Please cite this paper as: Geras A., Siudem G., Gagolewski M., Time to vote: Temporal clustering of user activity on Stack Overflow, Journal of the Association for Information Science and Technology 73(12), 1681–1691, 2022, DOI:10.1002/asi.24658

Time to vote: Temporal clustering of user activity on Stack Overflow

Agnieszka Geras[1], Grzegorz Siudem[2,*], Marek Gagolewski[3,4,1]

[1] Warsaw University of Technology, Faculty of Mathematics and Information Science
ul. Koszykowa 75, 00-662 Warsaw, Poland

[2] Warsaw University of Technology, Faculty of Physics,
ul. Koszykowa 75, 00-662 Warsaw, Poland

[3] Deakin University, School of IT, Geelong, VIC 3220, Australia

[4] Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland

[*] Corresponding author; email: grzegorz.siudem@pw.edu.pl

Abstract

Question-and-answer (Q&A) sites improve access to information and ease transfer of knowledge. In recent years, they have grown in popularity and importance, enabling research on behavioural patterns of their users. We study the dynamics related to the casting of 7M votes across a sample of 700k posts on Stack Overflow, a large community of professional software developers. We employ log-Gaussian mixture modelling and Markov chains to formulate a simple yet elegant description of the considered phenomena. We indicate that the inter-event times can naturally be clustered into 3 typical time scales: those which occur within hours, weeks, and months and show how the events become rarer and rarer as time passes. It turns out that the posts' popularity in a short period after publication is a weak predictor of its overall success, contrary to what was observed, e.g., in case of YouTube clips. Nonetheless, the sleeping beauties sometimes awake and can receive bursts of votes following each other relatively quickly.

Keywords: clustering, inter-event times, log-normal mixtures, Q&A networks, Stack Overflow, burstiness

Time to vote: Temporal clustering of user activity on Stack Overflow

Introduction

Stack Exchange (https://stackexchange.com) is very popular platform hosting many question-and-answer (Q&A) sites, where people come together to learn and share their knowledge. The spectrum of topics covered is very wide, including philosophy, computer science, astronomy, home improvement, personal finance, pets, and photography. Stack Exchange users play two different roles: they can ask new questions or answer questions already posed. Every question and answer published therein can be voted for (Up-Vote) or against (Down-Vote) by other registered users. Based on those positive and negative votes, one can assess their quality.

Such scores enable the convenient identification of the most common problems (relative to the site's topic) and the most effective solutions thereto. Users looking for a solution to an issue they find puzzling want to find a similar question (identifying similarity is often non-trivial, see (Mondal & Roy, 2022)) asked by someone else in the past and use the best answers given to it.

Usually, the top-rated content (displayed at the top of the results list) is checked first, therefore there is a strong preferential (rich and richer) component governing the popularity growth. Furthermore, some items can be interacted with in a random (accidental) order, for instance when they appear in the results of external search engines. Also, the results can be sorted with respect to the creation date, therefore newer items might be slightly more popular. These mechanisms governing the accumulation of votes' distribution is similar to those displayed in many other domains, e.g., when describing citations to scientific papers; compare (Janosov et al., 2020; Liu et al., 2021; Siudem et al., 2020).

Different aspects of the Stack Exchange Network community's contribution and activity have been of interest to a significant group of researchers. For a comprehensive list of publications based on the Stack Exchange data, please refer to (Vasilescu et al., 2020).

The majority of studies take a data mining or machine learning approach to

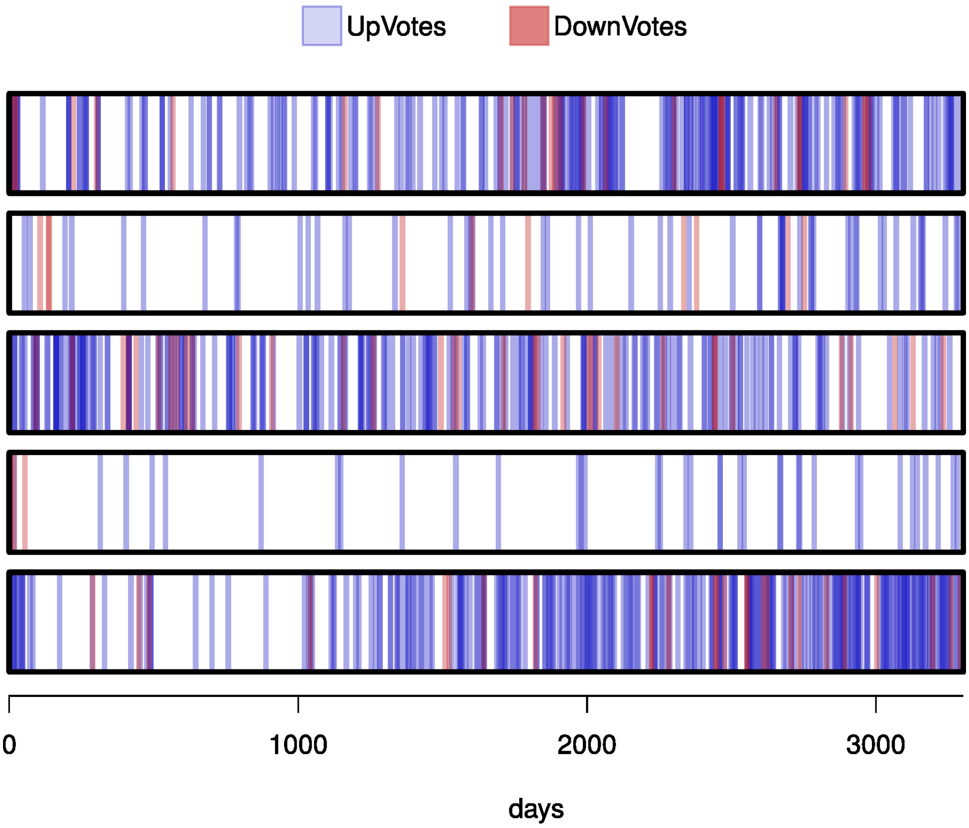

Figure 1. 5 randomly selected posts on Stack Overflow and the dynamics of the evolution of their popularity. Each vertical line represents one vote. We note that there is no particular structure, however, the system alternates between periods of low and high frequency events: they frequently occur in bursts, compare also (Karsai et al., 2018).

answer a variety of questions about their users and their behaviour (e.g., regarding the behaviour of novice software engineers (Chatterjee et al., 2020), gender differences (May et al., 2019), improving answer rates by gamification (Papoutsoglou et al., 2020; Zhou et al., 2020), how knowledge is generated and shared (Tausczik & Huang, 2020)). Some papers are interested in the modelling of the structure of the Stack Exchange network (e.g., (Moutidis & Williams, 2021)) or mechanisms governing its dynamics (e.g., describing the effects of negative votes (Geras et al., 2020) or predicting the emergence of highly cited (Zhao et al., 2021) or obsolete (Zhang et al., 2021) posts).

Our attention is drawn to the largest and the most popular community, Stack Overflow (https://stackoverflow.com), which as of 2022-01-19 features ca. 22M questions, 32M answers, and has 16M registered users. It is used on a daily basis by software professionals, researchers, and students, for instance to get answers to questions such as "how do I do X in language Y?" or "why the result of F is Z?".

In this paper we are interested in modelling the network's evolution over time and revealing the temporal landscape of the dynamics governing the posts' popularity. Even though, at first glance, the vote accumulation process seems to be quite unstructured (see Figure 1), we shall uncover some hidden patterns in the studied dataset. Ultimately, our goal is to tackle the question "what is the time until the next vote?" and "how does it depend on the post's popularity so far?".

While investigating time-varying phenomena and attempting to capture regularities therein, the inter-event time distribution (the distribution of time duration between two consecutive events) is an important trace of latent mechanisms generating the considered processes (Karsai et al., 2018). The accumulating of attention by items which are elements of complex evolving systems (e.g., papers receiving citations in scientific communities (Egghe, 2009; Nadarajah & Kotz, 2008) or posts being retweeted on microblogging platforms (Gao et al., 2015)) is a matter often investigated in the literature. Most authors name such processes bursty (starting from (Goh & Barabási, 2008), where the so-called burstiness parameter is introduced), compare monographs (Jo & Hiraoka, 2019) and (Karsai et al., 2018). They occur when short periods of very intensive activity is followed by long periods of inactivity (see Figure 1 again). Such phenomena are widely observed in various communication processes (Karsai et al., 2018) in particular in exchanging e-mails (Barabasi, 2005), phone calls (W. Wang et al., 2015), conflicts between Wikipedia editors (Yasseri et al., 2012), GitHub users activity (Yan et al., 2017), Youtube and Ding content popularity growth (Szabo & Huberman, 2010), paper updating intervals on arXiv (Jo et al., 2012), paper processing times in academic journals (Hartonen, 2013), to name a few. Bursty processes are not limited only to human behaviour; they were observed in brain activity (Thompson et al., 2017) communication events on supercomputer systems (Chen et al., 2019), BGP routing anomalies (Moriano et al., 2021), and in the description of the temporal networks

(Cencetti et al., 2021; Zou et al., 2021), especially in the context of epidemic contagions and diffusion of information (Unicomb et al., 2021).

Inter-event time distributions of various activities are usually reported to follow mostly exponential or heavy-tailed (e.g., power-law) distributions. To the best of our knowledge there are no results concerning inter-event time distributions of Stack Exchange users' activity. As it will turn out below, we have indicated that such processes follow mixtures of log-normal distributions. It seems as a novel observation, however there were some findings of log-normal distributions in waiting times related to the Physica A journal contents (Mryglod et al., 2012) or mixtures of exponential distributions in an Office data set (Okada et al., 2020). However, studies rarely take into account whether it is the 1st, 2nd, or n-th event in a series. Initial popularity was considered in (Szabo & Huberman, 2010) and different types of nonhomogeneous stochastic processes were employed in (Rizoiu et al., 2018). However, most relevant to our study is the reinforcement dynamics with memory model (Karsai et al., 2012). The main difference between their and our approach is that our model is data-driven and not postulated a priori: we shall observe two or three (log-normal) temporal clusters in the data.

The remainder of the paper is organised as follows. In the next section we describe the data set used in our research and note that the distribution of inter-event times is naturally multimodal. Then we analyse the burstiness and memory coefficients of the times between the casting of consecutive votes. Following this is a section devoted to the modelling of the inter-event time distributions and their evolution over time.

Data and Their Structure

We have obtained the Stack Overflow data dump dated 2021-03-01 from https://archive.org/details/stackexchange. From the `Posts.xml` file (uncompressed size of 82 GiB), we have gathered data on all 920,231 answers (`PostTypeId=2`, to which we will refer in the sequel as "posts") created in 2009 (the portal premiered in mid-2008, hence 2009 was the first complete calendar year).

Unfortunately, the data dumps do not feature the precise time when the user votes were cast: for privacy reasons, only calendar date (and not date-time) is included therein. Courtesy of the Stack Overflow Academic Partnership Program, we have obtained an unabridged version of the `Votes.xml` dataset (a 15 GiB file), from which we have extracted 6,982,687 Up-Votes (`VoteTypeId=2`) and 147,007 Down-Votes (`VoteTypeId=3`). The total number of answers that received at least 1 vote (of any kind: Up or Down) was 695,831. Note that, nevertheless, our results can be reproduced to some extent using the publicly available dataset.

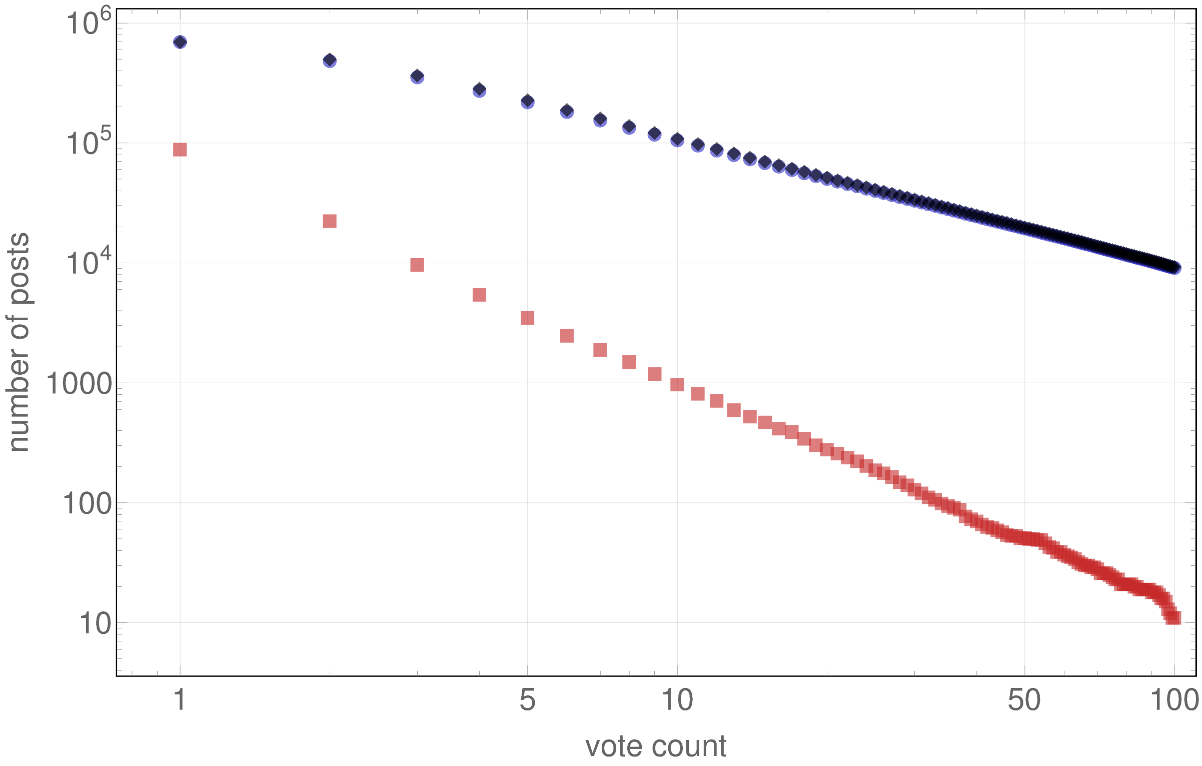

Figure 2. Number of posts with a given number of votes (for answers with no more than 100 votes); Down-Votes in red, Up-Votes in blue, sum thereof in grey; note the logarithmic log scale on both axes and that the grey plotting characters are virtually indistinguishable from the blue ones.

Let us take a closer look at the introduced data, especially in the context of vote accumulation dynamics. Figure 2 depicts the number of posts with a given number of Down- and Up-votes. The latter are significantly more abundant (147,007 vs 6,982,687,

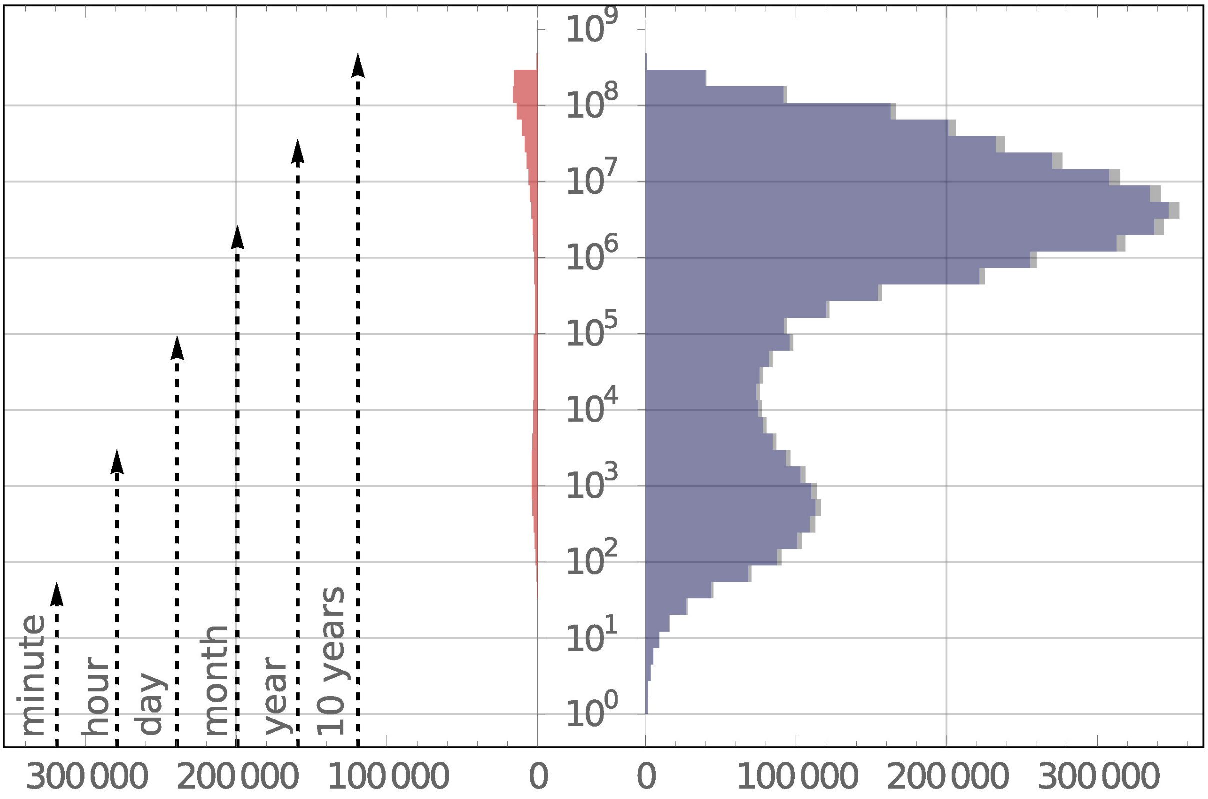

Figure 3. Times between consecutive votes, in seconds (up to 100 first votes for each post, altogether). The red bars represent the Down-Votes, blue bars give the Up-Votes, and grey ones present total vote counts. The Up-Votes and total vote count histograms are bimodal, indicating that fitting a mixture of 2 or more bell-shaped distributions (on the log scale) might be worth considering.

respectively). As usual with such kinds of processes, e.g., (Clauset et al., 2009), the counts follow a power-law distribution: a straight line can naturally be fitted on the log-log scale. To back up the hypothesis that data follow the Zipf distribution (which is a discrete distribution governed by the power law), we have performed the chi-squared goodness-of-fit test. Indeed, for the Up-Votes, the p-value was equal to 0.239 ($\chi^2 = 9900$) when testing that data are Zipfian with exponent $s = 0.9528$ (the maximum likelihood estimator). For the Down-votes, we have the p-value of 0.2667 ($\chi^2 = 7600$) and $s = 1.9439$.

For the sake of robustness, unless explicitly stated otherwise, we will restrict our analysis only to the first 100 votes for each post, as smaller samples are naturally more prone to variability. Still, there are only 10 posts which have received 100 Down-votes, hence in the further analysis we will restrict ourselves to Up-Votes only. Notably, recent findings (Geras et al., 2020) suggest that Down-votes do not change the dynamics of post score significantly anyway.

Furthermore, Figure 3 gives the distribution of times between the consecutive votes. The shape of the histogram for the Up-Votes is bimodal, with the modes corresponding to every $\sim 1/4$ hour and every ~ 2 months. Hence, below we shall consider modelling it with a mixture of at least two or more bell-curved distributions on the log-scale. Note that bibliometric data are often assumed (Shen et al., 2014; D. Wang et al., 2013) to follow Poisson processes, but in our case most runs of the Kolmogorov–Smirnov tests for each post strongly advise against the hypotheses that the times are exponentially distributed.

Actually, ca. 30% of the Up-Votes are received on the first day since a post's creation, and then the cumulative popularity increases rather steadily.

Burstiness and Memory

The inter-event time distribution is not a sufficient descriptor of event burstiness by itself, because there might be significant dependencies between consecutively cast votes. In such a case, we say that the processes are not memoryless.

Let us consider two memory-adjusted tools to quantify the level of burstiness in data sets (or appropriate models) (Karsai et al., 2018): the burstiness parameter B and the memory coefficient M (Goh & Barabási, 2008).

Let us represent an event sequence with n events as an ordered list $\mathbf{T} = (T_1, T_2, \ldots, T_{n-1}, T_n)$ where T_k denotes the timing of the k-th event; in our case, the appearance of an Up-Vote in seconds. Further, we define the inter-event times as $\tau_k = T_k - T_{k-1}$, where by convention T_0 is the time of the publication of a given post.

With $P_k(\tau)$ let us denote the probability that a post which already received k votes gains the next vote in τ seconds. We observe that considering different k values

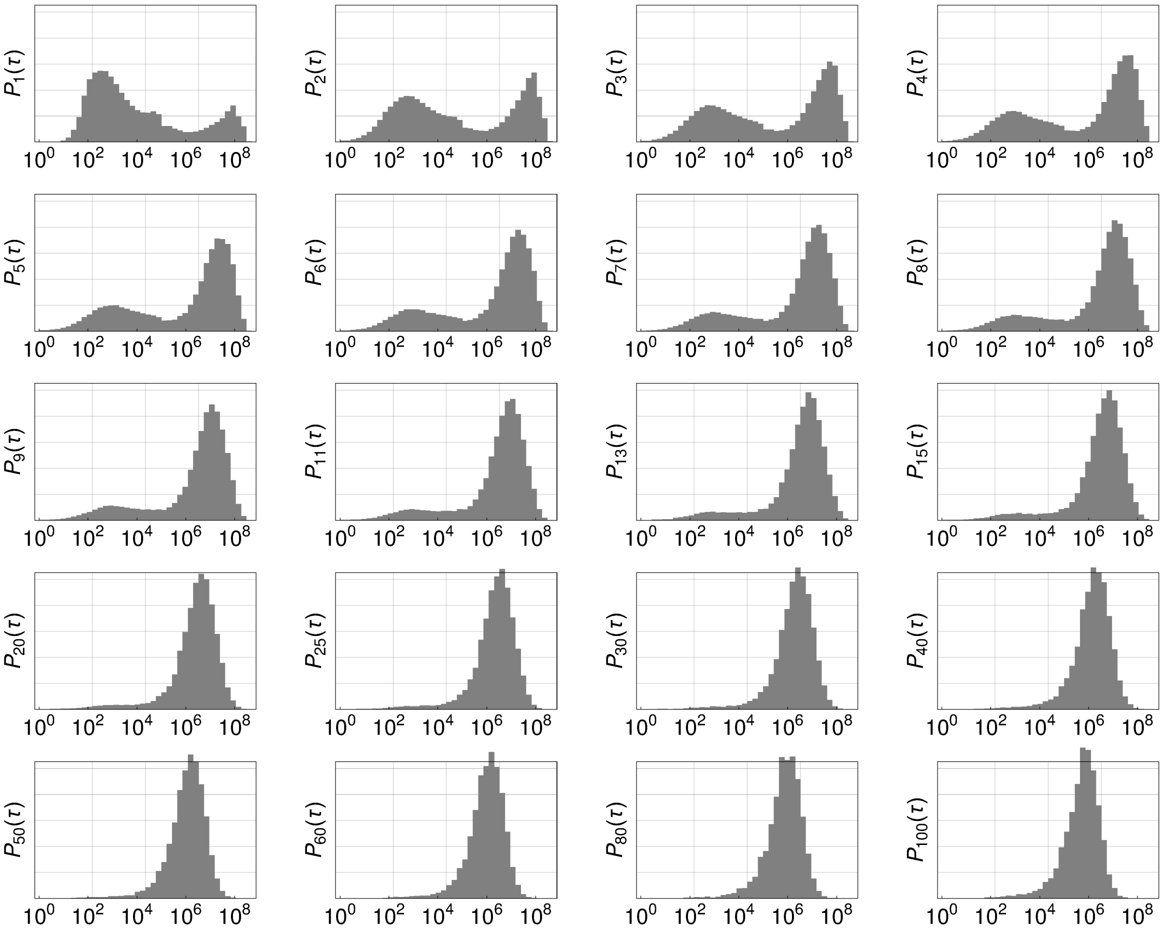

Figure 4. Evolution of the inter-event time distributions, P_k (to be read rowwisely): time to wait for the k-th vote in seconds, on a logarithmic scale. The more votes a post has, the longer, on average, one has to wait for the next event of this kind.

(Figure 4) reveals some non-trivial dynamics that are otherwise unrevealed in the aggregated perspective (Figure 3). For every k, $P_k(\tau)$ resembles a mixture of a few bell-shaped curves (note the log scale on the Ox axis though), whose contributions change in time. For small k, the bulk of probability mass is in the first cluster, representing short-time effects. Later on, it is transferred to the long-term group. We shall engage in a modelling task featuring a mixture of Gaussians in the sequel.

Firstly, let us consider the burstiness parameter B, which for a given vector of $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$ of inter-event times and $k \leq n$ is defined as

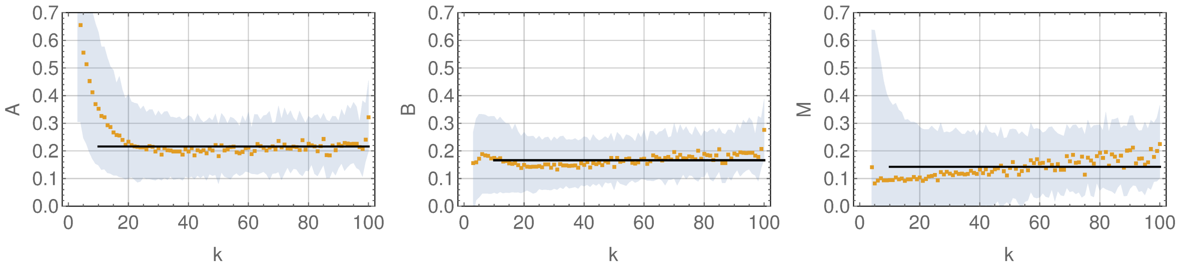

Figure 5. Distributions of the burstiness and memory coefficients as a function of k, i.e., when we consider the times between the first k events. The orange orange squares represent the medians over all posts and the blue bands span between the 1st and 3rd quartiles. The black line gives the global mean computed over all $k \in [10, 100]$.

$$B_\tau(k) = \frac{\mathrm{sd}(\tau_1, \dots, \tau_k) - \mathrm{avg}(\tau_1, \dots, \tau_k)}{\mathrm{sd}(\tau_1, \dots, \tau_k) + \mathrm{avg}(\tau_1, \dots, \tau_k)}, \tag{1}$$

where $\mathrm{avg}(\tau_1, \dots, \tau_k)$ and $\mathrm{sd}(\tau_1, \dots, \tau_k)$ denote the arithmetic mean and the standard deviation, respectively, of the first k inter-event times. This parameter has a very natural interpretation: $B = 1$ indicates the most bursty process, $B = 0$ represents a random (Poissonian) one, and $B = -1$ corresponds to a completely regular (periodic) realisation. However, it was noted in (Kim & Jo, 2016) that the parameter B can be misleading for small sample sizes, hence we will also consider its corrected version:

$$A_\tau(k) = \frac{\sqrt{k+1}r_k - \sqrt{k-1}}{(\sqrt{k+1} - 2)r_k + \sqrt{k-1}}, \qquad r_k = \frac{\mathrm{sd}(\tau_1, \dots, \tau_k)}{\mathrm{avg}(\tau_1, \dots, \tau_k)}. \tag{2}$$

Furthermore, the memory coefficient M was introduced in (Goh & Barabási, 2008). It is given by

$$M_\tau(k) = \frac{1}{k-1} \sum_{i=1}^{k-1} \frac{(\tau_i - \mathrm{avg}(\tau_1, \dots, \tau_{k-1}))(\tau_{i+1} - \mathrm{avg}(\tau_2, \dots, \tau_k))}{\mathrm{sd}(\tau_1, \dots, \tau_{k-1})\,\mathrm{sd}(\tau_2, \dots, \tau_k)}. \tag{3}$$

M varies in the range $[-1, 1]$. When it is positive, a short inter-event time tends to be followed by another short one, and long inter-event times repeat one after another. In contrast, when it has a negative value, long inter-event times are followed by short ones and vice versa.

In Figure 5 we depict the change in the median of the A, B, and M coefficients as a function of k – the cut-off threshold. For each fixed k, the width of the corresponding blue band (a vertical line segment) gives the interquartile range of the observed coefficients (based on $A_{\boldsymbol{\tau}_i}(k), B_{\boldsymbol{\tau}_i}(k), M_{\boldsymbol{\tau}_i}(k)$ corresponding to the inter-even times of each post $\boldsymbol{\tau}_i$).

The computed coefficients are typical of human-generated processes (compare (Goh & Barabási, 2008; Karsai et al., 2018)). All the coefficients stabilise at $k \approx 20$. In particular, many vectors have high M for small k, which indicates that there are either bursts of votes appearing quickly or slowly one after another. For $k \approx 20$ the median M stabilises at ≈ 0.2, which indicates that there is only a weak, yet positive correlation between the interval lengths.

Inter-Event Time Modelling

Fitting Log-Normal Mixtures

From Figure 4 we see that the inter-event distributions might be approximated by a mixture of bell-shaped curves on the log scale. Let us thus perform data modelling via a convex combination of log-normal distributions.

A mixture of m normal distributions $\mathcal{N}(\cdot; \mu_j, \sigma_j)$, $j = 1, \dots, m$, where μ_j is the expected value and σ_j is the standard deviation, is defined by the following density function:

$$P(x) = \sum_{j=1}^{m} \alpha_j \mathcal{N}(x; \mu_j, \sigma_j), \tag{4}$$

where $\alpha_j \in (0, 1)$ is the share (weight) of j-th component in the whole mixture, $\sum_{j=1}^{m} \alpha_j = 1$.

Notably, if τ_k is a random variable representing the time between the $(k-1)$-th and the k-th event whose logarithm follows a mixture of normal distributions, then τ_k follows the same mixture of log-normal distributions with the corresponding parameters.

We apply the standard expectation-maximisation-type algorithm, see, e.g., (Robert & Casella, 2004), to identify the maximum likelihood estimators of the

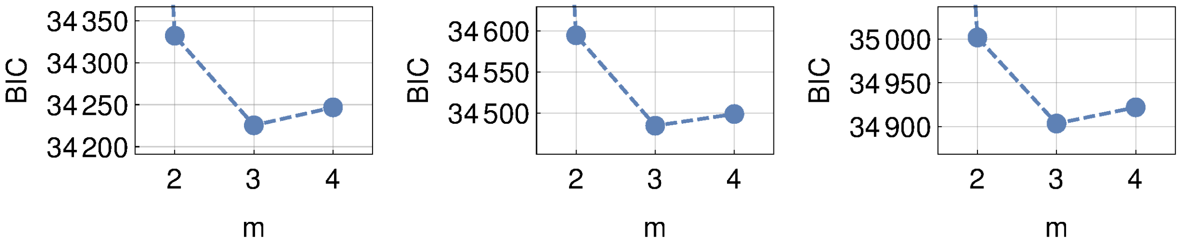

Figure 6. BIC scores for the first, second, and third inter-vote times, respectively, as a function of the number of components in the mixture. Clearly, $m = 3$ gives a good balance between the model complexity and accuracy.

$\alpha_j, \mu_j, \sigma_j$ parameters based on logarithmically transformed data.

The selection of the number of components in the mixture, m, is guided by the Bayesian Information Criterion (BIC) (Schwarz, 1978), which takes into account the trade-off between the number of parameters $(3m - 1)$ and goodness of fit (which naturally gets better as m increases). Figure 6 gives the BIC scores for τ_1, τ_2, and τ_3 (where the distributions' multimodality is most evident). Data suggest that $m = 3$ mixture components should be chosen.

Figure 7 depicts the fitted mixtures for different inter-event times τ_k. For the first-ever votes ($k = 1$, the subplot denoted as $P_1(\tau)$), the first (blue) cluster dominates (with $\alpha_1^{(1)} \approx 0.44$, $\mu_1^{(1)} \approx 5.78$, $\sigma_1^{(1)} \approx 1.32$). Let us note the orders of magnitude of the time scales in each cluster:

$$\begin{aligned}
\exp(\mu_1^{(1)}) &\approx 325\,\mathrm{s} &&\approx 5.4\,\text{minutes},\\
\exp(\mu_2^{(1)}) &\approx 1.8\cdot 10^4\,\mathrm{s} &&\approx 5.0\,\text{hours},\\
\exp(\mu_3^{(1)}) &\approx 4.3\cdot 10^7\,\mathrm{s} &&\approx 1.4\,\text{years}.
\end{aligned}$$

In other words, the first cluster (blue) is comprised of the votes cast just few minutes after a post's publication (note that the author themself can Up-vote their own answer), the second (orange) component corresponds to a few hours, while the third (green) represents long-term effects measured in years.

Further, for $k > 1$, the maximum weight is assigned to the third (green) mixture

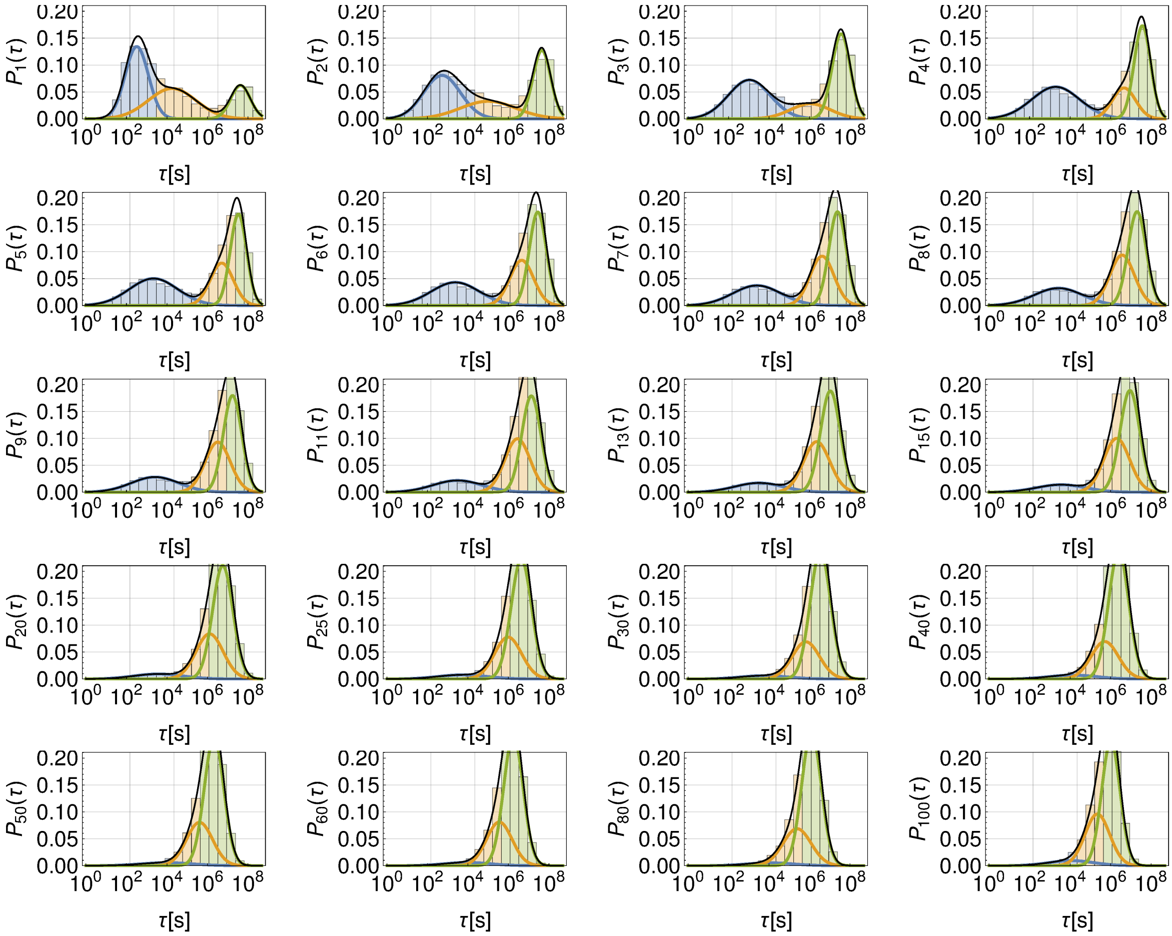

Figure 7. The fitted mixtures of $m = 3$ log-normal distributions. The discovered clusters correspond to short (hours), medium (days), and long (months) wait times. As time passes, the latter effects start to dominate: even for high-valued posts, most votes are accumulated slowly.

component (e.g., for $k = 4$ we get $\alpha_3^{(4)} \approx 0.41$, $\mu_3^{(4)} \approx 17.43$, $\sigma_3^{(4)} \approx 0.94$).

As i increases, the clusters evolve, see Figure 8 for an illustration of the change in the estimated mixture parameters. Once ca. 20 votes are cast, the models stabilise at:

$$\begin{aligned}
\alpha_1^{(\infty)} &\approx 0.04, & \mu_1^{(\infty)} &\approx 9.1, & \sigma_1^{(\infty)} &\approx 2.7, & \exp\left(\mu_1^{(\infty)}\right) &\approx 2.5\,\text{hours},\\
\alpha_2^{(\infty)} &\approx 0.29, & \mu_2^{(\infty)} &\approx 12.9, & \sigma_2^{(\infty)} &\approx 1.5, & \exp\left(\mu_2^{(\infty)}\right) &\approx 4.5\,\text{days},\\
\alpha_3^{(\infty)} &\approx 0.67, & \mu_3^{(\infty)} &\approx 14.5, & \sigma_3^{(\infty)} &\approx 1.1, & \exp\left(\mu_3^{(\infty)}\right) &\approx 22.1\,\text{days}.
\end{aligned} \tag{5}$$

This suggest that in the long run, there are three typical time scales: hours (1, blue), weeks (2, orange), and months (3, green). However, the third (long-term) group is

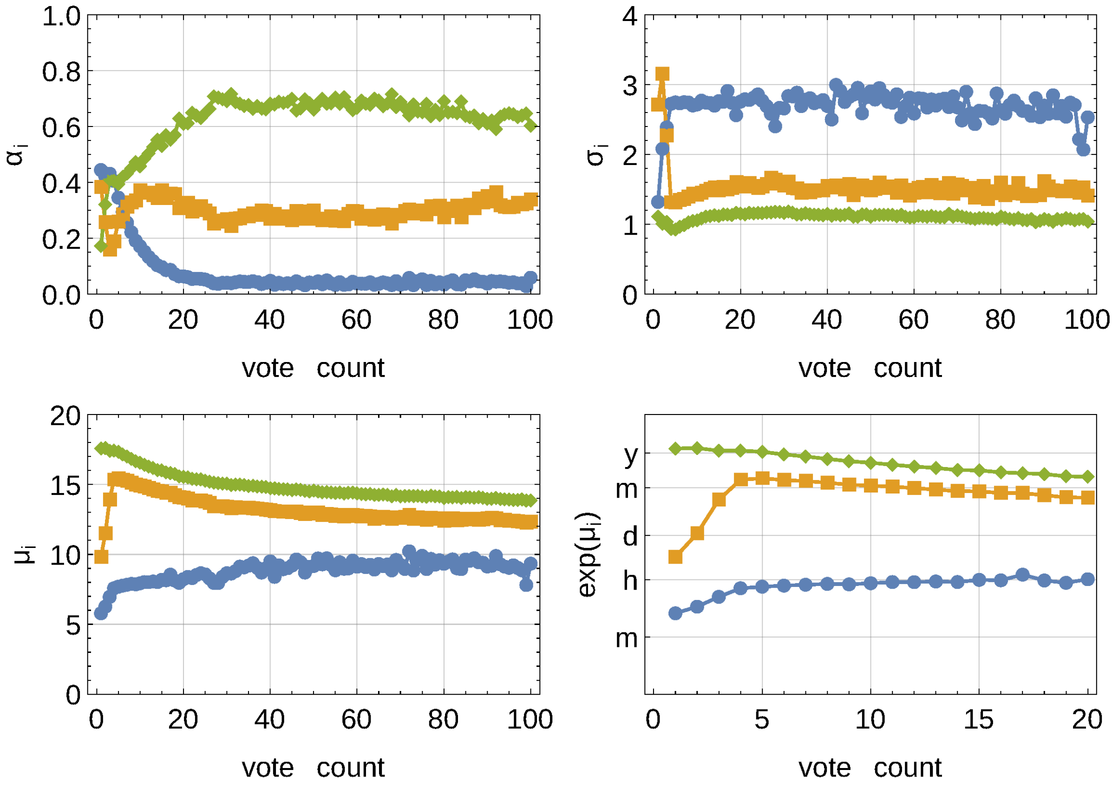

Figure 8. Evolution of the fitted parameters of the mixtures of $m = 3$ log-Gaussians, row-wisely: the component weights/importances $\alpha_i^{(k)}$, standard deviations $\sigma_i^{(k)}$, expected values $\mu_i^{(k)}$ on the linear and logarithmic scale as a function vote number k for each cluster $i = 1, 2, 3$. Colour coding is consistent with that in Figure 7. The process stabilises as k grows. The importance of the short-term effects' cluster decreases over time, but is never negligible.

allocated twice as much weight as the second one and the first component has the smallest share (only few posts are still active, but they are non-negligible; in fact – they are amongst the most interesting ones). Further, let us stress that the observations are in fact log-normally distributed, which implies they are heavy-tailed and thus extreme events (very long waiting times) are likely to occur.

Modelling Vote Dynamics

Above we have identified the temporal clusters for each time step separately. Let us now combine them so that we can arrive at a model for the votes' dynamics in a form of a three-state Markov chain. In such a model, the system is able to switch between clusters as time passes. For a similar dynamical aggregation for diffusion on networks of networks, see (Siudem & Hołyst, 2019); for journal citations, refer to (Delbianco et al., 2020); and for burstiness phenomena (in the case of two states), see (Karsai et al., 2012). In particular, as opposed to the latter, in our case we derive the model directly from data, with minimal a priori assumptions: namely, we only presuppose that there the post memory is of size ≤ 1.

In order to introduce the said model, let us define two values $x_{1.5}$ and $x_{2.5}$ which correspond to the points where the probability density functions of the fitted Gaussians intersect, i.e., points which fulfil the equality

$$\frac{\alpha_i}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(x_{i+0.5}-\mu_i)^2}{2\sigma_i^2}\right]=\frac{\alpha_{i+1}}{\sqrt{2\pi}\sigma_{i+1}}\exp\left[-\frac{(x_{i+0.5}-\mu_{i+1})^2}{2\sigma_{i+1}^2}\right],$$

for $i \in \{1, 2\}$. For example, in the top-left part of Figure 7 we have $x_{1.5} \simeq 2400$ (where the blue and the orange curves meet) and $x_{2.5} \simeq 4.3 \cdot 10^6$ (where the orange and green densities coincide). For brevity of notation, let us also assume $x_{0.5} = -\infty$ and $x_{3.5} = \infty$. This way, we can say than an observable x belongs to the i-th ($i = 1, 2, 3$) cluster, whenever $x \in (x_{i-0.5}, x_{i+0.5})$.

Now we can model the dynamics of the cluster change over time, i.e., modelling the probabilities $p_{j,i}^{(k)}$ of moving to cluster j from being in the i-th cluster at time step k. For instance, $p_{3,2}^{(1)}$ denotes the probability that, after the first ($k = 1$) vote being a medium-term (cluster $i = 2$) one, the wait time for the next vote will be much longer (cluster $j = 3$).

The estimated probabilities for $k = 1, \ldots, 6$ are depicted in Figure 9. Here, the vertex sizes are proportional to the number of corresponding posts. The sizes are of course decreasing, which is in line with our previous observations (see Figure 2).

Also, Figure 10 gives a snapshot of the transition probabilities for further events. This time, the vertex sizes are constants and the arrow sizes are scaled so that the out-flow from each node is preserved, as otherwise the picture would be illegible.

Considering the transition matrices as a function of k we observed that they tend

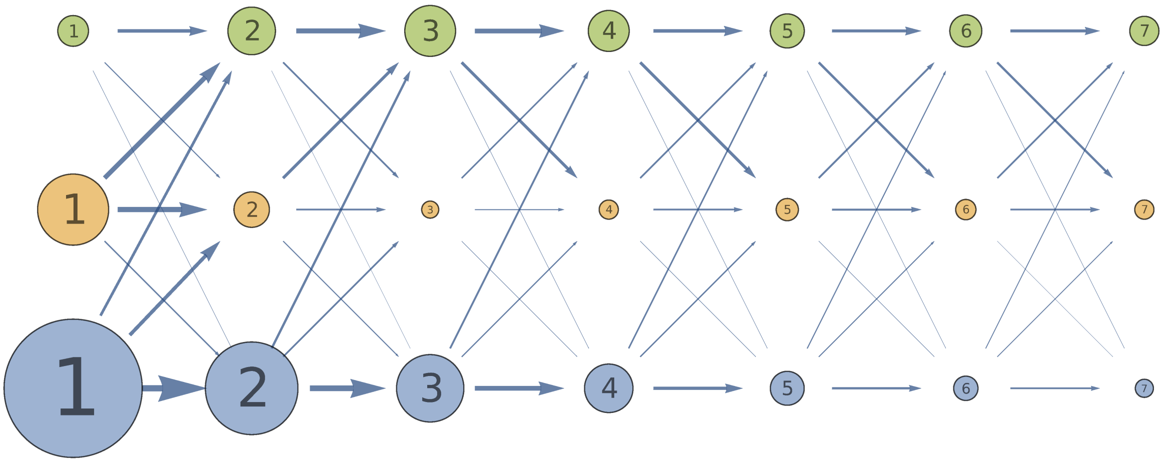

Figure 9. The aggregated dynamics of few first votes: vertex colours correspond to the temporal cluster IDs (waiting times for 1st, 2nd, . . . votes: short in blue, medium in orange and long ones in green), and their sizes are proportional to the number of posts. Arrow sizes are proportional to the transition probabilities $p_{j,i}^{(k)}$ (e.g., an arrow from orange 1 to green 2 represents $p_{3,2}^{(1)}$). Transitions from long to short waiting times are much less likely than vice versa.

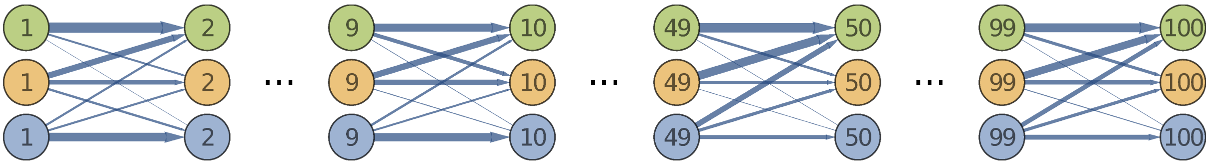

Figure 10. The normalised aggregated dynamics for later votes: this time, arrow sizes are scaled so that the flow from every node is constant. Note the dominating tendency to stay inside the blue (short-term) and green (long-term) clusters and to transition from the orange (medium-term) to the green ones.

to stabilise. We can summarise such inter-event time dynamics by the Markovian transition matrix

$$\mathbf{P} = (p_{j,i}) = \begin{pmatrix} 0.41 & 0.05 & 0.02 \\ 0.24 & 0.35 & 0.23 \\ 0.35 & 0.60 & 0.75 \end{pmatrix}, \tag{6}$$

where, as above, the entry in the j-th row and the i-th column gives the probability of transition from i to j.

Our observations are summarised in the next section.

Conclusions

This work introduces new insights with regards to human activity dynamics. We modelled the probability distributions of the random variables reflecting the temporal characteristics of the vote accumulation processes. Whilst exploring real-word data related to Stack Overflow user activity, we discovered three temporal clusters. The first cluster is comprised of the votes that appear almost immediately one after another; its relative size diminishes in time but does not become negligible. The second group represents the medium-term effects (measured in days). The third cluster groups all the rare events.

Importantly, the temporal structure of vote accumulation on Stack Overflow deviates from the simple Poisson processes. We observed the varying character of the inter-event time distribution and the dependence of its parameters upon whether it is an early or a late vote in the sequence.

Computing the eigenvector corresponding to eigenvalue 1 of the aggregated event transition probability matrix **P** (Eq. (6)) we get the ergodic distribution of the discussed Markov chain, which, once normalised, gives

$$(0.04, 0.26, 0.70)^T$$

which means that, in the long run, the system on average chooses the first cluster in 4%, second in 26% and third in 70% of the cases (also compare the α_i^∞ coefficients in Eq. (5)). Hence, the system most of the time has long breaks.

However, the probability of transition from a long- to a short-term cluster is not negligible: the so-called sleeping beauties (e.g., (Burrell, 2002, 2005; Egghe et al., 2011)) are expected to awake occasionally, although the data at hand did not allow us to identify any predictors why it might be so – it is an interesting topic for further research. Moreover, we noted that votes frequently arrive in bursts – short next-vote wait times tend to be followed by another immediate events.

Also, due to the above, we find that predicting whether a post will be successful or not based on its activity in its infancy cannot be done accurately (unlike in the case of YouTube and Ding (Szabo & Huberman, 2010)) without supplementary information, e.g., with regards to its thematic category, authorship, etc. In the future, we shall investigate to what extent such additions improve the predictive power.

The above requires extra attention when identifying the post's score with its substantive value. In our previous paper (Geras et al., 2020), we pointed out that negative votes are not so much a (negative) measure of quality. With this, however, one can notice that due to the bursty nature of the phenomena considered, high vote counts may be due to a short-lived popularity peak and not to the objective value of the answer or question.

Acknowledgements

All data used in this study were obtained through the Stack Overflow Academic Research Partnership Program; their publicly available, anonymised version is available at https://data.stackexchange.com/.

This research was supported by the Australian Research Council Discovery Project ARC DP210100227 (MG) and by the POB Research Centre Cybersecurity and Data Science of Warsaw University of Technology within the Excellence Initiative Program – Research University (ID-UB) (GS).

Agnieszka Geras would like to thank Catriona Geras and Aleksandra Krzynówek for their insightful comments.

References

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207–211.

Burrell, Q. L. (2002). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology*, *53*, 232–235.

Burrell, Q. L. (2005). Are »sleeping beauties« to be expected? *Scientometrics*, *65*.

Cencetti, G., Battiston, F., Lepri, B. & Karsai, M. (2021). Temporal properties of higher-order interactions in social networks. *Scientific Reports*, *11*(1), 1–10.

Chatterjee, P., Kong, M. & Pollock, L. (2020). Finding help with programming errors: An exploratory study of novice software engineers' focus in Stack Overflow posts. *Journal of Systems and Software*, *159*, 110454. https://doi.org/10.1016/j.jss.2019.110454

Chen, J., Zhou, W., Dong, Y., Wang, Z., Cui, C., Wu, F., Zhou, E. & Tang, Y. (2019). Analyzing time-dimension communication characterizations for representative scientific applications on supercomputer systems. *Frontiers of Computer Science*, *13*(6), 1228–1242.

Clauset, A., Shalizi, C. R. & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703.

Delbianco, F., Fioriti, A., Hernandez-Chanto, A. & Tohmé, F. (2020). A Markov-switching approach to the study of citations in academic journals. *Journal of Informetrics*, *14*(4), 101081.

Egghe, L. (2009). Time-dependent Lotkaian informetrics incorporating growth of sources and items. *Mathematical and Computer Modelling*, *49*.

Egghe, L., Guns, R. & Rousseau, R. (2011). Thoughts on uncitedness: Nobel laureates and Fields medalists as case studies. *Journal of the American Society for Information Science and Technology*, *62*.

Gao, S., Ma, J. & Chen, Z. (2015). Modeling and predicting retweeting dynamics on microblogging platforms. *Proceedings of the Eighth ACM International*

Conference on Web Search and Data Mining. https://doi.org/10.1145/2684822.2685303

Geras, A., Siudem, G. & Gagolewski, M. (2020). Should we introduce a dislike button for academic articles? *Journal of the Association for Information Science and Technology*, *71*(2), 221–229. https://doi.org/10.1002/asi.24231

Goh, K.-I. & Barabási, A.-L. (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, *81*(4).

Hartonen, A. (2013). How important tasks are performed: Peer review. *Scientific Reports*, *3*.

Janosov, M., Battiston, F. & Sinatra, R. (2020). Success and luck in creative careers. *EPJ Data Science*, *9*(1), 9.

Jo, H.-H. & Hiraoka, T. (2019). Bursty time series analysis for temporal networks. *Temporal network theory* (pp. 161–179). Springer.

Jo, H.-H., Pan, R. K. & Kaski, K. (2012). Time-varying priority queuing models for human dynamics. *Phys. Rev. E*, *85*.

Karsai, M., Jo, H.-H., Kaski, K. et al. (2018). *Bursty human dynamics.* Springer.

Karsai, M., Kaski, K., Barabási, A.-L. & Kertész, J. (2012). Universal features of correlated bursty behaviour. *Scientific Reports*, *2*(1), 1–7.

Kim, E.-K. & Jo, H.-H. (2016). Measuring burstiness for finite event sequences. *Phys. Rev. E*, *94*.

Liu, J., Baltes, S., Treude, C., Lo, D., Zhang, Y. & Xia, X. (2021). Characterizing search activities on stack overflow, 919–931. https://doi.org/10.1145/3468264.3468582

May, A., Wachs, J. & Hannák, A. (2019). Gender differences in participation and reward on Stack Overflow. *Empirical Software Engineering*, *24*(4), 1997–2019.

Mondal, S. & Roy, B. (2022). Reproducibility challenges and their impacts on technical q&a websites: The practitioners' perspectives. *15th Innovations in Software Engineering Conference*, 1–11. https://arxiv.org/pdf/2112.10056.pdf

Moriano, P., Hill, R. & Camp, L. J. (2021). Using bursty announcements for detecting bgp routing anomalies. *Computer Networks*, *188*, 107835.

Moutidis, I. & Williams, H. T. P. (2021). Community evolution on Stack Overflow. *PLOS ONE*, *16*(6), 1–23. https://doi.org/10.1371/journal.pone.0253010

Mryglod, O., Holovatch, Y. & Mryglod, I. (2012). Editorial process in scientific journals: Analysis and modeling. *Scientometrics*, *91*(1), 101–112.

Nadarajah, S. & Kotz, S. (2008). Models for citation behavior. *Scientometrics*, *72*(2).

Okada, M., Yamanishi, K. & Masuda, N. (2020). Long-tailed distributions of inter-event times as mixtures of exponential distributions. *Royal Society open science*, *7*(2), 191643.

Papoutsoglou, M., Kapitsaki, G. M. & Angelis, L. (2020). Modeling the effect of the badges gamification mechanism on personality traits of Stack Overflow users. *Simulation Modelling Practice and Theory*, *105*, 102157.

Rizoiu, M.-A., Lee, Y., Mishra, S. & Xie, L. (2018). Frontiers of multimedia research. Association for Computing Machinery; Morgan.

Robert, C. P. & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*.

Shen, H., Wang, D., Song, C. & Barabási, A.-L. (2014). Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Siudem, G. & Hołyst, J. A. (2019). Diffusion on hierarchical systems of weakly-coupled networks. *Physica A: Statistical Mechanics and its Applications*, *513*, 675–686.

Siudem, G., Żogała-Siudem, B., Cena, A. & Gagolewski, M. (2020). Three dimensions of scientific impact. *Proceedings of the National Academy of Sciences*, *117*(25), 13896–13900. https://doi.org/10.1073/pnas.2001064117

Szabo, G. & Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM*, *53*(8), 80–88.

Tausczik, Y. & Huang, X. (2020). Knowledge generation and sharing in online communities: Current trends and future directions. *Current Opinion in Psychology*, *36*, 60–64.

Thompson, W. H., Brantefors, P. & Fransson, P. (2017). From static to temporal network theory: Applications to functional brain connectivity. *Network Neuroscience*, *1*(2), 69–99. https://doi.org/10.1162/NETN_a_00011

Unicomb, S., Iñiguez, G., Gleeson, J. P. & Karsai, M. (2021). Dynamics of cascades on burstiness-controlled temporal networks. *Nature communications*, *12*(1), 1–10.

Vasilescu, B. et al. (2020). Academic papers using Stack Exchange data [https://meta.stackexchange.com/questions/134495, last accessed: 26th April 2022].

Wang, D., Song, C. & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science*, *342*. https://doi.org/10.1126/science.1237825

Wang, W., Yuan, N., Pan, L., Jiao, P., Dai, W., Xue, G. & Liu, D. (2015). Temporal patterns of emergency calls of a metropolitan city in china. *Physica A: Statistical Mechanics and its Applications*, *436*.

Yan, D.-C., Wei, Z.-W., Han, X.-P. & Wang, B.-H. (2017). Empirical analysis on the human dynamics of blogging behavior on github. *Physica A: Statistical Mechanics and its Applications*, *465*, 775–781.

Yasseri, T., Sumi, R., Rung, A., Kornai, A. & Kertész, J. (2012). Dynamics of conflicts in wikipedia. *PLOS ONE*, *7*(6), e38869.

Zhang, H., Wang, S., Chen, T.-H., Zou, Y. & Hassan, A. E. (2021). An empirical study of obsolete answers on Stack Overflow. *IEEE Transactions on Software Engineering*, *47*(4), 850–862. https://doi.org/10.1109/TSE.2019.2906315

Zhao, L. X., Zhang, L. & Jiang, J. (2021). Hot question prediction in Stack Overflow. *IET Software*, *15*(1), 90–106.

Zhou, J., Wang, S., Bezemer, C.-P. & Hassan, A. E. (2020). Bounties on technical Q&A sites: A case study of Stack Overflow bounties. *Empirical Software Engineering*, *25*(1), 139–177.

Zou, L., Wang, C., Zeng, A., Fan, Y. & Di, Z. (2021). Link prediction in growing networks with aging. *Social Networks*, *65*, 1–7.