

The dynamics of discovery and the Heaps-Zipf relationship

Célestin Zimmerlin,¹ Thomas Louail,¹ Manuel Moussallam,² and Marc Barthelemy^{3,4}

¹*CNRS, UMR Géographie-cités & LabCom MIXTAPES
Campus Condorcet*

FR-93322 Aubervilliers Cedex

²*Deezer research & LabCom MIXTAPES*

³*Université Paris-Saclay, CNRS, CEA, Institut de Physique Théorique, 91191 Gif-sur-Yvette, France*

⁴*Centre d'Analyse et de Mathématiques Sociales (CNRS/EHESS) 54 Avenue de Raspail, 75006 Paris, France*

(Dated: October 27, 2025)

When following a sequence — such as reading a text or tracking a user’s activity — one can measure how the ‘dictionary’ of distinct elements (types) grows with the number of observations (tokens). When this growth follows a power law, it is referred to as Heaps’ law, a regularity often associated with Zipf’s law and frequently used to characterize human innovation and discovery processes. While random sampling from a Zipf-like distribution can reproduce Heaps’ law, this connection relies on the assumption of temporal independence — an assumption often violated in real-world systems although frequently found in the literature. Here, we investigate how temporal correlations in token sequences affect the type–token curve. In systems like music listening and web browsing, domain-specific correlations in token ordering lead to systematic deviations from the Zipf–Heaps framework, effectively decoupling the type–token plot from the rank–frequency distribution. Using a minimal one-parameter model, we reproduce a wide variety of type–token trajectories, including the extremal cases that bound all possible behaviors compatible with a given frequency distribution. Our results demonstrate that type–token growth reflects not only the empirical distribution of type frequencies, but also the temporal structure of the sequence — a factor often overlooked in empirical applications of scaling laws to characterize human behavior.

Keywords: Heaps’ Law, Zipf’s Law, Type–token plot, Discovery processes, Music listening behavior

I. INTRODUCTION

When observing a sequence—such as reading a text, browsing websites, or listening to music—elements may either recur or appear for the first time, reflecting a continuous interplay between familiar elements and novel ones. Two empirical laws—*Heaps’ law* and *Zipf’s law*—have emerged as central tools for describing how novelty and frequency are distributed in these systems [1]. Heaps’ law characterizes the growth of the number of distinct types D with the number of observed tokens k in a sequence, typically following a sublinear power law.

$$D \propto k^\alpha, \quad (1)$$

with $\alpha \in [0.4, 0.7]$ in most empirical cases [2, 3]. It has been observed in systems ranging from natural language and source code to scientific and chemical databases [4–6], and is often interpreted as a signature of innovation [7–9].

Zipf’s law, in contrast, describes the distribution of type frequencies. When types are ranked by decreasing frequency, the frequency $f(r)$ of the type at rank r follows

$$f(r) \propto \frac{1}{r^\nu}, \quad (2)$$

with $\nu \approx 1$ in many systems [10, 11]. This pattern appears in diverse domains: historically in the case of city sizes [12], word frequencies in texts [10], genome expression [13], and web page popularity in online behavior [14]. A related formulation considers frequency f as a random

variable, with distribution

$$p(f) \propto f^{-\gamma}, \quad (3)$$

where $\gamma = 1 + 1/\nu$ [15, 16]. While both expressions are often used interchangeably, they are only equivalent asymptotically and at low ranks [17].

Heaps’ and Zipf’s laws often co-occur, with a widely reported relation between their exponents [18, 19]

$$\alpha = \frac{1}{\nu}. \quad (4)$$

The coexistence of these laws has been reproduced by theoretical models, such as the Yule–Simon process [20, 21], and has even inspired frameworks specifically designed to account for both, notably the adjacent possible model for innovation [7, 8]. Yet the equivalence between Zipf and Heaps remains analytically fragile. Studies have shown that the relation between the two exponents can weaken even under independent sampling [22, 23]. In particular Font-Clos et al. [23] argue that rank-based distributions $p(r)$ poorly represent low-frequency types and recommend using the frequency histogram $p(f)$, supported by further simulation results [17].

A key conceptual distinction is that Zipf’s law is static: it describes aggregate frequency distributions, irrespective of token order. Heaps’ law, however, is inherently dynamic: it captures how novelty accumulates over time. Any analytic link between the two thus relies on the strong assumption of temporal independence, which is often violated in real-world systems.

Interestingly, in the case of written texts Heap’s law can still be observed even though this assumption is known to be false. Despite known burstiness and thematic recurrence, random sampling from empirical word-frequency distributions still reproduces reasonable type–token curves [18]. This has been attributed to the fact that temporal correlations affect the recurrence of known words more than the introduction of new ones [23].

Outside human language, however, these assumptions may not hold. In digital environments — such as streaming platforms, or even the world wide web — individual users select content from vast, ever-growing catalogues, and their sampling strategies are not constrained by syntactic or semantic rules. While the analogy with natural language is tempting, it is reasonable to assume that the dynamics of exploration will be different. Nevertheless, plotting a type–token curve for a user’s sequence and fitting a Heaps-like law is straightforward. It results in an exponent value α that it is convenient to use as proxy for the user’s “average discovery rate”, with $\alpha < 1$ indicating sublinear growth (i.e., novelty declines with k as $dD/dk \sim k^{\alpha-1}$).

To test the robustness of Heaps’ law under real-world human discovery dynamics, we compare empirical sequences of text, individual music listening and web browsing with their reshuffled counterparts, where temporal correlations are removed. By estimating the scaling exponent α in both real and reshuffled versions, we isolate the impact of temporal correlations. Our results show that while written texts statistics are largely robust to reshuffling, digital exploration sequences are not, highlighting the role of system-specific temporal dynamics in shaping discovery processes. The top row of Figure 1 presents an illustrative example of type–token curves for individual sequences from each dataset.

II. INFLUENCE OF TEMPORAL CORRELATIONS ON HEAPS’ LAW

We compare here datasets documenting very different processes: individual music listening histories on a streaming platform; individual web browsing histories; and written texts from English language literature. Despite their differences, these systems have all been previously analyzed as instances of general processes of human exploration, discovery, or innovation [8, 9, 24]. In particular, for music listening and written text, these dynamics have been quantitatively described within the Heaps’ law framework, where the scaling exponent α of the type–token relation (see Eq. 1) is interpreted as a quantitative signature of the underlying discovery process.

More specifically, we contrast the classical case of vocabulary growth in written texts (English literary classics) with that of individual users who progressively listen to new tracks on a music streaming platform over the course of several years — that is, the growth of their

personal musical ‘vocabulary’. In addition, we analyze a complementary dataset of digital exploration consisting in individual web browsing histories over a month [25]. For these three case studies, we compute the scaling exponents obtained by fitting Heaps’ law (see Eq. (1)) to both the original token discovery sequences and their reshuffled counterparts. The reshuffled sequences correspond to a scenario in which each token is drawn independently at random from the empirical rank–frequency distribution $p(r)$.

The second row of Figure 1 shows the distribution of the exponent values obtained for both cases—ordered and reshuffled sequences—across the different datasets. We denote by α the scaling exponent estimated from the original (ordered) sequence, and by α^* the exponent obtained from the reshuffled sequence rank–frequency distribution $p(r)$.

For the case of texts (left panel of Figure 1), the exponent values computed from the observed (ordered) sequences are highly correlated with those obtained from reshuffled sequences. The overall distributions of exponent values are also very similar in the two cases, indicating that the temporal structure of token occurrences has little effect on the scaling behavior.

In contrast, individual music listening histories over several years show a markedly different pattern. The scaling exponents obtained from the observed sequences are, on average, significantly higher than those obtained from reshuffled data, and they exhibit greater variability (i.e., larger variance). Moreover, the correlation between the real and reshuffled exponent values is weak, as shown in the bottom row of the figure.

Web browsing trajectories display a similar pattern to music listening: the exponent values derived from the observed sequences $D(k)$ are consistently larger than those from reshuffled sequences. Here too, the correlation between the two sets of exponents is weak (bottom row), although the effect is somewhat less pronounced than in the music case.

These observations suggest that the scaling exponent α obtained by fitting Heaps’ law (Eq. 1) to individual exploration sequences $D(k)$ in web browsing or music listening is shaped not only by the empirical rank–frequency distribution $p(r)$ (which reflects how the user’s attention is distributed among items), but also by the underlying temporal dynamics. This is in contrast to vocabulary growth in texts, where the exponent obtained from the empirical sequence can be accurately reproduced by randomly sampling tokens from $p(r)$.

In particular, in the case of musical exploration, the exponent values obtained for the ordered case (observed discovery sequences) are much closer to $\alpha = 1$ than the ones obtained when reshuffling the same music listening sequences (for which we obtain $\alpha^* = 0.5$). $\alpha = 1$ corresponds to the linear case of a stationary integration of discoveries into the user’s catalogue. Interestingly, a substantial portion of our sample also exhibits exponent values $\alpha > 1$, which appears to contradict the

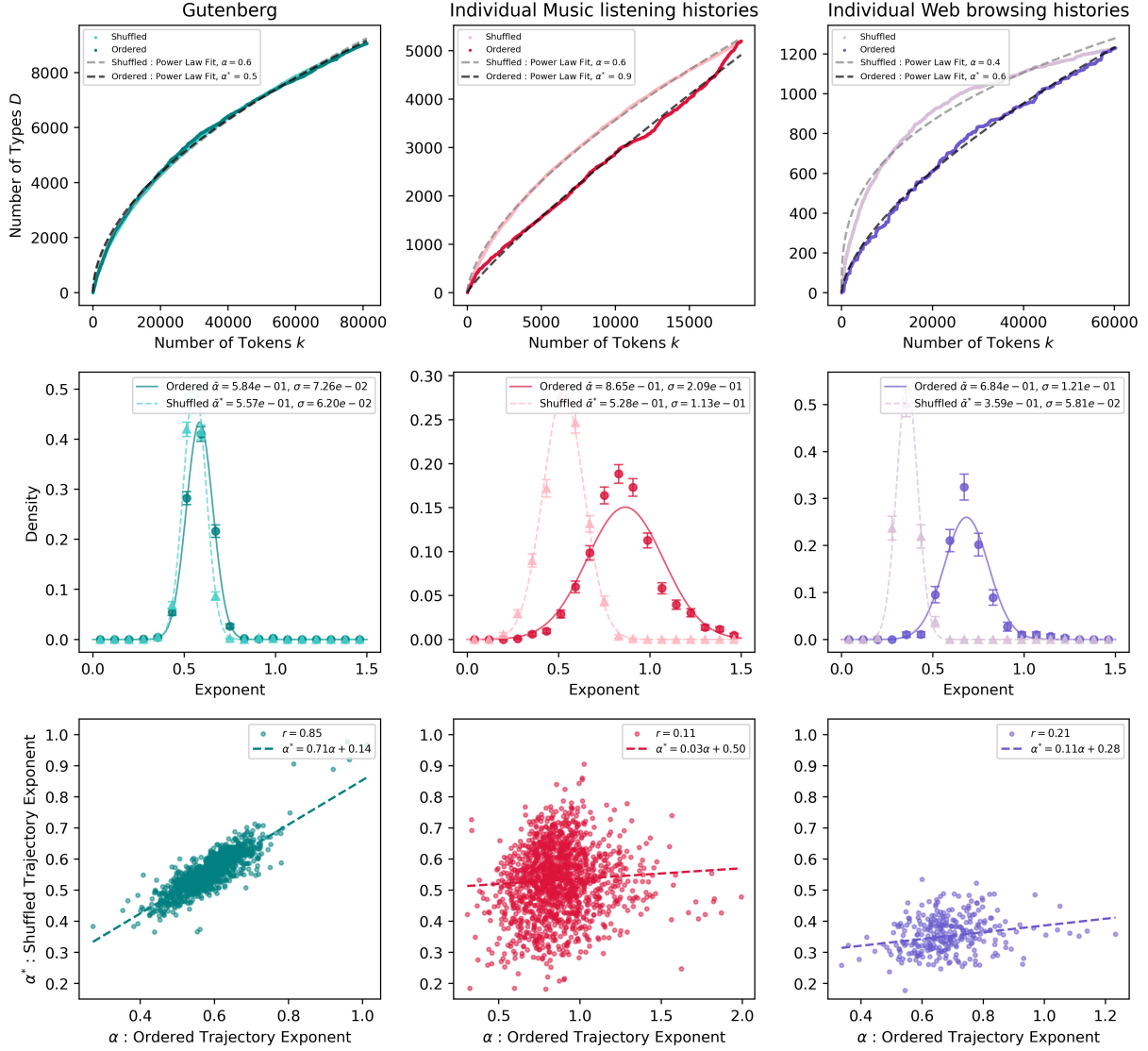


FIG. 1. **Top:** Type-token plots for a random sample from each dataset: texts from the Project Gutenberg corpus (left), individual music listening histories of Deezer users (middle), and individual web browsing histories (right). For each case, both the empirical ordered trajectory and a reshuffled trajectory are shown. The scaling relation $D = ck^\alpha$ (1) is fitted using two alternative methods; for each sequence, the exponent corresponding to the fit with the lowest Bayesian Information Criterion (BIC) is retained (see Appendix B). **Top:** Distributions of scaling exponents α obtained for both ordered and reshuffled sequences $D(k)$, across the three datasets. We consider more than 1000 sequences for both the Gutenberg and Deezer datasets, and 292 sequences for the Web Tracking dataset. Error bars are computed using the bootstrap method described in [26], with 1,000 bootstrap samples. **Bottom:** Relationship between exponent values estimated for the ordered (α) and reshuffled (α^*) sequences of the same individual user or text. Each plot reports the Bravais-Pearson correlation coefficient r , along with a linear regression $\alpha^* = a\alpha + b$ fitted by minimizing the absolute deviation to reduce sensitivity to outliers.

standard formulation of Heaps' law [22, 27]. Heaps' law is traditionally understood as an asymptotic result, and in the asymptotic regime it is indeed true that due to the constraint $D < k$, then necessarily $\alpha < 1$. However, in a more general framework applied to finite-length sequences, values of $\alpha > 1$ are entirely possible, as long as the prefactor c in the relation $D = ck^\alpha$ remains sufficiently small (in this case $ck^\alpha < k$ implies that $k < k_c$ with $k_c = 1/c^{1/(\alpha-1)}$). These $\alpha > 1$ cases correspond to situations in which the appetite of the individual for

discovery grows over time.

The observation of these super-linear exploration trajectories in musical contexts, as opposed to written texts, points to a fundamental asymmetry between the catalogues under study. On streaming platforms and the web, the set of available items is not only larger than in written corpora but also expands at a rate that greatly exceeds the pace of individual exploration. This fundamental constraint helps explain why the space of discoveries trajectories differs between contexts, even without

addressing the different patterns of human exploration across these contexts.

It is worth noting that some works on individual music discovery enforce fitting procedures that systematically prevent the emergence of exponents $\alpha > 1$ [9, 28], potentially overlooking this important dynamical feature.

To complete the analysis of the exponents, we compare the R^2 scores of power-law fits for each trajectory—both ordered and reshuffled—across the different datasets (see Appendix B and Fig. 4).

III. DEPENDENCE OF SCALING DEVIATIONS ON SEQUENCE LENGTH

We have shown that individual music listening (over the course of several years) and web browsing (over several weeks) exhibit exploration dynamics that differ markedly from the paradigmatic case of written texts, where Heaps’ law was originally observed. In the latter, the accumulation of novel elements follows a relatively smooth and universal trend, whereas in music or web navigation, discovery is intertwined with strong repetition and user-specific content preferences [29, 30]. One important consequence of these differences is that, for a given empirical rank–frequency distribution $p(r)$ —which gives the number of tokens associated with each type—very different scaling exponents α can emerge depending on the discovery process, that is on the temporal ordering of the token sequence $D(k)$. More precisely, we find that the exponent α measured on empirical (ordered) sequences is often substantially different from the value α^* obtained by reshuffling the same sequence, which corresponds to a scenario of random exploration drawn from the same catalog. This discrepancy indicates that the exponent reflects not just the frequency distribution $p(r)$, but also system-specific temporal correlations in the discovery process.

To better understand this deviation, we investigate how the absolute difference $|\alpha - \alpha^*|$ depends on the length of the sequence. To do this, we truncate each sequence at different values of $k_{\max}$, and compute the absolute difference $|\alpha - \alpha^*|$ for each truncation point. Figure 2 shows the results as a heatmap (darker colors indicate higher point density), along with a linear regression line to guide the eye.

As shown in this figure, we find that for music listening, the difference $|\alpha - \alpha^*|$ increases with sequence length, unlike in the case of written texts where the difference is essentially small and constant. This suggests that human exploration of large music catalogs increasingly deviates from random exploration as the sequence grows longer.

These results further indicate that estimating a single exponent α over a given sequence compresses multiple heterogeneous factors into a single parameter—namely, the attention distribution $p(r)$ and the temporal correlations of the sequence (e.g., repetition vs. novelty) compressed into the scaling exponent α . The relative con-

tributions of these factors themselves depend on the sequence length, which is often constrained arbitrarily by data availability.

IV. A TOY MODEL TO EXPLORE THE ROLE OF TEMPORAL CORRELATIONS

A. The envelope of all possible Heaps curves

When the assumption of temporal independence is lifted and all possible orderings of a sequence are allowed, the type–token trajectory $D(k)$ can vary widely—even for a fixed rank–frequency distribution $p(r)$. In particular, two extreme cases can be identified that form the envelope of all possible trajectories. These are shown as black and red dashed lines in Figure 3 (top). They correspond to ‘maximally accelerated’ and ‘maximally delayed’ discoveries of new types, depending solely on the temporal ordering of the tokens.

The black dashed line corresponds to the case where all $D_{\max}$ distinct types are discovered at the beginning of the sequence—one new type is introduced at each step until $k = D_{\max}$. The remaining $k_{\max} - D$ tokens are then repetitions of already observed types. The other extreme case, shown in red, corresponds to a sequence where discoveries are delayed as much as possible while remaining consistent with the frequency distribution $p(r)$. In this case, the most frequent type (rank 1, with frequency $f(1) = k_1$) appears k_1 times at the beginning, followed by k_2 repetitions of the second most frequent type, and so on, until the least frequent type is introduced at the very end. This strategy delays the growth of the dictionary to reach the final value D only near $k = k_{\max}$. These two limiting cases form the envelope of all possible type–token trajectories compatible with a given $p(r)$.

The variability of these curves is illustrated in Figure 3 using synthetic sequences generated from a Zipf-like distribution of type frequencies $p(r) \propto r^{-1.5}$, where different temporal orderings produce widely varying $D(k)$ curves. The bottom panel shows the corresponding rank–frequency distribution, fitted with a power-law using the method of [31].

B. A toy model

This envelope of possible behaviors can be reproduced using a simple toy model, controlled by a single parameter that modulates local temporal correlations. We start from a synthetic set of tokens, sampled at random from a rank–frequency distribution that follows Zipf’s law (see Eq. 2). The purpose of the model is to iteratively build a sequence from this set by using a stochastic mechanism that introduces temporal correlations between tokens of the same type.

The general principle is the following: the sequence is generated one step at a time, and at each step k , each

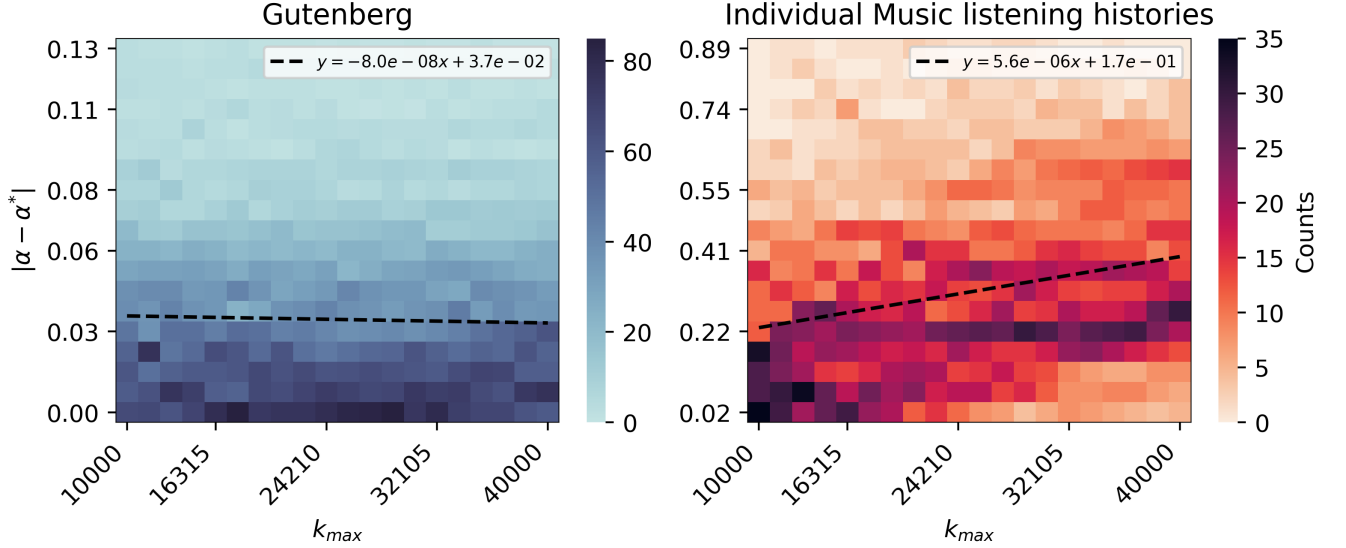


FIG. 2. Evolution of the absolute difference $|\alpha - \alpha^*|$ as a function of the sequence length $k_{\max}$. Each sequence (either a text or an individual listening history) is truncated at 20 different values of $k_{\max}$, ranging from 10,000 to 40,000 tokens. For each truncation point, we compute the absolute difference between the exponent obtained for the ordered sequence (α) and the one obtained for the reshuffled sequence (α^*). The resulting values of $|\alpha - \alpha^*|$ are displayed as a heatmap. To highlight the overall trend, we further compute, for each $k_{\max}$, the average $|\alpha - \alpha^*|$ across all sequences and fit a linear regression to these averages. The analysis is based on 500 texts and 200 individual listening histories. The web browsing dataset is excluded due to insufficient sequence length (see Appendix A).

token still available (i.e., not used yet) is selected with a probability that depends on the time elapsed since the last appearance of its type. For types that have not yet appeared in the sequence, we set the last appearance to step 0, so that their age is k .

The probability of selecting a token of type i at step k is chosen to be:

$$p(i) \propto z(i) \cdot \exp\left(-\frac{k - l(i)}{d_c}\right) \quad (5)$$

where: $z(i)$ is the number of remaining tokens of type i , $l(i)$ is the index of the last position of this type in the sequence, and d_c is a correlation length parameter controlling the memory of the process. By varying the value of d_c , the model can generate a wide spectrum of temporal correlation regimes:

- When $d_c \rightarrow 0^+$, the model favors recently used types, leading to bursty dynamics and delayed discovery (maximally delayed case).
- When $d_c \rightarrow 0^-$, the model favors types that have rarely (or never) appeared, accelerating discoveries as much as possible (maximally accelerated case).
- When $d_c \rightarrow \infty$, the memory vanishes and the sequence becomes an independent random sampling from $p(r)$, recovering the classical Heaps' law.

Although this model does not capture the full complexity of real-world exploration processes such as music discovery or web browsing (e.g. continuous arrival of new items, long-range dependencies), it demonstrates that local temporal correlations alone can generate a wide va-

riety of $D(k)$ trajectories, even when the underlying distribution $p(r)$ is held fixed.

V. DISCUSSION

We compared different systems — written texts in English literature, individual music listening histories, and web browsing histories — that can be represented as ordered sequences of discrete tokens taken from different types (words, songs or web pages). We compared these systems by fitting a power-law of the form $D = ck^\alpha$ to both the empirically observed (ordered) sequences $D(k)$ and to reshuffled versions of the same sequences, as shown in the type-token plot.

In written texts, our results indicate that the appearance dynamics of new types (i.e., new words), as captured by the curve $D(k)$, can be well-approximated by a random exploration of a fixed vocabulary. In this case, the type-token trajectory depends primarily on the empirical frequency distribution $p(r)$. This finding aligns with previous analytical results showing that Heaps' law can emerge from random sampling over a Zipfian distribution [19], and supports the idea that such an approach offers a reasonable starting point for modeling vocabulary growth in natural language.

However, we showed that this assumption breaks down in the two other systems we studied — individual online music listening and web browsing. In these cases, the individual discovery process — and consequently the al-

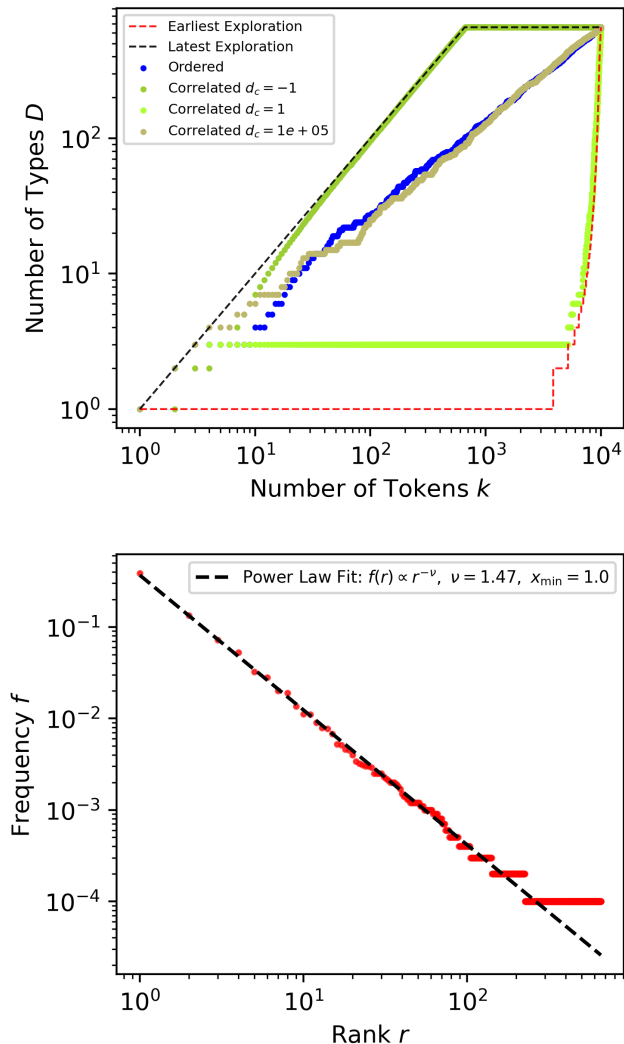


FIG. 3. **Top:** Type-token trajectories for different temporal arrangements of a sequence sampled from $p(r) \propto r^{-1.5}$, showing variability in discovery dynamics for the same underlying distribution. The black and red dashed lines correspond to the limiting cases of maximally accelerated and delayed discoveries. **Bottom:** Corresponding rank-frequency distribution of types in the sequence, plotted in log-log scale.

terrance between repetitions of already known tracks and pages, and the temporal correlations in the ordering of token appearances — play a central role in shaping the observed type-token trajectories. To further illustrate this point, we introduced a minimal toy model that generates correlated sequences by modifying the temporal structure of token occurrences. Despite its simplicity, the model reveals that local correlations can drastically reshape the type-token plot. By tuning a single parameter controlling the strength of correlations, the model reproduces a wide range of behaviors, including the envelope that bounds all possible trajectories in the type-token plane for a given set of tokens.

It should be noted that the frequency-rank plot is ap-

propriate for a ‘static’ study of discovery phenomena, as it captures the overall composition of a sequence. However, it does not convey the temporal dynamics of discovery, which are instead reflected in the Type-Token plot. Because these two objects are of fundamentally different nature, attempting to formally relate them is only possible under very specific conditions, which are generally not satisfied in real sequences. By default, the scaling exponent value α obtained when fitting a power-law on the Type-Token plot aggregates phenomena of different natures: it depends both on the type-frequency distribution $p(r)$ and on the system-specific temporal dynamics. Consequently, quantitative comparisons of exponent values across different systems are of limited interpretative value. In addition, the relative impact of temporal correlations on the estimated exponent can itself depend on the sequence length, further complicating comparisons. More generally, applying the Heaps/Zipf framework across a wide variety of phenomena is risky and should be done without assuming that all associated properties hold, such as the relationship between exponents as in (4). The system-specific dynamics, particularly temporal correlations, must be carefully analyzed.

Also, contrary to what is often stated in the literature, the Heaps exponent should not be interpreted as a direct measure of a system’s propensity for discovery (often referred to as the ‘discovery rate’). Instead, it characterizes the dynamics of discovery itself, such as the rate of growth, decay, or stationarity over the course of the sequence. The prefactor obtained in a power-law fit of the form Eq. (1) further quantifies the rate of discovery.

It is important to remain aware of the limitations of these tools, which are often overlooked. Beyond previously noted issues—such as the conceptual limitations of rank plots, the empirical fragility of Zipf’s and Heaps’ laws, and their instability under random sampling—we emphasize an additional point: temporal correlations in sequence dynamics can significantly influence the type-token curve. These correlations are rarely analyzed, yet ignoring them may lead to misattributing deviations in the curve to intrinsic content diversity rather than to the structure of exploration dynamics. These considerations support using the Zipf-Heaps framework primarily as a descriptive lens. While useful for identifying regularities, it does not offer rigorous statistical signatures, let alone universal laws of innovation. Quantitative comparisons across systems of different nature should therefore be approached with caution. More broadly, the difficulty of reducing heterogeneous phenomena to a single Heaps exponent highlights a deeper issue: collapsing diverse modes of exploration into a universal model risks erasing crucial domain-specific differences. An alternative is to examine the alternation between exploration and repetition within specific contexts, which can reveal internal semantics and the particular ways people engage with different forms of cultural content.

ACKNOWLEDGMENTS

This project has received financial support from the CNRS through the MITI interdisciplinary programs through its exploratory research program. T.L., M.M.,

and M.B. supervised this work. C.Z. performed the data analysis, coding, and initial draft writing. All authors contributed to the study's conceptualization, validation, final writing, review, and editing. C.Z. thanks the Deezer Research team and the Sony Computer Science Laboratory (Paris) for useful comments and support, and Emile Emery for his feedback.

-
- [1] E. G. Altmann, *Statistical Laws in Complex Systems: Combining Mechanistic Models and Data Analysis*, Understanding Complex Systems (Springer Nature Switzerland, Cham, 2024).
 - [2] G. Herdan, *Type-token mathematics : a textbook of mathematical linguistics*, Janua linguarum Series maior (Mouton, 'S-Gravenhage, 1960).
 - [3] H. S. Heaps, *The Library Quarterly* **50**, 153 (1980).
 - [4] A. Gelbukh and G. Sidorov, in *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, Vol. 2004, edited by A. Gelbukh (Springer, Berlin, 2001) pp. 332–335.
 - [5] H. Zhang, *Information Processing & Management* **45**, 477 (2009).
 - [6] R. W. Benz, S. J. Swamidass, and P. Baldi, *Journal of Chemical Information and Modeling* **48**, 1138 (2008).
 - [7] F. Tria, V. Loreto, and V. D. P. Servedio, *Entropy* **20**, 752 (2018).
 - [8] F. Tria, V. Loreto, V. D. P. Servedio, and S. H. Strogatz, *Scientific Reports* **4**, 1 (2014).
 - [9] G. Di Bona, A. Bellina, G. De Marzo, A. Petralia, I. Iacopini, and V. Latora, *Nature Communications* **16**, 393 (2025).
 - [10] G. K. Zipf, *Human Behavior And The Principle Of Least Effort : An Introduction to Human Ecology*, addison-wesley ed. (Cambridge MA, 1950).
 - [11] M. E. J. Newman, *Contemporary Physics* **46**, 323 (2005).
 - [12] F. Auerbach and A. Ciccone, *Environment and Planning B: Urban Analytics and City Science* **50**, 290 (2023).
 - [13] C. Furusawa and K. Kaneko, *Physical Review Letters* **90**, 088102 (2003).
 - [14] L. A. Adamic and B. A. Huberman, *Glottometrics* **3**, 143 (2002).
 - [15] W. Li, *Glottometrics* **5**, 14 (2002).
 - [16] L. A. Adamic, Zipf, Power-laws, and Pareto - a ranking tutorial (2000), (unpublished).
 - [17] A. Corral, I. Serra, and R. Ferrer-i Cancho, *Physical Review E* **102**, 052113 (2020).
 - [18] M. À. Serrano, A. Flammini, and F. Menczer, *PLOS ONE* **4**, e5372 (2009).
 - [19] D. Vanleijenhorst and T. Vanderweide, *Information Sciences* **170**, 263 (2005).
 - [20] G. U. Yule, *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* **213**, 21 (1925).
 - [21] H. A. Simon, *Biometrika* **42**, 425 (1955).
 - [22] L. Lü, Z.-K. Zhang, and T. Zhou, *PLoS ONE* **5**, e14139 (2010).
 - [23] F. Font-Clos and Á. Corral, *Physical Review Letters* **114**, 238701 (2015).
 - [24] H. S. Barbosa, F. B. De Lima Neto, A. Evsukoff, and R. Menezes, in *Complex Networks VII*, Studies in Computational Intelligence, Vol. 644, edited by H. Cherifi, B. Gonçalves, R. Menezes, and R. Sinatra (Springer International Publishing, Cham, 2016) pp. 173–184.
 - [25] J. Kulshrestha, M. Oliveira, O. Karaçalik, D. Bonnay, and C. Wagner, *Proceedings of the International AAAI Conference on Web and Social Media* **15**, 327 (2021).
 - [26] F. Sohil, M. U. Sohali, and J. Shabbir, *Statistical Theory and Related Fields* **6**, 87 (2022).
 - [27] A. Chacoma and D. H. Zanette, *Royal Society Open Science* **7**, 200008 (2020).
 - [28] G. D. Bona, E. Ubaldi, I. Iacopini, B. Monechi, V. Latora, and V. Loreto, *Social interactions affect discovery processes* (2022), (unpublished).
 - [29] B. Sguerra, V.-A. Tran, and R. Hennequin, in *Proceedings of the 17th ACM Conference on Recommender Systems* (2023) pp. 971–977.
 - [30] B. Sguerra, V.-A. Tran, and R. Hennequin, in *Proceedings of the 16th ACM Conference on Recommender Systems* (2022) pp. 556–561.
 - [31] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *SIAM Review* **51**, 661 (2009).
 - [32] Y. Renisio, A. Beaumont, J.-S. Beuscart, S. Coavoux, P. Coulangeon, R. Cura, B. L. Bigot, M. Moussallam, C. Roth, and T. Louail, *Revue française de sociologie* **65**, 129 (2024).
 - [33] J. C. Leitão, J. M. Miotto, M. Gerlach, and E. G. Altmann, *Royal Society Open Science* **3**, 150649 (2016).

TABLE I. Fitting methods

Noise type	Log-likelihood	Fitting technique
Additive $y_i = \hat{y}_i + \epsilon_i$	$\ln L(\hat{y}, \hat{\sigma}) = -\frac{n}{2} \ln(2\pi\hat{\sigma}^2 e)$	Least squares on y_i
Multiplicative $y_i = \hat{y}_i \exp(\epsilon_i)$	$\ln L(\hat{y}, \hat{\sigma}) = -\frac{n}{2} \ln(2\pi\hat{\sigma}^2 e) - \sum_{i=1}^n \ln y_i$	Least squares on $\ln y_i$

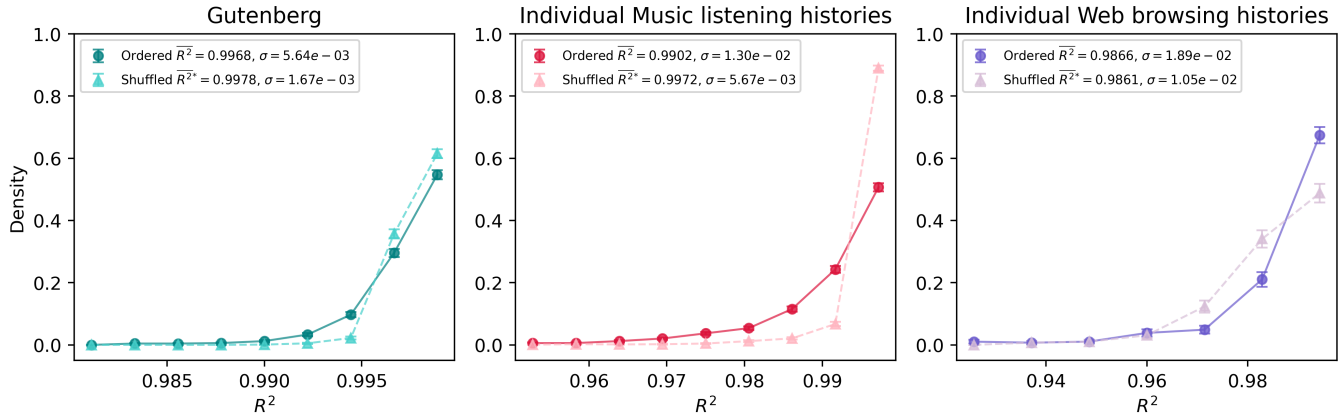


FIG. 4. Distribution of the coefficient of determination (R^2) when fitting the power-law $D = ck^\alpha$ on the type-token plot (1), for both ordered and reshuffled trajectories across the three datasets. Error bars are computed using a bootstrap procedure with 1000 resamples, following the approach described in [26].

Appendix A: Datasets

We use three different datasets covering textual, musical, and web browsing activities:

- **Gutenberg.** We rely on the Project Gutenberg corpus, consisting of 1400 texts listed in the catalog available here. From this set, we retain 1178 texts selected such that $k_{\max} > 10^4$.
- **Individual online music interaction data.** We analyze anonymized and timestamped listening histories of individual users of the streaming platform Deezer, in France, provided within the RECORDS project [32]. The data collected in this project include the listening history data on Deezer of about 16,000 users in France, along with their responses to a detailed online survey about their listening habits, music preferences and cultural practices beyond music. The survey includes a standard socioeconomic module which allowed to collect precise information about the social characteristics of the survey participants, and consequently to analyze the social structure of contemporary music listening practices. The public release of several datasets combining listening data with survey data is planned for 2026. A subset of 3000 users with complete listening histories over five years was first selected. Among these, 68% meet the criterion $k_{\max} > 10^4$. From this subgroup, 1400 users were sampled at random. These individual listening histories are available from the corresponding author upon reasonable request.
- **Web Tracking Dataset.** We use the dataset pub-

lished by [25], available on Zenodo, which contains one month (October 2018) of web tracking data for 2148 German users. We retain 292 users with $k_{\max} > 9000$. This threshold is comparatively more restrictive: it corresponds to individuals who opened more than 9000 web pages in a single month. This introduces a possible selection bias toward particularly active users. Furthermore, the relatively short sequence lengths limit the ability to study the longitudinal evolution of the gap between the fitted and reshuffled exponents (see 2).

Appendix B: R^2 Comparison

To evaluate the power-law fits, we compute the coefficient of determination (R^2) for all trajectories and plot their distributions in Fig. 4.

We find excellent R^2 values, which should nonetheless be interpreted with caution: this indicator measures the correlation between model and data rather than the statistical significance of the fit, and it tends to be overly optimistic when evaluating power-law scaling [33]. The purpose here is not to determine whether the data were truly generated by a power-law process—Eq. (1) is clearly a simplification of real world complexity—but to use this metric to compare fit quality across different trajectories.

As shown in Figure 4, power-law fits to Eq. (1) are generally better for text data than for music listening or web browsing sequences. For music and, to a lesser extent, texts, the real (ordered) sequences produce lower

and more variable R^2 values compared with reshuffled sequences. In contrast, web browsing sequences show slightly higher average R^2 for ordered trajectories than for reshuffled ones, although the dispersion is larger.

Overall, this reinforces two key points: 1) Correlations in the temporal ordering of events can systematically affect the empirical fit to Heaps' law. 2) This effect depends on the dataset considered, and is particularly pronounced in the case of music listening.

Appendix C: Fitting Procedure

We fit Heaps' law in type–token plots (1) using two alternative assumptions about the nature of the noise:

1. **Additive noise.** This corresponds to minimizing

the squared residuals between predictions and observations, i.e. using `scipy.curve_fit` on the original scale.

2. **Multiplicative noise.** This corresponds to minimizing the squared residuals between the logarithm of predictions and the logarithm of observations, i.e. using `sm.OLS` on the log-transformed data.

To guard against initialization effects, the first 1% of each fitted trajectory is omitted. For each trajectory, the retained exponent is the one obtained from the fit with the lowest Bayesian Information Criterion (BIC). The likelihoods corresponding to each noise assumption are summarized in Table I. This procedure was employed to determine the exponents of the complete trajectories (see Fig. 1). When examining the exponent evolution with $k_{\max}$ (see Fig. 2), we only used the additive noise hypothesis for convenience.