comparing the upper and the lower time series in Figure 2. Event clustering must be handled correctly when establishing the statistical significance of an observed correlation score. The second challenge is that the choice of threshold has a strong impact on the results of the analysis but is often not straightforward. In fact, the magnitude of the response may vary from event to event: a full picture of the association between events and peaks can only be obtained when considering exceedances at multiple thresholds.

We now proceed with a detailed exposition of the statistical methodology that we propose for the analysis. We embed our contributions within the existing framework of ECA. We begin with a discrete-time formulation of ECA for pairs of event series in Section 4.1 that corresponds to the original continuous-time formulation for point processes (Donges et al. (2016)). In Section 4.2 we address the challenge of serial dependencies by deriving a novel analytical null distribution for the ECA statistic that is valid for threshold exceedance series for a large class of strictly stationary time series. We then derive the joint null distribution for coincidences at multiple thresholds in Section 4.4 and describe two test procedures to assess statistical significance. We complement the analytical results with a novel visualization of the association via quantile-trigger rate (QTR) plots.

4.1. *Discrete-time event coincidence analysis.* ECA is a statistical methodology to assess whether two types of events are independent or whether one kind of event systematically triggers or precedes the other kind of event. The basic idea of ECA is to count how many times the two kinds of events coincide and to assess whether this number is statistically significant under the assumption of independence.

4.1.1. *Definition.* Let $\mathcal{A} = (E_t^{\mathcal{A}})_{t=1}^T$ and $\mathcal{B} = (E_t^{\mathcal{B}})_{t=1}^T$ be two event series of length T with $E_t^{\mathcal{A}}, E_t^{\mathcal{B}} \in \{0, 1\}$ for all t, and $N_{\mathcal{A}} = \sum_t E_t^{\mathcal{A}}$ and $N_{\mathcal{B}} = \sum_t E_t^{\mathcal{B}}$ event occurrences. Furthermore, let $\Delta \in \mathbb{N}_0$ be a user-defined time tolerance. ECA measures the extent to which B events precede A events with a time tolerance of Δ. Thus, we refer to $\mathcal{A}$ as the *lagging* and $\mathcal{B}$ as the *leading* event series. ECA considers two possibilities to measure this extent, *trigger coincidences* and *precursor coincidences*. A trigger coincidence occurs whenever a $\mathcal{B}$ event triggers an $\mathcal{A}$ event within the next Δ time steps, whereas a precursor coincidence occurs whenever an $\mathcal{A}$ event is preceded by a $\mathcal{B}$ event within the previous Δ time steps. In the special case $\Delta = 0$, the two concepts are identical. The two types of coincidences are illustrated in Figure 3 with $\Delta = 4$. In the example there are three trigger coincidences and four precursor

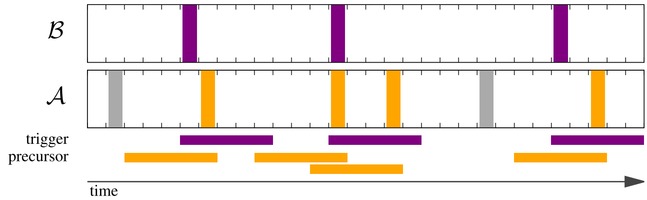

FIG. 3. *Trigger coincidences and precursor coincidences for two event series $\mathcal{A}$ and $\mathcal{B}$ with time tolerance* $\Delta = 4$.

coincidences. A significant number of trigger or precursor coincidences indicates a possible causal link from $\mathcal{B}$ to $\mathcal{A}$. The opposite direction can be analyzed analogously by exchanging the labels. Formally, the *number of trigger coincidences* is defined as

$$K_{\text{tr}} = K_{\text{tr}}^{\Delta}(\mathcal{B}, \mathcal{A}) := \sum_{t=1}^{T-\Delta} E_t^{\mathcal{B}} \cdot \Big(\max_{\delta=0,\dots,\Delta} E_{t+\delta}^{\mathcal{A}} \Big) \tag{2}$$

and the *number of precursor coincidences* as

$$K_{\text{pre}} = K_{\text{pre}}^{\Delta}(\mathcal{B}, \mathcal{A}) := \sum_{t=\Delta+1}^{T} E_t^{\mathcal{A}} \cdot \Big(\max_{\delta=0,\dots,\Delta} E_{t-\delta}^{\mathcal{B}} \Big). \tag{3}$$

The order of the function arguments $\mathcal{B}$ and $\mathcal{A}$ corresponds to the temporal ordering that is analyzed (and thus the potential causal direction). We omit the parameter Δ and the function arguments whenever they are clear from the context. The corresponding *coincidence rates* are given by $r_{\text{tr}} := K_{\text{tr}}/N_{\mathcal{B}}$ and $r_{\text{pre}} := K_{\text{pre}}/N_{\mathcal{A}}$. In the example from Figure 3, we have $r_{\text{tr}} = 1$ and $r_{\text{pre}} = \frac{2}{3}$. A high *trigger coincidence rate* indicates that a large fraction of $\mathcal{B}$ events is followed by an $\mathcal{A}$ event. In other words, $\mathcal{B}$ events systematically trigger $\mathcal{A}$ events. A high *precursor coincidence rate* indicates that a large fraction of $\mathcal{A}$ events is preceded by a $\mathcal{B}$ event, that is, the occurrence of $\mathcal{A}$ events can be explained to a large degree by $\mathcal{B}$ events. The two measures are complementary and should be selected based on the research question.

4.1.2. *Null distribution.* To assess whether an observed trigger coincidence rate is statistically significant, we need the null distribution of K_{tr} under the assumption that the processes are independent. For this purpose we introduce the binary random variables $Z_t^{\mathcal{A}} := \max_{\delta=0,\dots,\Delta} E_{t+\delta}^{\mathcal{A}}$ for all $t = 1, \dots, T - \Delta$ that indicate whether there is an $\mathcal{A}$ event in the window $t, \dots, t + \Delta$. This allows rewriting equation (2) as

$$K_{\text{tr}} = \sum_{t=1}^{T-\Delta} E_t^{\mathcal{B}} \cdot Z_t^{\mathcal{A}} = \sum_{t: E_t^{\mathcal{B}}=1} Z_t^{\mathcal{A}} \tag{4}$$

and reveals that the number of trigger coincidences is, effectively, a sum of Bernoulli trials, each associated with an event occurrence in $\mathcal{B}$. Sums over fixed numbers of independent and identically distributed Bernoulli trials follow binomial distributions. However, for an arbitrary event series $\mathcal{A}$, the random variables $Z_t^{\mathcal{A}}$ and $Z_{t'}^{\mathcal{A}}$ may be neither identically distributed nor independent. Additional assumptions are required to derive the null distribution analytically.

In a simple and analytically tractable case, the event series $\mathcal{A}$ and $\mathcal{B}$ are independent Bernoulli processes with $\mathbb{P}(E_t^{\mathcal{A}} = 1) = p_{\mathcal{A}}$ and $\mathbb{P}(E_t^{\mathcal{B}} = 1) = p_{\mathcal{B}}$ for all t. In this case all $Z_t^{\mathcal{A}}$ are identically distributed with success probability

$$\mathbb{P}(Z_t^{\mathcal{A}} = 1) = 1 - \mathbb{P}(E_t^{\mathcal{A}} = 0, \dots, E_{t+\Delta}^{\mathcal{A}} = 0) = 1 - (1 - p_{\mathcal{A}})^{\Delta+1}. \tag{5}$$

Furthermore, two variables $Z_t^{\mathcal{A}}$ and $Z_{t'}^{\mathcal{A}}$ are independent whenever they are separated by more than Δ time steps. Under the additional assumption that the $N_{\mathcal{B}}$ events in $\mathcal{B}$ are separated by more than Δ time steps, the null distribution of K_{tr} is the binomial

$$K_{\text{tr}} \mid N_{\mathcal{B}} \sim \text{Binomial}\big(N_{\mathcal{B}}, 1-(1-p_{\mathcal{A}})^{\Delta+1}\big). \tag{6}$$

4.1.3. *Statistical test procedure.* Following the framework of ECA, we use K_{tr} as a test statistic to decide between the null hypothesis of independence of $\mathcal{A}$ and $\mathcal{B}$ and the alternative hypothesis of a trigger relationship. If the number of trigger coincidences is unusually large, the null hypothesis is rejected in favor of the alternative hypothesis. The success probabilities are estimated as $\hat{p}_{\mathcal{A}} = N_{\mathcal{A}}/T$ and $\hat{p}_{\mathcal{B}} = N_{\mathcal{B}}/T$, respectively. The p-value for an observed number of trigger coincidences k_{tr} is obtained from the probability mass function of the binomial distribution in equation (6),

$$\mathbb{P}(K_{\text{tr}} \geq k_{\text{tr}} \mid N_{\mathcal{B}}) = \sum_{k=k_{\text{tr}}}^{N_{\mathcal{B}}} \binom{N_{\mathcal{B}}}{k} \cdot \pi^k \cdot (1-\pi)^{N_{\mathcal{B}}-k}, \tag{7}$$

where $\pi = 1-(1-\hat{p}_{\mathcal{A}})^{\Delta+1}$. The null hypothesis is rejected at the desired significance level α if $\mathbb{P}(K_{\text{tr}} \geq k_{\text{tr}} \mid N_{\mathcal{B}}) < \alpha$. The null distribution and test for significance for the number of precursor coincidences can be derived completely analogously. Equation (7) is valid for event series that follow Bernoulli processes; they correspond to the null distributions derived for homogeneous Poisson processes in continuous-time ECA (Donges et al. (2016)). For other processes such analytical results have not been obtained, and, to date, Monte Carlo methods are required to simulate the null distribution.

4.2. *Coincidences with threshold exceedances.* Threshold exceedance series are typically not Bernoulli processes. We now derive novel analytical results when the lagging event series is a threshold exceedance series. Our key observation is that K_{tr} can be reformulated such that the Extremal Types Theorem (Coles (2001)) is applicable. We first define the number of trigger coincidences for a leading event series $\mathcal{E}$, lagging time series $\mathcal{X}$, time tolerance $\Delta \in \mathbb{N}_0$ and a threshold $\tau \in \mathbb{R}$ by substituting the threshold exceedance series into equation (2),

$$K_{\text{tr}} = K_{\text{tr}}^{\Delta,\tau}(\mathcal{E}, \mathcal{X}) := \sum_{t=1}^{T-\Delta} E_t \cdot \Big(\max_{\delta=0,\dots,\Delta} \mathcal{I}(X_{t+\delta} > \tau)\Big). \tag{8}$$

Observe that we can now swap the order of the max operator and the indicator function

$$\max_{\delta=0,\dots,\Delta} \mathcal{I}(X_{t+\delta} > \tau) = \mathcal{I}\Big(\Big(\max_{\delta=0,\dots,\Delta} X_{t+\delta}\Big) > \tau\Big) \tag{9}$$

and introduce the helper variables

$$Z_t^{\Delta,\tau} := \mathcal{I}\Big(\Big(\max_{\delta=0,\dots,\Delta} X_{t+\delta}\Big) > \tau\Big) \tag{10}$$

such that

$$K_{\text{tr}} = \sum_{t=1}^{T-\Delta} E_t \cdot Z_t^{\Delta,\tau} = \sum_{t:E_t=1} Z_t^{\Delta,\tau}. \tag{11}$$

The number of trigger coincidences is again a sum of Bernoulli trials, each associated with one of the $N_{\mathcal{E}}$ event occurrences in $\mathcal{E}$. If the time series $\mathcal{X}$ is strictly stationary, the Bernoulli trials are all identically distributed with the same marginal distribution $\mathbb{P}(Z_t^{\Delta,\tau})$, but they are, in general, not independent. The benefit of swapping the max-operator and the indicator function is that, now, the success probability of the Bernoulli trials

$$\mathbb{P}(Z_t^{\Delta,\tau}=1)=\mathbb{P}\Big(\max_{\delta=0,\dots,\Delta} X_{t+\delta}>\tau\Big)=1-\mathbb{P}\Big(\max_{\delta=0,\dots,\Delta} X_{t+\delta}\le\tau\Big) \tag{12}$$

is defined directly on the time series and not on the event series as in equation (5). Probabilities in the form of equation (12) have been studied extensively in Extreme Value Theory. In fact, according to the Extremal Types Theorem (ETT), the maximum of a large number of i.i.d. random variables is approximated by the Generalized Extreme Value (GEV) distribution, under mild constraints on the underlying distribution of the random variables:

EXTREMAL TYPES THEOREM (ETT, Coles (2001)). *Let $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} F$ and $M_n=\max_{i=1,\dots,n} X_i$. If there exist sequences of constants $\{a_n>0\}$ and $\{b_n\}$ such that*

$$\mathbb{P}\Big(\frac{M_n-b_n}{a_n}\le z\Big)\longrightarrow G(z)\quad \textit{as } n\longrightarrow\infty$$

for a nondegenerate distribution function G, then G is a member of the GEV family

$$G(z)=\exp\Big\{-\Big[1+\xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}\Big\},$$

defined on $\{z:1+\xi(z-\mu)/\sigma>0\}$, where $-\infty<\mu<\infty$, $\sigma>0$ and $-\infty<\xi<\infty$.

The ETT was shown to also apply more generally to the maxima of strictly stationary time series, if they fulfill a regularity condition that eliminates long-range dependencies; see Coles ((2001), Chapter 5.2) for technical details. Under the conditions of the ETT and for large Δ, the variables $Z_t^{\Delta,\tau}$ are thus identically distributed with

$$\mathbb{P}(Z_t^{\Delta,\tau}=1)\approx 1-G(\tau;\boldsymbol{\theta}_\Delta). \tag{13}$$

The normalizing constants from the ETT disappear in the GEV parameter vector $\boldsymbol{\theta}_\Delta=(\xi,\mu,\sigma)$ that depends on Δ. The parameters are estimated by splitting the time series $\mathcal{X}$ into consecutive blocks of size $\Delta+1$ and fitting the GEV distribution to the maxima of each block, for example, by maximum likelihood. The larger Δ, the better the approximation by the GEV distribution.[1]

This allows us to state the key result of this section: If the conditions of the ETT hold and events in $\mathcal{E}$ are sparse (such that the variables $Z_t^{\Delta,\tau}$ associated with the $N_{\mathcal{E}}$ event occurrences in $\mathcal{E}$ are approximately independent), then the null distribution of the number of trigger coincidences for large Δ is approximated by the binomial

$$K_{\text{tr}}^{\tau,\Delta}\mid N_{\mathcal{E}}\sim \text{Binomial}\big(N_{\mathcal{E}},1-G(\tau;\boldsymbol{\theta}_\Delta)\big). \tag{14}$$

Hence, the statistical significance of an observed number of trigger coincidences k_{tr} can be assessed using the p-value from equation (7) with $\pi=1-G(\tau;\boldsymbol{\theta}_\Delta)$.

[1] In Section B.1 we demonstrate that a value of $\Delta=7$ can be large enough for good approximations, if the threshold τ is a high quantile of the data.

4.3. *Examples.* With these results we are able to apply discrete-time ECA to analyze the triggers of exceedances of a *single* threshold. The threshold may be derived from domain-specific hypotheses. For example, we might want to test whether Islamist terrorist attacks systematically trigger bursts of more than 1000 posts per day on Twitter that contain the hashtag `#stopislam`. We observe $K_{\text{tr}} = 9$ trigger coincidences within a time tolerance of $\Delta = 7$ days after the events which gives a trigger coincidence rate of $r_{\text{tr}} = 0.53$. Assuming that the raw Twitter time series fulfills the conditions of the ETT, we obtain a p-value of $p = 0.0014$ using our GEV-based null distribution. Lacking a specific hypothesis for the value of the threshold, generic values can be tested, such as the empirical 95%-quantile. For example, we can test the hypothesis that Islamist terrorist attacks systematically trigger bursts of `#notinmyname` usage that exceed the volume of 95% of all days. We observe $K_{\text{tr}} = 4$ trigger coincidences with $\Delta = 7$ which gives a trigger coincidence rate of $r_{\text{tr}} = 0.24$. Again, assuming that the raw Twitter time series fulfills the ETT conditions, we obtain a p-value of $p = 0.2427$. These examples are only illustrative, since the raw Twitter time series are not strictly stationary and, thus, do not fulfill the ETT conditions. In Appendix A we describe the preprocessing scheme that we use for the results in Section 5 to make the time series stationary.

4.4. *Exceedances of increasing thresholds.* In case a reasonable threshold is unknown or if a full picture of the association with peaks of various magnitudes is required, exceedances at multiple thresholds have to be considered. Threshold exceedances at multiple levels are highly dependent: if an observation exceeds any threshold τ, it also exceeds all lower thresholds. The numbers of trigger coincidences at multiple thresholds are thus dependent as well. To enable joint analyses of multiple threshold exceedances and, thereby, eliminate the need of selecting a single fixed threshold, we now derive the joint null distribution of trigger coincidences at multiple thresholds and provide a novel visualization for this statistical association.

4.4.1. *Trigger coincidence processes.* Let $\boldsymbol{\tau} = (\tau_1, \dots, \tau_M)$ be a sequence of increasing thresholds $\tau_1 < \cdots < \tau_M$. The *trigger coincidence process*

$$\mathcal{K}_{\text{tr}} = \mathcal{K}_{\text{tr}}^{\Delta,\boldsymbol{\tau}}(\mathcal{E}, \mathcal{X}) = \left(K_{\text{tr}}^{\Delta,\tau_1}, \dots, K_{\text{tr}}^{\Delta,\tau_M}\right) \tag{15}$$

is the corresponding sequence of the numbers of trigger coincidences for all given thresholds $\boldsymbol{\tau}$. A trigger coincidence process is always monotonically decreasing. The *canonical trigger coincidence process* is given by the specific threshold sequence $\boldsymbol{\tau} = (\tau_1, \dots, \tau_M) = (X_{(1)}, \dots, X_{(T)})$, where $X_{(t)}$ denotes the order statistic of the time series such that $X_{(1)} < \cdots < X_{(T)}$. Trigger coincidence processes for other sequences of thresholds approximate the canonical trigger coincidence process. Figure 4 (bottom left) illustrates the concept in a simulated example; simulation details are in the caption. At low thresholds large numbers of trigger coincidences are observed both for the dependent and the independent event series. For higher thresholds the numbers of trigger coincidences for the dependent event series dramatically exceed the numbers of the independent event series. By construction all 32 events in the dependent series trigger an exceedance of the threshold 4; see the marker ($*$). The threshold 5 is exceeded after 13 out of 32 events from the dependent event series, while it is only exceeded after a single event from the independent event series; see the marker (†).

4.4.2. *Quantile-trigger rate plots.* Plots of trigger coincidence processes, as in Figure 4 (bottom left), help to visually assess whether events in $\mathcal{E}$ systematically trigger peaks of various magnitudes in a time series $\mathcal{X}$. However, the scales of the axes depend on the range of

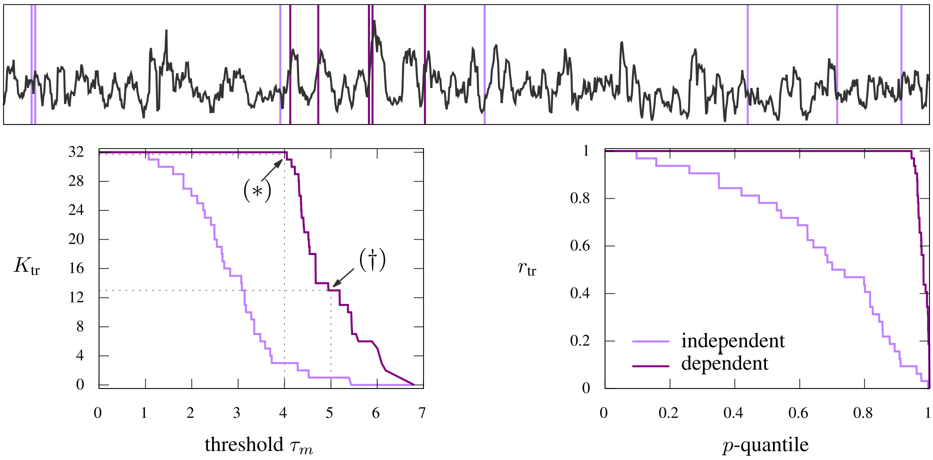

FIG. 4. *Canonical trigger coincidence processes (left) and corresponding QTR plot (right), for a simulated time series paired with two event series (independent and dependent, excerpts shown on top). We generated the time series of length $T = 4096$ from i.i.d. exponential random variables, applied a moving average (MA) filter of order* 8, *standardized and subtracted the minimum to obtain a nonnegative time series $\mathcal{X}$ with serial dependencies. We then generated two event series, an independent and a dependent one. To simulate a peak trigger relationship, we randomly sampled $N_{\mathcal{E}} = 32$ time steps t from the time series where $X_t > 4$ and set $E_{t-4} = 1$ for these t in the dependent event series. In the independent event series we distributed the* 32 *events completely at random. The time tolerance is set to $\Delta = 7$.*

values in $\mathcal{X}$ and the number of events $N_{\mathcal{E}}$ which makes it hard to compare these plots across multiple pairs of time series and event series. Furthermore, the absolute threshold value is not informative about the actual extremeness of a peak with respect to the bulk of the data. Therefore, we propose *quantile-trigger rate (QTR) plots* as a standardized visualization of trigger coincidence processes with normalized axes. In a QTR plot the x-axis is normalized by using empirical p-quantiles from $\mathcal{X}$ instead of the absolute thresholds τ_m, while the y-axis is normalized by using the trigger coincidence rate r_{tr} instead of the absolute number of trigger coincidences K_{tr}. The QTR plot for the example above is shown in Figure 4 (bottom right). The most striking difference is that, now, the dependent curve appears more extreme, since the thresholds larger than 4 correspond to high empirical p-quantiles. Intuitively, the closer an observed trigger coincidence process to the top-right corner of the QTR plot, the more events coincide with threshold exceedances, at more extreme levels.

However, QTR plots have to be interpreted with care. The shape of a trigger coincidence process for an *independent* pair of event series and time series in a QTR plot depends on the statistical properties of the input data. For example, if $\mathcal{X}$ is an i.i.d. time series and $\mathcal{E}$ an i.i.d. Bernoulli process, the fraction of events that coincide with an exceedance of the empirical p-quantile of $\mathcal{X}$ (with time tolerance $\Delta = 0$) is exactly $1 - p$, and the trigger coincidence process is a straight line from $(0, 1)$ to $(1, 0)$ in the QTR plot. Figure 5 illustrates the impact of serial dependencies in $\mathcal{X}$ and increasing time tolerance Δ on the shape of the trigger coincidence process under independence in a QTR plot. With increasing time tolerance Δ, there are more trigger coincidences under independence, and the lines in the QTR plot move toward the top-right corner of the plot. This effect is strongest for i.i.d. time series but also occurs for time series with serial dependencies. Thus, a line that bends toward the top-right corner of the QTR plot is necessary but not sufficient to conclude a trigger relationship. We need a statistical test that operates on the trigger coincidence process to assess whether the shape in a QTR plot is unusual under an independence assumption. For this purpose we now derive the statistical properties of the trigger coincidence process.

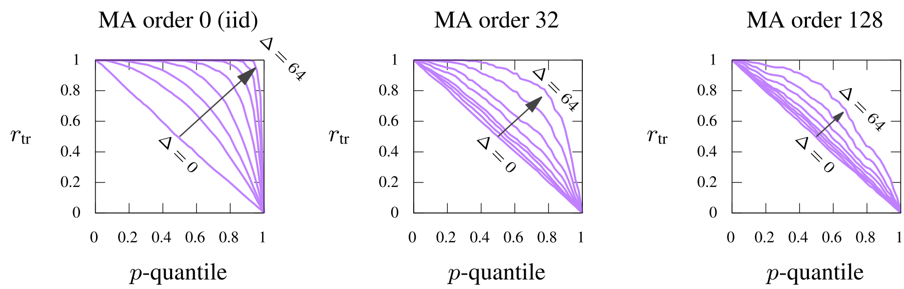

FIG. 5. *Expected QTR plots for three time series with different levels of serial dependencies* (*MA orders* 0, 32, 128) *and independent event series. For every MA order we simulate a single time series of length* $T = 4096$ *from the exponential moving average model described above and select* 50 *thresholds at equally spaced* p*-quantiles between* 0 *and* 1. *For every threshold* τ *and every* $\Delta \in \{0, 1, 2, 4, 8, 16, 32, 64\}$, *we estimate the expected trigger coincidence rate* $r_{\text{tr}}^{\Delta,\tau} = K_{\text{tr}}^{\Delta,\tau}/N_{\mathcal{E}}$ *for an independent event series by simulating* 100 *independent event series with* $N_{\mathcal{E}} = 32$ *events and averaging the trigger coicidence rates over the* 100 *runs. Note that, for large* Δ *and* τ, *this expectation can be approximated by the expected value of our GEV-based binomial distribution from equation* (14).

4.4.3. *Markov model.* To assess whether a trigger coincidence process is so unusual that the null hypothesis of independence has to be rejected, we consider the joint distribution $\mathbb{P}(\mathcal{K}_{\text{tr}}) = \mathbb{P}(K_{\text{tr}}^{\Delta,\tau_1}, \dots, K_{\text{tr}}^{\Delta,\tau_M})$. The product rule yields

$$\mathbb{P}(\mathcal{K}_{\text{tr}}) = \mathbb{P}\big(K_{\text{tr}}^{\Delta,\tau_1}\big) \cdot \prod_{i=2}^{M} \mathbb{P}\big(K_{\text{tr}}^{\Delta,\tau_i} \mid K_{\text{tr}}^{\Delta,\tau_1}, \dots, K_{\text{tr}}^{\Delta,\tau_{i-1}}\big). \tag{16}$$

We have already derived the marginal distribution $\mathbb{P}(K_{\text{tr}}^{\Delta,\tau})$ in equation (14) and now focus on the conditionals. Suppose there is an exceedance of τ_{i-1} in $\mathcal{X}$ within the window $t, \dots, t+\Delta$, that is, $Z_t^{\Delta,\tau_{i-1}} = 1$. The probability that there is also an exceedance of the higher threshold τ_i is

$$\mathbb{P}\big(Z_t^{\Delta,\tau_i} = 1 \mid Z_t^{\Delta,\tau_{i-1}} = 1\big) = \frac{\mathbb{P}(Z_t^{\Delta,\tau_i} = 1)}{\mathbb{P}(Z_t^{\Delta,\tau_{i-1}} = 1)} \approx \frac{1 - G(\tau_i; \boldsymbol{\theta}_\Delta)}{1 - G(\tau_{i-1}; \boldsymbol{\theta}_\Delta)}, \tag{17}$$

where we used equation (13) for the approximation. Equation (17) is valid whenever the marginal approximation by the GEV is admissible. We now rewrite the conditional random variable $K_{\text{tr}}^{\Delta,\tau_{i-1}} \mid K_{\text{tr}}^{\Delta,\tau_i}$ by restricting the summation from equation (11) to time steps with both $E_t = 1$ and $Z_t^{\Delta,\tau_{i-1}} = 1$ to incorporate our additional knowledge,

$$K_{\text{tr}}^{\Delta,\tau_i} \mid K_{\text{tr}}^{\Delta,\tau_{i-1}} = \sum_{t: E_t = 1, Z_t^{\Delta,\tau_{i-1}} = 1} Z_t^{\Delta,\tau_i}. \tag{18}$$

As in Section 4.2 we have a sum of identically distributed Bernoulli trials, with success probability now given by equation (17) due to the additional knowledge. Under the absence of long-range dependencies, the individual variables Z_t^{Δ,τ_i} are approximately independent, and the conditional number of trigger coincidences follows the binomial distribution

$$K_{\text{tr}}^{\Delta,\tau_i} \mid K_{\text{tr}}^{\Delta,\tau_{i-1}} \sim \text{Binomial}\bigg(K_{\text{tr}}^{\Delta,\tau_{i-1}}, \frac{1 - \mathrm{G}(\tau_i; \boldsymbol{\theta}_\Delta)}{1 - \mathrm{G}(\tau_{i-1}; \boldsymbol{\theta}_\Delta)}\bigg). \tag{19}$$

Thus, we rewrite the conditional distributions in equation (16) to a first-order Markov structure $\mathbb{P}(K_{\text{tr}}^{\Delta,\tau_i} \mid K_{\text{tr}}^{\Delta,\tau_1}, \dots, K_{\text{tr}}^{\Delta,\tau_{i-1}}) = \mathbb{P}(K_{\text{tr}}^{\Delta,\tau_i} \mid K_{\text{tr}}^{\Delta,\tau_{i-1}})$. The joint probability $\mathbb{P}(\mathcal{K}_{\text{tr}})$ of the trigger coincidence process under the null hypothesis of independence is then fully described by equation (14) for the smallest threshold and by equation (19) for all larger thresholds.

4.4.4. *Statistical test procedures.* With the results from Sections 4.2 and 4.4.3, we can now devise two test procedures to collect empirical evidence against the null hypothesis of independence and in favor of the alternative hypothesis that events systematically trigger exceedances at various thresholds:

1. We can employ our test procedure for *pointwise* exceedances of individual thresholds *multiple times* at all given thresholds and adjust the resulting p-values using methods for multiple hypothesis testing (Dudoit and van der Laan (2008)). A potential shortcoming of this procedure is that the dependency structure of trigger coincidence processes is ignored.
2. We can compute the joint probability of the observed trigger coincidence process and reject the null hypothesis if the whole process is unusually unlikely under the null distribution in the sense specified below. This approach takes the full dependency structure into account, but requires Monte Carlo simulations. We refer to it as the *multiple threshold test*.

For the second approach, observe that the trigger coincidence process is a high-dimensional discrete random variable, where every single realization, even the mode of the distribution, has a very small likelihood. We have to assess whether the observed likelihood is *unusually small* with respect to the *distribution of the likelihood values* under independence, that is, we treat the likelihood as a random variable. For numerical reasons we work with the negative log-likelihood $S(\mathcal{K}_{\mathrm{tr}}) = -\log \mathbb{P}(\mathcal{K}_{\mathrm{tr}})$ instead of the likelihood. Formally, we use S as a test statistic and reject the null hypothesis of independence at significance level α if the p-value $\mathbb{P}(S \geq s) < \alpha$, where $s = S(\mathcal{K}_{\mathrm{tr}}(\mathcal{E}, \mathcal{X}))$ is the observed value. We use Monte Carlo simulations to approximate this p-value. For this purpose we generate R independent event series $\mathcal{E}'$ with the same number of events as $\mathcal{E}$ by randomly permuting $\mathcal{E}$. For each independent event series we determine the test statistic value $s' = S(\mathcal{K}_{\mathrm{tr}}(\mathcal{E}', \mathcal{X}))$ and compute the Monte Carlo p-value (Davison and Hinkley (1997)) via $\hat{p} = \frac{1+|\{s' | s' \geq s\}|}{R+1}$.

5. Results and discussion. We now utilize our statistical methodology to analyze whether severe Islamist terrorist attacks in Western Europe and North America systematically trigger bursts of hate speech or counter-hate speech on Twitter. Additional simulations that support the methodological contributions described above can be found in Appendix B.

5.1. *Setup.* We use the data described in Section 3 which spans a total of $T = 1096$ days with $N_{\mathcal{E}} = 17$ events. We choose a time tolerance of $\Delta = 7$ days to allow enough time for the news about the incidents to spread globally. The simulation study in Section B.1 shows that this time tolerance is also large enough for our GEV-based null distributions to be accurate. For every social media time series $\mathcal{X}_i$, we estimate a GEV distribution G_i by splitting $\mathcal{X}_i$ into consecutive blocks of size $\Delta + 1$ and fitting the parameters of G_i to the block maxima by maximum likelihood estimation. We then select $M = 32$ thresholds $\boldsymbol{\tau}_i = (\tau_{i,1}, \dots, \tau_{i,32})$ at equidistant p-quantiles between 0.75 and 1 from $\mathcal{X}_i$ and use the GEV distribution G_i to obtain the parameters of the binomial distributions described by equations (14) and (19). We compute the observed trigger coincidence processes between the terrorist attack event series $\mathcal{E}$ and all social media time series $\mathcal{X}_i$ and obtain the respective test statistic values s_i for the multiple threshold test. We compute Monte Carlo p-values with $R = 10{,}000$ simulations. QTR plots with the Monte Carlo p-values are depicted in Figure 6. The plots are augmented with the marginally expected trigger coincidence processes under independence and the marginal 95% confidence intervals to additionally assess *pointwise* exceedances of individual thresholds.

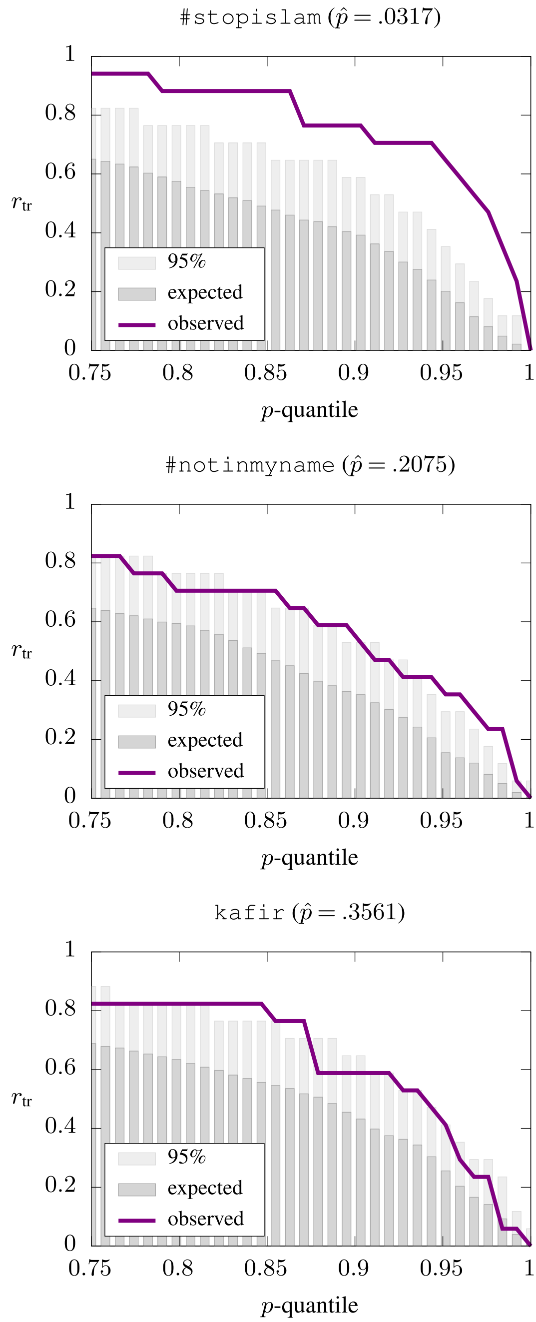

FIG. 6. *QTR plots for severe Islamist terrorist attacks and their online social media response.*

5.2. *Results*. The analysis shows that Islamist terrorist attacks in Western Europe and North America *systematically* trigger bursts of anti-Muslim hate speech on Twitter (#stopislam, $\hat{p} = 0.0317$). 90% of Islamist terrorist attacks triggered an exceedance of the 0.85-quantile, and 60% of Islamist terrorist attacks even triggered an exceedance of the 0.95-quantile. Our results confirm the findings of previous quantitative studies Burnap et al. (2014), Magdy, Darwish and Abokhodair (2015), Olteanu et al. (2018) with a novel statistical methodology and a larger study period: *there is a clear systematic relationship between Islamist extremist violence offline and anti-Muslim hate speech online*.

On the other hand, our analysis *does not* provide evidence for a *systematic* association between Islamist terrorist attacks and peaks in jihadist hate speech (`kafir`, $\hat{p} = 0.2075$) or Muslim counter-hate speech (`#notinmyname`, $\hat{p} = 0.3561$) in the study period. We stress that individual terrorist attacks may still have triggered such a social media response. Visual inspection of the data in Figure 1 suggests peaks in the hashtag `#notinmyname` for Islamist terrorist attacks before July 2016. Hashtag usage is typically subject to trends, so a systematic relationship can only be established for hashtags that are used consistently throughout the study period. The impact of an *individual* terrorist attack on the social media time series can be assessed, for example, with counterfactual analyses (Brodersen et al. (2015)), regardless of trends.

Figure 6 also shows that even for jihadist hate speech and Muslim counter-hate speech, the observed numbers of trigger coincidences fall outside the *pointwise* 95% confidence intervals for some thresholds. Pointwise tests at these specific thresholds would reject the null hypothesis of independence on the basis of only a narrow perspective on the total association. The multiple threshold test thus decreases the dangers of data dredging at the cost of a lower sensitivity. To validate the results from the multiple threshold test, we computed all p-values for the pointwise tests at all thresholds and used different adjustment methods that control the familywise error rate at level $\alpha = 0.05$: Bonferroni, single-step Šidák, step-down Holm in its original variant and in the Šidák variant (Dudoit and van der Laan (2008)). For all multiple test adjustment methods the results agree with our multiple threshold test.

5.3. *Sensitivity analysis.* To assess the stability of the results, we further experimented with different choices for the time tolerance $\Delta = 4 \dots 16$ (ceteris paribus). We found that for $\Delta = 4 \dots 8$, the results of all tests on all time series are unchanged. For $\Delta = 9 \dots 14$ our multiple threshold test fails to reject the null hypothesis for the `#stopislam` time series, while the multiple pointwise test procedures still reject. For $\Delta = 15$ only the Šidák procedures reject the null hypothesis on `#stopislam`, while for $\Delta = 16$ no test procedure rejects the null hypotheses on any time series. Choosing a time tolerance Δ that is longer than necessary thus reduces the sensitivity of the tests. We also varied the number of thresholds M between 8 and 64 (ceteris paribus) which did not change the outcome of any test. At last, we moved the thresholds upward to more extreme levels by choosing equidistant p-quantiles from the ranges 0.85 to 1 and 0.95 to 1, respectively (ceteris paribus). The outcomes on the `#stopislam` and `kafir` time series remain unchanged, while our multiple threshold test now detects an additional trigger relationship for `#notinmyname` that is not detected by the multiple pointwise test procedures. Overall, the trigger relationship for `#stopislam` is very stable across all test procedures with different parameterizations, whereas the results on `#notinmyname` are inconclusive.

6. Conclusions. We have refined the statistical methodology to infer potential causal links between an event series and peaks in a time series. Based on the framework of event coincidence analysis, the tests focus only on the *timing* of events and peaks and no other distributional characteristics. We have derived analytical expressions for the null distributions of the ECA statistic for coincidences with exceedances of a single threshold and multiple thresholds. Our results are valid if the lagging time series satisfies the regularity conditions of the Extremal Types Theorem. We further require event occurrences in the leading event series to be separated by a sufficient number of time steps such that the binomial null distributions are valid. Our analysis is, therefore, most suitable for sparse event series. For a complete causal analysis, confounding factors that influence both series must still be ruled out. This is a challenging direction for future work, as it either requires the specification of a joint model for the confounding factors, the time series and the event series in the spirit of Granger causality (Granger (1969)) or nontrivial changes in the nonparametric procedure of ECA. A first step in this direction is conditional ECA and joint ECA (Siegmund et al. (2016)).

APPENDIX A: DESCRIPTION OF SOCIAL MEDIA DATA

The hashtag `#stopislam` has been observed in anti-Muslim hate speech before (Magdy, Darwish and Abokhodair (2015), Olteanu et al. (2018)) and has also received some media attention (Dewey (2016), Hemmings (2016)). Many posts that contain the hashtag actually condemn its usage, so spikes in the volume should not be seen as pure bursts of hate speech. Yet, such condemnation is typically triggered by initial anti-Muslim posts. Due to the mixed usage, the magnitude of a spike is no indicator for the *extent* of online hate, only the *presence* of a spike is informative.

The phrase "not in my name" is used by members of a group to express their disapproval of actions that are associated with that group or (perceived or actual) representatives of the group (Tormey (2006), Čehajić and Brown (2008)). It was observed, for example, during global protests against the 2003 war of the U.S.-led coalition against Iraq (Bennett (2005)), or more recently during protests sparked by the murder of a Muslim boy by Hindu nationalists in India 2017 (Krishnan (2017)). Most importantly for the present study, Muslim social media users have repeatedly used the hashtag after Islamist terrorist attacks (Davidson (2014)). Due to the generic nature of the phrase, it cannot solely be viewed as Muslim counter-hate speech. Nonetheless, online social media posts that contain `#notinmyname` right after Islamist terrorist attacks are likely to convey a Muslim counter-hate message.

At last, the Arabic word `kafir` translates to the English word "nonbeliever." It is traditionally used by Muslim fundamentalists against other Muslims that do not adhere to the fundamentalist ideology (Alvi (2014)) but also against non-Muslims (Bartlett and Miller (2012)) in both cases to justify their killing. The occurrence of the keyword `kafir` within online social media posts was recently shown to be a strong indicator for jihadist hate speech (De Smedt, De Pauw and Van Ostaeyen (2018)). We use male, female and plural forms (`kafir`–`kafirah`–`kuffar`) in Arabic script for the query.

We used the ForSight platform provided by Crimson Hexagon[2] to retrieve daily time series of the global Twitter volume for our keywords. We excluded posts with the keyword `RT` to ignore retweets. The time series are based on the full Twitter stream which makes the numbers exact. We preprocessed the original time series by taking the logarithm to base 2 and subtracting the running mean over the past 30 days to make them stationary.

APPENDIX B: SIMULATION STUDY

B.1. Comparison of the null distributions. The central result from Section 4.2 is that, under the null hypothesis of independence (and some constraints on the time series), the number of trigger coincidences for a single threshold approximately follows the binomial distribution from equation (14), where the success probability is obtained from a GEV distribution. This approximate result is useful, specifically, for the case of time series with serial dependencies, where the Bernoulli-based null distribution from equation (6) cannot be applied. We now demonstrate that the Bernoulli-based null distribution indeed fails to describe the empirically observed numbers of trigger coincidences for time series with serial dependencies, while our GEV-based null distribution accurately describes the observed data.

For this purpose we simulate three time series with MA orders of 0, 32 and 64. For every time series we simulate 1000 independent pairs of event series with $N_{\mathcal{E}} = 32$ events and record the numbers of trigger coincidences at the three thresholds $\tau \in \{3, 4, 5\}$ with time tolerance $\Delta = 7$. For every time series and choice of threshold, we compare the empirically obtained (Monte Carlo) null distribution with the two analytical null distributions. All cumulative probability mass functions are visualized in Figure 7. The visualizations clearly show

[2] https://www.crimsonhexagon.com/

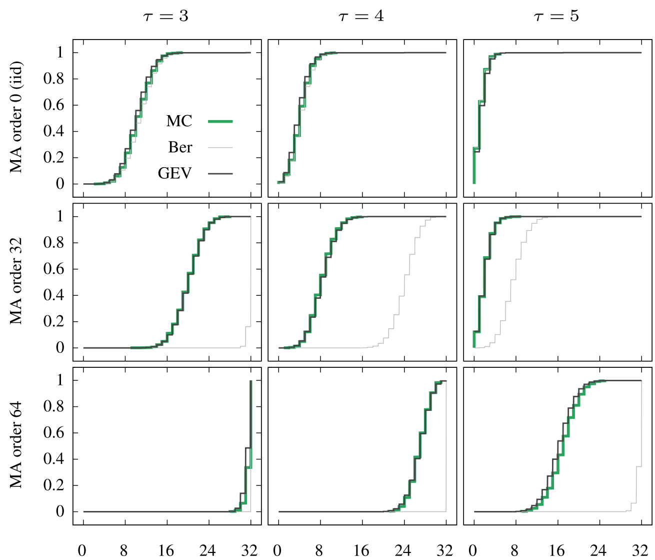

FIG. 7. *Cumulative probability mass functions for the number of trigger coincidences under independence, obtained empirically by Monte Carlo simulations* (*MC*) *and analytically with the Bernoulli-based binomial* (*Ber*) *and the GEV-based binomial* (*GEV*).

that our GEV-based estimate closely follows the empirical distribution in all runs, while the Bernoulli-based estimate is only correct for i.i.d. time series. These results also demonstrate that a value of $\Delta = 7$ is already large enough for the GEV approximation to be valid.

B.2. Analysis of the Markov model. Our second central result is the Markov model for trigger coincidence processes from Section 4.4 with the associated test statistic for the multiple threshold test. Larger values of the test statistic should correspond with visually more "extreme" trigger coincidence processes in a QTR plot. To confirm this assumption, we illustrate the test statistic values for a single simulated time series (MA order 8) and 1000 independent event series (with 32 events). We use a time tolerance $\Delta = 7$ and 32 thresholds at equally spaced p-quantiles between 0.75 and 1 from the time series. All resulting trigger coincidence processes are plotted in Figure 8 and colorized by their test statistic values. We also

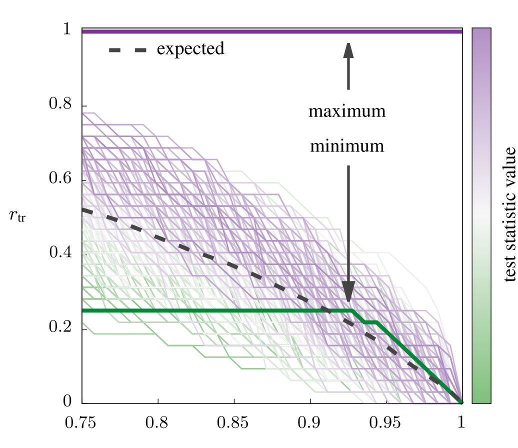

FIG. 8. *Simulated trigger coincidence processes under independence, colorized by the test statistic value along with the processes that attain the theoretical minimum and maximum test statistic value.*

plot the trigger coincidence process with the highest (lowest) test statistic value that is theoretically possible; we obtain them by maximizing (minimizing) the test statistic over all possible processes with a dynamic programming approach. At last, we show the marginally expected trigger coincidence process at every threshold τ_m, that is, the value of $\mathrm{E}[K_{\mathrm{tr}}^{\Delta,\tau_m} \mid N_{\mathcal{E}} = 32]$ obtained from equation (14). All simulated trigger coincidence processes are close to the marginally expected sequence; the more they bend toward the top-right corner of the plot, the higher the test statistic value. The trigger coincidence process with the highest possible test statistic value closely traces the top-right corner which corresponds to our intuitive notion of the most unusual outcome: all events trigger exceedances of the highest quantiles.

SUPPLEMENTARY MATERIAL

Source codes and data (DOI: 10.1214/20-AOAS1338SUPP; .zip). To reproduce simulations and hate speech results.

REFERENCES

ALVI, H. (2014). The diffusion of intra-Islamic violence and terrorism: The impact of the proliferation of Salafi/Wahhabi ideologies. *Middle East Rev. Int. Aff.* **18** 38–50.

BARTLETT, J. and MILLER, C. (2012). The edge of violence: Towards telling the difference between violent and non-violent radicalization. *Terrorism Polit. Violence* **24** 1–21.

BENNETT, W. L. (2005). Social movements beyond borders: Organization, communication, and political capacity in two eras of transnational activism. In *Transnational Protest and Global Activism* (D. Della Porta and S. Tarrow, eds.) 203–226. Rowman & Littlefield Publishers, Inc., Lanham, MD, USA.

BESSERVE, M., LOGOTHETIS, N. K. and SCHÖLKOPF, B. (2013). Statistical analysis of coupled time series with kernel cross-spectral density operators. In *Advances in Neural Information Processing Systems* (*NIPS*).

BOWMAN-GRIEVE, L. (2009). Exploring "Stormfront": A virtual community of the radical right. *Stud. Confl. Terrorism* **32** 989–1007.

BOX, G. E. P. and TIAO, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *J. Amer. Statist. Assoc.* **70** 70–79. MR0365957

BRESSLER, S. L. and SETH, A. K. (2011). Wiener–Granger causality: A well established methodology. *NeuroImage* **58** 323–329.

BRODERSEN, K. H., GALLUSSER, F., KOEHLER, J., REMY, N. and SCOTT, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.* **9** 247–274. MR3341115 https://doi.org/10.1214/14-AOAS788

BURNAP, P. and WILLIAMS, M. L. (2014). Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. In *Proceedings of the Internet, Policy and Politics Conference* (*IPP*).

BURNAP, P. and WILLIAMS, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Sci.* **5** 1–15.

BURNAP, P., WILLIAMS, M. L., SLOAN, L., RANA, O., HOUSLEY, W., EDWARDS, A., KNIGHT, V., PROCTER, R. and VOSS, A. (2014). Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Soc. Netw. Anal. Min.* **4** 1–14.

ČEHAJIĆ, S. and BROWN, R. (2008). Not in my name: A social psychological study of antecedents and consequences of acknowledgment of in-group atrocities. *Genocide Stud. Prev.* **3** 195–211.

CHANDRASEKHARAN, E., PAVALANATHAN, U., SRINIVASAN, A., GLYNN, A., EISTENSTEIN, J. and GILBERT, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Human-Comput. Interact.* **1** 31:1–31:22.

CHWIALKOWSKI, K. and GRETTON, A. (2014). A kernel independence test for random processes. In *Proceedings of the* 31*st International Conference on Machine Learning* (*ICML*).

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. *Springer Series in Statistics*. Springer London, Ltd., London. MR1932132 https://doi.org/10.1007/978-1-4471-3675-0

CUNHA, T. O., WEBER, I. and PAPPA, G. L. (2017). A warm welcome matters! The link between social feedback and weight loss in /r/loseit. In *Proceedings of the* 26*th International Conference on World Wide Web* (*WWW*).

DAVIDSON, H. (2014). 'Not in my name' campaign organiser warns of danger of political rhetoric. *The Guardian*. Available at https://www.theguardian.com/australia-news/2014/oct/03/not-in-my-name-campaign-organiser-warns-of-danger-of-political-rhetoric, 2014-10-03.

DAVIDSON, T., WARMSLEY, D., MACY, M. and WEBER, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the* 11*th International Conference on Web and Social Media* (*ICWSM*).

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. *Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673 https://doi.org/10.1017/CBO9780511802843

DE KOSTER, W. and HOUTMAN, D. (2008). 'Stormfront is like a second home to me.' On virtual community formation by right-wing extremists. *Inf*. *Commun*. *Soc*. **11** 1155–1176.

DE SMEDT, T., DE PAUW, G. and VAN OSTAEYEN, P. (2018). Automatic detection of online jihadist hate speech. Technical report, CLiPS Research Center, Univ. Antwerp.

DEWEY, C. (2016). #StopIslam Twitter-trended for all the right reasons. *Washington Post*. Available at https://www.washingtonpost.com/news/the-intersect/wp/2016/03/22/stopislam-twitter-trended-for-all-the-right-reasons/, 2016-03-22.

DJURIC, N., ZHOU, J., MORRIS, R., GRBOVIC, M., RADOSAVLJEVIC, V. and BHAMIDIPATI, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of Th* 24*th International Conference on World Wide Web* (*WWW*).

DONGES, J. F., SCHLEUSNER, C. F., SIEGMUND, J. F. and DONNER, R. V. (2016). Event coincidence analysis for quantifying statistical interrelationships between event time series: On the role of flood events as triggers of epidemic outbreaks. *Eur. Phys*. *J*. *Spec*. *Top*. **487** 471–487.

DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. *Springer Series in Statistics*. Springer, New York. MR2373771 https://doi.org/10.1007/978-0-387-49317-6

ELSHERIEF, M., NILIZADEH, S., NGUYEN, D., VIGNA, G. and BELDING, E. (2018a). Peer to peer hate: Hate speech instigators and their targets. In *Proceedings of the* 12*th International AAAI Conference on Web and Social Media* (*ICWSM*).

ELSHERIEF, M., KULKARNI, V., NGUYEN, D., WANG, W. Y. and BELDING, E. (2018b). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the* 12*th International AAAI Conference on Web and Social Media* (*ICWSM*).

NATIONAL CONSORTIUM FOR THE STUDY OF TERRORISM (2018). Global terrorism database [2014–2017]. Available at https://www.start.umd.edu/gtd.

FRAHM, G., JUNKER, M. and SCHMIDT, R. (2005). Estimating the tail-dependence coefficient: Properties and pitfalls. *Insurance Math*. *Econom*. **37** 80–100. MR2156598 https://doi.org/10.1016/j.insmatheco.2005.05.008

GAGLIARDONE, I., GAL, D., ALVES, T. and MARTINEZ, G. (2015). *Countering Online Hate Speech*. UNESCO, Paris, France.

GRANGER, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37** 424–438.

HEMMINGS, C. (2016). From #StopIslam to Allison Pearson and Katie Hopkins, the social media response to Brussels has been shocking. *The Independent*. Available at https://www.independent.co.uk/voices/from-stopislam-to-allison-pearson-and-katie-hopkins-the-social-media-response-to-brussels-has-been-a6946116.html, 2016-03-22.

KRISHNAN, M. (2017). 'Not in my name'—Indians protest vigilante attacks on Muslims. *Deutsche Welle*. Available at https://p.dw.com/p/2fYJP, 2017-06-29.

KWOK, I. and WANG, Y. (2013). Locate the hate: Detecting tweets against blacks. In *Proceedings of the* 27*th AAAI Conference on Artificial Intelligence* (*AAAI*).

MAGDY, W., DARWISH, K. and ABOKHODAIR, N. (2015). Quantifying public response towards Islam on Twitter after Paris attacks. Preprint. Available at arXiv:1512.04570.

MATIAS, J. N., JOHNSON, A., BOESEL, W. E., KEEGAN, B., FRIEDMAN, J. and DETAR, C. (2015). *Reporting*, *Reviewing*, *and Responding to Harassment on Twitter*. Women, Action, and the Media.

MEDDAUGH, P. M. and KAY, J. (2009). Hate speech or "Reasonable racism?" The other in stormfront. *J. Mass Media Ethics* **24** 251–268.

MONDAL, M., SILVA, L. A. and BENEVENUTO, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the* 28*th ACM Conference on Hypertext and Social Media* (*HT*).

MÜLLER, K. and SCHWARZ, C. (2019). Fanning the flames of hate: Social media and hate crime. *SSRN Electron*. *J*. 1–84. https://doi.org/10.2139/ssrn.3082972

OLTEANU, A., CASTILLO, C., BOY, J. and VARSHNEY, K. R. (2018). The effect of extremist violence on hateful speech online. In *Proceedings of the* 12*th International AAAI Conference on Web and Social Media* (*ICWSM*).

QUIROGA, R. Q., KREUZ, T. and GRASSBERGER, P. (2002). Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Phys*. *Rev*. *E* (3) **66** 041904, 9. MR1935193 https://doi.org/10.1103/PhysRevE.66.041904

RIBEIRO, M. H., CALAIS, P. H., SANTOS, Y. A., ALMEIDA, V. A. F. and MEIRA, W. (2018). Characterizing and detecting hateful users on Twitter. In *Proceedings of the* 12*th International AAAI Conference on Web and Social Media* (*ICWSM*).

SAHA, K., CHANDRASEKHARAN, E. and DE CHOUDHURY, M. (2019). Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the* 11*th ACM Conference on Web Science* (*WebSci*).

SCHARWÄCHTER, E. and MÜLLER, E. (2020). Two-sample testing for event impacts in time series. In *Proceedings of the SIAM International Conference on Data Mining* (*SIAM SDM*). https://doi.org/10.1137/1.9781611976236.2

SCHARWÄCHTER, E. and MÜLLER, E. (2020). Supplement to "Does terrorism trigger online hate speech? On the association of events and time series." https://doi.org/10.1214/20-AOAS1338SUPP

SCHLEUSSNER, C.-F., DONGES, J. F., DONNER, R. V. and SCHELLNHUBER, H. J. (2016). Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries. *Proc. Natl. Acad. Sci. USA* **113** 9216–9221.

SCHMIDT, A. and WIEGAND, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the* 5*th International Workshop on Natural Language Processing for Social Media* (*SocialNLP*).

SCHREIBER, T. (2000). Measuring information transfer. *Phys. Rev. Lett.* **85** 461–464.

SIEGMUND, J. F., SANDERS, T. G. M., HEINRICH, I., VAN DER MAATEN, E., SIMARD, S., HELLE, G. and DONNER, R. V. (2016). Meteorological drivers of extremes in daily stem radius variations of beech, oak, and pine in northeastern Germany: An event coincidence analysis. *Front. Plant Sci.* **7** 733. https://doi.org/10.3389/fpls.2016.00733

SILVA, L., MONDAL, M., CORREA, D., BENEVENUTO, F. and WEBER, I. (2016). Analyzing the targets of hate in online social media. In *Proceedings of the* 10*th International Conference on Web and Social Media* (*ICWSM*).

TORMEY, S. (2006). 'Not in my name': Deleuze, Zapatismo and the critique of representation. *Parliam. Aff.* **59** 138–154.

WARNER, W. and HIRSCHBERG, J. (2012). Detecting hate speech on the World Wide Web. In *Proceedings of the* 2*nd Workshop on Language in Social Media* (*LSM*).

WILLIAMS, M. L. and BURNAP, P. (2016). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *Br. J. Criminol.* **56** 211–238.

YAN, X., WU, Q. and ZHANG, W. (2019). Cross-sectional learning of extremal dependence among financial assets. In *Neural Information Processing Systems*.