PLOS ONE

RESEARCH ARTICLE

Local Variation of Hashtag Spike Trains and Popularity in Twitter

Ceyda Sanlı*, Renaud Lambiotte

CompleXity and Networks, naXys, Department of Mathematics, University of Namur, Namur, Belgium

* cedaysan@gmail.com

Citation: Sanlı C, Lambiotte R (2015) Local Variation of Hashtag Spike Trains and Popularity in Twitter. PLoS ONE 10(7): e0131704. doi:10.1371/journal.pone.0131704

Editor: Eduardo G. Altmann, Max Planck Institute for the Physics of Complex Systems, GERMANY

Received: March 20, 2015

Accepted: June 4, 2015

Published: July 10, 2015

Data Availability Statement: All data files are publicly available from the Twitter API (Application Programming Interface).

Funding: This work was supported by grant number F.N.R.S MIS F4527.12 48888F3 (grant holder: RL, funding receiver: CS, http://www.fnrs.be/), the EU 7th Framework OptimizR Project: 48909A2 CE OPTIMIZR (grant holder: RL, funding receiver: CS, http://optimizr.eu/), and the National Institute of Informatics Tokyo (http://www.nii.ac.jp/en/) for partial traveling support. This funder had a role in preparation of the manuscript (visit for scientific exchange of ideas), but did not have a further role in

Abstract

We draw a parallel between hashtag time series and neuron spike trains. In each case, the process presents complex dynamic patterns including temporal correlations, burstiness, and all other types of nonstationarity. We propose the adoption of the so-called local variation in order to uncover salient dynamical properties, while properly detrending for the time-dependent features of a signal. The methodology is tested on both real and randomized hashtag spike trains, and identifies that popular hashtags present regular and so less bursty behavior, suggesting its potential use for predicting online popularity in social media.

Introduction

In this paper, we focus on the statistical properties of Twitter and, in particular, on the dynamics and popularity of hashtags. Twitter is a micro-blogging service allowing users to post short messages and to follow those published by other users. Messages often incorporate hashtags, keywords identified by the symbol #, which users can track and respond to the message content and makes the platform interactive. Hashtags play a significant role in information diffusion by enhancing information and rumor spreading and consequently increase the impact of news. Discussions on protests [1, 2] and political elections, advertisement of new products in marketing, announcements of scientific innovations [3], panic events such as earthquakes [4], and comments on TV shows are some examples where hashtags are widely used. Additionally, hashtags can be even used to track and locate crisis [5] and can spread under the influences of both endogeneous factors, that is the propagation between Twitter users following each others, and exogeneous sources such as TV and newspapers [6].

The statistical properties of Twitter and, more generally, of human activity, are characterized by a strong heterogeneity in different dimensions. First, human behavior is known to generate bursty temporal patterns, significantly deviating from independent Poisson processes, as a majority of events take place over short time scales while a few events take place over very large times. This property translates into fat-tailed distributions for the timings $\Delta\tau$ between occurrences of a certain type of events, e.g. between two phone calls or two emails emitted by an individual. For instance, the inter-event time distribution $P(\Delta\tau)$ for the timings between two tweets of a user, or the use of a hashtag is well fitted by a power law such as $P(\Delta\tau) \approx \Delta\tau^{\alpha}$ [3]. The deviation from an exponential (uncorrelated) distribution may be either driven by

study design, data collection and analysis or decision to publish.

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

complex decision-making and cascading mechanisms [7–9] or by the time dependency of the underlying process, partly because of its intrinsic circadian and weekly rhythms [10, 11], as described in Fig 1, or by a combination of these factors [12–15]. Importantly, the nonstationarity of the signal is known to broaden $P(\Delta\tau)$ and therefore to artificially increase the value of standard metrics, such the variance or the Fano factor, originally defined for stationary processes.

In addition to temporal heterogeneity in $\Delta\tau$, online human activity often generates a heterogeneity in popularity [16]. The popularity p of a hashtag is measured by the number of times that it appears in an observation time window. While a majority of hashtags attracts no attention only very few of them propagate heavily [8, 17]. Understanding the mechanisms by which certain hashtags or messages gain attention is a central topic of research in the study of online social media [18]. Potential mechanisms for the emergence of this heterogeneity include forms of preferential attachment and competition-induced forces [19–22] driven by the limited amount of attention of users.

Our main purpose is to explore connections between temporal heterogeneity and heterogeneity in popularity. As a first contribution, we introduce a temporal measure for online human dynamics, suited for the analysis of nonstationary time series to quantify bursts, regularity, and temporal correlations. Originally defined for the study of inter-spike intervals of neurons [23–27], the so-called local variation L_V is then shown to identify deviations from Poisson (uncorrelated) processes and to help characterize successful hashtags.

Data mining and basic analysis

Data collection and basic overview

The data set has been collected via the publicly open Twitter streaming API between April 30, 2012, 10 pm and May 10, 2012, 10 pm. Only the geographical constraint has been applied as follows: The actions of all Twitter users located in France have been considered to avoid the existence of time differences between countries and regions, and no language filtering has been applied. The time resolution is 1 second and multiple activity can be recorded in the same second. During this time period, two major public events took place: An important political debate held on May 2 and the French presidential election-2012 held on May 6. These events are not the topic of this work, but they are clearly visible in the time series, as shown in Fig 1.

The total number of tweets, including retweets, captured during the data collection is 9,747,351. The total number of tweets including at least one hashtag is 2,942,239. Around 30% of the tweets therefore contain a hashtag. The fact that hashtags are used in regular tweets or in retweets is not specified. Moreover, any message (identical or not) considering at least one hashtag is recorded. Due to the debate and the election taking place during the data collection, the most popular hashtags are related to politics, as seen in Table 1. The time series of the hashtag study in this paper are provided in Supporting Information (S1 File). A total number of 473,243 individual users has been identified. Among those, 228,525 users published at least one hashtag, e.g. almost half of the social network is associated with hashtag diffusion. To further characterize the importance of hashtags in Twitter activity, we compare the total number of seconds when any action is performed in the data set, 763,262 s $\approx$ 8.8 days and thus 88% of the total duration, to the number of seconds when at least one hashtag is published, 667,996 s $\approx$ 7.7 days, that is 77% of the total duration. In any case, the hashtag data cover a majority of the time window, even during off-peak hours. These numbers confirm the importance of hashtags in the Twitter ecosystem and their prevalence in a variety of contexts.

Any type of human activity is influenced by circadian and weekly cycles. This observation has been verified in recent years in a variety of social data sets, going from mobile phone [12]

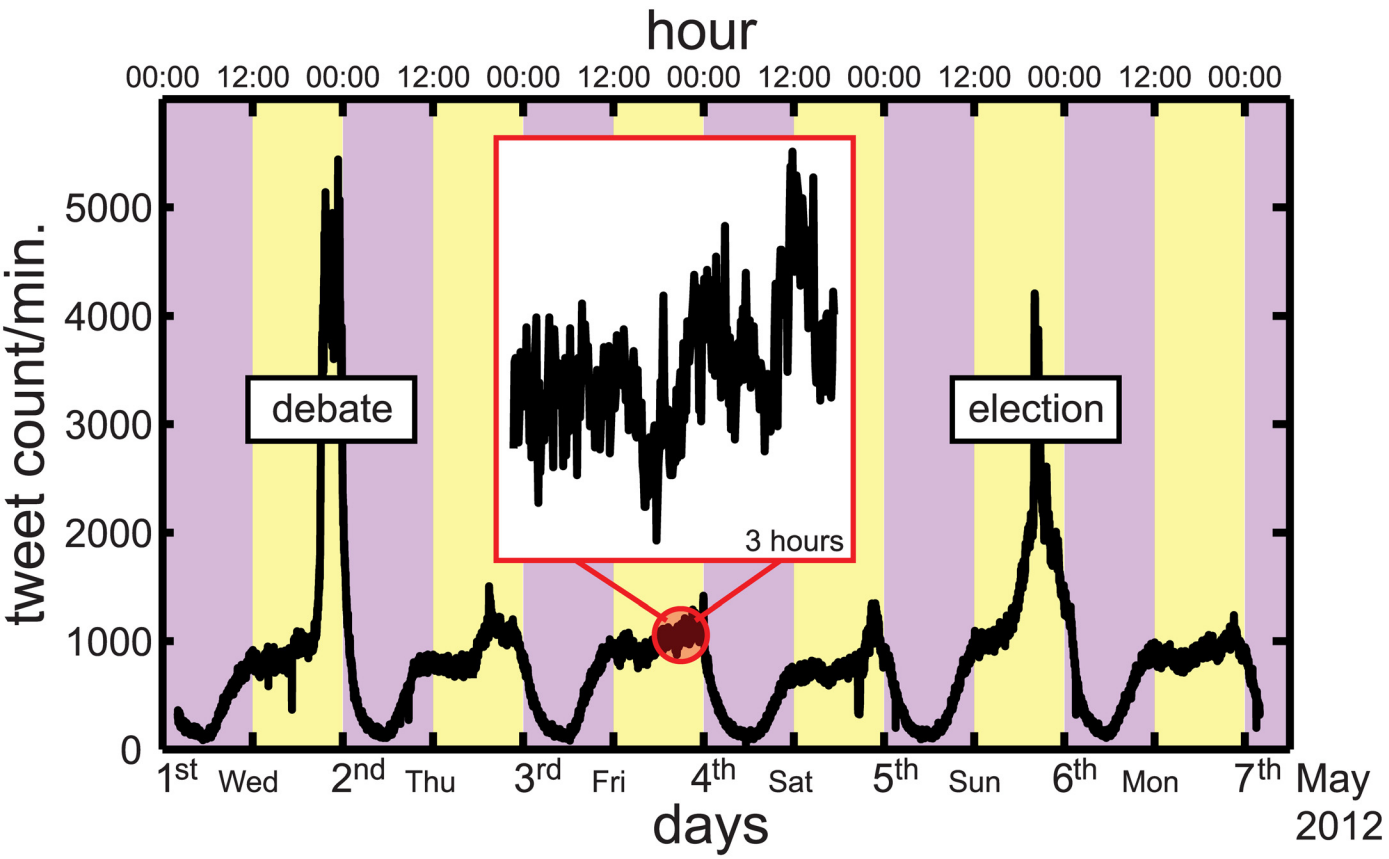

Fig 1. Circadian pattern of tweeting activity. Increasing amount of tweets from midday (12:00) to midnight (00:00) is shown in the yellow shaded regions. Significant decays of the activity are observed during nights. The activity increases during mornings as shown in purple shaded rectangles. In the inset, we show the temporal evolution at a finer scale, where fluctuations are visible. The data exhibit two peaks: The first one is in the evening of a political debate, on May 2 2012 and the second is on the French presidential election day, May 6 2012.

doi:10.1371/journal.pone.0131704.g001

to online social media [13–15]. In addition, deviations from these cycles can help at detecting atypical events such as responses to catastrophes [3–5]. Fig 1 in Introduction shows the total number of tweets per minute over a sub-period of 6 days and confirms these findings, with clear circadian patterns and two peaks during major public events related to the French presidential election-2012. Besides this smooth periodic behavior, the data also exhibit a noisy signal at a finer time scale, as shown in the inset of Fig 1. In the following, we will analyze the properties of these complex time series, by decomposing it into groups of hashtags depending on their popularity, and uncover temporal statistical differences between these groups.

Heterogeneity in popularity of hashtags

The success of a hashtag can be measured by its popularity p, defined as its number of occurrences, and equivalent to its frequency. Fig 2 presents the Zipf-plot and the probability density function (PDF) of p, for the 295,697 unique hashtags observed in the data set. The Zipf-plot [Fig 2(a)] indicates that more than half of the hashtags ($\approx 60\%$) appears just once in the data set, with $p = 1$. Moreover, around 83% of the hashtags has $p < 5$, in the pink-colored region in the last (right) rectangle of Fig 2(a). For moderate values of p, if we set a threshold of p to 1000

Table 1. Ranking of popular hashtags. The first 40 most used hashtags are listed with the corresponding popularity p. The hashtags related to the debate and the presidential election such as ledebat, hollande, sarkozy, votehollande, france2012, and présidentielle are recognized.

rank	hashtag	popularity p	rank	hashtag	popularity p
1	ledebat	180946	21	ns	18715
2	hollande	143636	22	ps	18492
3	sarkozy	116906	23	teamfollowback	18476
4	votehollande	99908	24	ggi	17734
5	radiolondres	97622	25	bastille	16056
6	bahrain	71571	26	présidentielle	13799
7	fh2012	67759	27	afp	13710
8	avecsarkozy	67549	28	france2	12906
9	ledébat	66668	29	syria	11594
10	ff	49499	30	psg	10566
11	ns2012	40337	31	sarko	10503
12	ump	25125	32	tf1	10201
13	thevoice	24696	33	mutualite	10093
14	fr	24249	34	egypt	9970
15	bayrou	23029	35	lavictoire	9949
16	fh	22369	36	fn	9763
17	rt	21598	37	franceforte	9626
18	france2012	20635	38	placeaupeuple	9211
19	reseaufdg	19488	39	jemesouviens	9098
20	france	19268	40	bfmtv	9010

doi:10.1371/journal.pone.0131704.t001

with an upper-bound to 25000, only 0.15% of the hashtags fits in the yellow-colored rectangle. Finally, the top hashtags with $p > 25000$, in the red-colored rectangle, are very rare ($\approx$ 0.0001%), but more frequent than would be expected for values so large as compared to the median. These observations are confirmed in Fig 2(b), where we show the probability distribution of p, $P(p)$ in a log-log plot. $P(p)$ is a clear example of a fat-tailed distribution associated with a strong heterogeneity in the system.

The heterogeneity in p has been already observed [8, 11, 16, 17]. A mechanism proposed for its emergence is the competition between information overload and the limited capacity of each user [19–22], sometimes coupled with cooperative effects [8, 9]. It has been also shown that hashtags having unique textual features become more popular than hashtags presenting common textual features [28]. In this paper, we are not interested in the origin of the heterogeneity, but in its relation with the temporal characteristics of hashtags.

Hashtag spike trains

Temporal heterogeneity

We will draw an analogy between hashtag dynamics and neuron spike trains. To this end, we introduce standard methods from the spike train analysis into the field of hashtag dynamics. Hashtags are keywords associated to different topics, which can be created, tracked and reused by users. Their popularity and unambiguity make them an essential object for information diffusion in Twitter. The statistical description of neuron spike sequences is crucial for extracting underlying information about the brain [29]. It was originally believed that in vivo cortical

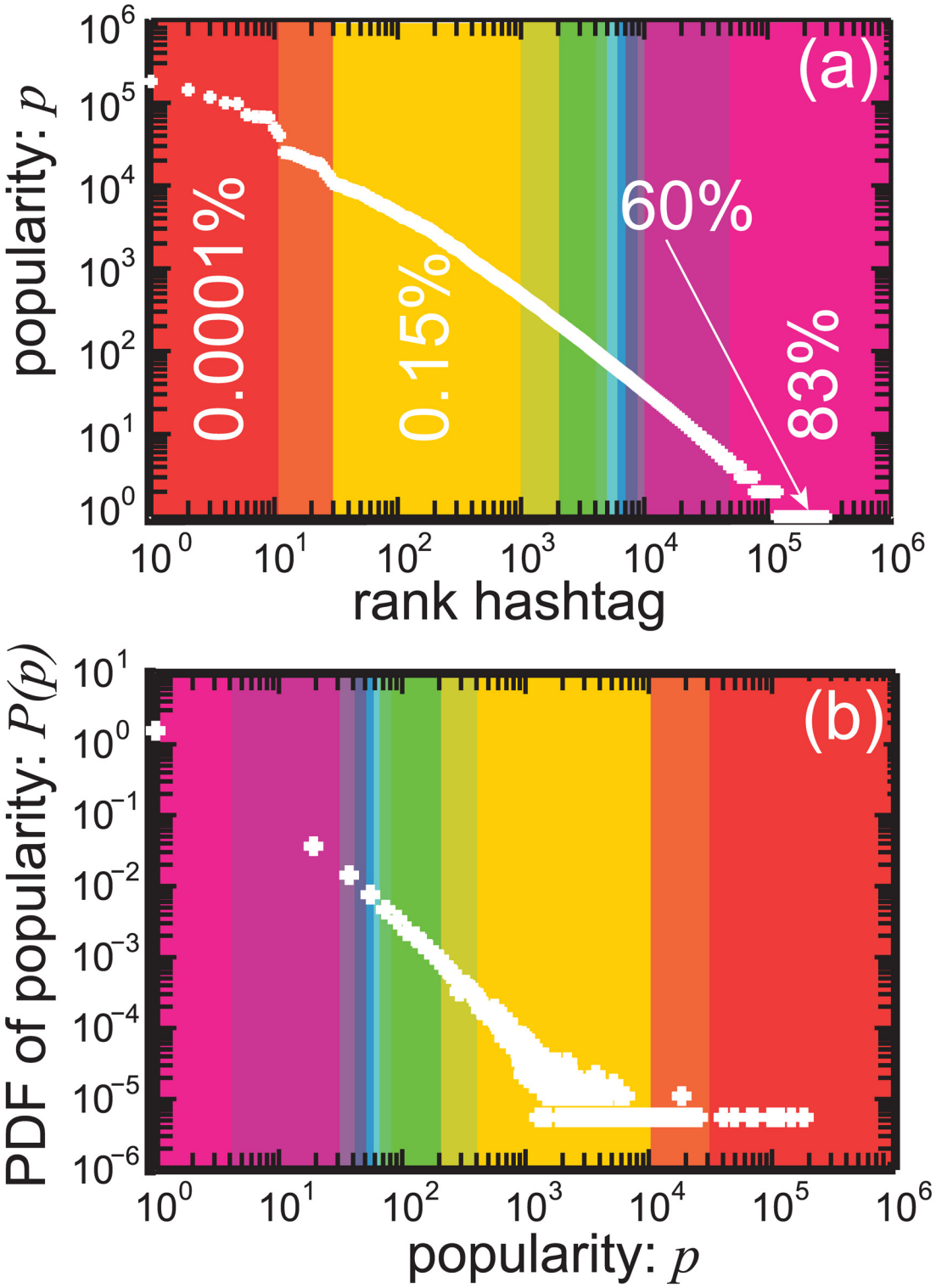

Fig 2. Heterogeneity in the hashtag popularity *p* is shown in (a) Zipf-plot and (b) probability density function (PDF), *P*(*p*). (a) Diversity in *p* (frequency) is visible in a power-law scaling in the log-log plot. We rank hashtags from high *p* (left) to low *p* (right). Different colored shaded rectangles highlight the value of *p* from red and orange (high *p*) to purple and pink (low *p*). The percentages describe the overall contributions of the corresponding rectangles. (b) Similarly, *P*(*p*) obeys a slowly decaying function and presents a power-law distribution with a fat tail. The same colored schema in (a) is applied to visualize the contributions of different values of *p*.

doi:10.1371/journal.pone.0131704.g002

neurons behave as time-dependent Poisson random spike generators, where successive inter-spike intervals are independently chosen from an exponential distribution with a time-dependent firing rate [30]. However, more recent observations have shown that the inter-spike interval distribution exhibits significant deviations from the exponential distribution, which has led to the construction of appropriate tools to describe neuron signals [23–27].

Similarly, a hashtag spike train is defined as the sequence of timings at which the concerned hashtag is observed in Twitter. In this framework, we do not specify the type of dynamics of hashtags, endogeneous or exogeneous [6], i.e. endogeneous, hashtag diffusion among members of the social network, or exogeneous, the diffusion driven by external factors such as TV and newspapers, but only in the timings. Each hashtag thus generates a unique hashtag spike train with a characteristic popularity p. As a first basic indicator, in Fig 3(a) and 3(b) we show the inter-hashtag spike interval cumulative and probability distributions, $CDF(\Delta\tau)$ and $P(\Delta\tau)$, respectively. To avoid deforming the distributions artificially because of the heterogeneity in p, we classify $CDF(\Delta\tau)$ and $P(\Delta\tau)$ in classes depending on p, illustrated by different colors in Fig 2. We observe similar behavior across the classes, as $P(\Delta\tau)$ deviates strongly from an exponential distribution (Poisson), $P(\Delta\tau) = \xi e^{-\xi\Delta\tau}$, where ξ is a firing rate (frequency and so p in our concept) at which hashtags appear. Instead, we observe fat-tailed distributions [3, 7, 12, 16, 31–33] as shown in Fig 3(b) for high and moderate p. As mentioned in Introduction, this deviation may either originate from temporal correlations or non-stationary patterns, making the system different from a stationary and an uncorrelated random signal [34–37]. Recently and unlikely, a stochastic model considering Poisson processes also suggests a broad distribution of the dynamics of brand names in Twitter [15].

Real and randomized data sets

We will analyze two sets of data, which we now describe: The empirical data set, directly coming from the data, and a randomized data set, serving as a null model in our analysis.

The *real data set* contains one spike train per hashtag, as illustrated in Fig 4(a). The time resolution of the spikes is the same as that of the data set, that is 1 second. In situations when multiple spikes of the same hashtag take place at the same time only one event is considered. The statistics of such events are provided at the end of this subsection. In each spike train, the appearance time of the spikes is ordered from the earliest time to the latest time.

The *random data set* is randomized version of the real data set, where each spike train of size p generates a spike train of the same size with random times. In practice, we first combine all hashtag spike trains and obtain one merged hashtag spike train as illustrated in Fig 4(b). This train carries the full history of all hashtags and, importantly, reproduces the nonstationary features of the original data in the presence of temporal correlations, burstiness, and the cyclic rhythm. As before, if two or more spikes generated in the same time, only one spike is shown in that time in the merged spike train, e.g. see the black spikes in Fig 4(b).

Randomization is performed by permuting elements, as shown in Fig 4(c), for instance by using randperm(T, p) in Matlab. Here, T represents the full matrix of times in the merged spike train and p is the desired popularity, number of total spikes in a train. The permutation procedure generates p times uniformly distributed unique numbers out of T and these numbers define the artificial spike train, e.g. $\ldots, \tau^r_{i-1}, \tau^r_i, \tau^r_{i+1}, \ldots$, as shown in Fig 4(c). In our data set, $p \ll T$ is always verified, as the maximum p is 180,900 and the length of T is 667,996. This procedure is applied to each spike train of size p [Fig 4(d)]. Generating independent, yet time-dependent events, the procedure is expected to create time-dependent Poisson random processes, $P(\Delta\tau, t) = \xi(t)e^{-\xi(t)\Delta\tau}$, where the firing rate $\xi(t)$ in this case explicitly depends on the time of the day and of the week.

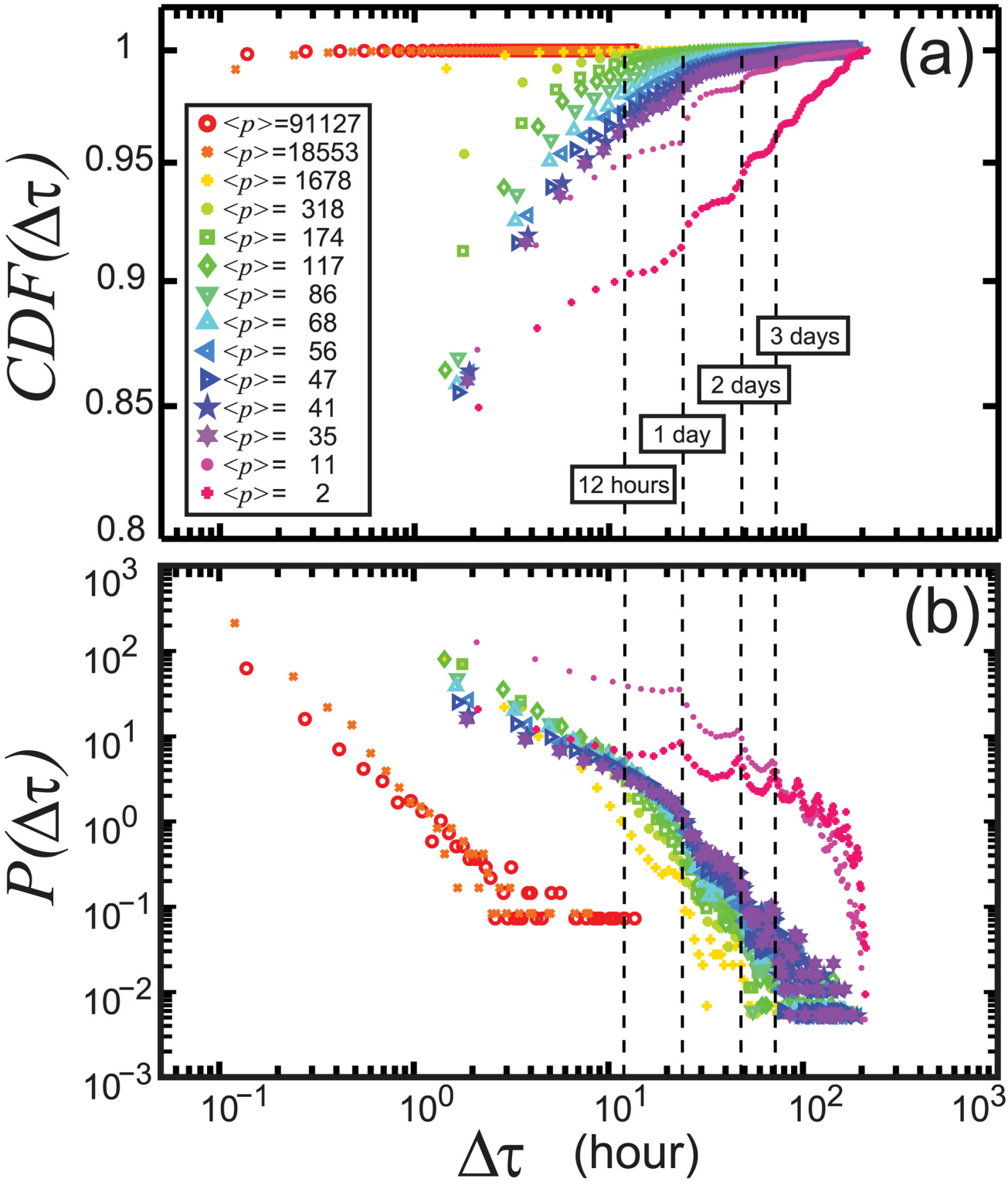

Fig 3. The cumulative (a), *CDF*(Δτ), and probability (b), *P*(Δτ), distributions of the inter-hashtag spike intervals. We observe that $P(\Delta\tau)$, for different classes of hashtags distinguished by their popularity, exhibits non-exponential features. The different colors correspond to those in Fig 2. The legend provides the average popularity $\langle p \rangle$ in each hashtag class. The dash lines indicate the positions of 1 day, 2 days, and 3 days, where $P(\Delta\tau)$ gives peaks for low p (pink symbols). The binning is varied from 8 minutes to 2 hours depending on p, e.g. 8 min. for high p (red-orange), 1.5 hour for moderate p (yellow-green-blue-purple), and 2 hours for low p (pink). All $P(\Delta\tau)$ present maxima at 1 second, which is not shown to describe tails in a larger window.

doi:10.1371/journal.pone.0131704.g003

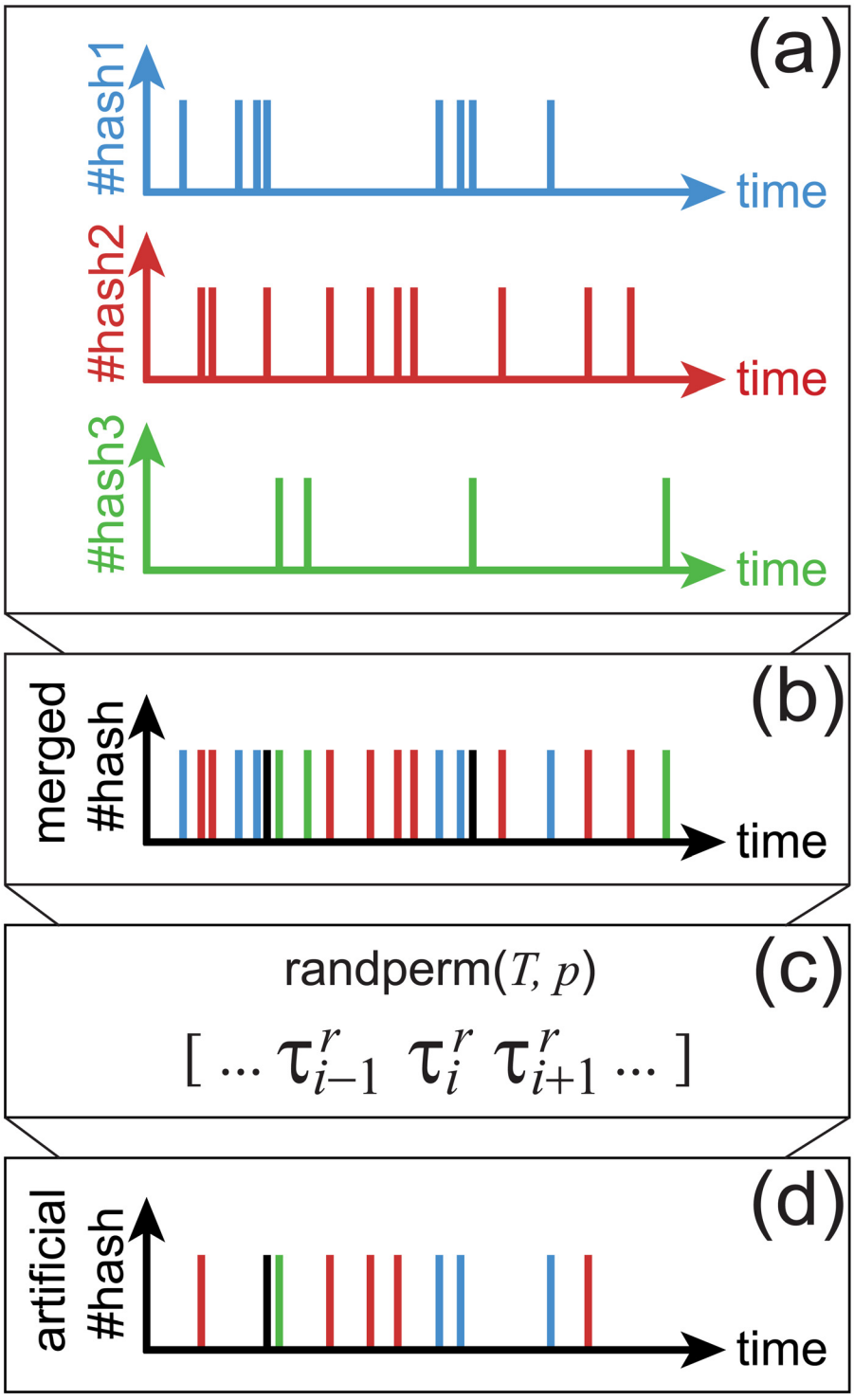

Fig 4. Real and artificial hashtag spike trains. (a) As an illustration of different hashtag spike trains representing different types of hashtag propagation of the data set. (b) Merging hashtag spike trains from the real data. The black spikes describe that only one activity is counted if multiple activities occur at the same time. (c) Randomization procedure by randperm (Matlab). *T* contains full hashtag activity of the data set. The randperm gives a matrix with *p* elements, *p* unique independent numbers out of *T*, and constructing random time series . . ., τ^r_{i-1}, τ^r_i, τ^r_{i+1}, . . . from full hashtag activity matrix *T*. (d) The resultant artificial hashtag spike train.

doi:10.1371/journal.pone.0131704.g004

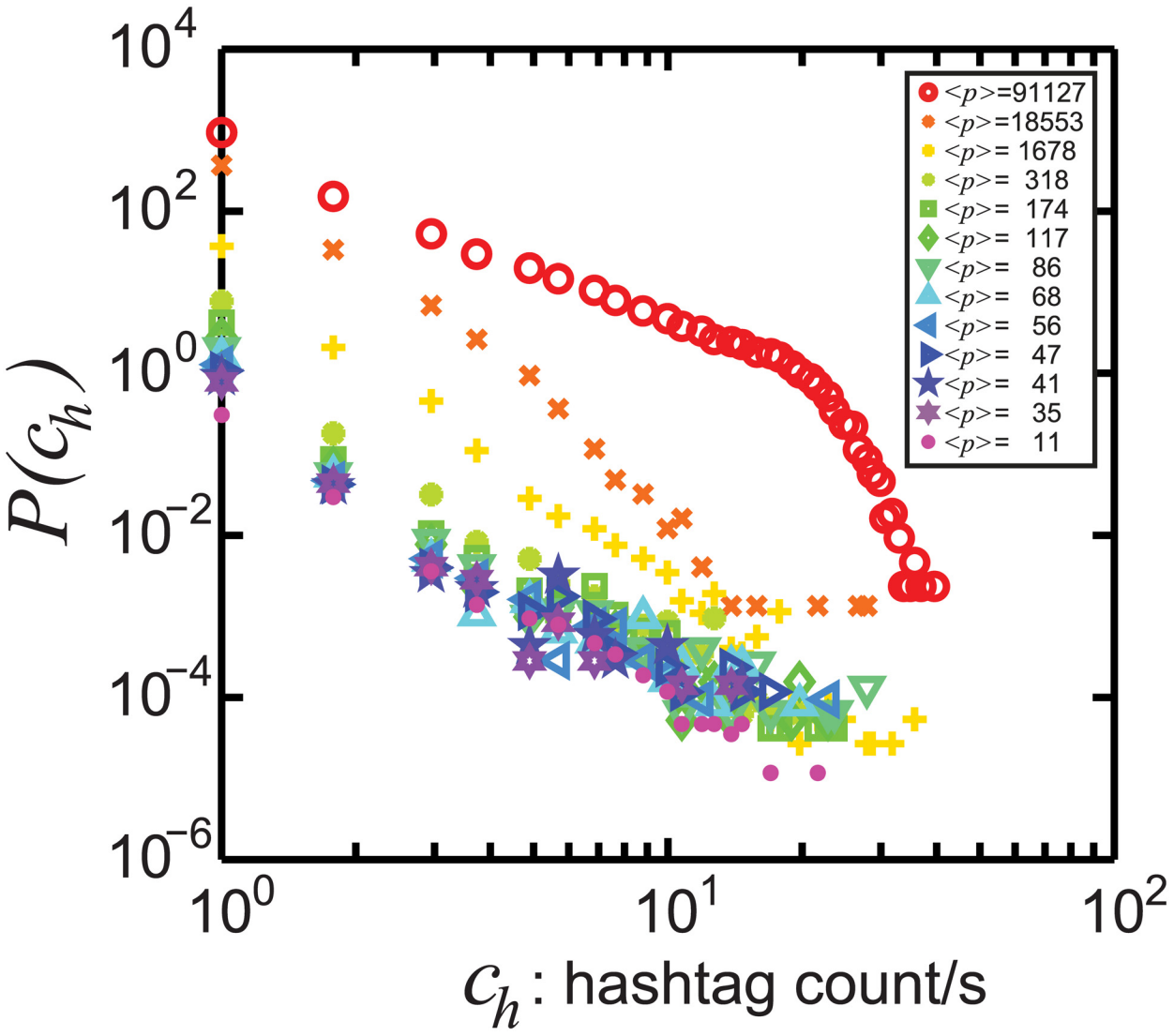

Fig 5. The probability distribution of count of hashtag activity per second $P(c_h)$. We show that, except for the top most popular hashtags listed in Table 1 with ranking 1–11 and presented here in red symbols, multiple activity in 1 second is very rare. The different colors correspond to those in Figs 2 and 3. The legend provides the average popularity $\langle p \rangle$ in each hashtag class.

doi:10.1371/journal.pone.0131704.g005

Statistics of multiple tweets in 1 second. We detect multiple occurrences in 1 second for 6661 hashtags. Fig 5 presents the probability distribution $P(c_h)$ of observing c_h occurrences of a hashtag during one second for different hashtag popularity class. Even though $c_h > 1$ occurs rarely, we observe that this possibility is more probable for popular hashtags (red open circles), as expected. For the most popular hashtag, ledebat, one finds $max(c_h) = 40$.

Local variation

The time series of spike trains are inherently nonstationary, as shown in Fig 1. For this reason, metrics defined for stationary processes are inadequate and might lead to incorrect conclusions. For instance, the non-exponential shapes of the inter-event time distribution $P(\Delta\tau)$ in

Fig 3 might originate from either correlated (and maybe even collective) dynamics or nonstationarity of the hashtag propagation. Similarly, statistical indicators based on this distribution, such as its variance or Fano factor, might be affected in a similar way. For this reason, we consider here the so-called local variation L_V, originally defined to determine intrinsic temporal dynamics of neuron spike trains [23–27].

Unlike quantities such as $P(\Delta\tau)$, L_V compares temporal variations with their local rates and is specifically defined for nonstationary processes [27]

$$L_V = \frac{3}{N-2}\sum_{i=2}^{N-1}\left(\frac{(\tau_{i+1}-\tau_i)-(\tau_i-\tau_{i-1})}{(\tau_{i+1}-\tau_i)+(\tau_i-\tau_{i-1})}\right)^2 \quad (1)$$

Here, N is the total number of spikes and . . ., τ_{i-1}, τ_i, τ_{i+1}, . . . represents successive time sequence of a single hashtag spike train. Eq 1 also takes the form [27]

$$L_V = \frac{3}{N-2}\sum_{i=2}^{N-1}\left(\frac{\Delta\tau_{i+1}-\Delta\tau_i}{\Delta\tau_{i+1}+\Delta\tau_i}\right)^2 \quad (2)$$

where $\Delta\tau_{i+1} = \tau_{i+1}-\tau_i$ and $\Delta\tau_i = \tau_i-\tau_{i-1}$. $\Delta\tau_{i+1}$ quantifies the forward delay and $\Delta\tau_i$ represents the backward waiting time for an event at τ_i. Importantly, the denominator normalizes the quantity such as to account for local variations of the rate at which events take place. By definition, L_V takes values in the interval [0:3].

The local variation L_V presents properties making it an interesting candidate for the analysis of hashtag spike trains [23–27]. In particular, L_V is on average equal to 1 when the random process is either a stationary or a non-stationary Poisson process [23], with the only condition that the time scale over which the inverse firing rate $1/\xi(t)$ fluctuates is slower than the typical time between spikes. Deviations from 1 originate from local correlations in the underlying signal, either under the form of pairwise correlations between successive inter-event time intervals, e.g. $\Delta\tau_{i+1}$ and $\Delta\tau_i$ which tend to decrease L_V, or because the inter-event time distribution is non-exponential. An interesting case is given by Gamma processes [23, 25]

$$P(\Delta\tau, t; \xi, \kappa) = (\xi\kappa)^{\kappa}\Delta\tau^{(\kappa-1)}e^{-\xi\kappa\Delta\tau}/\Gamma(\kappa) \quad (3)$$

where κ is called a shape parameter and determines the shape of the distribution, ξ is a firing rate (frequency) as previously defined, and Γ is the Gamma function. Here, ξ and κ are the two parameters of the Gamma process and both can be time-dependent. While ξ determines the speed of the dynamics, κ controls for the burstiness (irregularity) of the spike trains. Assuming that events are independently drawn, the shape factor is related to L_V as follows [23, 25]

$$\langle L_V\rangle = \frac{3}{2\kappa+1} \quad (4)$$

Here, the brackets describe the average taken over the given distribution [23]. When $\kappa = 1$, an exponential is recovered, and one finds $\langle L_V\rangle = 1$ as expected. Smaller values of κ increase the variance in $\Delta\tau$ and therefore its burstiness, making L_V larger than 1. On the other hand, larger values of κ decrease the variance of $\Delta\tau$ and the burstiness of the process, making $\langle L_V\rangle \approx 0$ smaller than 1.

We measure L_V of hashtag spike trains and group the values depending on the popularity p of their hashtags as was done in Figs 2 and 3. Fig 6 shows scatter plots of L_V for the real data set (a), the empirical sequence . . ., τ_{i-1}, τ_i, τ_{i+1}, . . ., and the random data set (b), the random sequence . . ., τ^r_{i-1}, τ^r_i, τ^r_{i+1}, . . ., on linear-log plots. Different colors are used to distinguish the different groups and the inset legend provides the average popularity $\langle p\rangle$ in the groups.

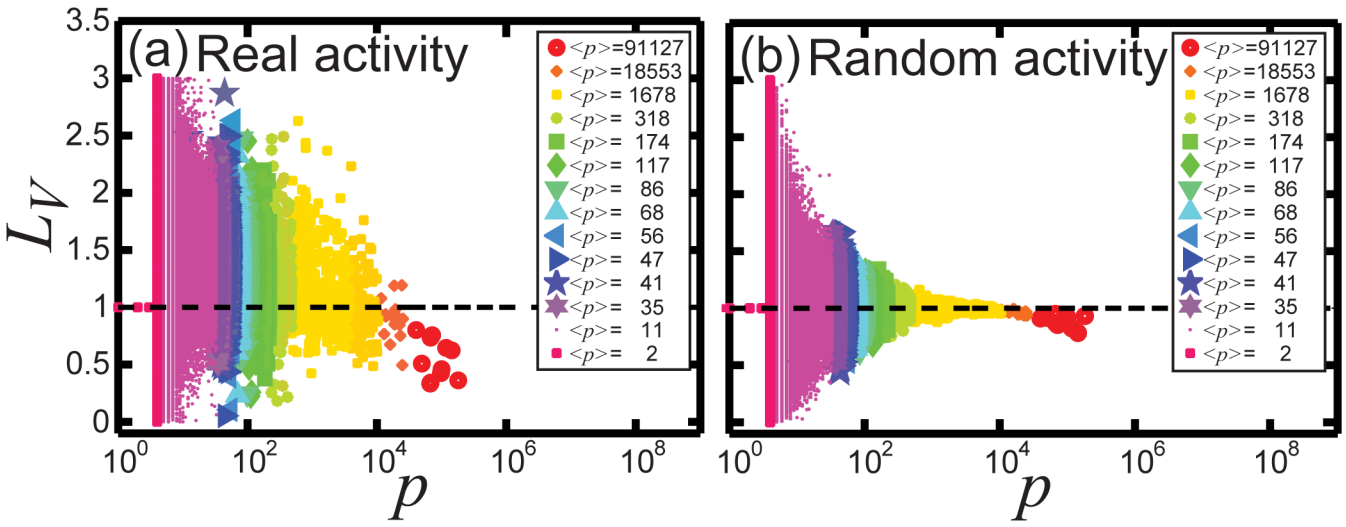

Fig 6. The local variation L_V of hashtag spike trains versus popularity p on a linear-log plot. Each color and symbol summarized in the legend present different range of p: Low p, pink and purple colors, and moderate p, blue, green, and yellow colors, and then high p, orange and red colors. In addition, the average p, $\langle p \rangle$, indicated in the legend ranks colors and symbols quantitatively. (a) Hashtag spike trains of the data set. (b) Artificial (randomized) hashtag spike trains.

doi:10.1371/journal.pone.0131704.g006

A more readable representation is provided in Fig 7, where we show histograms $P(L_V)$ of the values of L_V, for the two data sets and for the distinguished hashtag groups in p. The results clearly show that L_V fluctuates around 1 in the random data set [Fig 7(b)], as expected for a time-dependent Poisson process. On the other hand, L_V systematically deviates from 1 in the original data set [Fig 7(a)], where temporal correlations and bursts are expected to be present.

These observations are confirmed in Fig 8(a), where we plot the mean $\mu(L_V)$ of L_V, with error bars, as a function of $\langle p \rangle$. L_V of the original data (blue circles) indicates that high impact hashtags (high p) are associated with lower values of L_V suggesting more homogeneous and regular time distributions. The results encourage the potential use of L_V as a metric not only to capture deviations from Poisson temporarily uncorrelated processes (red squares), but also to identify distinct statistical properties generated specifically in high p. Moreover, Fig 8(b) presents the statistical differences between the real and the random spike trains in detail. The deviations from Poisson processes where $\mu_0(L_V) = 1$ are calculated by $z = \mu(L_V) - \mu_0(L_V)/\sigma(L_V)/\sqrt{n}$ with the standard deviations of L_V, $\sigma(L_V)$, and the number of the data points given in the distributions in Fig 7, n. We observe that z–values for the random spikes (red squares) are almost equal to 0, excluding in high p, indicating the agreement between Poisson signals and our random spike trains, which is not the case for the real trains (blue circles) giving $z \ncong 0$ in any of $\langle p \rangle$.

To conclude, we perform an analysis to test the persistence of the temporal characteristics of the hashtags, as measured by L_V, through time. To do so, we divide each hashtag time series into two equal time series. The resulting values of local variations are $L_V(t_1)$ for the first half of a spike train and $L_V(t_2)$ for the second half of the train, and then we calculate the Pearson correlation coefficient $r(L_V(t_1), L_V(t_2))$ between these values [38]. In Fig 9(a), we show the linear relations between $L_V(t_1)$ and $L_V(t_2)$ for different p classes and Fig 9(b) presents $r(L_V(t_1), L_V(t_2))$ as a function of the average popularity $\langle p \rangle$ on a linear-log plot. Both indicate that the values of L_V for the same hashtags at different times are significantly and temporarily correlated.

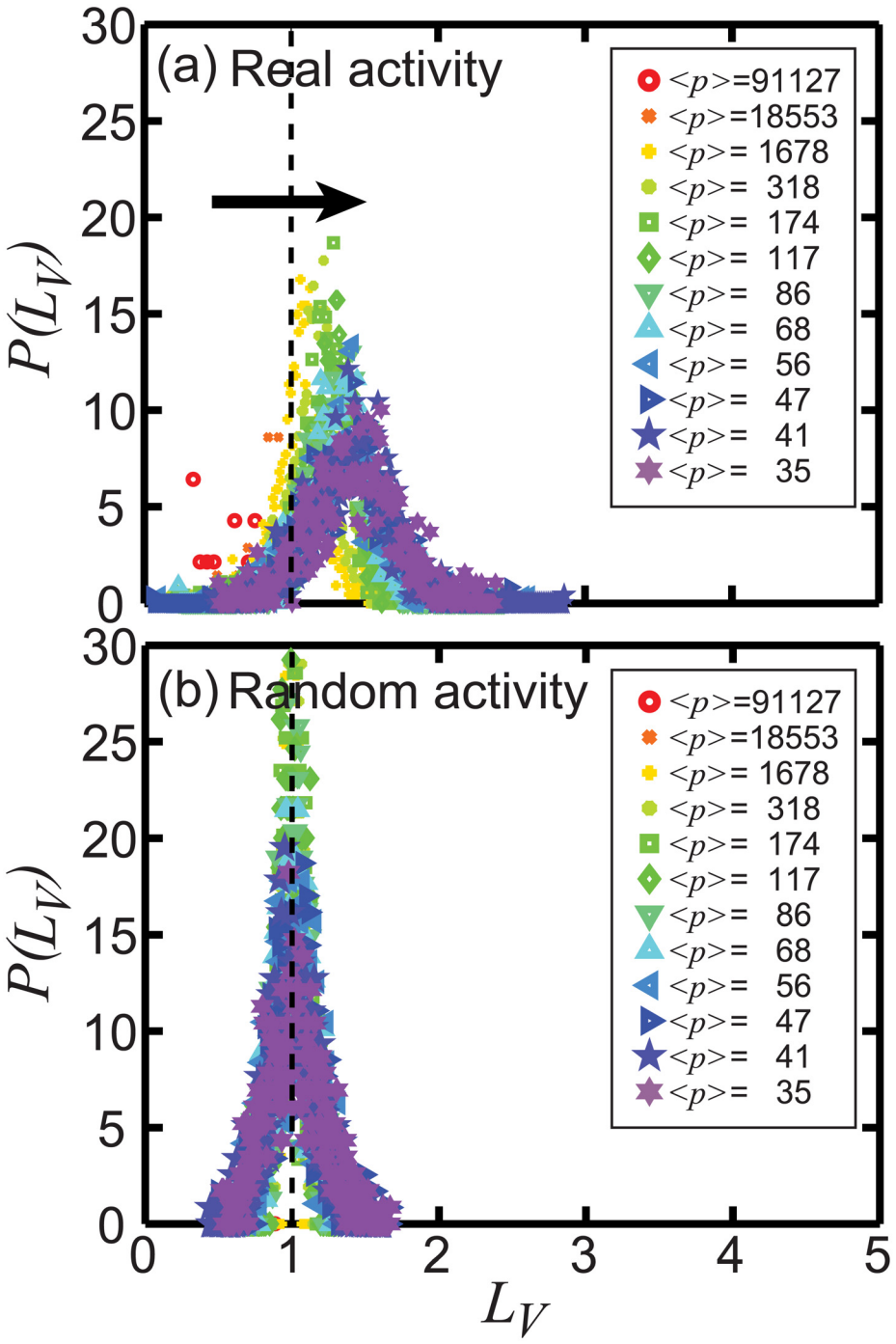

Fig 7. Probability density function (PDF) of the local variation L_V of real hashtag propagation (a) and random hashtag time sequences (b). Two distinct shapes are visible: (a) From high p to low p, the peak position of $P(L_V)$ shifts from low values of L_V to higher values of L_V. (b) $P(L_V)$ always peaks around 1 for the random sequences generated by artificial hashtag spike trains. The same color coding is applied as already used in Fig 6.

doi:10.1371/journal.pone.0131704.g007

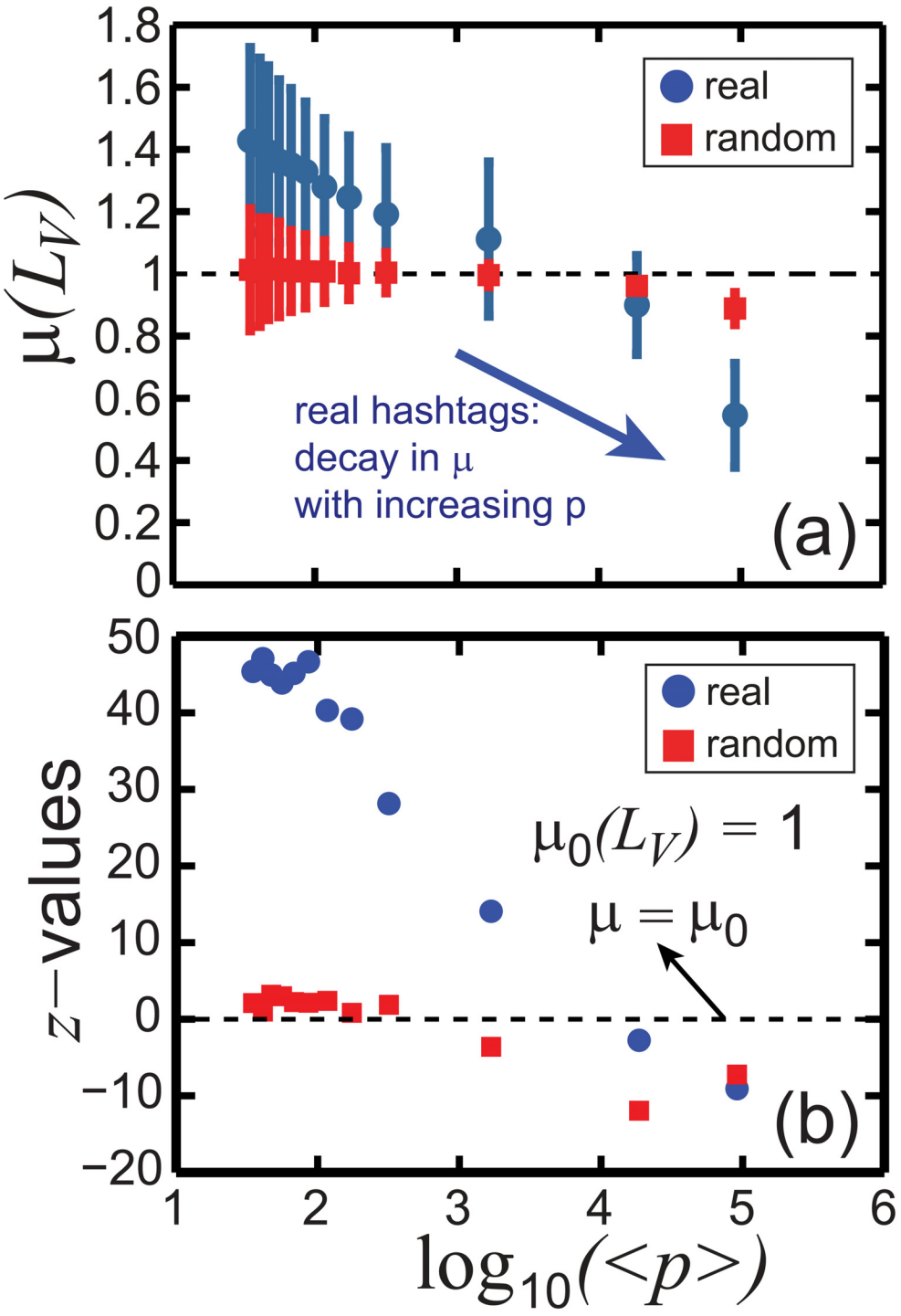

Fig 8. Statistical inference of L_V and comparison between the real and the random hashtag spike trains. (a) Mean μ of the local variation L_V of single hashtag time series versus the logarithmic average popularity $\log_{10}\langle p\rangle$. The real hashtag propagation is described in blue circles, whereas red squares represent randomly selected hashtag activity from the real data set. The arrow indicates the decay of $\mu(L_V)$ when $\langle p\rangle$ increases, which shows that popular hashtags propagate regularly on the contrary to moderately popular hashtags presenting bursty time sequences. The bars indicate the corresponding standard deviations $\sigma(L_V)$. (b) A standard z–values versus $\log_{10}\langle p\rangle$. While the random trains (red squares) with $z \approx 0$ show the evidence of Poisson signals with mean $\mu_0(L_V) = 1$, large and non-zero values of z for the real trains (blue circles) suggest the presence of temporal correlations.

doi:10.1371/journal.pone.0131704.g008

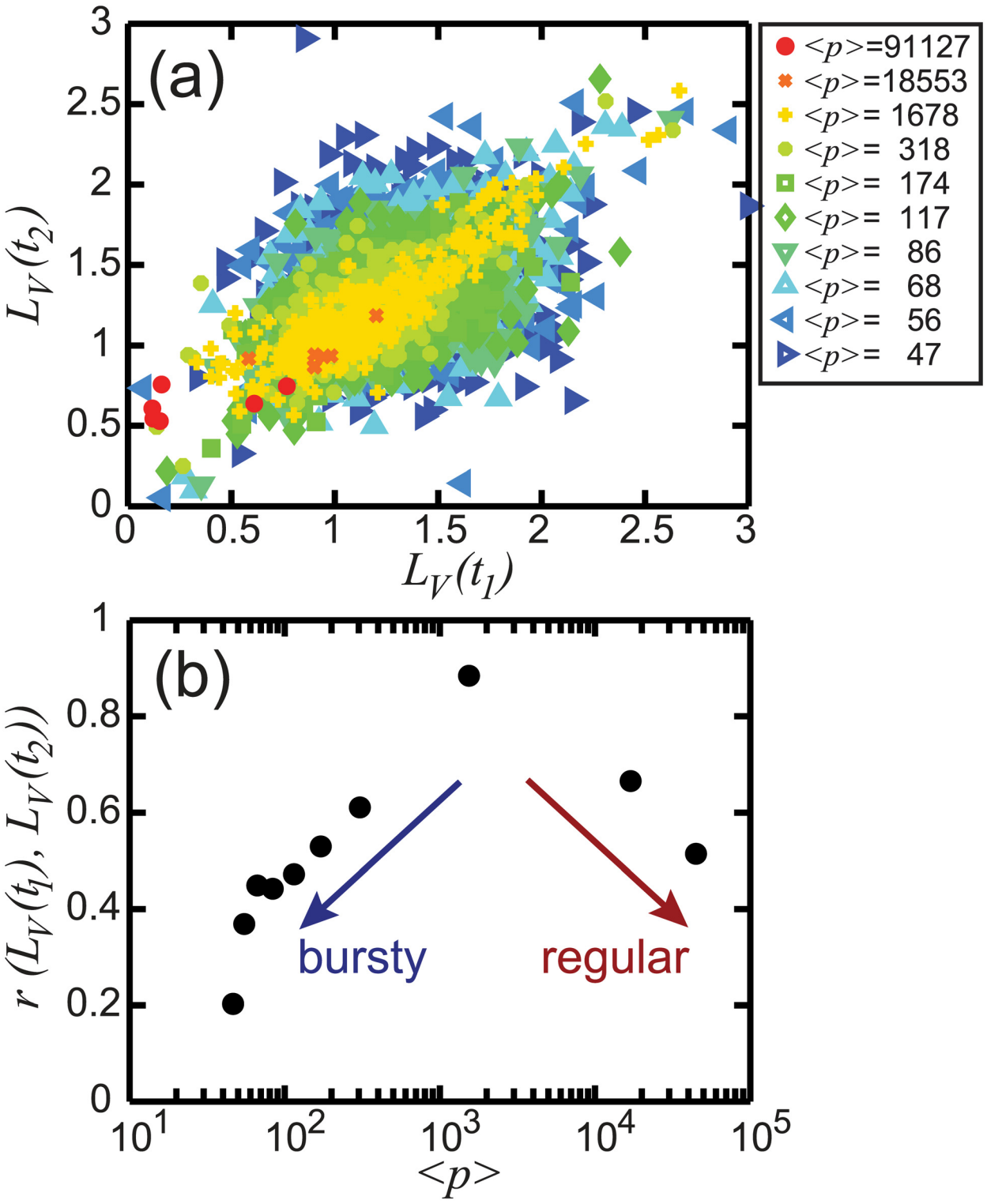

Fig 9. Linear correlation of L_V through real hashtag spike trains. (a) The linear relations of the first and the second halves of the empirical spike trains, $L_V(t_1)$ and $L_V(t_2)$, respectively, are investigated. The legend ranks $\langle p \rangle$ in different colors and symbols. (b) The Pearson correlation coefficient $r(L_V(t_1), L_V(t_2))$ between these quantities shows that while the linear correlations through moderately popular spike trains give maximum values, r reaches the minimum values for both bursty (high L_V and low p) and regular (low L_V and high p) spike trains.

doi:10.1371/journal.pone.0131704.g009

Interestingly, we observe that while bursty (low p) and regular (high p) signals give small r, the spike trains with moderate p provide the largest values of r, indicating more uniform temporal behavior through the individual trains in moderate p.

Discussion

The main purpose of this paper is to introduce a statistical measure suitable for the analysis of non-stationary time series, as they often take place in online social media and communications in social systems. As a test case, we have focused on the dynamics of hashtags in Twitter. However, the same methodology could be also applied to the other types of correlated, bursty, and non-stationary signals, for instance the dynamics of cascades in Twitter and Facebook or phone call activity.

Instead of measuring standard statistical properties of noisy hashtag signals such as the inter-event time distribution, the variance or the Fano factor, conventionally applied to characterize non-stationarity of a signal, we have focused on the local variation L_V, a metric capturing the fluctuations of a signal as compared to a local characteristic time. This measure, previously defined for neuron spike train analysis, nicely uncovers the regularity and the firing rate of the trains [23–27] and so helps to identify local temporal correlations. It is important to stress that the current analysis exclusively focuses on properties of time series and considers neither the mechanisms leading to the observed statistical dynamic properties nor the effects of the underlying topology, e.g. through following-follower relations. Interesting lines of research would study the relation between L_V and the underlying topology [39] and would consider diffusive models, for instance the Hawkes process [40, 41]. In addition, both neurons [30] and hashtags can be driven by multiple firing rates and L_V analysis associated to Gamma distributions would provide more concrete results on hashtag spike trains, as done for neuron spikes [25].

We should also note that the finite temporal resolution of the data (1 sec), which induces the fact that multiple events per time window are neglected, makes L_V artificially small for popular hashtags. In an extreme case, the time series is indeed regular, with events taking place every second. In this work, we have therefore carefully verified that the fluctuations in L_V are not artificially driven by these limitations. To this end, we have compared the values of L_V in the empirical data with those of a null model. We observe a small decay of L_V for popular hashtags in the null model (see Fig 8), but this decay is much more limited than the one observed in the empirical data, e.g. $L_V = 0.89$ for $\langle p \rangle \approx 10^5$ in the null model while it is equal to $L_V = 0.54$ for the real data. In addition, a decay of L_V in the real hashtag data is also present in moderately popular hashtags, where multiple events per second are very rare. An interesting research direction would be to generalize the definition of local variation to allow for the analysis of multiple events per time window, thereby evaluating the dense time series more precisely. Finally, in a finite time window, as observed in the empirical data, the statistics of high frequency hashtags is much better than that of low frequency hashtags, simply because the former occurs many more times than the latter. For this reason, the measurements of L_V for less popular hashtags are more subject to noise.

The empirical analysis also reveals an interesting pattern observed in the data, as more popular hashtags tend to present more regular temporal behavior. This lack of burstiness ensures that popular hashtags do not disappear from the social network for very long periods of time, consequently allowing for a regular activation of the interest of Twitter users. These findings are reminiscent of a recent observation in numerical simulations showing that burstiness hinders the size of cascades [42], and should be incorporated into the modeling of theoretical information diffusion models, in particular threshold [43] and stochastic [44] models, on temporal networks.

Supporting Information

S1 File.
(ZIP)

Acknowledgments

We thank Takaaki Aoki and Taro Takaguchi for their useful comments and Lionel Tabourier for providing preliminary data set. This work was supported by grant number: F.N.R.S MIS F4527.12 48888F3 (Grant holder: RL, Funding receiver: CS—http://www.fnrs.be/), the EU 7th Framework OptimizR Project: 48909A2 CE OPTIMIZR (Grant holder: RL, Funding receiver: CS—http://optimizr.eu/), and the National Institute of Informatics Tokyo (http://www.nii.ac.jp/en/) for partial traveling support. This funder had a role in preparation of the manuscript, but did not have a further role in study design, data collection and analysis or decision to publish.

Author Contributions

Conceived and designed the experiments: CS RL. Performed the experiments: CS. Analyzed the data: CS. Contributed reagents/materials/analysis tools: RL. Wrote the paper: CS RL.

References

1. Borge-Holthoefer J, Rivero A, García I, Cauhé E, Ferrer A, Ferrer D, et al. Structural and Dynamical Patterns on Online Social Networks: The Spanish May 15th Movement as a Case Study. PLoS ONE. 2011 08; 6(8):e23883. doi: 10.1371/journal.pone.0023883 PMID: 21886834
2. González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y. The Dynamics of Protest Recruitment through an Online Network. Sci Rep. 2011 12; 1:197. doi: 10.1038/srep00197 PMID: 22355712
3. Domenico MD, Lima A, Mougel P, Musolesi M. The Anatomy of a Scientific Rumor. Sci Rep. 2013 10; 3:2980. Available from: http://dx.doi.org/10.1038/srep02980.
4. Sasahara K, Hirata Y, Toyoda M, Kitsuregawa M, Aihara K. Quantifying Collective Attention from Tweet Stream. PLoS ONE. 2013 04; 8(4):e61823. doi: 10.1371/journal.pone.0061823 PMID: 23637913
5. Kenett DY, Morstatter F, Stanley HE, Liu H. Discovering Social Events through Online Attention. PLoS ONE. 2014 07; 9(7):e102001. doi: 10.1371/journal.pone.0102001 PMID: 25076410
6. Deschâtres F, Sornette D. Dynamics of book sales: Endogenous versus exogenous shocks in complex networks. Phys Rev E. 2005 Jul; 72:016112. doi: 10.1103/PhysRevE.72.016112
7. Barabási AL. The origin of bursts and heavy tails in human dynamics. Nature. 2005 05; 435:207–211. doi: 10.1038/nature03459 PMID: 15889093
8. Coscia M. Competition and Success in the Meme Pool: A Case Study on Quickmeme.com; 2013. Available from: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/5990.
9. Myers SA, Leskovec J. Clash of the Contagions: Cooperation and Competition in Information Diffusion. In: Data Mining (ICDM), 2012 IEEE 12th International Conference on; 2012. p. 539–548.
10. Malmgren RD, Stouffer DB, Motter AE, Amaral LAN. A Poissonian explanation for heavy tails in e-mail communication. Proceedings of the National Academy of Sciences. 2008; 105(47):18153–18158. Available from: http://www.pnas.org/content/105/47/18153.abstract. doi: 10.1073/pnas.0800332105
11. Lambiotte R, Ausloos M, Thelwall M. Word statistics in Blogs and {RSS} feeds: Towards empirical universal evidence. Journal of Informetrics. 2007; 1(4):277–286 Available from: http://www.sciencedirect.com/science/article/pii/S1751157707000582. doi: 10.1016/j.joi.2007.07.001
12. Jo HH, Karsai M, Kertész J, Kaski K. Circadian pattern and burstiness in mobile phone communication. New Journal of Physics. 2012; 14(1):013055. Available from: http://stacks.iop.org/1367-2630/14/i=1/a=013055. doi: 10.1088/1367-2630/14/1/013055
13. Myers SA, Leskovec J. The Bursty Dynamics of the Twitter Information Network. In: Proceedings of the 23rd International Conference on World Wide Web. WWW '14. New York, NY, USA: ACM; 2014. p. 913–924. Available from: http://doi.acm.org/10.1145/2566486.2568043.
14. França U, Sayama H, McSwiggen C, Daneshvar R, Bar-Yam Y. Visualizing the "Heartbeat" of a City with Tweets. ArXiv e-prints. 2014 Nov;.

15. Mollgaard A, Mathiesen J. Emergent user behavior on Twitter modelled by a stochastic differential equation. ArXiv e-prints. 2015 Feb;.
16. Ratkiewicz J, Fortunato S, Flammini A, Menczer F, Vespignani A. Characterizing and Modeling the Dynamics of Online Popularity. Phys Rev Lett. 2010 Oct; 105:158701. Available from: doi: 10.1103/PhysRevLett.105.158701 PMID: 21230945
17. Weng L, Menczer F, Ahn YY. Virality Prediction and Community Structure in Social Networks. Sci Rep. 2013 08; 3:2522. doi: 10.1038/srep02522 PMID: 23982106
18. Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J. Can Cascades Be Predicted? In: Proceedings of the 23rd International Conference on World Wide Web. WWW'14. New York, NY, USA: ACM; 2014. p. 925–936. Available from: http://doi.acm.org/10.1145/2566486.2567997.
19. Weng L, Flammini A, Vespignani A, Menczer F. Competition among memes in a world with limited attention. Sci Rep. 2012 03; 2:335. doi: 10.1038/srep00335 PMID: 22461971
20. Gleeson JP, Ward JA, O'Sullivan KP, Lee WT. Competition-Induced Criticality in a Model of Meme Popularity. Phys Rev Lett. 2014 Jan; 112:048701. doi: 10.1103/PhysRevLett.112.048701 PMID: 24580496
21. Cetin U, Bingol HO. Attention competition with advertisement. Phys Rev E. 2014 Sep; 90:032801. doi: 10.1103/PhysRevE.90.032801
22. Gleeson JP, O'Sullivan KP, Baños RA, Moreno Y. Determinants of Meme Popularity. ArXiv e-prints. 2015 Jan;.
23. Shinomoto S, Shima K, Tanji J. Differences in Spiking Patterns Among Cortical Neurons. Neural Comput. 2003 12; 15:2823–2842. doi: 10.1162/089976603322518759 PMID: 14629869
24. Koyama S, Shinomoto S. Empirical Bayes interpretations of random point events. Journal of Physics A: Mathematical and General. 2005 07; 38(29):L531–L537. Available from: http://stacks.iop.org/0305-4470/38/i=29/a=L04. doi: 10.1088/0305-4470/38/29/L04
25. Miura K, Okada M, Amari S. Estimating Spiking Irregularities Under Changing Environments. Neural Comput. 2006 10; 18:2359–2386. doi: 10.1162/neco.2006.18.10.2359 PMID: 16907630
26. Shimazaki H, Shinomoto S. A Method for Selecting the Bin Size of a Time Histogram. Neural Comput. 2007 04; 19:1503–1527. doi: 10.1162/neco.2007.19.6.1503 PMID: 17444758
27. Omi T, Shinomoto S. Optimizing Time Histograms for Non-Poissonian Spike Trains. Neural Comput. 2011 12; 23:3125–3144. doi: 10.1162/NECO_a_00213 PMID: 21919781
28. Coscia M. Average is Boring: How Similarity Kills a Meme's Success. Sci Rep. 2014 09; 4:6477. doi: 10.1038/srep06477 PMID: 25257730
29. Tuckwell HC. Introduction to Theoretical Neurobiology. vol. 2. Cambridge University Press; 1988.
30. Softky WR, Koch C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random. J Neurosci. 1993 1; 13:334–350. Available from: http://www.jneurosci.org/content/13/1/334.abstract. PMID: 8423479
31. Takaguchi T, Masuda N. Voter model with non-Poissonian interevent intervals. Phys Rev E. 2011 Sep; 84:036115. doi: 10.1103/PhysRevE.84.036115
32. Vestergaard CL, Génois M, Barrat A. How memory generates heterogeneous dynamics in temporal networks. Phys Rev E. 2014 Oct; 90:042805. doi: 10.1103/PhysRevE.90.042805
33. Miotto JM, Altmann EG. Predictability of Extreme Events in Social Media. PLoS ONE. 2014 11; 9(11): e111506. doi: 10.1371/journal.pone.0111506 PMID: 25369138
34. Karsai M, Kaski K, Barabási AL, Kertész J. Universal features of correlated bursty behaviour. Sci Rep. 2012 05; 2:397. doi: 10.1038/srep00397 PMID: 22563526
35. Szabolcs V, Tóth B, Kertész J. Modelling bursty time series. New J Phys. 2013 10; 15:103023. doi: 10.1088/1367-2630/15/10/103023
36. Lambiotte R, Tabourier L, Delvenne JC. Burstiness and spreading on temporal networks. The European Physical Journal B. 2013; 86(7). doi: 10.1140/epjb/e2013-40456-9
37. Jo HH, Perotti JI, Kaski K, Kertesz J. Correlated bursts and the role of memory range. ArXiv e-prints. 2015 May;.
38. Shinomoto S, Kim H, Shimokawa T, Matsuno N, Funahashi S, Shima K, et al. Relating Neuronal Firing Patterns to Functional Differentiation of Cerebral Cortex. PLoS Comput Biol. 2009 07; 5(7):e1000433. doi: 10.1371/journal.pcbi.1000433 PMID: 19593378
39. Rodriguez MG BDSB Leskovec J. Uncovering the structure and temporal dynamics of information propagation. Network Science. 2014 4; 2:26–65. Available from: http://journals.cambridge.org/article_S2050124214000034. doi: 10.1017/nws.2014.3

40. Onaga T, Shinomoto S. Bursting transition in a linear self-exciting point process. Phys Rev E. 2014 Apr; 89:042817. doi: 10.1103/PhysRevE.89.042817

41. Jovanović S, Hertz J, Rotter S. Cumulants of Hawkes point processes. Phys Rev E. 2015 Apr; 91:042802. doi: 10.1103/PhysRevE.91.042802

42. Backlund VP, Saramäki J, Pan RK. Effects of temporal correlations on cascades: Threshold models on temporal networks. Phys Rev E. 2014 Jun; 89:062815. doi: 10.1103/PhysRevE.89.062815

43. Karimi F, Holme P. Threshold model of cascades in empirical temporal networks. Physica A: Statistical Mechanics and its Applications. 2013; 392(16):3476–3483. Available from: http://www.sciencedirect.com/science/article/pii/S0378437113002835. doi: 10.1016/j.physa.2013.03.050

44. Kawamoto T. A stochastic model of tweet diffusion on the Twitter network. Physica A: Statistical Mechanics and its Applications. 2013; 392(16):3470–3475. Available from: http://www.sciencedirect.com/science/article/pii/S0378437113002811. doi: 10.1016/j.physa.2013.03.048