

# Universities in the Avalanche Dynamics of Novelties and Non-Novelties

Filippo Santoro

*CREF - Centro Ricerche Enrico Fermi, Via Panisperna, 89a, 00184 Roma RM and  
Sapienza University of Rome, Physics Department, Piazzale Aldo Moro 5, 00185 Rome, Italy*

Alberto Petri

*CNR - Institute of Complex Systems, Department of Physics, Piazzale Aldo Moro 5, 00185, Italy and  
CREF - Centro Ricerche Enrico Fermi, Via Panisperna, 89a, 00184 Roma RM*

Francesca Tria\*

*Sapienza University of Rome, Physics Department, Piazzale Aldo Moro 5, 00185 Rome, Italy  
Complexity Science Hub Vienna, Austria and  
CREF - Centro Ricerche Enrico Fermi, Via Panisperna, 89a, 00184 Rome, Italy*

Unprecedented events intertwine with the repetition of the past in natural phenomena and human activities. Key statistical patterns, such as Heaps' and Taylor's laws and Zipf's law, have been identified as characterizing the dynamical processes that govern the emergence of novelties and the abundance of repeated elements. Observing these statistical regularities has been pivotal in motivating the search for modeling schemes that can explain them and clarify key mechanisms underlying the appearance of new elements and their subsequent recurrence. In this study, we analyze sequences of novel and non-novel elements, referred to as avalanches, in real-world systems. We show that avalanche statistics provide a complementary characterization of innovation dynamics, extending beyond the three fundamental laws mentioned above. Although arising from collective dynamics, some systems behave as a single instance of a stochastic process. Others, such as natural language, exhibit features that we can only explain by a superposition of different dynamics. This distinction is not apparent when considering Heaps' law alone, while it clearly emerges in the avalanche statistics. By interpreting these empirical observations, we also advance the theoretical understanding of urn-based models that successfully reproduce the observed behaviors associated with Heaps', Zipf's, and Taylor's laws. We derive analytical expressions that accurately describe the probability distributions of avalanches and the Heaps law beyond its asymptotic regime. Building on these results, we derive a scaling relation that we show also holds in real-world systems, indicating a form of universality in the dynamics of novelty.

## INTRODUCTION

Innovations and novelties are increasingly recognized as valuable perspectives for gaining insights into systems related to human activities, social behaviors, and biological evolution. Diverse efforts have been devoted to understanding and explaining the mechanisms driving societal transformations and creativity, from economic growth [1, 2] to the building of collective knowledge [3–5], to the emergence of language structures [6]. We all recall how the emergence and diffusion of a new viral variant triggered novel societal habits and attitudes, which in turn impacted environmental awareness [7, 8]. This cascade effect often marks less dramatic events: a conversation with a new collaborator can lead to the opening of a new research line, listening to a song by an unknown artist can spark interest in their entire body of work, or a single word dropped in conversation can initiate a new topic of discussion. We immediately note that we sometimes mention genuine innovations and sometimes novelties – the former referring to something new to the world on a global scale, and the latter to the adoption or experience of something new for an individual or a group on a local scale. In what follows, we will treat innovation

and novelties on a common ground since their dynamics exhibit shared behaviors [9], even though they are distinct concepts. Let us describe an innovation system as an ordered sequence of elements or events. This perspective encompasses a wide variety of different systems, from a sequence of genes sampled in a population to the sequence of songs listened to by users of a website, to the sequence of words that compose a book, to the sequence of contacts in social networks [9–11]. In this view, three fundamental laws hold: the Heaps [12], Zipf [13], and Taylor [14] laws. The Heaps law describes the emergence rate of novelties or innovations and is related to the Zipf law, which describes the frequency distribution of the distinct elements in a system [15, 16]. The Taylor law [14] was initially introduced in ecology, relating the variance of a population density to its mean. A departure from a proportionality relation indicates a correlation in the distribution of individuals; specifically, variance growing more rapidly than the mean is a signal of mutual attraction among individuals in a natural population. In the framework of language [17] and of innovation systems at large [18, 19], Taylor's law is applied to the relation between the variance and the mean of the number of innovations that emerged from the onset of the system until

a given moment. It turns out that the variance grows approximately with the square of the mean, pointing to positive correlations in the emergence of novelties [18, 19].

Here, we explicitly address the correlation in the temporal occurrence of innovations from its statistical perspective by considering the occurrence of consecutive sequences of novelties and non-novelties, i.e., *avalanches* of novelties and non-novelties.

Avalanches are referred to as phenomena that display randomly distributed events interspersed by inactivity or events of a different nature. They have received much attention because of their widespread presence in systems of a very different nature, from the paradigmatic case of earthquakes [20–22], to magnetic hysteresis [23], structural phase transitions [24–26], plastic deformation [27], granular shearing [28–32], internal sliding [33], as well as in neural activity [34–37], and in the mobility of humans [38] and of ants [39]. Distinctive quantities of avalanches, such as size and duration, are generally observed to display statistical distributions in the form of power laws in non-trivial systems. This behavior is often connected to the presence of some critical point [40–45] and of some form of scaling, i.e., in a given system, distributions related to different system parameters can collapse onto a single, universal, invariant shape if expressed as a function of a suitable variable. This points to the presence of universal behaviors [44, 46–48]. Moreover, it has been pointed out [49] that systems that display similar avalanche distributions sometimes feature different inter-event time distributions, suggesting that the latter can provide a complementary characterization that helps understand the process's dynamics and discriminate among different hypotheses and models.

Here, we investigate the avalanche statistics of novelties and non-novelties in several innovation systems. We define an avalanche of novelties as a sequence of only consecutive novelties. Of particular interest are the intervals (or *inter-times*) between two successive novelties, which we also name avalanches of non-novelties, i.e., sequences of only consecutive non-novelties. While avalanches of novelties are very short, the lengths of the avalanches of non-novelties span different orders of magnitude, so that their distribution provides helpful insights into understanding the underlying dynamics and assessing the goodness of generative models.

We find that systems that behave similarly regarding the three fundamental laws discussed above—the Heaps, Zipf, and Taylor laws—exhibit different behaviours concerning the avalanche statistics of novelties and non-novelties. In particular, among the considered systems, some exhibit a form of the inter-times distribution featuring an exponential-like decay. In contrast, others feature a fatter tail. Further, the Corpus of English written texts displays a clear crossover between two power law regimes.

To gain insights into these results, we refer to a class of Polya's urn-based models proposed in [9] that suc-

cessfully account for the three fundamental laws mentioned above and acts as reference models for various research in innovation systems and complex networks analysis [5, 11, 50]. We will focus in particular on the urn model with triggering (UMT) and a specific version of it, exhibiting exchangeability [51, 52] and here dubbed UMT-E, that provides an urn characterisation of the two-parameter Poisson-Dirichlet process or Pitman-Yor process [19, 53] and was recently used in inference problems [54]. The focus on the UMT-E model is motivated by its analytical convenience. However, we show in SI that the same results can be obtained by referring to the general UMT model, as well as to a generalization of it, accounting for semantic correlations, the urn model with semantic triggering (UMST) [9].

Through the lens of the UMT model, we find that the avalanche size statistics allow us to distinguish between processes that, although resulting from the contribution of many agents, can be described as a single collective dynamics and processes that require an explicit description in terms of a superposition of individual dynamics. In particular, we find a perfect agreement between observations and model predictions for the datasets featuring a closed-to-exponential tail. At the same time, we argue that the fat-tail distributions are an output of collective dynamics.

Moreover, we observe that, where a superposition of individual dynamics can account for the observed avalanche statistics, a single realization of the UMT can still reproduce it, though for different parameters than those that reproduce the Heaps law of the process. This observation could help understand the relation between the individual and the collective innovation rate.

As a support for the analysis, and as a value per se, we provide analytical expressions both for the Heaps law in the UMT model, beyond the asymptotic exponent, and for the avalanche size distribution of novelties and non-novelties. We firstly discuss predictions for the UMT-E, where a closed form for the average number of distinct elements as a function of the sequence length, expressing the Heaps law, can be written. Further, we provide the predictions of the general non-exchangeable version of the UMT, where it is possible to write an inverse relation for the Heaps law, namely the average time corresponding to a given number of distinct elements. One can recover the asymptotic relation provided in [9] by inverting this exact relation up to the leading order. In addition, we show that the inverse relation is sufficient to obtain the avalanche size distribution of novelties and non-novelties. As a final result, we predict a scaling relation for the avalanche distributions from the UMT model and successfully apply it to the considered systems.

We organize the paper as follows. The Results section consists of three main parts. First, we report and comment on the results for the novelty and non-novelty avalanche size statistics in several systems featuring in-

novation. We contrast these results with the predictions of the UMT model. Second, we present analytical results for the main observables we consider, namely the number of distinct elements as a function of the sequence length, which expresses Heaps' law, and the probability distribution of avalanche lengths for both novelties and non-novelties. We finally derive the scaling relation for the avalanche distributions from the UMT-E model, and we test it in all the considered systems, as well as in the general UMT and the generalized UMST models. We find a striking agreement with the theoretical predictions in all the cases. A Methods section follows, detailing the techniques adopted in the data analyses and numerical simulations. The Supplementary Information provides more details.

## RESULTS

### Avalanches of novelties and non-novelties in real data

Innovation is an intermittent process, for which real systems lie between the drive to explore the adjacent possible and the need to exploit the already known. An ordered sequence can be divided into sub-sequences according to their nature, composed of new or not new elements, so avalanches of novelties and non-novelties can be defined.

An avalanche of novelties is a sub-sequence composed of elements new to the system, bounded on the left and right by two elements that are not new. An avalanche of non-novelties is a sub-sequence composed of elements not new to the system, bounded on the left and right by two new elements. Fig. 1 schematizes sub-sequences of consecutive novelties or non-novelties. Avalanches of non-novelties can be understood as intervals between two novelties, and vice versa. In the following, we will use these expressions interchangeably.

We here analyze the size distribution of novelties and non-novelties avalanches in four databases related to web-mediated human activity (LastFM, Github repositories, Twitter Hashtags, Wikipedia) and one dataset related to non-web-mediated cultural production (The Gutenberg corpus of digitized English books). Moreover, in the light of the results in [11], which show that the expansion into the adjacent possible can account for the human dynamics in social networks, we also analyse the avalanche statistics in the sequences of users' activity in the Twitter and Github datasets. All the systems considered share common statistical features expressed by the Zipf, Heaps, and Taylor laws [9, 19, 55].

We can formalize each dataset as a sequence  $S$  of elements  $s_i$ ,  $i = 1, \dots, t_{\max}$ , where we can think of each index as a discrete time and where the values that the  $s_i$  can take depend on the dataset: an element is a word

in the case of books and Wikipedia pages, an hashtag in the case of Twitter hashtags, a repository in the case of Github, a person in the Twitter and Github users datasets, a song in LastFM.

Let us label each element  $s_i$  with a label  $x_i = 0$  if it is a novelty, i.e., if its value did not appear in the sequence at times  $t < i$ , and  $x_i = 1$  otherwise. For instance, in the Gutenberg and Wikipedia datasets, a novelty is a word read for the first time by a reader who reads the sequences of pages or books in a given order set once and for all; in the sequences of users' activity in Twitter and Github, a novelty is a user who appears in the dataset for the first time; and so on. We provide details on the datasets in the section Methods.

We call an avalanche of novelties of size (or length)  $l$  a subsequence of  $l$  contiguous elements with label 0, preceded and followed by at least one element with label 1; vice versa for an avalanche of non-novelties.

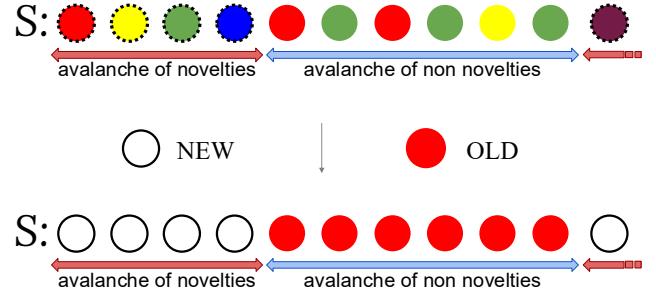


FIG. 1. Cartoon of a sequence  $S$  of events (flowing from left to right) where we highlight avalanches of novelties and avalanches of non-novelties. Top: coloured balls encircled with dotted lines represent novelties, i.e., elements that have never appeared in the sequence until that moment; coloured balls with no contour line represent non-novelties, i.e., elements already present. Each color represents a distinct element that, depending on the context, can be: a word, a song, an edit, a hashtag, etc. Bottom: same as in the top, but we only characterize elements by being a novelty (white balls) or a non-novelty (red balls).

As could be expected, the avalanches of novelties are small, spanning only an order of magnitude. We report the avalanche size distributions for the considered datasets and the corresponding results from the UMT-E model in Fig. 2.

More interesting are the avalanches of non-novelties, or equivalently, the inter-time distributions between two successive novelties, since they span several orders of magnitude. We show them in Fig. 3, along with the theoretical prediction from the UMT-E model and the corresponding numerical simulations. To set the model's parameters, we fit the Heaps law by using the theoretical prediction discussed below (please refer to the fit in S1 of SI). We use the obtained parameter values to gen-

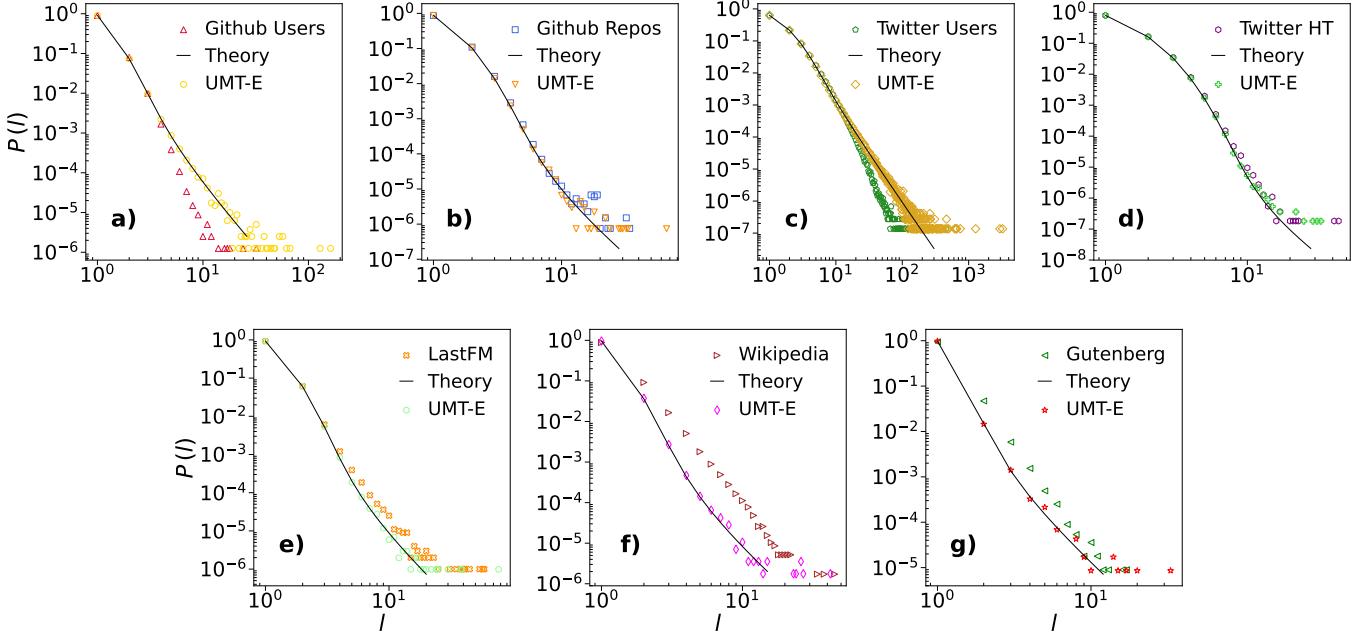


FIG. 2. **Avalanche size distribution for novelties.** We show results for the empirical datasets, contrasting them with predictions from the UMT-E model, as obtained through analytical computation and numerical simulations. The parameters used for the UMT-E model predictions are those obtained from the fit of the Heaps law in the empirical datasets, and we report them in brackets for each dataset. **a)** Github Users ( $\nu = 4, \rho = 7, N_0 = 45468$ ). **b)** Github Repositories ( $\nu = 22, \rho = 29, N_0 = 14640$ ). **c)** Twitter Users ( $\nu = 16, \rho = 29, N_0 = 42183043$ ). **d)** Twitter Hashtags( $\nu = 19, \rho = 22, N_0 = 1995$ ). **e)** LastFM( $\nu = 28, \rho = 41, N_0 = 27200$ ). **f)** The Wikipedia Corpus ( $\nu = 31, \rho = 50, N_0 = 25000$ ). **g)** the Gutenberg Corpus ( $\nu = 9, \rho = 19, N_0 = 5890$ ). We have truncated the sequences of both Gutenberg and Wikipedia at length  $2 \cdot 10^7$ .

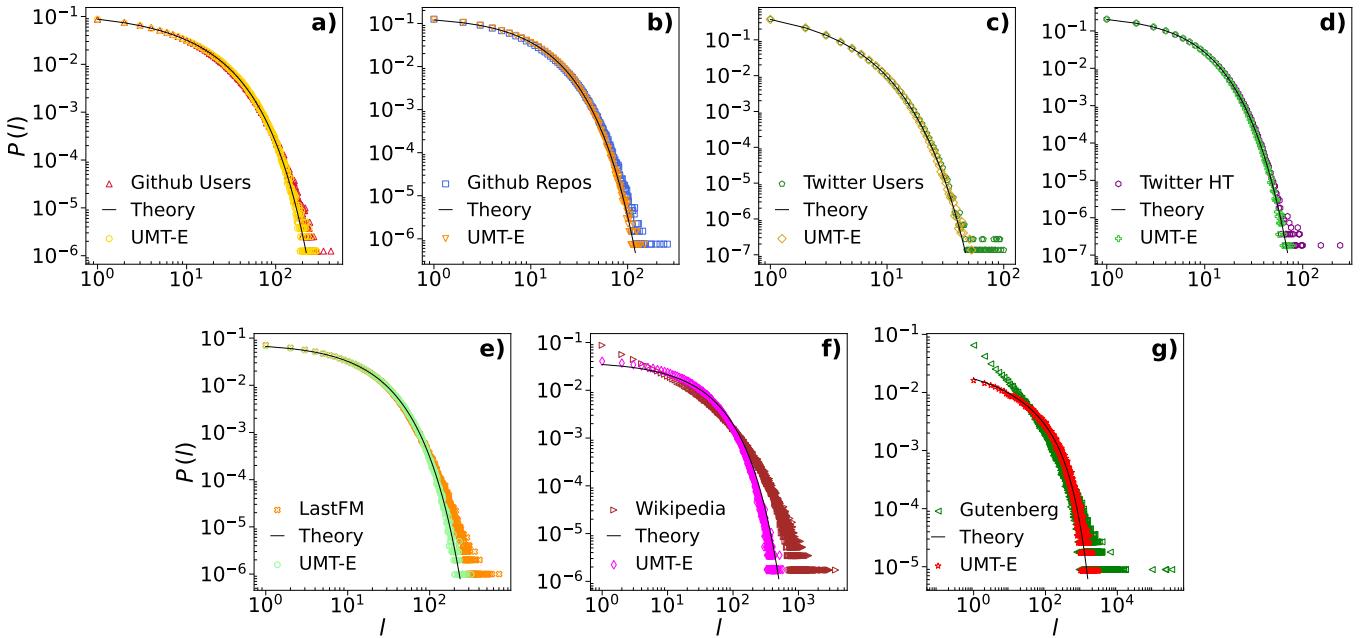
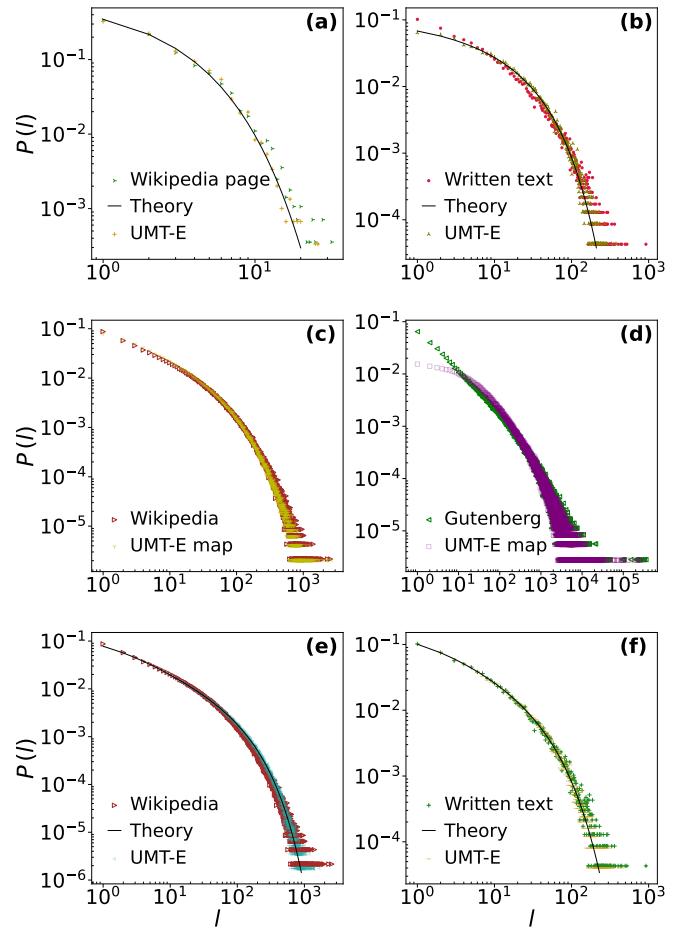


FIG. 3. **Avalanche size distribution for non-novelties.** We show results for the empirical datasets, contrasting them with predictions from the UMT-E model, as obtained through analytical computation and numerical simulations. The parameters used for the UMT-E model predictions are those obtained from the fit of the Heaps law in the empirical datasets, and we report them in brackets for each dataset. **a)** Github Users ( $\nu = 4, \rho = 7, N_0 = 45468$ ). **b)** Github Repositories ( $\nu = 22, \rho = 29, N_0 = 14640$ ). **c)** Twitter Users ( $\nu = 16, \rho = 29, N_0 = 42183043$ ). **d)** Twitter Hashtags ( $\nu = 19, \rho = 22, N_0 = 1995$ ). **e)** LastFM ( $\nu = 28, \rho = 41, N_0 = 27200$ ). **f)** The Wikipedia Corpus ( $\nu = 31, \rho = 50, N_0 = 25000$ ). **g)** the Gutenberg Corpus ( $\nu = 9, \rho = 19, N_0 = 5890$ ). We have truncated the sequences of both Gutenberg and Wikipedia at length  $2 \cdot 10^7$ .

erate sequences from the UMT-E and in the analytical estimate for the avalanche size statistics.

We observe that the model perfectly reproduces the behavior displayed by the Github and Twitter datasets, related both to the users' production and activity (Fig. 3(a-d)). The inter-time distribution in LastFM (Fig. 3(e)) features a slightly larger tail than that reproduced by the model, while in the Gutenberg and the Wikipedia datasets it features a fat tail (Fig. 3(f-g)). Although both referred to the written English language, the last two corpora show a clear difference. The inter-times distribution of the Gutenberg corpus displays a power law behavior with a double slope, which is absent in the Wikipedia corpus. To better understand the origin of this difference, we investigate the behavior of single books and single Wikipedia pages. We find that the UMT-E model perfectly reproduces the avalanche size statistics of single Wikipedia pages (Fig. 4(a)). Conversely, single books exhibit a slight excess in short avalanche sizes and a slightly fatter tail than the corresponding sequences from the model, but thinner than for the whole Corpus (Fig. 4(b)). Please refer to S3 of SI for further examples). We hypothesize, therefore, that the fat tail in the inter-times distribution in the Wikipedia and Gutenberg corpora derives from the convolution of different dynamics of single Wikipedia pages or single books, respectively. To test this hypothesis, we devised an experiment in which we separately fit the Heaps law for all the books in the Gutenberg corpus. We then generate sequences from the UMT-E model with the fitted parameters and stack them one after the other to form a unique sequence, as with real data. We performed the same experiment on the Wikipedia pages. This procedure requires a suitable mapping between elements in the generated sequences and words (we leave a complete hierarchical model for further research, beyond the scope of this work). Here, we use a direct mapping by identifying the most popular element in the generated sequence with the most popular word in the corresponding book or Wikipedia page, the second most popular element with the second most popular word, and so on (please refer to the Method section for details). We show the results of these experiments in Fig. 4(c-d), respectively for the Wikipedia and the Gutenberg corpus. We can see that the synthetic data perfectly reproduce the fat tail of the distributions of both datasets. The power law behaviour at short lengths observed in the Gutenberg corpus is, on the contrary, not explained by the bare UMT model.

As a last remark, we note that the inter-time distributions for the whole Wikipedia corpus and for single books can be well reproduced by a single UMT-E process with suitable parameter values, if we relax the constraint that the model should also fit the Heaps law (Fig. 4(e-f)). This analogy between the whole Wikipedia corpus and single books suggests that the coexistence of different topics within a single book can play a key role in shaping the



**FIG. 4. Avalanche size distribution for non-novelties in the natural language corpora.** Top: We contrast results for the empirical datasets with predictions from the UMT-E model obtained through analytical computation and numerical simulations. The parameters used for the UMT-E model predictions, reported in brackets, are those obtained from the fit of the Heaps law in the corresponding empirical data: **a**) A single Wikipedia page ( $\nu = 4, \rho = 7, N_0 = 45468$ ); **b**) a single book ( $\nu = 50, \rho = 97, N_0 = 22912$ ). Center: We contrast results for the empirical datasets, i.e., **c**) the Wikipedia corpus and **d**) the Gutenberg corpus, with predictions from numerical simulations of multiple realizations of the UMT-E model, stacked one after the other, as described in the main text. In the case of Wikipedia, we have considered sequences with more than 500 words. Bottom: We contrast results for the empirical datasets with predictions from the UMT-E model obtained through analytical computation and numerical simulations. The parameters used for the UMT-E model predictions, reported in brackets, are those obtained from the fit of the avalanche size distribution for non-novelties in the corresponding empirical data: **e**) The Wikipedia corpus ( $\nu = 2, \rho = 9, N_0 = 400000$ ); **f**) a single book ( $\nu = 23, \rho = 73, N_0 = 127020$ ).

avalanche statistics in the literary Corpus.

In the next section, we derive the theoretical prediction for the Heaps law and the avalanche statistics for the UMT-E and the general UMT model.

### Predictions from the urn model with triggering

In the urn-based models proposed in [9], sequences of events are constructed through repeated extraction from an urn containing all the possible outcomes. The content of the urn changes along the evolution of the system, always allowing the entrance of brand new elements. Depending on the context, elements in the urn represent: words in a book or in a Wikipedia page, songs listened to on a music website, edits in collaborative software projects, hashtags on Twitter, and so on. The sequence of elements drawn from the urn represents the realized history.

#### *The model*

The urn model with triggering (UMT) is a generative model introduced in [9]. It is inspired by the adjacent possible expansion [56] as the framework for the emergence of novelties, and predicts the power laws for the Heaps, Zipf, and Taylor laws [9, 18, 19].

The process starts with an urn containing  $N_0$  balls of different colors. A realization is obtained through the sequential random extraction of balls from the urn. Each time a ball is extracted, it is appended to the sequence  $S$  and its color is reinforced, putting in the urn  $\rho$  copies of it. If, in addition, the color is a novelty, that is, it appears for the first time in the sequence, then also  $\nu+1$  different balls of brand new colors are added to the urn (*triggered innovation* process). Let us now consider a slightly more general version of the model by modifying the reinforcing parameter  $\rho$  for novelties. In particular, one can use the reinforcing parameter  $\rho$  if the extracted element is a non-novelty, and a reinforcing parameter  $\tilde{\rho}$  if the extracted element is a novelty. This change does not impact the asymptotic exponents of the Heaps, Zipf, and Taylor laws [9]. When  $\rho = \tilde{\rho}$ , we end up in the original UMT model. By fixing instead  $\tilde{\rho} = \rho - (\nu + 1)$ , in the case  $\nu < \rho$ , so that the number of balls inserted in the urn is constant at each step, independently on the drawn of a new or old element, the model becomes exchangeable, meaning that the probability of any generated sequence depends only on the numbers  $n_i$  of each color  $i$  in the sequence (and independently of the identity of the color  $i$ ) and not on the order in which they appear. The exchangeable version of the model, UMT-E, was shown [18, 19] to be an urn representation of the two-parameter Poisson-Dirichlet process. In Fig. 5 we highlight the difference between the UMT and the UMT-E

models.

For any value of  $\tilde{\rho} \geq 0$ , the model predicts asymptotic power laws for the Heaps and Zipf laws, of the form respectively  $D(t) \simeq t^\beta$  and  $f(R) \simeq R^{-\alpha}$ , where  $D(t)$  is the average number of distinct elements in a sequence of length  $t$  and  $f(R)$  is the fraction of elements of rank  $R$ , where elements are ordered according to their frequency, from the most to the least frequent. The exponents depend on the relative importance of innovation versus reinforcement, through the ratio  $\frac{\nu}{\rho}$ , and read  $\alpha = \frac{\rho}{\nu}$ , and  $\beta = \frac{\nu}{\rho}$  when  $\nu < \rho$ ,  $\beta = 1$  when  $\nu > \rho$  [9]. Furthermore, the model also predicts the Taylor law with a relation  $\sigma \simeq \mu^\gamma$ , where  $\gamma = 1$  when  $\nu < \rho$ , as approximately observed in real data, and spanning all the exponents between 1 and 0.5 when  $\nu > \rho$ , the value 0.5 characterizing processes where innovation enter one after the other independently [18, 19].

The asymptotic exponents are enough to characterize the Heaps', Zipf, and Taylor laws. However, to provide an analytical expression for the avalanche distribution of novelties and non-novelties, we need to step further and seek an expression of  $D(t)$  that is also valid at small time  $t$ . The initial number of balls  $N_0$  is a crucial parameter for the exact form of  $D(t)$  and for the shape of the avalanche statistics.

#### *The Heaps law in the UMT-E*

The exchangeable version of the model (UMT-E) allows for an explicit expression of the average number of distinct elements  $D(t)$  as a function of the total number of elements  $t$ . We can solve the continuous-time differential equation

$$\frac{dD(t)}{dt} = p_{\text{new}}(t) = \frac{N_0 + \nu D(t)}{N_0 + \rho t} \quad (1)$$

with initial conditions  $D(t = 0) = 0$ . We obtain:

$$\begin{aligned} D(t) &= \frac{N_0}{\nu} \left( 1 + \frac{\rho t}{N_0} \right)^{\frac{\nu}{\rho}} - \frac{N_0}{\nu} = \\ &= \frac{\theta}{\alpha} \left( 1 + \frac{t}{\theta} \right)^\alpha - \frac{\theta}{\alpha} \end{aligned} \quad (2)$$

where in the last expression we set the identities  $\frac{N_0}{\rho} \equiv \theta$  and  $\frac{\nu}{\rho} \equiv \alpha$  to stress the dependence of the exchangeable UMT-E model on only two parameters and make more evident the connection with the two-parameter Poisson-Dirichlet process, as discussed in [18, 19]. We note that the result in Eq. (2) could be derived, less straightforwardly, by taking the long sequence limit, from the exact solution of the two-parameter Poisson-Dirichlet process [57], or also from the results on triangular urn models in [58]. In Fig. 6, we show the theoretical result in

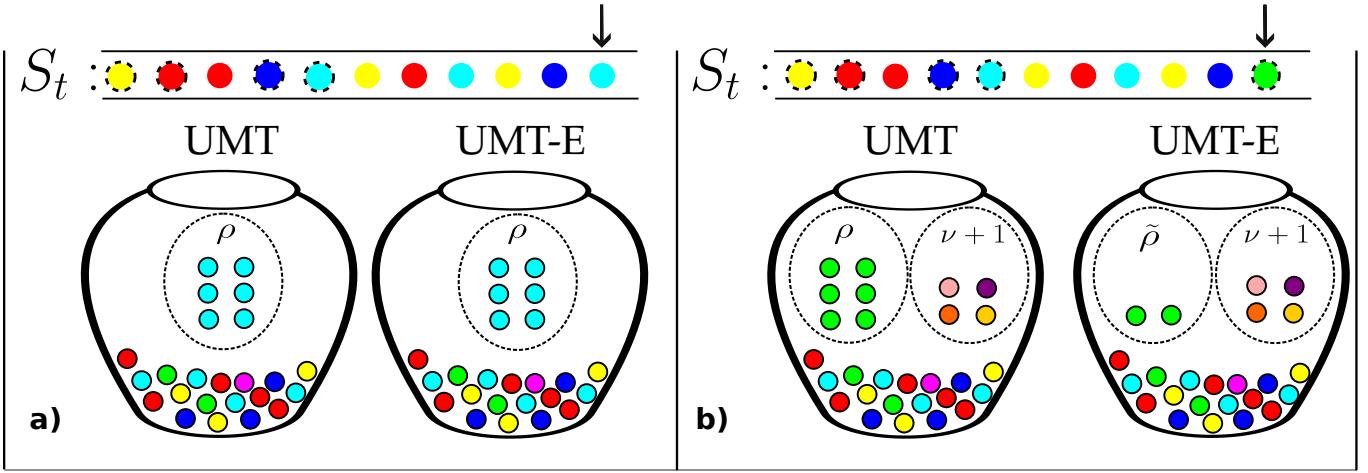


FIG. 5. **Cartoon of the UMT process and its exchangeable version, UMT-E.** **a)** The last extracted element (the cyan ball) has already appeared in the sequence, i.e., is a non-novelty. Only the reinforcement process occurs: we insert  $\rho$  copies of the cyan ball into the urn in both models. **b)** The last extracted element (the lime ball) has never appeared in the sequence, i.e., it is a novelty. In this case, both reinforcement and triggering processes occur. The triggering process is identical for both models: we insert  $\nu + 1$  new distinct elements, whose colors were not present in the urn, into it. Conversely, the reinforcement process differentiates the two models: in UMT, we introduce  $\rho$  copies of the last appended ball, while in UMT-E, we only introduce  $\tilde{\rho} = \rho - (\nu + 1)$  copies of it. Note that in the UMT-E model, we introduce the same number of balls in the urn at each step, independently of the last extracted ball.

Eq. (2) contrasted with results from simulations for different values of the ratios  $\frac{\nu}{\rho}$  (Fig. 6 (a)) and  $\frac{N_0}{\rho}$  (Fig. 6 (b)).

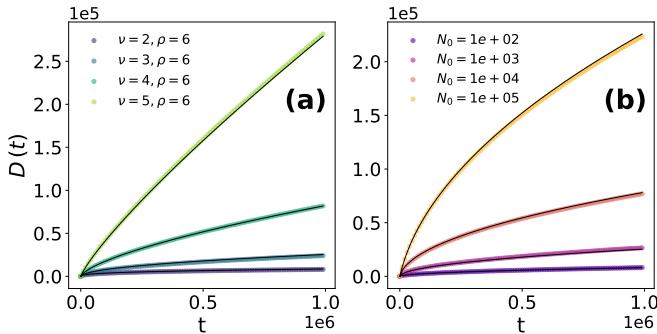


FIG. 6. **The Heaps law in the UMT-E.** We show results for the Heaps law from numerical simulations of the UMT-E, along with the corresponding analytical predictions. **a)** We fix  $N_0 = 1000$  and  $\rho = 6$ , and we consider different ratios  $\frac{\nu}{\rho}$  by moving  $\nu$  in the integer range [2, 5]. **b)** We fix  $\nu = 3$ ,  $\rho = 6$ , and  $N_0$  runs across different orders of magnitude from  $10^2$  to  $10^5$ . All the sequences generated from numerical simulations have length  $10^6$ .

#### *The Heaps law in the general UMT*

We here provide the inverse relation  $t(D)$  for the general UMT, with any definition of  $\tilde{\rho} \geq 0$ . This relation cannot be inverted to provide an exact expression of  $D(t)$

in the general case. However, the results in [9] can be obtained by inverting  $t(D)$  up to the first-order approximation in the power of  $t$ . Through  $t(D)$ , we can provide an analytical expression for the avalanche size distributions of novelties and non-novelties.

Let us then consider the general UMT model with  $\tilde{\rho} \geq 0$ . The continuous time equation for  $D(t)$  reads:

$$\frac{dD(t)}{dt} = p_{\text{new}}(t) = \frac{N_0 + \nu D(t)}{N_0 + \rho t + aD(t)} \quad (3)$$

with  $a \equiv \tilde{\rho} + \nu + 1 - \rho$ , and for  $a = 0$  we recover the UMT-E, while for  $a = \nu + 1$  we recover the original UMT model. We solve instead the inverse equation:

$$\frac{dt(D)}{dD} = \frac{N_0 + \rho t(D) + aD}{N_0 + \nu D} \quad (4)$$

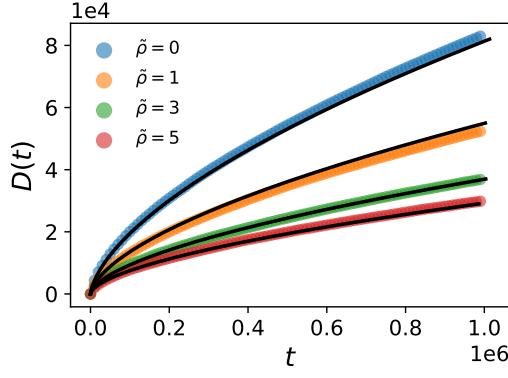
that we can write as

$$\frac{dt(D)}{dD} - \frac{\rho}{N_0 + \nu D} t(D) = \frac{N_0 + aD}{N_0 + \nu D} \quad (5)$$

putting in evidence its nature as a non-homogeneous first-order linear differential equation. Solving by standard methods and imposing the initial condition  $t(0) = 0$ , we obtain the solution:

$$\begin{aligned} t(D) &= \frac{N_0}{\rho} \left( 1 + \frac{a}{\rho - \nu} \right) \left( \frac{N_0 + \nu D}{N_0} \right)^{\frac{\rho}{\nu}} + \\ &- a \left( \frac{N_0 + \nu D}{\nu(\rho - \nu)} \right) - \frac{N_0}{\rho} \left( \frac{\nu - a}{\nu} \right) \end{aligned} \quad (6)$$

We show in Fig. 7 the theoretical result in Eq.(6) contrasted with simulated results of the model for different values of  $\tilde{\rho} \geq 0$ , finding perfect agreement.



**FIG. 7. The Heaps law in the general UMT.** We show results for the Heaps law from numerical simulations of the UMT, along with the corresponding analytical predictions. We set the parameters values to ( $\nu = 3$ ,  $\rho = 5$ , and  $N_0 = 1000$ ), and we vary  $\tilde{\rho}$  in the range  $[0, 1, 3, 5]$ . Continuous lines represent the theoretical predictions as discussed in the main text. All the sequences generated from numerical simulations have length  $10^6$ .

#### Avalanches statistics in the UMT-E

Let us first focus on the probability distribution of the non-novelties avalanche size, and we then derive the corresponding probability for the novelties in an analogous way. Let us label as above  $x_i = 0$  if the element is a novelty, and  $x_i = 1$  if it is not a novelty (it is a non-novelty). We call an avalanche of non-novelties of size (or length)  $l$  a subsequence of  $l$  contiguous elements with label 1, preceded and followed by at least one element with label 0.

We start by noting that the probability of finding that subsequence depends on the time at which we seek it, so we need to define a time-dependent probability  $p_t(x_{t-1}, x_t, x_{t+1}, \dots, x_{t+l-1}, x_{t+l})$  such that  $x_{t-1} = x_{t+l} = 0$  and  $x_t, x_{t+1}, \dots, x_{t+l-1} = 1$ . We can write:

$$\begin{aligned} p_t(x_{t-1}, x_t, x_{t+1}, \dots, x_{t+l-1}, x_{t+l}) &= \\ &= p_{new}(t-1)p_{new}(t+l)\prod_{i=0}^{l-1}(1-p_{new}(i)) \simeq \\ &\simeq p_{new}(t)^2(1-p_{new}(t))^l \end{aligned} \quad (7)$$

where  $p_{new}(t)$  is the probability of finding a new element at time  $t$ . We then have to sum up over all the times from zero to  $t_{\max}$ , where we can use the sequence length as  $t_{\max}$  [59]. Furthermore, the non-novelties avalanche size distribution is obtained by conditioning the above probability on the actual finding of a non-novelty avalanche,

that is, on the probability of finding the pair  $\{novelty, non-novelty\}$ . The latter can be expressed by  $p_t(0,1) = p_{new}(t-1)(1-p_{new}(t)) \simeq p_{new}(t)(1-p_{new}(t))$ . By using Bayes' theorem, we finally obtain the non-novelties avalanche size distribution:

$$P_{\text{non-nov}}(l) = \frac{\int_0^{t_{\max}} p_{new}(t)^2(1-p_{new}(t))^l dt}{\int_0^{t_{\max}} p_{new}(t)(1-p_{new}(t))dt}. \quad (8)$$

By substituting, from Eqs.(1) and (2):

$$p_{new}(t) = \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho}-1}. \quad (9)$$

We can thus write explicitly:

$$P_{\text{non-nov}}(l) = \frac{\int_0^{t_{\max}} \left(1 + \frac{\rho t}{N_0}\right)^{\frac{2\nu}{\rho}-2} \left(1 - \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho}-1}\right)^l dt}{\int_0^{t_{\max}} \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho}-1} \left(1 - \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho}-1}\right) dt}. \quad (10)$$

The curves in Figs. 3,4(a-b-e-f), 8 and 9 are obtained by performing a numerical integration of Eq. (8). In the Supplementary Information (S4), we also provide an explicit series expansion for  $P_{\text{non-nov}}(l)$ .

The avalanche size distribution for novelties can be found in a completely analogous way, obtaining:

$$P_{\text{nov}}(l) = \frac{\int_0^{t_{\max}} p_{new}(t)^l(1-p_{new}(t))^2 dt}{\int_0^{t_{\max}} p_{new}(t)(1-p_{new}(t))dt}, \quad (11)$$

and, substituting the expression for  $p_{new}(t)$ :

$$P_{\text{nov}}(l) = \frac{\int_0^{t_{\max}} \left(1 + \frac{\rho t}{N_0}\right)^{\frac{2\nu}{\rho}-l} \left(1 - \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho}-1}\right)^2 dt}{\int_0^{t_{\max}} \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho}-1} \left(1 - \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho}-1}\right) dt}. \quad (12)$$

We note that the avalanche size distributions depend on the ratios  $\frac{N_0}{\rho}$  and  $\frac{\nu}{\rho}$ , and on the sequence length  $t_{\max}$ . However, we can rescale the avalanche size  $l$  to let curves obtained at different sequence lengths  $t_{\max}$  collapse into a unique curve, as shown in the next subsection.

In Fig. 8, we contrast simulations and analytical results of the inter-times distribution of the UMT-E for different values of the ratio  $\nu/\rho$  at fixed  $N_0/\rho$ , (a), and for different values of  $N_0/\rho$  at fixed ratio  $\nu/\rho$  (b), at a fixed sequence length in both cases. We note that, as one can intuitively anticipate, the distribution features a support that enlarges when those ratios decrease. We report the corresponding analysis of the avalanches of novelties in the S2 of the Supplementary Information.

#### Avalanches statistics in the general UMT

Building on the results of the previous section, to express the non-novelties avalanche size distribution, we

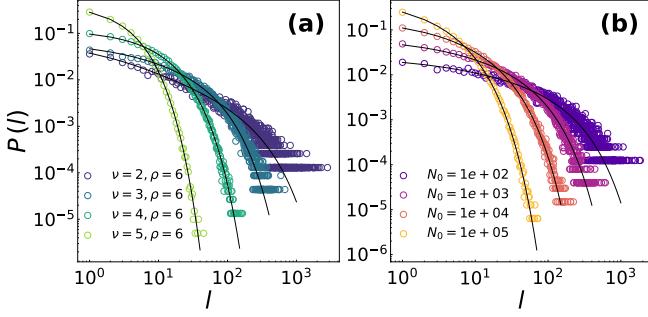


FIG. 8. **The Avalanche size distribution for non-novelties in the UMT-E.** We show results from numerical simulations of the UMT-E, along with the corresponding analytical predictions. **a)** We fix  $N_0 = 1000$  and  $\rho = 6$ , and we consider different ratios  $\frac{\nu}{\rho}$  by moving  $\nu$  in the integer range [2, 5]. **b)** We fix  $\nu = 3$ ,  $\rho = 6$ , and  $N_0$  runs across different orders of magnitude from  $10^2$  to  $10^5$ . All the sequences generated from numerical simulations have length  $10^6$ .

consider  $D$  as the independent variable in Eq. (8). We obtain  $p_{\text{new}}(D)$  by substituting 6 in 3, and we can thus write:

$$\begin{aligned} P_{\text{non-nov}}(l) &= \frac{\int_0^{D_{\max}} p_{\text{new}}(D)^2 (1 - p_{\text{new}}(D))^l \frac{dt}{dD} dD}{\int_0^{D_{\max}} p_{\text{new}}(D) (1 - p_{\text{new}}(D)) \frac{dt}{dD} dD} = \\ &= \frac{\int_0^{D_{\max}} p_{\text{new}}(D) (1 - p_{\text{new}}(D))^l d(D)}{\int_0^{D_{\max}} (1 - p_{\text{new}}(D)) dD} \quad (13) \end{aligned}$$

Here  $D_{\max} = D(t_{\max})$ , and again we can perform a numerical integration. We can proceed in an analogous way to obtain the expression for  $P_{\text{nov}}(l)$ . In the Supplementary Information (S4), we show the agreement of the above expression with simulated results.

#### Scaling relation in the model and in real data

The results above show that the avalanche size distributions depend on the sequence length  $t_{\max}$ . We seek now a scaling relation to compare the avalanche size statistics at different sequence lengths. We observe that an interval between two successive novelties corresponds to an increase by one of the total number of distinct elements  $D(t)$ . We can thus write, setting the identity  $l \equiv \Delta t$ :

$$\frac{1}{l} \simeq \frac{\Delta D(t)}{\Delta t} \simeq \frac{dD(t)}{dt} = \left(1 + \frac{\rho}{N_0} t\right)^{\frac{\nu}{\rho}-1} \quad (14)$$

We thus make the ansatz:

$$\frac{1}{l_{\max}} \sim \left(1 + \frac{\rho}{N_0} t_{\max}\right)^{\frac{\nu}{\rho}-1} \sim t_{\max}^{\frac{\nu}{\rho}-1} \quad (15)$$

In Fig. 9(a) we show the inter-times distribution of sequences of different lengths from the UMT-E. In Fig. 9(b)

we show the same data as in Fig. 9(a) as a function of the rescaled variable  $l \times t_{\max}^{-\gamma}$ , by defining  $\gamma \equiv \frac{\nu}{\rho} - 1$ . We can see that the curves at different sequence lengths perfectly collapse on a unique curve. From the results in the above section, we can further observe that sequences generated by the UMT-E with the same Heaps' exponent, i.e., with the same ratio  $\frac{\nu}{\rho}$ , and different  $N_0$ , can be rescaled on a unique curve by means of the variable  $l \times z^{-\gamma}$ , with  $z \equiv \frac{\rho}{N_0} t_{\max}$ . In Fig. 9(c), we show the same data as in Fig. 8(b) as a function of the rescaled variable  $l \times z^{-\gamma}$ . We can see that the curves at different values of the ratio  $\frac{N_0}{\rho}$  perfectly collapse on a unique curve. We further show in Fig. 8 9(d) that the dependence on the ratio  $\frac{\nu}{\rho}$  is not eliminated by rescaling the avalanche's length, but remains weak.

In Fig. 10, we report the same scaling applied to all the datasets, showing that all the innovation systems considered obey the scaling relation for different sequence lengths. For each dataset, we truncated the whole sequence at various lengths. We set  $\frac{\nu}{\rho}$  to the value measured over the Heaps law of the entire sequence (we note that, for sufficiently long sequences, the fitted value  $\frac{\nu}{\rho}$  does not depend on the sequence length).

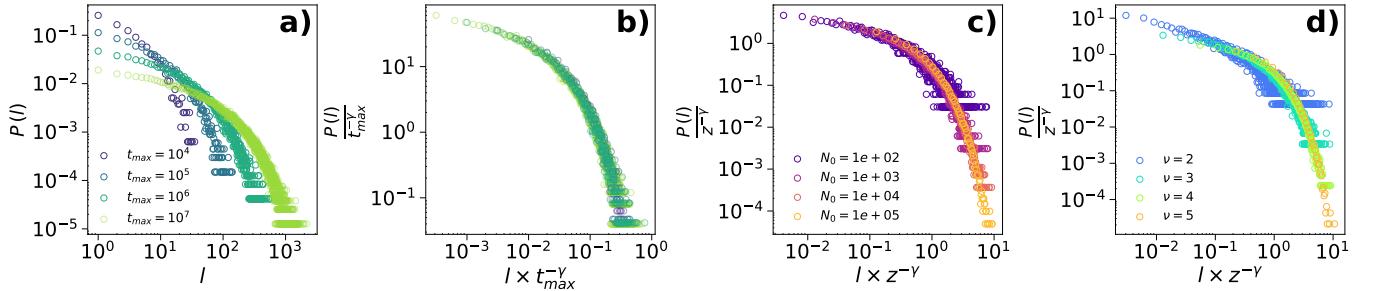
The seven considered datasets do not feature the same Heaps' exponents. However, all the exponents of the five datasets whose avalanche size statistics are well reproduced by a simple UMT model (LastFM, Github repositories, Twitter Hashtags, Github, and Twitter users) span a sufficiently small range, and the curves of the inter-time distributions approximately superimpose when plotted as a function of the variable  $l \times z^{-\gamma}$  (Fig. 11). As for the ratio  $\frac{\nu}{\rho}$ , the ratio  $\frac{N_0}{\rho}$  is estimated for each dataset by fitting the Heaps' law in Eq. (2), which gives the parameters reported in the captions of Figs. 2 and 3.

## METHODS

### Data

Our analysis takes into account seven different data sets.

(i-ii) **Github** is a repository hosting service that allows many software developers to work on the same project. Users can register for this online platform and create new projects or interact with existing ones (provided they are public or have access to them). The interaction can range from creating a new branch of the code to creating or modifying the Wiki page related to the project, and so on. In our analysis, we do not distinguish between the different possible interactions within the repositories, and we focus instead on the time sequence of such interactions, so that the elements of the temporal sequence are the repositories themselves. The data we used contains the



**FIG. 9. Scaling laws in the avalanche size distribution for non-novelties in the UMT-E.** **a)** Avalanche size distribution of non-novelties for an instance of the same UMT-E process cut at different lengths. We set the parameters' values to  $\nu = 3, \rho = 6, N_0 = 1000$ . **b)** The same distributions in **a)** rescaled so that they superimpose on a unique curve. **c)** The same distributions as in Fig. 8,b), ( $\nu = 3, \rho = 6$ , and  $N_0$  running across different orders of magnitude from  $10^2$  to  $10^5$ ) rescaled so that they superimpose on a unique curve. **d)** The same distributions as in Fig. 8,a), ( $N_0 = 1000, \rho = 6, \nu$  ranging in the integer values [2, 5]), rescaled as in **c)**.

activity of GitHub users in January 2015, with  $1.4 \times 10^7$  actions performed over  $1.5 \times 10^6$  repositories by  $8.8 \times 10^5$  users. In this work, we have considered both the time series of the actions taken over the repositories and the time series of the users performing those actions.

(ii-iv) **Twitter**, now referred to as **X**, is a social networking service. It is one of the world's largest social media websites and the fourteenth-most visited website. Users can share short text messages, images, and videos in posts (formerly "tweets") and like or repost/retweet other users' content. X includes direct messaging, video and audio calling, bookmarks, lists and communities, and Spaces, a social audio feature. The dataset we used was collected daily in January 2013, recording 10% of the users' activity. Tweets without "hashtags" have not been considered. The dataset thus consists of a sequence of time-ordered hashtags adopted by users, each identified by its own ID. Every hashtag is considered a separate element. We have considered the time series of hashtags within the same tweet in the order in which they appear. We restricted our analysis to the first week of data, and we considered both the time series of the hashtags and the time series of the users performing the corresponding tweets. In total, we have  $3.47 \times 10^7$  tweets containing  $6.9 \times 10^6$  distinct hashtags shared by  $1.28 \times 10^7$  different users.

(v) **LastFM** is a music website with a recommendation system. LastFM builds a detailed profile of each user's musical taste by recording details of the songs the user listens to, either from Internet radio stations, the user's computer, or many portable music devices. The data set we used contains the whole listening habits of 1000 users till May 5th, 2009, recorded in plain text form. It contains about  $1.9 \times 10^7$  listened tracks with information on user, time stamp, artist, track-id, and track name, resulting in  $1.08 \times 10^6$  different songs.

(vi) The **Gutenberg Project** is a Corpus of English literary texts derived from the Gutenberg Project e-book

collection till February 2007, resulting in a set of about 4600 non-copyrighted e-books from which we have selected the 1000 longest, of various topics and including both prose and poetry. We preprocessed the Corpus in our analysis, adopting the following choices: we ignored capitalization. We took as distinct words sharing the same lexical root, i.e., we considered the word *flower* different from *flowers*. We treated homonyms as the same word, for example, the verbal past perfect saw and the substantive saw. We removed punctuation marks. The sequence obtained by stacking all 1000 books together comprises  $1.58 \times 10^8$  words with a total vocabulary of  $3.86 \times 10^5$  words. We stacked books in a random order, since information about each text's creation time is not available.

(vii) The **English Wikipedia** dataset we analyzed consists of the texts of 40000 different Wiki pages, randomly selected from the whole Wikipedia Corpus. We perform the same preprocessing as in the Gutenberg Project corpus. The sequence is obtained by stacking all the pages in a unique text of  $2.90 \times 10^7$  words, where the number of distinct words is  $6.73 \times 10^5$ .

### The mapping

To test the hypothesis that the avalanche size statistics of the Gutenberg and the Wikipedia corpora reflect a *convolution* between different semantic contexts, we reproduce this setting within our modeling scheme. We do that most simply, not aiming at giving a full modeling framework, but to show that accounting for the superposition of semantic contexts is crucial to understanding the dynamics observed in those datasets. We then proceeded as follows:

1. For each Corpus, we fit the Heaps' law of each textual sequence, say  $S_C^i$  with vocabulary size  $D_C^i$  (each book for the Gutenberg corpus and each

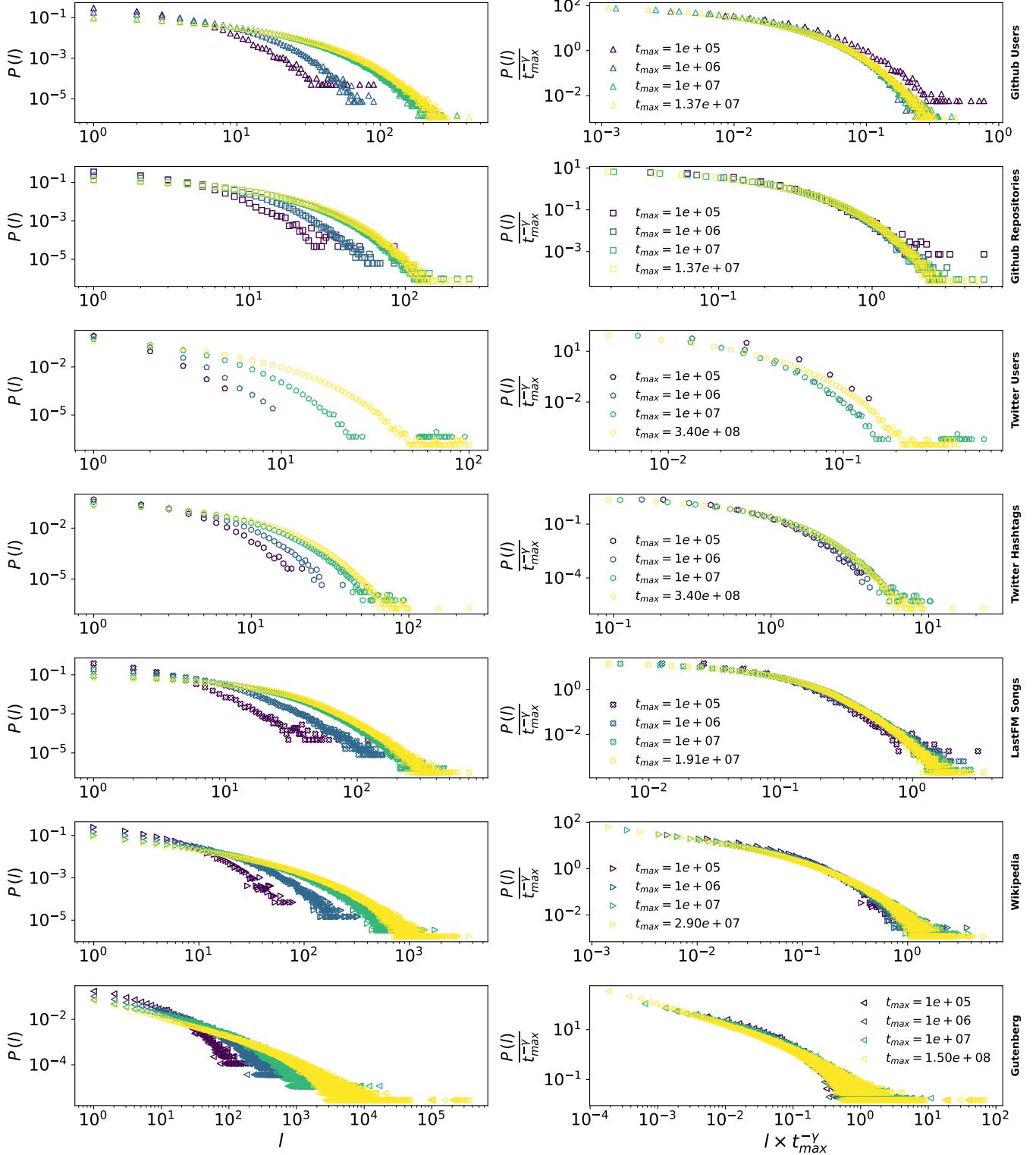
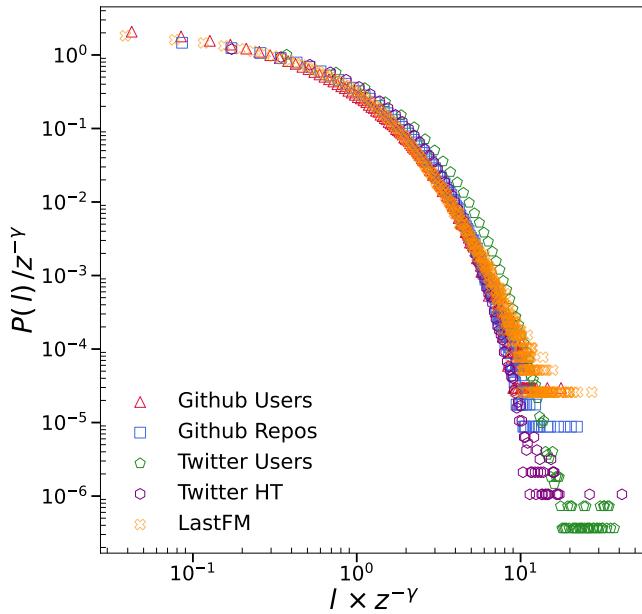


FIG. 10. **Scaling laws in the non-novelties avalanche size distributions for different sequences' lengths of each real-world dataset.** We consider the sequence of each dataset truncated at different lengths  $t_{\max}$ . Left: Avalanche size distributions of non-novelties in each dataset for sequences of different lengths. Right: same distributions as on the right, where we adopted the scaling law derived for the UMT-E. The  $\gamma$  exponent is fitted from the longest sequence in each dataset, as described in the main text.



**FIG. 11. Scaling laws in the non-novelties avalanche size distributions in the real-world datasets.** We here rescale the length  $l$  of the non-novelties avalanches according to the variable  $z^{-\gamma}$ . This rescaling allows the curves for different systems to almost collapse into one unique shape. We fit the  $\gamma$  exponent and the ratio  $N_0/\rho$  from the longest sequence in each dataset.

Wikipedia page for the Wikipedia corpus), with the analytical formula predicted for the UMT-E (eq. 2) and estimate the ratios  $\frac{\nu}{\rho} \equiv \alpha$  and  $\frac{N_0}{\rho} \equiv \theta$ . From the parameters  $\theta$  and  $\alpha$ , we obtained  $\nu$  and  $\rho$  as the two integers whose ratio  $\frac{\nu}{\rho}$  is the closest to  $\alpha$ , in the ranges  $\nu \in [1, 99]$  and  $\rho \in [\nu + 1, 100]$ . Once we fixed the value of  $\rho$ , we set  $N_0$  as the closest integer to  $\theta \times \rho$ .

2. We generate a sequence from UMT-E, say  $S_{UMT-E}^i$ , with the estimated parameters and of the same length of  $S_C^i$ , resulting in a vocabulary size  $D_{UMT-E}^i$ .
3. If  $D_{UMT-E}^i > D_R^i$ , we cut  $S_{UMT-E}^i$  at the time  $t$  such that  $D_{UMT-E}^i(t+1) = D_C^i + 1$ , to have  $D_{UMT-E}^i = D_C^i$ .
4. We sort each word in  $S_C^i$  and each element in  $S_{UMT-E}^i$  according to their rank. We replace each distinct element of  $S_{UMT-E}^i$  with a word in  $S_C^i$  according to their frequencies: we replace the most frequent element of  $S_{UMT-E}^i$  with the most frequent word in  $S_C^i$ , the second most frequent element in  $S_{UMT-E}^i$  with the second most frequent word in  $S_C^i$  and so on.
5. Finally, for each Corpus, we stack together all the

sequences  $S_{UMT-E}^i$  in the same order as the corresponding sequences  $S_C^i$ .

### The UMST model

The UMST model introduced in [9] is a generalization of the UMT, where the concept of semantic groups is added. The dynamics is such to favor the stay within a semantic group, though a fourth parameter  $\eta$  belonging to the real interval  $[0, 1]$ . The process works as follows: at  $t = 0$  the urn contains  $N_0$  different elements divided in  $\frac{N_0}{\nu+1}$  different semantic groups. We randomly extract the first element. At each time step  $t$  an element  $s_t$  is drawn from the urn according to a probability proportional to its occurrences inside the urn if it belongs to the same semantic group of  $s_{t-1}$ ; it is instead drawn with a probability proportional to its occurrences inside the urn times  $\eta$  if it belongs to a semantic group different from  $s_{t-1}$ . The reinforcement and triggering rules are analogous to the ones adopted in the UMT model. If  $s_t$  is a novelty, it is reinforced with  $\rho$  copies inside the urn, and further enlarges the adjacent possible by introducing  $\nu + 1$  new distinct elements inside the urn. These  $\nu + 1$  new elements share a new semantic group. Only the reinforcement process occurs if  $s_t$  is a non-novelty. In the  $\nu < \rho$  case, in analogy with the UMT, if the last extracted element is a novelty, we can choose the reinforcement parameter to be  $\tilde{\rho}$ . Note, however, that for  $\tilde{\rho} = \rho - (\nu + 1)$  the UMST model is still non exchangeable.

### CONCLUSIONS

Over the past years, several studies have advanced the identification of significant statistical signatures of innovation processes [9, 10, 17]. This work introduces the avalanche statistics as key observables for characterizing their dynamics, complementing the well-established Zipf's, Heaps', and Taylor's laws [12–14].

Within the framework of the urn model with triggering (UMT) [9], we derive exact expressions for the avalanche size distributions of novelties and non-novelties, which reproduce empirical observations across diverse innovation-related datasets, including collaborative platforms, social media, cultural production, and natural language corpora. At the same time, we extend the existing characterization of Heaps' law within the UMT framework [9, 18], emphasizing the role of the initial size of the adjacent possible [60]. We establish an equivalence between the UMT, its exchangeable version (UMT-E), and the urn model with semantic triggering (UMST) [9], i.e., when tuned to generate the same Heaps law, these models produce identical avalanche statistics. This result sets a direct connection between Heaps' law and avalanche dynamics. However, the avalanche size statistics, espe-

cially for non-novelties, contain more information than Heaps' law alone. Empirical analyses also reveal deviations from model predictions, suggesting that avalanche size statistics can distinguish between dynamics that can be effectively described as a single collective process (and thus obey the UMT predictions) and those that emerge from the superposition of individual contributions, uncovering dynamical features that go beyond those described by the Heaps law. Indeed, both the Wikipedia and Gutenberg corpora exhibit heavy-tailed interevent time distributions, consistent with the aggregation of individual contributions. However, only Gutenberg shows a clear power-law regime for small avalanche sizes, reflecting the higher internal complexity of books, where multiple topics and dynamics coexist. The similarity between the avalanche statistics of Wikipedia and those of individual books within the Gutenberg corpus supports this interpretation.

From the UMT-E analytical results, we provide a scaling relation that allows us to superimpose avalanche distributions from the same system but corresponding to different sequence lengths. All the datasets considered obey this, revealing a form of universality in the dynamics of innovation. Moreover, we derive a scaling relation that extends the collapse on a unique curve to nearly all considered datasets, except for the two natural language corpora, which deviate from the model predictions. This result highlights the roles of the initial repertoire size and the rate of the adjacent possible expansion for each system.

Overall, the avalanche framework provides insights that deepen our understanding of the mechanisms driving innovation. We believe it offers a foundation for future theoretical and empirical studies across innovation systems.

## Data Availability

The data used in this manuscript are available at [https://figshare.com/articles/dataset/datasets\\_avalanches/30186733](https://figshare.com/articles/dataset/datasets_avalanches/30186733) and have been obtained and used according to their terms and conditions.

## Code Availability

All the codes used to process and analyze the data and numerically realize the models can be found at [61].

- [2] F. Saracco, R. Di Clemente, A. Gabrielli, and L. Pietronero, From innovation to diversification: a simple competitive model, *PloS one* **10**, e0140420 (2015).
- [3] M. M. Dankulov, R. Melnik, and B. Tadić, The dynamics of meaningful social interactions and the emergence of collective knowledge, *Scientific Reports* **5**, 12197 (2015).
- [4] I. Iacopini, S. Milojević, and V. Latora, Network dynamics of innovation processes, *Physical Review Letters* **120**, 048301 (2018).
- [5] I. Iacopini, G. Di Bona, E. Ubaldi, V. Loreto, and V. Latora, Interacting discovery processes on complex networks, *Phys. Rev. Lett.* **125**, 248301 (2020).
- [6] M. Gerlach and E. G. Altmann, Stochastic model for the vocabulary growth in natural languages, *Physical Review X* **3**, 021006 (2013).
- [7] J. Sheth, Impact of covid-19 on consumer behavior: Will the old habits return or die?, *Journal of Business Research* **117**, 280 (2020).
- [8] J. Han, J. Yin, X. Wu, D. Wang, and C. Li, Environment and covid-19 incidence: A critical review, *Journal of Environmental Sciences* **124**, 933 (2023).
- [9] F. Tria, V. Loreto, V. D. P. Servedio, and S. H. Strogatz, The dynamics of correlated novelties, *Scientific reports* **4**, 5890 (2014).
- [10] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto, Vocabulary growth in collaborative tagging systems (2007), [arXiv:0704.3316 \[cs.IR\]](https://arxiv.org/abs/0704.3316).
- [11] E. Ubaldi, R. Burioni, V. Loreto, and F. Tria, Emergence and evolution of social networks through exploration of the adjacent possible space, *Communications Physics* **4**, 28 (2021).
- [12] H. S. Heaps, *Information retrieval: Computational and theoretical aspects* (Academic Press, Inc., 1978).
- [13] G. K. Zipf, *The Psychobiology of Language* (Houghton-Mifflin, New York, NY, USA, 1935).
- [14] L. R. Taylor, Aggregation, variance and the mean, *Nature* **189**, 732 (1961).
- [15] L. Lü, Z.-K. Zhang, and T. Zhou, Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems, *PloS one* **5**, e14139 (2010).
- [16] V. Loreto, V. D. P. Servedio, S. H. Strogatz, and F. Tria, Dynamics on expanding spaces: Modeling the emergence of novelties, in *Creativity and Universality in Language*, edited by M. Degli Esposti, E. G. Altmann, and F. Pachet (Springer International Publishing, Cham, 2016) pp. 59–83.
- [17] M. Gerlach and E. G. Altmann, Scaling laws and fluctuations in the statistics of word frequencies, *New Journal of Physics* **16**, 113010 (2014).
- [18] F. Tria, V. Loreto, and V. D. Servedio, Zipf's, Heaps' and Taylor's Laws are Determined by the Expansion into the Adjacent Possible, *Entropy* **20**, 752 (2018).
- [19] F. Tria, I. Crimaldi, G. Aletti, and V. D. Servedio, Taylor's law in innovation processes, *Entropy* **22**, 573 (2020).
- [20] I. Main, Statistical physics, seismogenesis, and seismic hazard, *Reviews of Geophysics* , 433 (1996).
- [21] A. Bizzarri, S. Das, and A. Petri, eds., *Proceedings of the International School of Physics “Enrico Fermi”, Course 202 “Mechanics of Earthquake Faulting”*, Vol. 202 (IOS Press, 2019).
- [22] A. Bizzarri, A. Petri, and A. Baldassarri, Earthquake dynamics constrained from laboratory experiments: new insights from granular materials, *Annals of Geophysics* **64**, SE44 (2021).

\* Corresponding Author: francesca.tria@uniroma1.it

[1] S. Thurner, P. Klimek, and R. Hanel, Schumpeterian economic dynamics as a quantifiable model of evolution, *New Journal of Physics* **12**, 075029 (2010).

- [23] S. Zapperi, P. Cizeau, G. Durin, and H. E. Stanley, Dynamics of a ferromagnetic domain wall: Avalanches, depinning transition, and the Barkhausen effect, *Physical Review B* **58**, 6353 (1998).
- [24] G. Cannelli, R. Cantelli, and F. Cordero, Self-organized criticality of the fracture processes associated with hydrogen precipitation in niobium by acoustic emission, *Phys. Rev. Lett.* **70**, 3923 (1993).
- [25] A. Petri, G. Paparo, A. Vespignani, A. Alippi, and M. Costantini, Experimental evidence for critical dynamics in microfracturing processes, *Phys. Rev. Lett.* **73**, 3423 (1994).
- [26] G. Caldarelli, F. D. Tolla, and A. Petri, Self-organization and annealed disorder in fracturing processes, *Phys. Rev. Lett.* **77**, 2503 (1996).
- [27] D. M. Dimiduk, C. Woodward, R. LeSar, and M. D. Uchic, Scale-free intermittent flow in crystal plasticity, *Science* **312**, 1188 (2006).
- [28] A. Baldassarri, F. Dalton, A. Petri, S. Zapperi, G. Pontuale, and L. Pietronero, Brownian forces in sheared granular matter, *Phys. Rev. Lett.* **96**, 118002 (2006).
- [29] A. Petri, A. Baldassarri, F. Dalton, G. Pontuale, L. Pietronero, and S. Zapperi, Stochastic dynamics of a sheared granular medium, *The European Physical Journal B - Condensed Matter and Complex Systems* **64**, 531 (2008).
- [30] M. Pica Ciamarra, F. Dalton, L. de Arcangelis, C. Godano, E. Lippiello, and A. Petri, The role of interstitial impurities in the frictional instability of seismic fault models, *Tribology Letters* **48**, 89 (2012).
- [31] A. Baldassarri, M. Annunziata, A. Gnoli, G. Pontuale, and A. Petri, Breakdown of scaling and friction weakening in intermittent granular flow, *Scientific reports* **9**, 16962 (2019).
- [32] A. Petri, Statistics of intermittent granular flow from confined tabletop experiments, *Rivista del Nuovo Cimento* **47**, 353 (2024).
- [33] J. Antonaglia, W. J. Wright, X. Gu, R. R. Byer, T. C. Hufnagel, M. LeBlanc, J. T. Uhl, and K. A. Dahmen, Bulk metallic glasses deform via slip avalanches, *Phys. Rev. Lett.* **112**, 155501 (2014).
- [34] L. De Arcangelis, C. Perrone-Capano, and H. J. Herrmann, Self-organized criticality model for brain plasticity, *Physical review letters* **96**, 028107 (2006).
- [35] F. Lombardi, H. J. Herrmann, C. Perrone-Capano, D. Plenz, and L. de Arcangelis, Balance between excitation and inhibition controls the temporal organization of neuronal avalanches, *Phys. Rev. Lett.* **108**, 228703 (2012).
- [36] M. K. Nandi, A. Sarracino, H. J. Herrmann, and L. de Arcangelis, Scaling of avalanche shape and activity power spectrum in neuronal networks, *Phys. Rev. E* **106**, 024304 (2022).
- [37] S. Scarpetta, N. Morisi, C. Mutti, N. Azzi, I. Trippi, R. Ciliento, I. Apicella, G. Messuti, M. Angioletti, F. Lombardi, *et al.*, Criticality of neuronal avalanches in human sleep and their relationship with sleep macro-and micro-architecture, *Iscience* **26** (2023).
- [38] D. Chialvo, A. G. Torrado, E. Gudowska-Nowak, J. Ochab, P. Montoya, M. Nowak, and E. Tagliazucchi, How we move is universal: Scaling in the average shape of human activity, *Papers in Physics* **7** (2015).
- [39] R. Gallotti and D. R. Chialvo, How ants move: individual and collective scaling properties, *Journal of The Royal Society Interface* **15** (2018).
- [40] J. P. Sethna, K. A. Dahmen, and C. R. Myers, Crackling noise, *Nature* **410**, 242 (2001).
- [41] L. De Arcangelis and H. J. Herrmann, Activity-dependent neuronal model on complex networks, *Frontiers in Physiology* **3**, 62 (2012).
- [42] S. Scarpetta and A. de Candia, Neural avalanches at the critical point between replay and non-replay of spatiotemporal patterns, *PLoS One* **8**, e64162 (2013).
- [43] G. Pontuale, F. Colaiori, and A. Petri, Slow crack propagation through a disordered medium : Critical transition and dissipation, *Europhysics Letters* **101** (2013).
- [44] A. Petri and G. Pontuale, Morphology and dynamics in soc universality classes, *Journal of Statistical Mechanics: Theory and Experiment* **2018**, 063201 (2018).
- [45] S. Zapperi, *Crackling Noise: Statistical Physics of Avalanche Phenomena* (Oxford University Press, 2022).
- [46] L. Laurson, X. Illa, S. Santucci, K. Tore Tallakstad, K. J. Måloy, and M. J. Alava, Evolution of the average avalanche shape with the universality class, *Nat. Commun.* **4** (2013).
- [47] L. de Arcangelis, C. Godano, E. Lippiello, and M. Nicodemi, Universality in solar flare and earthquake occurrence, *Physical review letters* **96**, 051102 (2006).
- [48] L. de Arcangelis, C. Godano, J. R. Grasso, and E. Lippiello, Statistical physics approach to earthquake occurrence and forecasting, *Physics Reports* **628**, 1 (2016).
- [49] P. Kumar, E. Korkolis, R. Benzi, D. Denisov, A. Niemeijer, P. Schall, F. Toschi, and J. Trampert, On interevent time distributions of avalanche dynamics, *Scientific reports* **10**, 626 (2020).
- [50] G. Di Bona, A. Bellina, G. De Marzo, A. Petralia, I. Iacopini, and V. Latora, The dynamics of higher-order novelties, *Nature Communications* **16**, 393 (2025).
- [51] B. de Finetti, La prévision: ses lois logiques, ses sources subjectives, *Annales de l'Institut Henri Poincaré* **7**, 1 (1937).
- [52] J. F. C. Kingman, Random partitions in population genetics, *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* **361**, 1 (1978).
- [53] J. Pitman, The two-parameter poisson-dirichlet distribution derived from a stable subordinator, *Annals of Probability* **25**, 855 (1997).
- [54] G. Tani Raffaelli, M. Lalli, and F. Tria, Inference through innovation processes tested in the authorship attribution task, *Communications Physics* **7**, 300 (2024).
- [55] B. Monechi, A. Ruiz-Serrano, F. Tria, and V. Loreto, Waves of novelties in the expansion into the adjacent possible, *PloS one* **12**, e0179303 (2017).
- [56] S. A. Kauffman, *Investigations* (Oxford University Press, New York/Oxford, 2000).
- [57] J. Pitman *et al.*, *Combinatorial stochastic processes*, Tech. Rep. (Technical Report 621, Department of Statistics, University of California, 1996).
- [58] P. Zhang, C. Chen, and H. Mahmoud, Explicit characterization of moments of balanced triangular Pólya urns by an elementary approach, *Statistics & Probability Letters* **96**, 149 (2015).
- [59] We should stop the integration at time  $t = t_{\max} - l$ , but neglecting this detail corresponds to the same order of approximation as that made in Eq. (7) and does not affect the results.
- [60] S. A. Kauffman, *Investigations: The nature of autonomous agents and the worlds they mutually create*

(Santa Fe Institute, 1996).  
[61] <https://github.com/filidibava/avalanches> (2025).

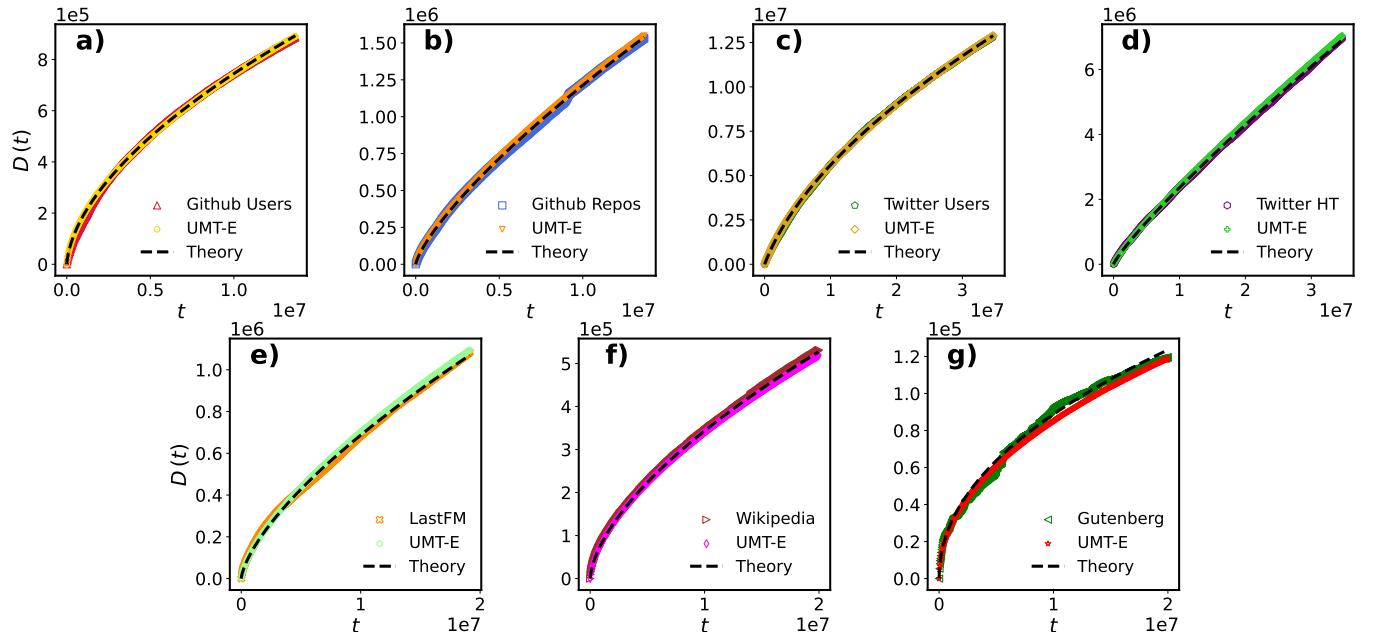
**SUPPLEMENTARY INFORMATION FOR "UNIVERSALITIES IN THE AVALANCHE DYNAMICS OF NOVELTIES AND NON-NOVELTIES"**

**S1. HEAPS' LAW FIT FOR REAL DATASETS**

We predict the avalanche size distribution of novelties and non-novelties in each real-world system from the UMT-E with the parameters  $(\nu, \rho, N_0)$  that best fit the Heaps law of each considered system. With these parameters, we derived the theoretical analytical expression for the avalanche distributions, which we call 'theory' in all the figures, and produced numerical realizations of the corresponding process from which we computed the empirical avalanche distributions (what we call 'UMT-E' in all the figures). For each dataset, we have fitted its Heaps law with the formula:

$$D(t) = \frac{\theta}{\alpha} \left(1 + \frac{t}{\theta}\right)^{\alpha} - \frac{\theta}{\alpha}$$

where  $\theta = \frac{N_0}{\rho}$  and  $\alpha = \frac{\nu}{\rho}$ . We have then chosen  $\nu$  and  $\rho$  as the integers whose ratio is the closest to  $\alpha$ . This procedure can lead to an arbitrary precision, and we have chosen  $\nu \in [1, 99]$  and  $\rho \in [\nu + 1, 100]$ . Once  $\nu$  and  $\rho$  are obtained, we have estimated  $N_0$  as the integer closest to the quantity  $\theta \times \rho$ . In Fig. 1 we show the Heaps law of all the datasets, the theoretical UMT-E Heaps law, and the Heaps law of the realization produced according to the estimated parameters. For the sake of completeness, we report all the parameters in Table I.



**FIG. 1. The Heaps law in real datasets.** Comparison between the Heaps Law of real datasets and the analytical UMT-E Heaps law that best fits each dataset, along with the corresponding numerical realization of the process.

TABLE I. Datasets parameters

Dataset	$\nu$	$\rho$	$N_0$	$t_{max}$
Github Users	4	7	45,468	$1.37 \times 10^7$
Github Repositories	22	29	14,640	$1.37 \times 10^7$
Twitter Users	16	29	42,183,043	$3.4 \times 10^8$
Twitter Hashtags	19	22	1,995	$3.4 \times 10^8$
LastFM	28	41	27,200	$1.91 \times 10^8$
Wikipedia Corpus*	31	50	25,000	$2.91 \times 10^7$
Gutenberg Corpus*	9	19	5,890	$1.58 \times 10^8$

\*For computational reasons, UMT-E sequences were limited to  $2.0 \times 10^7$  elements for these parameter sets.

## S2. AVALANCHES OF NOVELTIES DISTRIBUTIONS FOR THE UMT-E

In Fig. 2 we show the analysis on the avalanches of novelties of the UMT-E model. By fixing ( $\rho = 6, N_0 = 1000$ ) we see that increasing  $\nu$  widens the support of the distribution, since larger avalanches become more probable to occur. In this sense,  $\nu$  has the same role as  $\rho$  in the avalanches of non-novelty distributions: increasing  $\nu$  increases the probability of extracting a novel element, allowing also larger avalanches of novelties to form. Increasing  $N_0$  also widens the support of the distribution. However, we remark that even though both increasing  $N_0$  and  $\nu$  favor innovations, their tuning reflects two different mechanisms: while increasing  $\nu$  directly expands the adjacent possible, increasing  $N_0$  sets the initial condition of the space of possibilities.

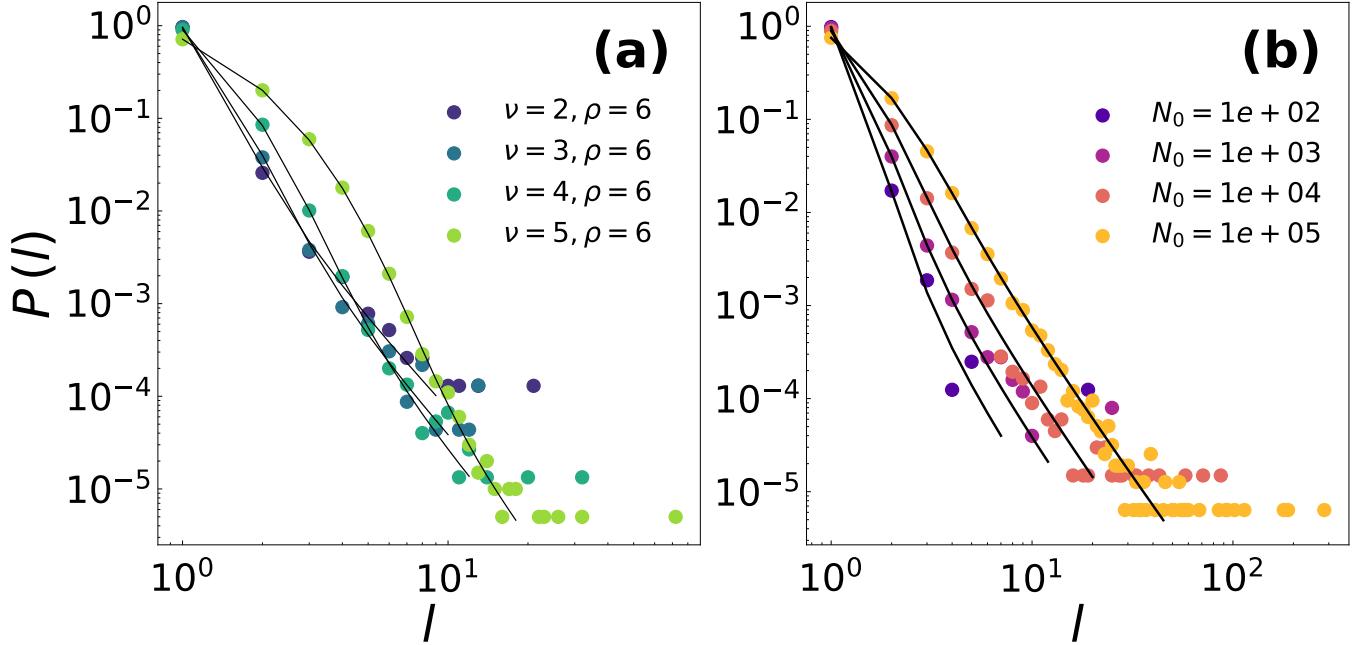


FIG. 2. **Avalanche size distributions for Novelties in the UMT-E.** a) Realizations of length  $10^6$  along with the theoretical predictions with  $\rho = 6$  and  $\nu$  in the integer range  $[2, 5]$ .  $N_0 = 10^3$  for all the sequences. b) Realizations of length  $10^6$  along with the theoretical predictions for  $(\nu = 3, \rho = 6)$ .  $N_0$  spans across the orders of magnitudes in the range  $[10^2, 10^5]$ .

### S3. WRITTEN TEXTS AND WIKIPEDIA PAGES ANALYSIS

In this section, we present supplementary analyses performed on individual written texts and Wikipedia pages to compare and understand how the analytical results obtained for the UMT-E model reproduce the Heaps law and the avalanches of novelties and non-novelties distributions observed in real data.

#### Written Texts Analysis

Here we report the analysis made on three written texts, randomly chosen among the 100 longest in the Gutenberg corpus, and their comparison with the UMT-E predictions. In Fig. 3, we compare the experimental Heaps and the UMT-E Heaps that best fit it. In Fig. 4 we show the comparison between the experimental intermission times distributions and their corresponding UMT-E predictions with the values of  $(\nu, \rho, N_0)$  derived from the fits of the Heaps law. In all the cases we have analyzed, the theoretical distribution predicts lower probability values for small inter-event times and slightly less fat tails than those observed in the written texts. In Fig. 6 we show the comparison between the predictions for the avalanche size distributions for novelties and the experimental observations with the same values of  $(\nu, \rho, N_0)$  mentioned above: in all the considered cases we find a good agreement between theoretical predictions and the experimental observations. In Table II we resume the estimated parameters.

TABLE II. Parameters for literary works

	<b>Work</b>	$\nu$	$\rho$	$N_0$	<b>Length</b>
a)	Narrative and Legendary poems by John Greenleaf Whittier	50	97	22,911	588765
b)	Literary Friends and Acquaintance by William Dean Howells	30	85	50,553	1485384
c)	The Translation of a Savage by Gilbert Parker	35	82	29,258	2234347

In Fig. 5 we show that it is always possible to find a triple  $(\nu, \rho, N_0)$ , different from the one which fits the Heaps law, that satisfactorily reproduces the intermission times distribution. In Table III we resume these parameters.

TABLE III. Parameters for literary works fitted on the delays

	<b>Work</b>	$\nu$	$\rho$	$N_0$	<b>Length</b>
a)	Narrative and Legendary poems by John Greenleaf Whittier	23	73	127020	588765
b)	Literary Friends and Acquaintance by William Dean Howells	33	100	135000	1485384
c)	The Translation of a Savage by Gilbert Parker	29	98	151900	2234347

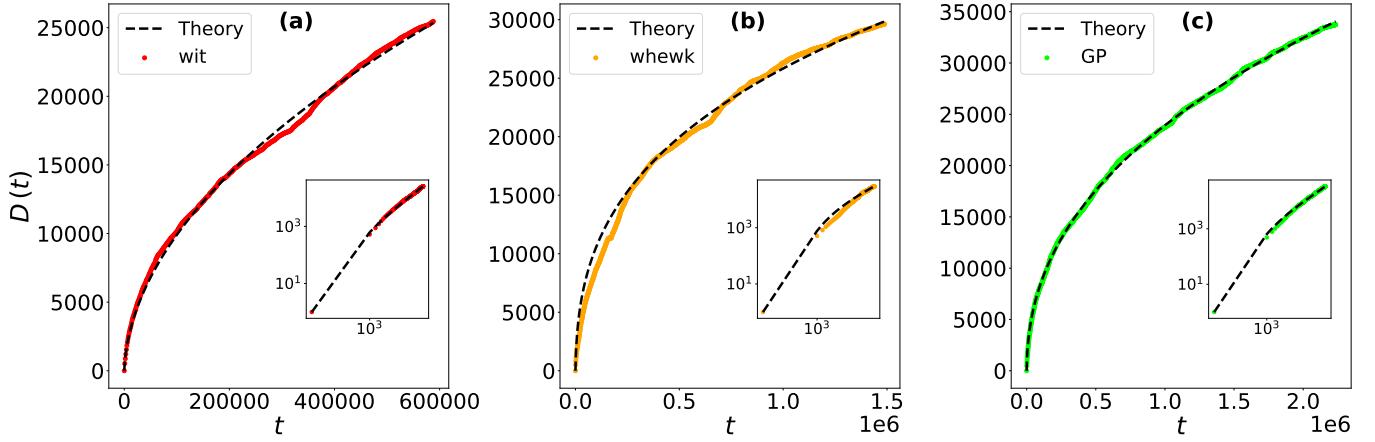


FIG. 3. **Heaps Law of the UMT-E that best fits each written text.** **a)** Narrative and Legendary Poems by John Greenleaf Whittier (red plot). **b)** Literary Friends and Acquaintances by William Dean Howells (orange plot). **c)** The Translation of a Savage of Gilbert Parker (lime plot).

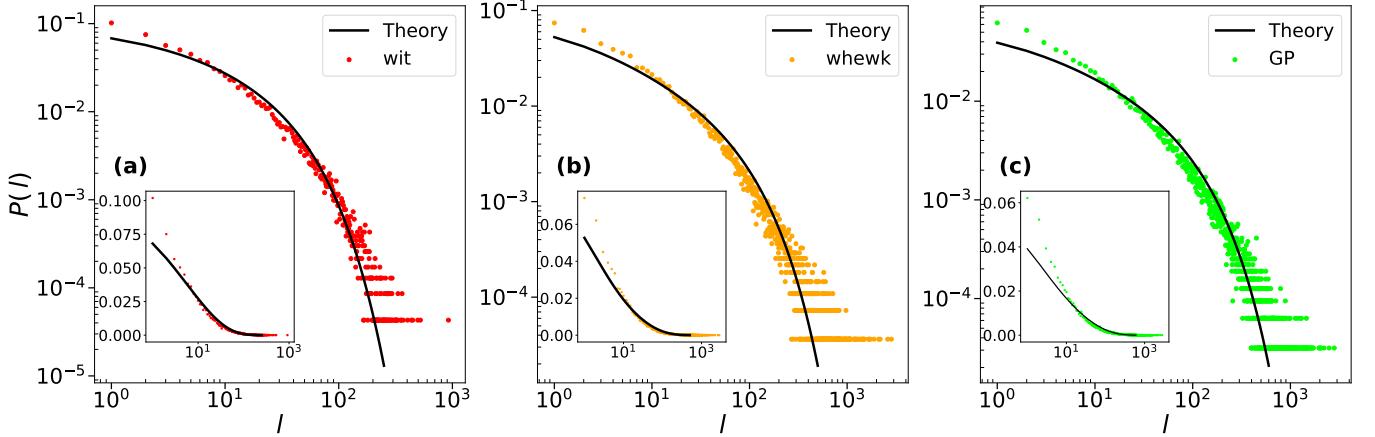


FIG. 4. **Avalanche size distributions for non-novelties in Written Texts.** Theoretical curves (continuous black lines) are obtained by fitting the UMT-E model Heaps law to the Heaps law of each written text. Scattered colored plots represent the avalanche size distributions of non-novelties for each considered written text. **a)** Narrative and Legendary poems by John Greenleaf Whittier (red scattered plot). **b)** Literary Friends and Acquaintances by William Dean Howells (orange scattered plot). **c)** The Translation of a Savage of Gilbert Parker (lime scattered plot).

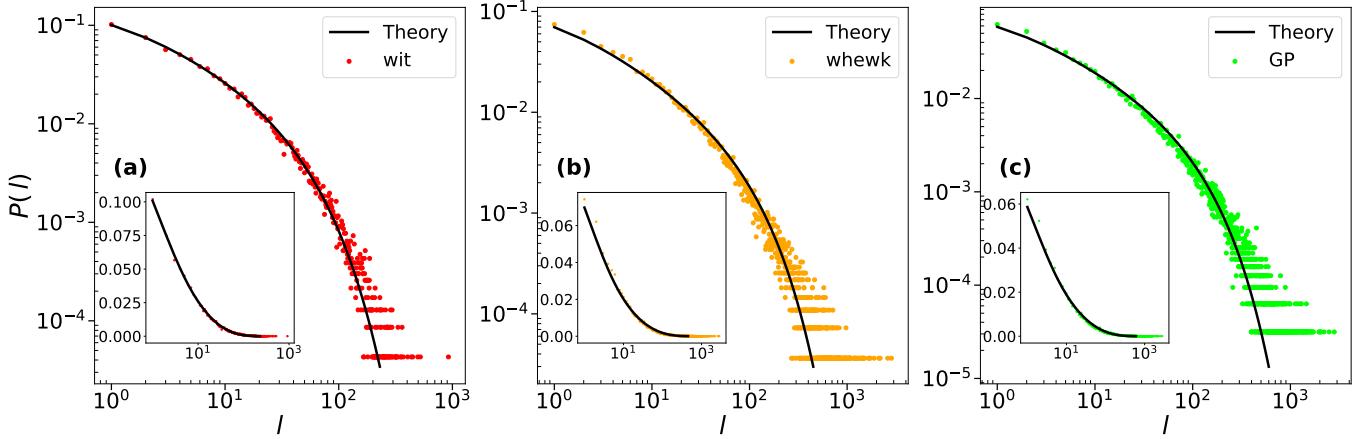


FIG. 5. **Avalanche size distributions for non-novelties in Written Texts and corresponding theoretical predictions of the UMT-E model fitted on them.** Theoretical curves (continuous black lines) are obtained by fitting the theoretical UMT-E model avalanches of non-novelties size distribution directly on the experimental avalanches size distributions for non-novelties (scattered colored plots). **a)** Narrative and Legendary poems by John Greenleaf Whittier (red scattered plot). **b)** Literary Friends and Acquaintances by William Dean Howells (orange scattered plot). **c)** The Translation of a Savage of Gilbert Parker (lime scattered plot).

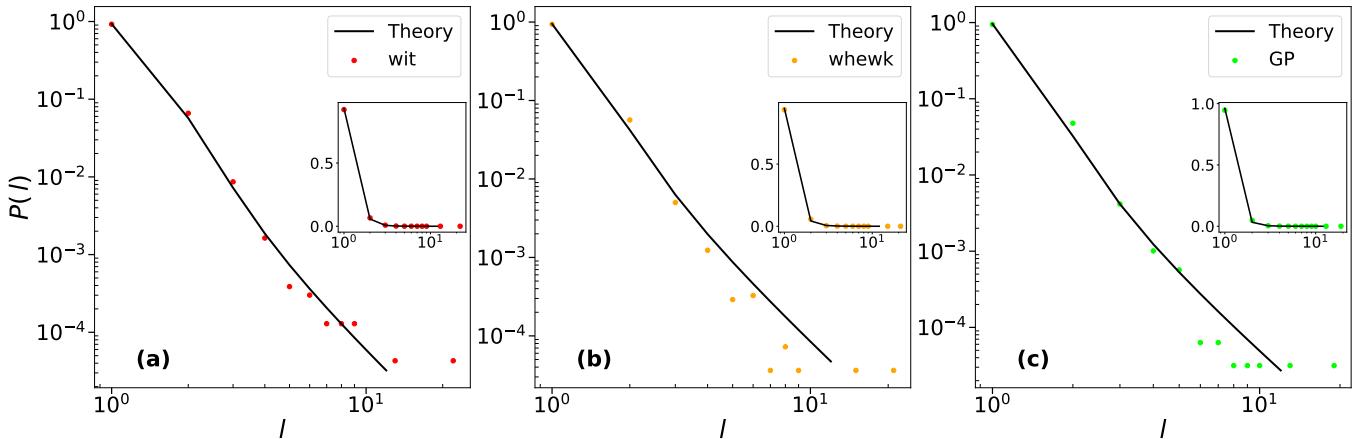


FIG. 6. **Avalanches size distributions for novelties in Written Texts.** Theoretical curves (continuous black lines) are obtained by fitting the UMT-E model Heaps law to the Heaps law of each written text. Scattered colored plots represent the avalanche size distributions of Novelties for each considered written text. **a)** Narrative and Legendary Poems by John Greenleaf Whittier (red scattered plot). **b)** Literary Friends and Acquaintances by William Dean Howells (orange scattered plot). **c)** The Translation of a Savage of Gilbert Parker (lime scattered plot).

### Wikipedia Pages

In this section, we follow the same methodology adopted for the written texts: we have randomly considered three pages among the 100 longest Wikipedia pages, and for each page, we compare the experimental Heaps law with the UMT-E Heaps law that best fits it in 7. In Fig. 8 we show the comparison between the experimental intermission times distributions and their corresponding UMT-E predictions with the values of  $(\nu, \rho, N_0)$  derived from the fits of the Heaps law. In all the cases we have analyzed, the theoretical predictions from UMT-E agree with the observed empirical behavior. In Fig. 9 we show the corresponding avalanche size distributions for novelties. Table IV resumes the estimated parameters.

TABLE IV. Parameters for Wikipedia Pages

	<b>Page</b>	$\nu$	$\rho$	$N_0$	<b>Length</b>
a)	text 31939	27	50	11,716	18150
b)	text 13859	56	75	5283	14160
c)	text 7372	51	86	36062	12068

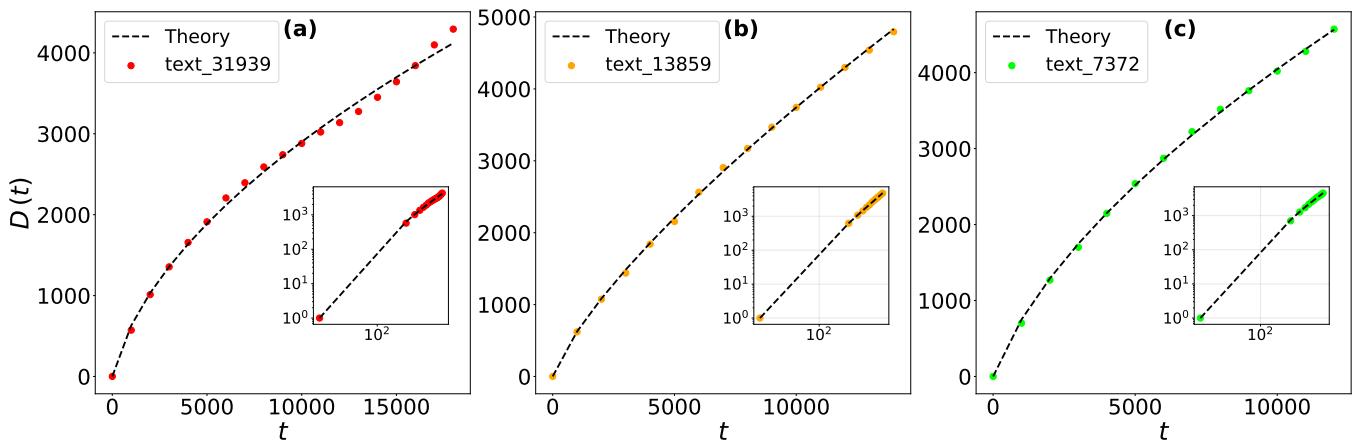


FIG. 7. **Heaps law of the UMT-E that best fits each Wikipedia page.** a) text 31939 (red scattered plot). b) text 13859 (orange scattered plot). c) text 7372 (lime scattered plot).

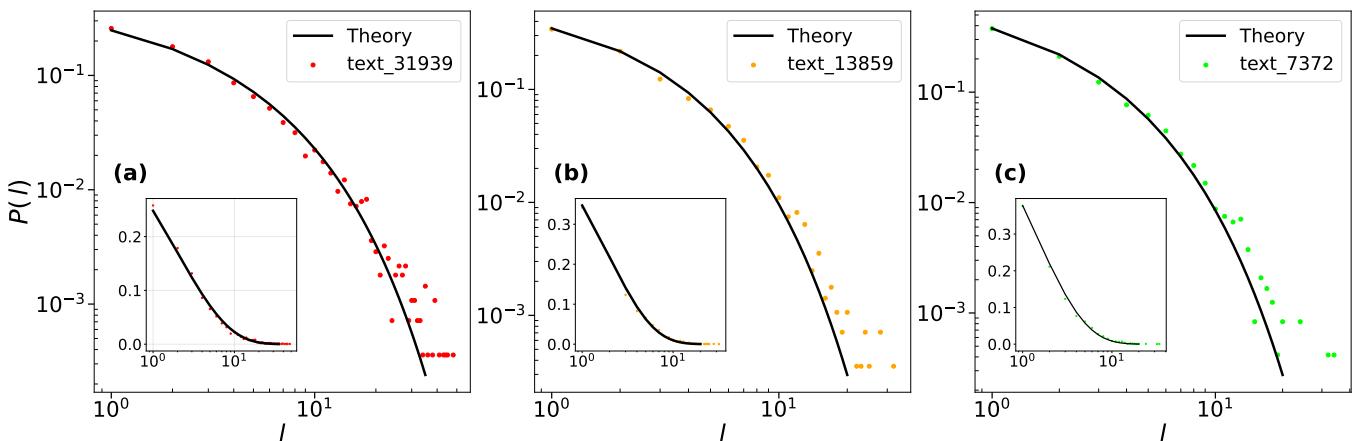


FIG. 8. **Avalanche size distributions for non-novelties in Wikipedia Pages.** Theoretical curves (continuous black lines) are obtained by fitting the UMT-E model Heaps law to the Heaps law of each written text. Scattered colored plots represent the avalanche size distributions of non-novelties for each considered Wikipedia page. a) text 31939 (red scattered plot). b) text 13859 (orange scattered plot). c) text 7372 (lime scattered plot).

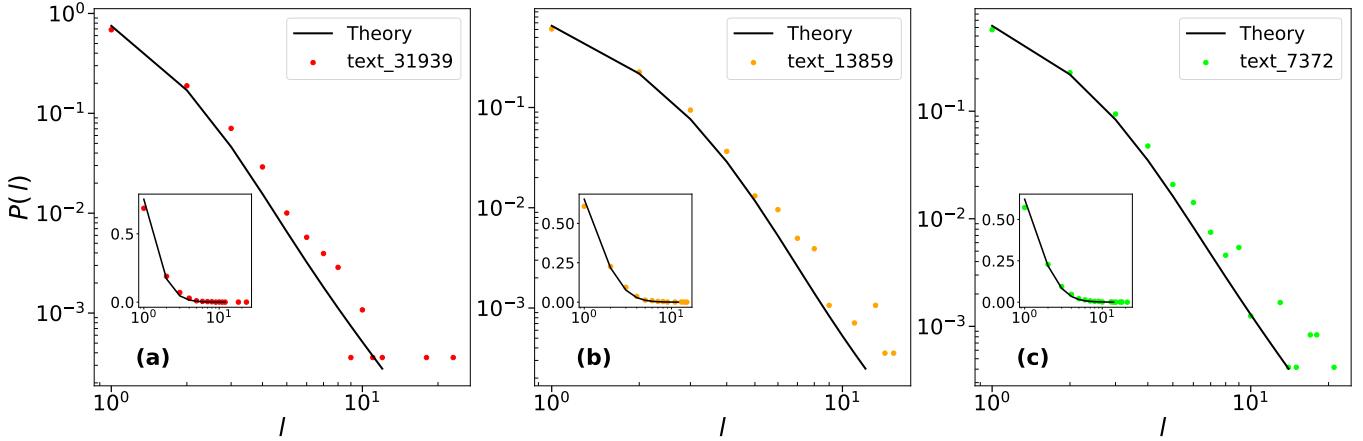


FIG. 9. **Avalanche size distributions for novelties in Wikipedia Pages.** Theoretical curves (continuous black lines) are obtained by fitting the UMT-E model Heaps law to the Heaps law of each written text. Scattered colored plots represent the avalanche size distributions of novelties for each considered Wikipedia page. **a)** text 31939 (red scattered plot). **b)** text 13859 (orange scattered plot). **c)** text 7372 (lime scattered plot).

#### S4. CALCULATION OF THE HEAPS LAW IN THE UMT MODEL AND THE INTERMISSION TIMES DISTRIBUTION

We consider the general UMT model with  $\tilde{\rho} \geq 0$ . The continuous time equation for  $D(t)$  reads:

$$\frac{dD(t)}{dt} = \frac{N_0 + \nu D(t)}{N_0 + \rho t + aD(t)} \quad (1)$$

with  $a \equiv \tilde{\rho} + \nu + 1 - \rho$ , and for  $a = 0$  we recover the UMT-E, while for  $a = \nu + 1$  we recover the original UMT model. Instead of solving this equation, we solve

$$\frac{dt(D)}{dD} = \frac{N_0 + \rho t(D) + aD}{N_0 + \nu D} \quad (2)$$

that we can write as

$$\frac{dt(D)}{dD} - \frac{\rho}{N_0 + \nu D} t(D) = \frac{N_0 + aD}{N_0 + \nu D} \quad (3)$$

putting in evidence its nature as a non-homogeneous first-order linear differential equation. Solving by standard methods, we write the solution for the homogeneous equation:

$$t_{\text{hom}} = C \left( \frac{N_0 + \nu D}{N_0} \right)^{\frac{\rho}{\nu}} \quad (4)$$

and a particular solution for the non-homogeneous:

$$t_p = -a \left( \frac{N_0 + \nu D}{\nu(\rho - \nu)} \right) - \frac{N_0}{\rho} \left( \frac{\nu - a}{\nu} \right) \quad (5)$$

By finally imposing the initial condition  $t(0) = 0$ , we obtain the solution

$$\begin{aligned} t(D) &= \frac{N_0}{\rho} \left( 1 + \frac{a}{\rho - \nu} \right) \left( \frac{N_0 + \nu D}{N_0} \right)^{\frac{\rho}{\nu}} + \\ &- a \left( \frac{N_0 + \nu D}{\nu(\rho - \nu)} \right) - \frac{N_0}{\rho} \left( \frac{\nu - a}{\nu} \right) \end{aligned} \quad (6)$$

This solution is not invertible in a closed form, but it gives the known approximated asymptotic regimes, for instance  $D(t) \simeq t^{\frac{\nu}{\rho}}$  when  $\nu < \rho$ . The solution in this form is suitable for predicting the avalanche size distributions of novelties

and non-novelties. For instance, for the avalanches of non-novelty distributions, it is only needed the numerical integration of

$$\begin{aligned} P_{\text{non-nov}}(l) &= \frac{\int_0^{D_{\max}} p_{\text{new}}(D)^2 (1 - p_{\text{new}}(D))^l \frac{dt}{dD} dD}{\int_0^{D_{\max}} p_{\text{new}}(D) (1 - p_{\text{new}}(D)) \frac{dt}{dD} dD} = \\ &= \frac{\int_0^{D_{\max}} p_{\text{new}}(D) (1 - p_{\text{new}}(D))^l d(D)}{\int_0^{D_{\max}} (1 - p_{\text{new}}(D)) dD} \end{aligned} \quad (7)$$

We show the perfect agreement between the solution and corresponding realizations of the process in Fig. 10.

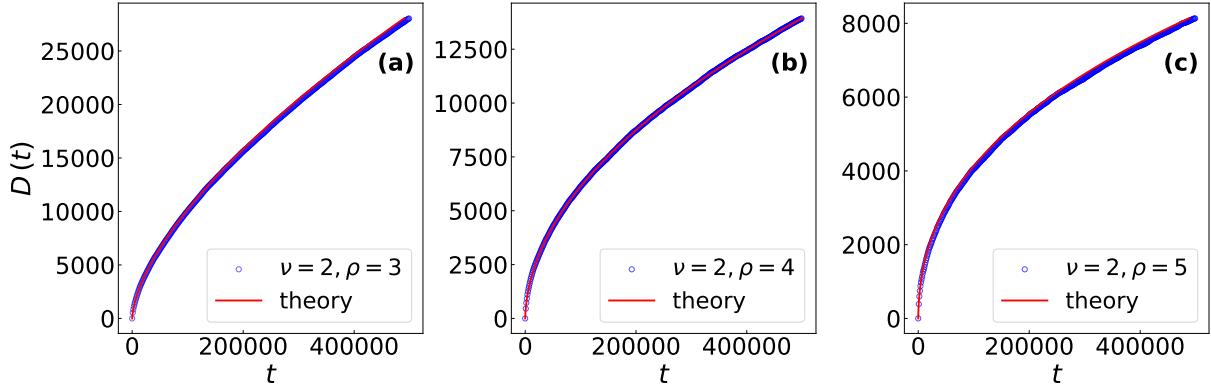


FIG. 10. **Heaps Law for the UMT model.** Comparison between the theoretical predictions and corresponding realizations. a) ( $\nu = 2, \rho = 3, N_0 = 1000$ ). b) ( $\nu = 2, \rho = 4, N_0 = 1000$ ). c) ( $\nu = 2, \rho = 5, N_0 = 1000$ ). All the realizations have length  $5 \times 10^5$ .

In Fig. 11 we consider the same realizations of the process and show the perfect agreement with the analytical avalanches of non-novelties distribution for the UMT model.

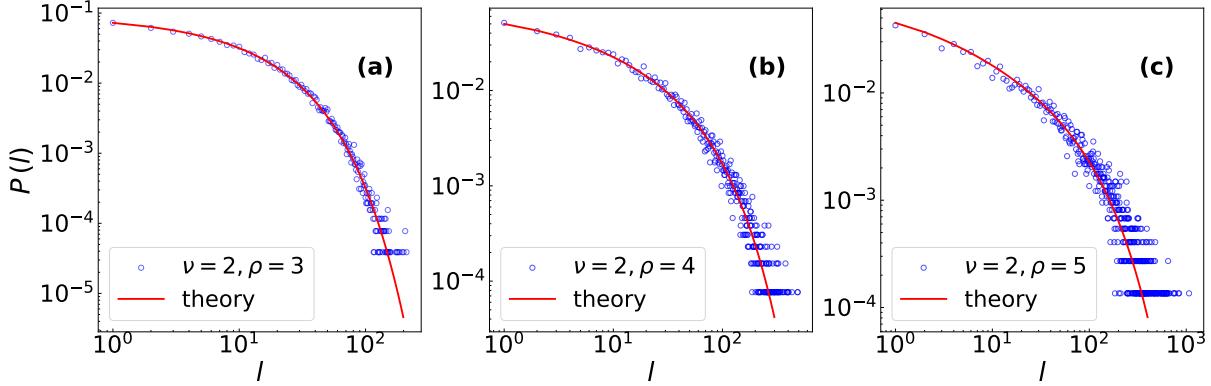


FIG. 11. **Avalanches size distributions for non-novelties for the UMT model.** Comparison between the theoretical predictions and corresponding realizations. a) ( $\nu = 2, \rho = 3, N_0 = 1000$ ). b) ( $\nu = 2, \rho = 4, N_0 = 1000$ ). c) ( $\nu = 2, \rho = 5, N_0 = 1000$ ). All the realizations have length  $5 \times 10^5$ .

#### S4 SERIES EXPANSION OF THE AVALANCHE SIZE DISTRIBUTION IN THE UMT-E

Here, we give an alternative expression for the avalanche size distribution of non-novelties in terms of an explicit series expansion.

In deriving this, it is convenient to consider separately the numerator and the denominator in Eq.(10) in the main text. The numerator has the form:

$$\left(1 + \frac{\rho t}{N_0}\right)^{\frac{2\nu}{\rho} - 2} \left(1 - \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho} - 1}\right)^l \quad (8)$$

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} (a)^{n-k} b^k, \quad (9)$$

so we obtain

$$\left(1 + \frac{\rho t}{N_0}\right)^{\frac{2\nu}{\rho} - 2} \sum_{s=0}^L \binom{L}{s} (-1)^s \left(1 + \frac{\rho t}{N_0}\right)^{(\frac{\nu}{\rho} - 1)s} \quad (10)$$

$$\sum_{s=0}^L \binom{L}{s} (-1)^s \left(1 + \frac{\rho t}{N_0}\right)^{(\frac{\nu}{\rho} - 1)(s+2)} \quad (11)$$

and by integrating between  $t_{max}$  and 0 we get:

$$\sum_{s=0}^L \binom{L}{s} (-1)^s \frac{\rho}{(s+2)(\nu - \rho) + \rho} \frac{N_0}{\rho} \left(1 + \frac{\rho t}{N_0}\right)^{(\frac{\nu}{\rho} - 1)(s+2)+1} \Big|_0^{t_{max}} = \quad (12)$$

$$\sum_{s=0}^L \binom{L}{s} (-1)^s \frac{N_0}{(s+2)(\nu - \rho) + \rho} \left[\left(1 + \frac{\rho t_{max}}{N_0}\right)^{(\frac{\nu}{\rho} - 1)(s+2)+1} - 1\right] \quad (13)$$

The denominator instead has the form:

$$\left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho} - 1} \left(1 - \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho} - 1}\right) = \left(1 + \frac{\rho t}{N_0}\right)^{\frac{\nu}{\rho} - 1} - \left(1 + \frac{\rho t}{N_0}\right)^{\frac{2\nu}{\rho} - 2} = \quad (14)$$

that integrating between 0 and  $t_{max}$  is

$$\frac{N_0}{\nu} \left[\left(1 + \frac{\rho t_{max}}{N_0}\right)^{\frac{\nu}{\rho}} - 1\right] - \frac{N_0}{2\nu - \rho} \left[\left(1 + \frac{\rho t_{max}}{N_0}\right)^{\frac{2\nu}{\rho} - 1} - 1\right] \simeq \frac{N_0}{\nu} \left(1 + \frac{\rho t_{max}}{N_0}\right)^{\frac{\nu}{\rho}} \quad (15)$$

By putting everything together, we get:

$$P(L) \simeq \sum_{s=0}^L \binom{L}{s} (-1)^s \frac{\nu}{(s+2)(\nu - \rho) + \rho} \left(1 + \frac{\rho t_{max}}{N_0}\right)^{(\frac{\nu}{\rho} - 1)(s+1)} \quad (16)$$

## S5. EQUIVALENCE BETWEEN THE UMT-E, UMT, AND UMST WITH RESPECT TO THE AVALANCHE STATISTICS

In this section, we demonstrate the equivalence of the UMT-E model with the general UMT, the UMST and the UMST model in its exchangeable version in terms of Heaps' law and avalanches of non-novelty distributions. We show that it is always possible to identify a triple  $(\nu, \rho, N_0)$  for the UMT-E model that reproduces the statistical behavior observed in the other models. The fitting procedure follows the same methodology adopted for real-world datasets. Given a realization from a specific model, we fit its Heaps' law using the analytical UMT-E Heaps law expression and estimate the corresponding parameters  $(\nu, \rho, N_0)$ . Subsequently, we compute the theoretical avalanche size distribution of non-novelties.

Figure 12 demonstrates the equivalence between the UMT-E and UMT model (in the  $\nu < \rho$  regime). Figure 13 shows the equivalence between the UMT-E and UMST model. In Figure 14, we additionally present the UMST-E variant, which corresponds to the UMST model with  $\tilde{\rho} = \rho - (\nu + 1)$ . This variant is an analogue of the UMT-E but lacks the exchangeability property.

In tables (V, VI, VII) we resume the equivalent parameters between each model.

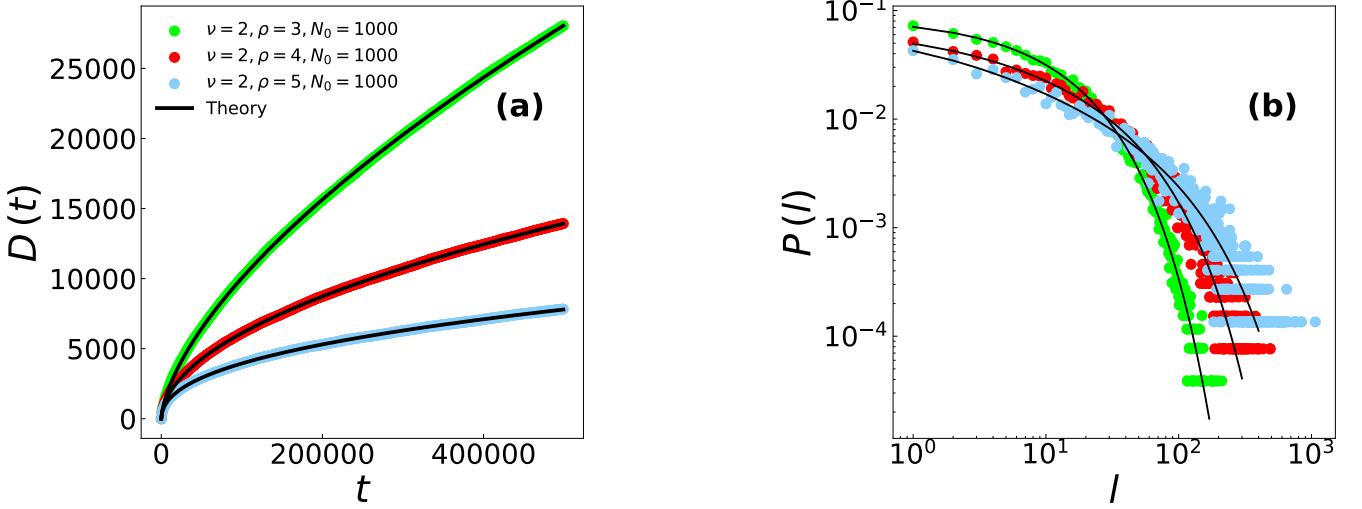


FIG. 12. **Equivalence between the UMT-E and the UMT.** a) Heaps' law for three UMT model realizations and corresponding analytical UMT-E Heaps law predictions. The UMT parameter sets  $(\nu = 2, \rho = 3, N_0 = 1000), (\nu = 2, \rho = 4, N_0 = 1000), (\nu = 2, \rho = 5, N_0 = 1000)$  are contrasted with the UMT-E parameter sets  $(\nu = 7, \rho = 11, N_0 = 583), (\nu = 1, \rho = 2, N_0 = 194), (\nu = 21, \rho = 52, N_0 = 5593)$ , respectively. b) Avalanche size distributions for non-novelties of the UMT realizations, along with the theoretical UMT-E avalanches size distributions for non-novelties. All the sequences have length  $5 \times 10^5$  and we use the same parameter values as in plot a).

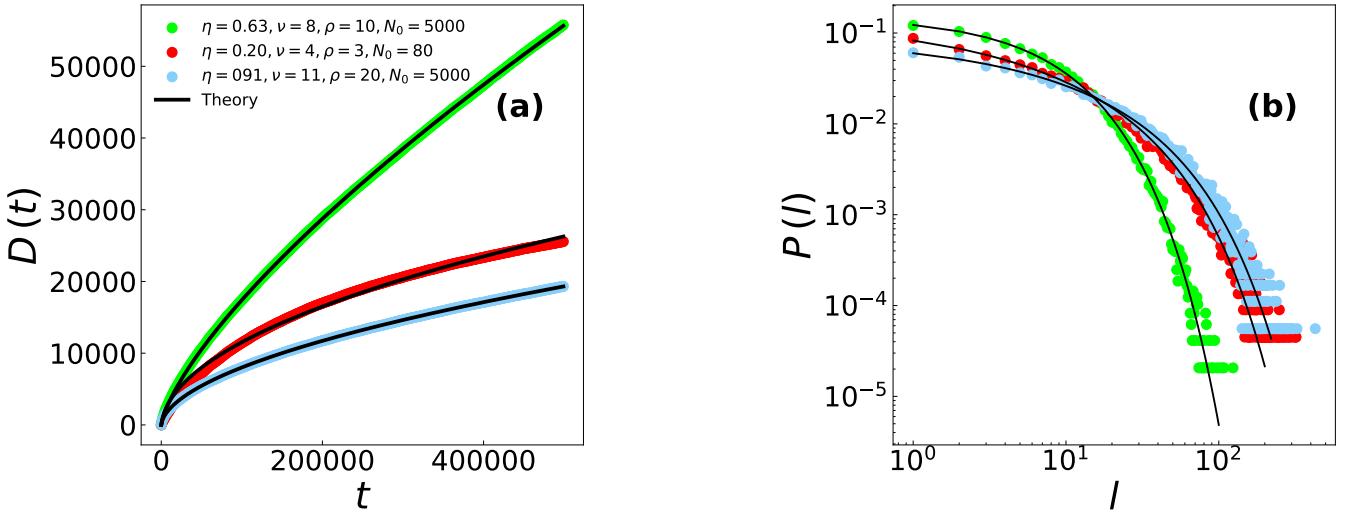


FIG. 13. **Equivalence between the UMT-E and the UMST.** a) Heaps' law for three UMST model realizations, and the corresponding analytical UMT-E Heaps law predictions. The UMST parameter sets  $(\eta = 0.63, \nu = 8, \rho = 10, N_0 = 500), (\eta = 0.20, \nu = 4, \rho = 30, N_0 = 80), (\eta = 0.91, \nu = 11, \rho = 20, N_0 = 1000)$  are contrasted with the UMT-E parameter sets  $(\nu = 31, \rho = 43, N_0 = 2570), (\nu = 22, \rho = 45, N_0 = 28509), (\nu = 50, \rho = 93, N_0 = 10937)$ , respectively . b) Avalanche size distributions for non-novelties of the UMST realizations, along with the theoretical UMT-E avalanches size distributions for non-novelties. All the sequences have length  $5 \times 10^5$  and we use the same parameter values as in plot a).

UMT Parameters			Equivalent UMT-E Parameters		
$\nu$	$\rho$	$N_0$	$\nu$	$\rho$	$N_0$
2	3	1000	7	11	583
2	4	1000	1	2	194
2	5	1000	21	52	5593

TABLE V. Equivalent parameter sets between UMT and UMT-E models

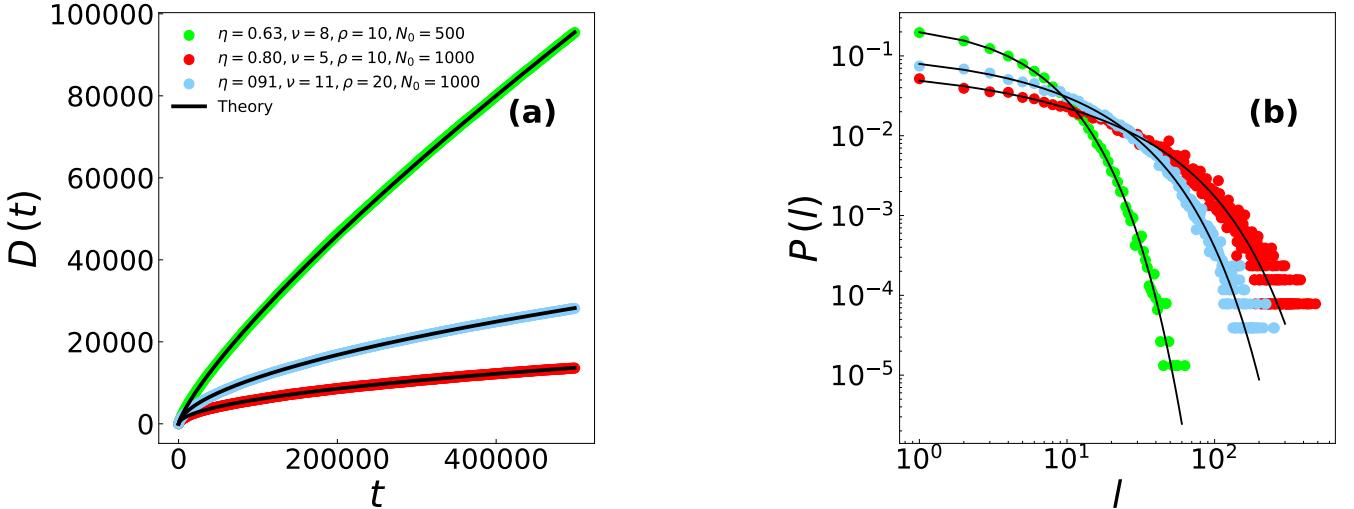


FIG. 14. **Equivalence between UMT-E and UMST-E models.** a) Heaps' law for three UMST-E model realizations along with the corresponding analytical UMT-E Heaps Law predictions. The UMST-E parameter sets ( $\eta = 0.63, \nu = 8, \rho = 10, N_0 = 500$ ), ( $\eta = 0.80, \nu = 5, \rho = 10, N_0 = 1000$ ), and ( $\eta = 0.91, \nu = 11, \rho = 20, N_0 = 1000$ ) are contrasted with the UMT-E parameter sets ( $\nu = 79, \rho = 99, N_0 = 4346$ ), ( $\nu = 50, \rho = 99, N_0 = 9128$ ), and ( $\nu = 29, \rho = 52, N_0 = 10804$ ), respectively. b) Avalanche size distributions for non-novelties of the UMT realizations, along with the theoretical UMT-E avalanche size distributions for non-novelties. All the sequences have length  $5 \times 10^5$  and we use the same parameter values as in plot a).

UMST Parameters				Equivalent UMT-E Parameters		
$\eta$	$\nu$	$\rho$	$N_0$	$\nu$	$\rho$	$N_0$
0.63	8	10	500	31	43	2570
0.20	4	3	80	22	45	28509
0.91	11	20	1000	50	93	10937

TABLE VI. Equivalent parameter sets between the UMST model and the UMT-E model.

UMST-E parameters				Equivalent UMT-E Parameters		
$\eta$	$\nu$	$\rho$	$N_0$	$\nu$	$\rho$	$N_0$
0.63	8	10	500	79	99	4346
0.80	5	10	1000	50	99	9128
0.91	11	20	1000	29	52	10804

TABLE VII. Equivalent parameter sets between the UMST-E model and the UMT-E model.