communications physics

Article

https://doi.org/10.1038/s42005-024-01714-6

Inference through innovation processes tested in the authorship attribution task

Giulio Tani Raffaelli [1], Margherita Lalli[2] & Francesca Tria [3]

Urn models for innovation capture fundamental empirical laws shared by several real-world processes. The so-called urn model with triggering includes, as particular cases, the urn representation of the two-parameter Poisson-Dirichlet process and the Dirichlet process, seminal in Bayesian non-parametric inference. In this work, we leverage this connection to introduce a general approach for quantifying closeness between symbolic sequences and test it within the framework of the authorship attribution problem. The method demonstrates high accuracy when compared to other related methods in different scenarios, featuring a substantial gain in computational efficiency and theoretical transparency. Beyond the practical convenience, this work demonstrates how the recently established connection between urn models and non-parametric Bayesian inference can pave the way for designing more efficient inference methods. In particular, the hybrid approach that we propose allows us to relax the exchangeability hypothesis, which can be particularly relevant for systems exhibiting complex correlation patterns and non-stationary dynamics.

Innovation enters a wide variety of human activities and natural processes, from artistic and technological production to the emergence of new behaviours or genomic variants. At the same time, the encounter with novelty permeates our daily lives more extensively than we typically realise. We continuously meet new people, learn and incorporate new words into our lexicon, listen to new songs, and embrace new technologies. Although innovation and novelties (i.e., new elements at the individual or local level) operate at different scales, we can describe their emergence within the same framework, at least in certain respects[1]. Shared statistical features, including the well-known Heaps'[2], Taylor's[3–6] and Zipf's[7,8] laws, suggest a common underlying principle governing their emergence. In this respect, an intriguing concept is the expansion into the adjacent possible[9]. The adjacent possible refers to the set of all the potential innovations or novelties attainable at any given time. When one of these possibilities is realised, the space of the actual enlarges, making additional possibilities achievable and thus expanding the adjacent possible. The processes introduced in[1] provide a mathematical formalisation of these concepts, extending Polya's urn model[10] to accommodate infinitely many colours. They generate sequences of items exhibiting Heaps', Zipf's, and Taylor's laws. The most general formulation of the modelling scheme proposed in ref. 1, the urn model with semantic triggering, also captures correlations in the occurrences of novelties, as observed in real-world systems. Further generalisations have been explored to capture the empirical phenomenology in diverse contexts: network growth and evolution[11], the varied destinies of different innovations[12], and mutually influencing events[13]. Additionally, the proposed modelling scheme can be cast within the framework of random walks on graphs, offering further intriguing perspectives and broadening its scope of applications[14–17].

We now want to address the question of whether these generative models can also be successfully used in inference problems. This question is further motivated by the precise connection that has been established[5,6] between the urn models in ref. 1 and seminal processes in Bayesian nonparametrics. The latter is a powerful tool for inference and prediction in innovation systems, where possible states or realisations are not predefined and fixed once and for all. Nonparametric Bayesian inference enables us to assign probabilities to unseen events and to deal with an ever-increasing number of new possibilities. Various applications have been proposed in diverse fields, including (but not limited to) estimation of diversity[18–22], classification problems[23,24], Bayesian modelling of complex networks[25,26] and they take a considerable role in Natural Language Processing[27,28].

The simplest model described in[1], the urn model with triggering (UMT), reproduces, with a specific parameter setting, the conditional probabilities that define the two-parameter Poisson-Dirichlet process[29], referred to as PD hereafter, that generalises the Dirichlet process[30]. The PD and the Dirichlet processes have gained special relevance as priors in Bayesian nonparametrics due to their generality and manageability[31], and the PD process predicts the Heaps, Zipf and Taylor laws, making its use more convenient in linguistically motivated problems.

Here, we aim to explore the potential of the outlined connection between urn models for innovation and priors for Bayesian nonparametric

[1]Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic. [2]Scuola Normale Superiore, Pisa, Italy. [3]Sapienza University of Rome, Physics Department, Rome, Italy. e-mail: francesca.tria@uniroma1.it

inference. As a sample application, we address the authorship attribution task[32].

The PD and Dirichlet processes have already been considered as underlying models for natural language processing and for authorship attribution purposes. The proposed procedures interpret the outputs of PD (or Dirichlet) processes as sequences of identifiers for distributions over words (i.e., topics)[33] and measure similarity among texts or authors based on topics' similarity[34,35]. We briefly discuss topic models in the Methods section. It is worth stressing here that these approaches have led to hierarchical formulations that require efficient sampling algorithms for solving the problem of computing posterior probabilities[28,33,36,37]. Moreover, these methods strongly rely on exchangeability, mainly due to the property of conditional independence it implies, through the de Finetti and Kingman theorems[38,39], and for guaranteeing the feasibility of the Gibbs sampling procedure[27,28]. Exchangeability refers to the property of the joint probability of a sequence of random variables being invariant under permutations of the elements. Notwithstanding the powerful tools it provides, this assumption is often unrealistic when modelling real-world processes.

We take a different perspective by interpreting the outputs of the underlying stochastic processes directly as sequences of words in texts or, more generally, tokens. Language serves as a paradigmatic example where novelty enters at different scales, ranging from true innovation–creation and diffusion of new words or meanings–to what we denote as novelties–the first time an individual adopts or encounters (or an author uses in their production) a word or expression. We thus borrow from information theory[40,41] the conceptualisation of a text as an instance of a stochastic process and consider urn models for innovation processes as underlying generative models. Specifically, we here consider the UMT model in its exchangeable version, which is equivalent to the PD process. We opt out of a fully Bayesian approach and use a heuristic method to determine the base distribution of the process–that is, the prior distribution of the items expected to appear in the sequence.

The overall change in perspective we adopt allows us to avoid the Monte Carlo sampling required in hierarchical methods. Moreover, while we consider here an exchangeable model, exchangeability is not crucial in our approach, paving the way for an urn-based inferential method that considers time-dependent correlations among items.

When comparing our method to various approaches used in authorship attribution tasks, we find promising results across different datasets (ranging from literary texts to blogs and emails), demonstrating that the method can scale to large, imbalanced datasets and remains robust to language variation.

Results

The authorship attribution task

To demonstrate a possible application of the UMT generative model for an inference problem, we used the probabilities of token sequences derived from the process to infer the authorship of texts. In the authorship attribution task, one is presented with a set of texts with known attribution – the reference corpus – along with a text T from an unknown author. The goal is to attribute T to one of the authors represented in the corpus (closed attribution task) or more generally, to recognise the author as one of those represented in the corpus or possibly as a new, unidentified author (open attribution task)[42]. Here we explicitly consider the case of the closed attribution task, although several strategies can be adopted to apply the method in open attribution problems as well.

Following the framework of Information Theory[40,41], we can think of an author as a stochastic source generating sequences of characters. In particular, a written text is regarded as a sequence of symbols, which can be dictionary words or, more generally, short strings of characters (e.g., n-grams if such strings have a fixed length n), with each symbol appearing multiple times throughout the sequence. Each symbol constitutes a novelty the first time it is introduced.

We evaluate the similarity between two symbolic sequences by computing the probability that they are part of a single realisation from the same source. More explicitly, let x_1^n and x_2^m be two symbolic sequences with length n and m respectively. Given their generative process—their source—we can compute the conditional probability $P(x_1^n|x_2^m)$, that is, the probability that x_1^n is the continuation of x_2^m. In the authorship attribution task, the anonymous text T is represented by a symbolic sequence x_T, while an author A by the symbolic sequence x_A obtained by concatenating the texts of A in the reference corpus. It is worth noting that an author A affects the probability of T both by defining the source and through the sequence x_A. We will use the notation $P(T|A) \equiv P_A(x_T|x_A)$ for the conditional probability of T to continue the production of A. The anonymous text T is attributed to the author $\tilde{A}$ that maximises such conditional probability: $\tilde{A} = \max_A P(T|A)$. We thus need to specify the processes generating the texts and the elements x_i of the symbolic sequences, i.e., the tokens.

The tokens. We can make several choices for defining the variables—or tokens—x_i. In what follows, we consider two alternatives: first, we consider Overlapping Space-Free N-Gram[43] (OSF). These are strings of characters of fixed length N as tokens, including spaces only as the first or last characters, thereby discarding words shorter than $N-2$. This choice has often yielded the best results. Secondly, we explore a hybrid approach where we exploit the structures captured by the Lempel and Ziv compression algorithm (LZ77)[44]. We define LZ77 sequence tokens as the repeated sequences extracted through a modified version of the Lempel and Ziv algorithm, which has been previously used for attribution purposes[45]. For each dataset, we select the token specification that provides the best performance. In the Supplementary Results, we compare the achieved accuracy when using the token definitions discussed above as well as when using simple dictionary words as tokens.

The generative process and the posterior probabilities. We consider the UMT model in its exchangeable version, which provides an urn representation of the PD process. The latter is defined by the conditional probabilities of drawing at time $t+1$ an old (already seen) element y and a new one (not seen until time t). They are given, respectively, by:

$$\begin{aligned} P(x_{t+1} = y|x^t) &= \frac{n_{y,t} - \alpha}{\theta + t}, \qquad \text{if} \quad n_{y,t} > 0 \\ P(x_{t+1} = y|x^t) &= \frac{\theta + \alpha D_t}{\theta + t} P_0(y), \quad \text{if} \quad n_{y,t} = 0 \end{aligned} \tag{1}$$

where $n_{y,t}$ is the number of elements of type y at time t and D_t is the total number of distinct types appearing in x^t; $0 < \alpha < 1$ and $\theta > -\alpha$ are two real-valued parameters and $P_0(\cdot)$ is a given distribution on the variables' space, called the base distribution. The UMT model does not explicitly define the prior probability for the items' identity, i.e., the base distribution P_0. The latter can be independently defined on top of the process, in the same way as for the Chinese restaurant representation of the Dirichlet or PD processes[46] (please refer to section UMT and PD processes in the Methods for a thorough discussion on the urn models for innovation and their relation with the PD process).

Crucially, Eqs. (1) are only valid when P_0 is non-atomic, which implies that each new token can be drawn from P_0 at most once with probability one. On the contrary, when P_0 is a discrete probability distribution (it has atoms), an already seen value y can be drawn again from it, and the conditional probabilities no longer have the simple form shown in Eq. (1) (as detailed in the Methods). In a problem of language processing, the tokens are naturally embedded in a discrete space, which has led to the development of hierarchical formulations of the PD process[47,48]. In these approaches, the P_0 is the (almost surely) discrete outcome of another PD process with a non-atomic base distribution. Here we follow a different approach. We regard P_0 as a prior probability on the space of new possibilities. In this view, the tokens take values from an uncountable set, and thus the probability of drawing the same token y from P_0 more than once is null. As a consequence, we can use the simple Eq. (1), where we need to make some arbitrary choices for the actual definition of the base distribution. In the

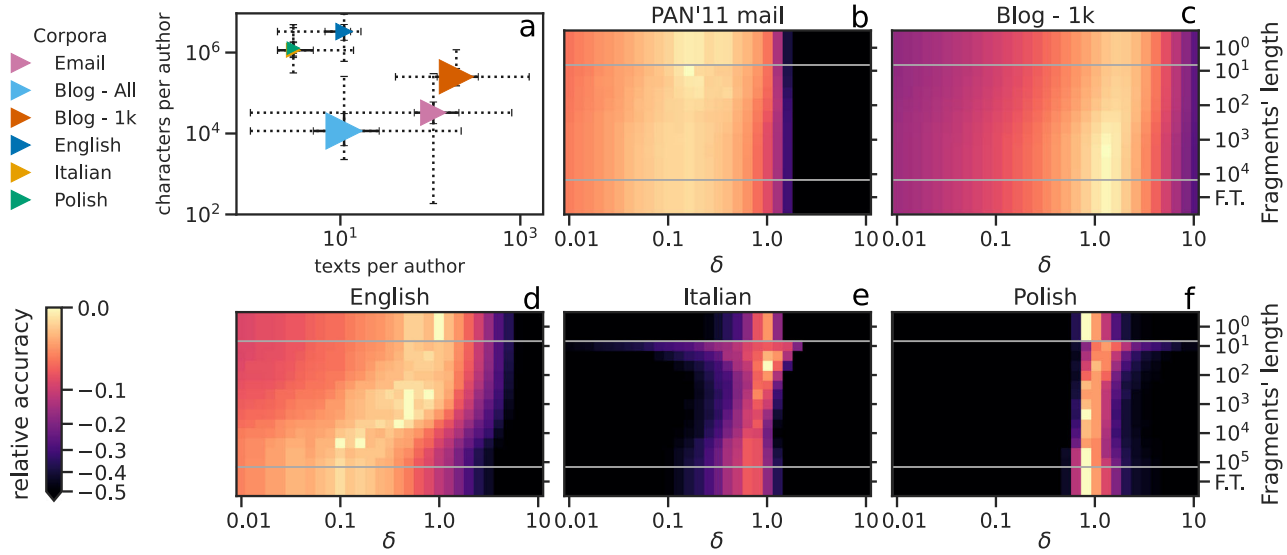

Fig. 1 | Corpora sizes and the impact of model parameters on attribution accuracy. In panel (**a**) we offer a pictorial view of various characteristics related to the size of the considered corpora. The size of the triangles is proportional to the logarithm of the corpus size, measured as number of documents. In the x and y axes we represent for each corpus the distribution of the numbers of texts (x axis) and of the numbers of characters (y axis) per author. Specifically, the continuous line bars represent the interquartile range of the distributions and the dotted lines show the 95% interval, to highlight their long tails. Panels (**b**–**f**) report the attribution accuracy varying the length of the fragments and the δ value. The colour scale refers to the difference relative to the maximum attribution accuracy obtained in each dataset. In the upper band, the considered length of fragments is of a single token. In the lower band, the text is not partitioned in fragments (full text).

following, we identify $P_0(y)$ with the frequency of y in each dataset, while still treating P_0 as a non-atomic distribution by ensuring that each item can be drawn at most once from it. However, this raises a tricky question of normalisation, which strongly depends on the dataset, resulting in the arbitrary modulation of the relative importance of innovations and repetitions. We have addressed this problem heuristically by introducing an additional parameter $\delta > 0$ that multiplies P_0: it suppresses ($\delta < 1$) or enhances ($\delta > 1$) the probability of introducing a novelty in T. In addition, we consider an author-dependent base distribution by discounting the vocabulary already appearing in A (details are given in section The strategy of P_0 in the "Methods" section). To summarise, the conditional probabilities $P(T|A)$ are derived from Eqs. (1), where the base distribution $P_0(y)$ is defined as discussed above. Different values of α and θ characterise the specific distribution associated with each author. We fix α_A and θ_A for each author A to the values that maximise her likelihood (refer to the Supplementary Methods for details). We denote by D_K (with $K = A, T$) the number of types (i.e., distinct tokens) in A and T, and by $D_{T\cup A} - D_A$ the number of types in T that do not appear in A. The conditional probability of a text T to be the continuation of the production of an author A reads:

$$P(T|A) = \frac{(\theta_A + \alpha_A D_A|\alpha_A)_{D_{T\cup A}-D_A}}{(\theta_A + m)_n} \prod_{j=1}^{D_T} Q_j,$$
$$Q_j \equiv \begin{cases} (1-\alpha_A)_{n_j^T - 1} P_0(y_j) & \text{if } y_j \notin A \\ (n_j^A - \alpha_A)_{n_j^T} & \text{otherwise}. \end{cases} \quad (2)$$

where n_j^K is the number of occurrences of y_j in K (with $K = A, T$), such that $\sum_j n_j^A = m$ and $\sum_j n_j^T = n$. The Pochhammer symbol and the Pochhammer symbol with increment k are defined respectively by $(z)_n \equiv z(z+1)\ldots(z+n-1) = \Gamma(z+n)/\Gamma(z)$ and $(z|k)_n \equiv z(z+k)\ldots(z+(n-1)k)$.

In practice, when attributing the unknown text, we adopt the procedure of dividing it into fragments and evaluating their conditional probability separately. The entire document is then attributed either to the author that maximises the probabilities of most fragments or to the author that maximises the whole document probability computed as a joint distribution over independent fragments (i.e., as a product of the probabilities of its fragments). We optimise this choice for each specific dataset, as described in the Supplementary Methods.

Results

We test our approach on literary corpora and informal corpora. To challenge the generality of our method versus language variation[49], we consider three corpora of literary texts in three different languages, English, Italian, and Polish, belonging to distinct Indoeuropean families and bearing a diverse degree of inflection (refer to the Supplementary Note 1 for details). We further consider informal corpora mainly composed of English texts. They are particularly challenging for the attribution task due to the strong unbalance in the number of samples per author and the texts' lengths (refer to Fig. 1, panel a). We consider, in particular, an email corpus and a blog corpus. The first is part of the Enron Email corpus proposed during the PAN'11 contest[50]. It is still used as a valuable benchmark, and we compare the accuracy of our method with those reported in refs. 34,35. The Blog corpus is one of the largest datasets used to test methods for authorship attribution[51]. This is a collection of 678,161 blog posts by 19,320 authors taken from ref. 52. Additionally, in line with refs. 53,54, we test our method on the subset of 1000 most prolific authors of this corpus. For more details on the corpora, please refer to the Supplementary Note 1.

In Fig. 1b–f, we illustrate the dependency of the attribution accuracy on the value of two free parameters of our model, specifically the normalisation δ and the length of the fragments in which we partition the text to be attributed. In particular, we report the accuracy achieved on each dataset in a leave-one-out experiment, where we select each text in turn and attribute it by training the model on the rest of the corpus (refer to the Supplementary Methods for more details). We note that, although simply setting $\delta = 1$ often gives the most or nearly the most accurate results, in a few datasets using a different value of δ significantly improves the accuracy. Indeed an effect of δ is also to correct for a non-optimal choice of the length of the fragments, as is evident in the literary English dataset. When attributing an anonymous text, we optimise these two parameters—as well as the selection of P_0, the definition of the tokens, and the strategy to attribute the whole document from the likelihood of single fragments—on the training and validation sets, as detailed in the Supplementary Methods.

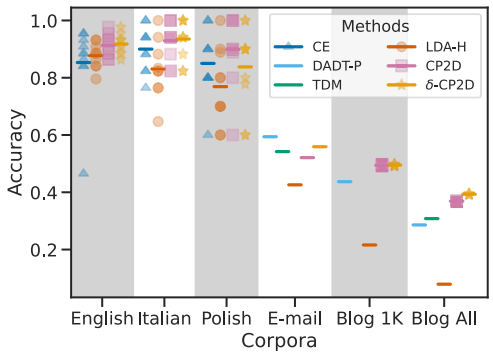

Fig. 2 | Attribution accuracy. For each of the considered datasets and attribution methods, thick lines show the average accuracy in the ten-fold stratified cross-validation experiment, while shaded circles refer to the attribution accuracy on each of the ten test sets separately. An exception is the E-mail dataset, where a unique test set is considered (see main text). We compare the accuracy achieved by our method (the Constrained Probability 2-parameters Poisson-Dirichlet, in both its versions with and without including the parameter δ: the CP2D and δ-CP2D) with the Cross-Entropy based approach (CE), the Latent Dirichlet Allocation plus Hellinger distance (LDA-H), the Disjoint Author-Document Topic model in its Probabilistic formulation (DADT-P), and the Topic Drift Model (TDM). On the literary corpora, the LDA-H accuracy is computed using our implementation; please refer to Supplementary Methods for details. For the informal corpora, the results are are available from a previous study [ref. 54, Table 1]. Results for the DADT-P and the TDM algorithms were available in the works by Seroussi et al.[34] [Tables 4 and 5] and Yang et al.[35] [Table 1], respectively.

Table 1 | Attribution results

	Eng	Ita	Pol	Email	Blog1K	Blog
LDA-H	0.877[a]	0.830[a]	0.769[a]	0.426	0.216	0.079
CE	0.853	0.900	0.870	—	—	—
DADT-P	—	—	—	**0.594**	0.437	0.286
TDM	—	—	—	0.542	—	0.308
CP2D	0.913	0.929	**0.899**	0.521	**0.494**	**0.369**
δ-CP2D	**0.918**	**0.935**	0.838	0.558	**0.495**	**0.393**
CP2D^{TR}	0.924	0.927	0.949	0.497	0.489	0.358
δ-CP2D^{TR}	0.926	0.929	0.956	0.518	0.490	0.386

[a]Our implementation.

Numerical values of the average accuracy, depicted in Fig. 2, are here listed for all considered methods: our Constrained Probability 2-parameters Poisson-Dirichlet, in both its versions with and without including the parameter δ (the CP2D and δ-CP2D), the Cross-Entropy based approach (CE), the Latent Dirichlet Allocation plus Hellinger distance (LDA-H), the Disjoint Author-Document Topic model in its Probabilistic formulation (DADT-P), and the Topic Drift Model (TDM). In addition, we report the average accuracy obtained on the training sets (TR) by our method, as discussed in the text. We highlight in bold the best performance on each dataset.

In the case of informal corpora, we compare our method with state-of-the-art methods in the family of topic models[33]. Topic models are among the most established applications of nonparametric Bayesian techniques in natural language processing, and different authors' attribution methods rely on this approach. The underlying idea is to consider each document as a mixture of topics and to compute the similarity between two documents in terms of a measure of overlap between them (as detailed in the Methods section and the Supplementary Methods). Those methods were proposed to address challenging situations, particularly in informal corpora with many reference authors and typically short texts. Moreover, they have a similar ground to the method we propose. We consider the Latent Dirichlet Allocation plus Hellinger distance (LDA-H)[53], the Disjoint Author-Document Topic model in its Probabilistic version (DADT-P)[34] and the Topic Drift Model (TDM)[35] since their performances are available on the informal corpora. LDA-H is a straightforward application of topic models to the authorship attribution task. The DADT-P algorithm is a generalisation of the LDA-H characterising both the topics associated with texts and with authors. TDM merges topic models with machine learning methods[55,56] to account for dynamical correlations between words.

For the literary corpora, there is no direct comparison available in the literature. In the family of topic models, we considered the LDA-H approach, whose implementation is available with the need for minor intervention (please refer to the Supplementary Methods for details on our implementation). In addition, we consider a cross-entropy (CE) approach[57,58] in the implementation used in previous research[45]. Compression-based methods are general and powerful tools to assess similarity between symbolic sequences and have been at the forefront of authorship attribution for considerable time[59].

When comparing the aforementioned methods and ours, we optimise the free parameters of our model (i.e., δ, length of fragments, attribution criterion, type of tokens, and P_0) on the training set, as detailed in the Supplementary Methods. The email corpus already provides training, validation, and test sets. For the remaining corpora, we use ten-fold stratified cross-validation[34,53,54]: in turn, one-tenth of the dataset is treated as a test set and the other nine-tenths as training, and the number of samples per author is kept constant across the different folds. In Fig. 2, we report the accuracy obtained on each of the ten partitions, as well as the average value over them. We show the results obtained by either switching off the parameter δ (that is, by fixing it to 1) or optimising it on each specific corpus. The first scenario is denoted by CP2D (Constrained Probability 2-parameters Poisson-Dirichlet), the latter by δ-CP2D. The second procedure yields better performances in all the datasets except for the Polish literary dataset, where the number of texts per author is too low to prevent overfitting in this simple training setting. In the literary corpora, the attribution accuracy is overall high, and that of our method consistently higher than that of the other techniques. In the informal corpora, our method achieves an accuracy slightly lower than the best-performing algorithm on the email corpus, while it is the most accurate on the blog corpus. This latter corpus presents a very large number of candidate authors, and our approach appeared more robust in these extreme conditions. In Table 1, we present the numerical value of the average accuracy over the ten partitions, as shown in Fig. 2 (additional evaluation metrics can be found in the Supplementary Results). We also add the attribution accuracy on the training set. We observe that in the literary corpora, only in the Polish dataset, the accuracy on the test set is significantly lower than that in the training set, pointing to overfitting, as discussed above. For the informal corpora, we conversely notice an increase in attribution rate from the training to the test corpora. For the email corpus, also other methods exhibit a similar behaviour[34,54]. This is probably related to the particular partition considered. For the Blog corpus, the attribution accuracy on the test set is not available for the other methods. Our method features a slightly greater accuracy on the test set than on the training, suggesting that, on the one hand, the corpus is sufficiently large to prevent overfitting. On the other hand, the method increases accuracy when increasing the length of the reference authors' sequences.

Conclusion

We present a method for authorship attribution based on urn models for innovation processes. We interpret texts as instances of stochastic processes, where the generative stochastic process represents the author. The attribution relies on the posterior probability of the anonymous text being generated by a particular author and continuing their production. We consider the UMT model[1] in its exchangeable version[5,6], which is equivalent to the two-parameter Poisson-Dirichlet process. While the latter process is widely used in Bayesian nonparametric inference, it is often employed in a hierarchical formulation. In the case of attribution tasks, this approach has led to topic models, where the output of the stochastic process is a sequence of topics, i.e., distribution over words. Here, we follow a more direct approach, where the stochastic process directly generates words. By relying on a heuristic approach, we can explicitly write posterior probabilities that

can be computed exactly. Besides its computational convenience, the method we propose is easily adaptable to incorporate more realistic models for innovation processes.

For instance, one avenue we intend to explore in future research is leveraging the urn model with semantic triggering[1].

We evaluate the performance of our approach by employing the simple UMT exchangeable model against various related approaches in the field. Specifically, we compare it with information theory-based methods[45,57,58] and probabilistic methods based on topic models[34,35]. Our method achieves overall better or comparable performance in datasets with diverse characteristics, ranging from literary texts in different languages to informal texts.

We acknowledge that our method may not compete with deep learning-based models (DL) when large pre-training datasets are available[60,61]. Nonetheless, it exhibits robustness in challenging situations for DL, for example, when only a few texts are available for many authors[61] or in languages where pre-training is less extensive[62]. A deeper comparison with deep learning-based approaches, perhaps by concurrently exploring more sophisticated urn models in our approach, is in order but beyond the scope of the present work (refer to the Supplementary Results for a more detailed discussion and a preliminary analysis).

As a final remark, we also note that we have here considered the so-called closed-set attribution[32], where the training set contains part of the production of the author of the anonymous text. In open-set attribution[63,64], the anonymous text may be of an author for which no other samples are available in the dataset. Despite the conceptual differences and nuances between the two tasks, approaches based on closed-set attribution[64] are sometimes used also in open-set problems, for instance, by assigning the text to an unknown author if a measure of confidence falls below a given threshold. Similar strategies can be employed with our method by leveraging the conditional probabilities of documents.

We finally note that the method presented here is highly general and can be valuable beyond authorship attribution tasks. Although we expect it to be particularly suitable when elements take values from an open set and follow an empirical distribution close to that produced by the model, it can be applied to assess the similarity between any class of symbolic sequences.

Methods

UMT and PD processes

In[1], a family of urn models with infinitely many colours was proposed to reproduce shared statistical properties observed in real-world systems featuring innovations. In this context, a realisation of the process is a sequence $x^t = x_1, \ldots, x_t$ of extractions of coloured balls, where x_t is the colour of the element drawn at time t, and the space of colours available at a given time – the urn – represents the adjacent possible space. The urn model with triggering (UMT)[1] (and in a more general setting in refs. 5,6) operates as follows: the system evolves by drawing items from an urn initially containing a finite number N_0 of balls of distinct colours. At each time step t, a ball is randomly selected from the urn, its colour registered into the sequence, and returned to the urn. If the colour of the drawn ball is not in the sequence x^t, $\tilde{\rho}$ balls of the same colour and $\nu+1$ balls of entirely new colours, i.e., not yet present in the urn, are added to the urn. Thus, the occurrence of new events facilitates others by enlarging the set of potential novelties. Conversely, if the colour of the drawn ball already exists in x^t, ρ balls of the same colour are added to the urn. Given the history of extractions x^t, the probabilities b_t and $q_{c,t}$ that the drawing at time t results in a new colour or yields a colour c already present in x^t are easily specified for this model:

$$b_t = \frac{N_0 + \nu D_t}{N_0 + \rho t + a D_t}$$
$$q_{c,t} = \frac{\rho n_{c,t} + a - \nu}{N_0 + \rho t + a D_t} \tag{3}$$

where D_t and $n_{c,t}$ are the number of distinct colours and the number of extractions of colour c in the sequence x^t, respectively, and $a = \tilde{\rho} - \rho + \nu + 1$. Different choices of the parameters $(\rho, \tilde{\rho}, \nu)$ lead to different scenarios, enabling the UMT model to capture the empirical properties summarised by Heap's, Zipf's and Taylor's laws. In the original formulation[1], only two values for the parameter $\tilde{\rho}$ were discussed: $\tilde{\rho} = \rho$ or $\tilde{\rho} = 0$; the special setting $\tilde{\rho} = \rho - (\nu + 1)$, which makes the model exchangeable, was later pointed out[5]. We remind that exchangeability refers to the property that the probability of drawing any sequence $x^t \equiv x_1, \ldots, x_t$ of any finite length t does not depend on the order in which the elements occur: $P(x_1, \ldots, x_t) = P(x_{\pi(1)}, \ldots, x_{\pi(t)})$ for each permutation π and each sequence length t. In this case, upon a proper redefinition of the parameters, namely $\nu/\rho \equiv \alpha$ and $N_0/\rho \equiv \theta$, the UMT model reproduces the conditional probabilities associated with the PD process (expressed in Eqs. (1)). We note here that such probabilities include the Dirichlet process as a special case, where $\alpha = 0$ and D_t grows logarithmically with t. In the framework of urn models, the Dirichlet process finds its counterpart in the Hoppe model[65] and in the exchangeable version of the UMT model with the additional choice $\nu = 0$. The PD process is defined by $0 < \alpha < 1$ and predicts the asymptotic behaviour $D_t \sim t^{\alpha}$ [46]. We note that the probabilities in Eqs. (3) coincide, when renaming the parameters as stated above, with those in Eq. (1).

The strategy for P_0

When P_0 is a discrete probability distribution (it has atoms), an already seen value y can be drawn again from it, and the conditional probabilities no longer have the simple form as in Eq. (1). In this case, the conditional probabilities depend not only on the sequence x^t of observable values but also on latent variables indicating, for each element in x^t, whether it has been drawn from P_0 or arose from the reinforcement process[66]. In particular, we can define, for each type y_i $(i = 1, \ldots, D_t)$ in x^t, a latent variable $\lambda_{i,t}$ that counts the number of times y_i is drawn from the base distribution P_0. The probabilities conditioned on the observable sequence x^t and on the latent variables sequence λ^{D_t} read:

$$\begin{aligned} P(x_{t+1} = y | x^t, \lambda^{D_t}) &= \frac{n_{y,t} - \lambda_{i,t}\alpha}{\theta + t} + \frac{\theta + \alpha\Lambda_t}{\theta + t} P_0(y) \quad \text{if} \quad n_{y,t} > 0 \\ P(x_{t+1} = y | x^t, \lambda^{D_t}) &= \frac{\theta + \alpha\Lambda_t}{\theta + t} P_0(y) \quad \text{if} \quad n_{y,t} = 0 \end{aligned} \tag{4}$$

Where $\Lambda_t \equiv \sum_i^{D_t} \lambda_{i,t}$ is the total number of extractions from P_0 till time t. To compute the probabilities conditioned to the observable sequence x^t, we must integrate out the latent variables. This is an exponentially hard problem and efficient sampling algorithms[33,36,37] have been developed for an approximate solution.

By taking the perspective of the urn model, we investigate the possibility of bypassing the problem by imposing that each element can be extracted only once from $P_0(\cdot)$, which is equivalent to fixing all the latent variables $\lambda_{i,t} = 1$ and set to zero the last term in the first equation in Eq. (4).

The latter procedure effectively replaces $P_0(y)$ with a history-dependent probability, normalised at each time over all the elements y not already appeared in x^t. It reads:

$$P_0^t(y) \equiv P_0(y | y \notin x^t) = \begin{cases} \frac{P_0(y)}{1 - \sum_{\bar{y} \in x^t} P_0(\bar{y})} & \text{if } y \notin x^t \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where the sum is over all the elements already drawn at time t. Note that this choice breaks the exchangeability of the process with respect to the order in which novel elements are introduced. In the implementation of our algorithm, we follow an even simpler and fast procedure, which yielded equivalent results. We simply introduce an author-dependent base distribution by considering, for each author A, the frequency of the tokens that do not appear in A. Such procedure translates into replacing $P_0(y)$ with $P_0^{(A)}(y) = \frac{P_0(y)}{P(A^C)}$, where A^C denotes the set of all distinct tokens that do not appear in A. This author-dependent base distribution proved to be preferable to simply using the original frequency, especially in datasets with short texts and few samples for each author.

LDA and topic models

LDA is a generative probabilistic model[67], which generates corpora of documents. A document is a finite sequence of words $w_1, w_2, \ldots, w_N$ and it is represented as a random mixture over latent topics. Each topic corresponds to a categorical probability distribution over the set of all possible words. Topics can be shared by different documents. The total number k of topics is fixed a priori and to each topic i in each document d is associated a probability $\theta_{i,d}$, extracted independently for each document from a k-dimensional Dirichlet distribution $D(\alpha_1, \ldots, \alpha_k)$. Each document d is generated as follows: first, its length N_d is extracted from a Poissonian distribution with a given mean. Then, the document is populated with words using the following procedure: a topic i is extracted with probability $\theta_{i,d}$ and a word w is extracted from i with the probability associated to it in topic i. The probabilities $p_i(w)$ of a word w in the topic i is in turn extracted independently from a W-dimensional Dirichlet distribution $D(\beta_1, \ldots, \beta_W)$, where W is the total number of words W in the corpus.

As in Eqs. (4), we can introduce latent variables[67], now with a different meaning. To each word $w_{i,d}$ in document d, $i = 1, \ldots, N_d$, we associate a latent variable $\lambda_{i,d}$ that is the identifier of the topic j from which the word $w_{i,d}$ is extracted. The joint distribution of the sequence of words $w^{N_d} \equiv w_{1,d}, \ldots, w_{N_d,d}$ and latent variables $\lambda^{N_d} \equiv \lambda_{1,d}, \ldots, \lambda_{N_d,d}$ in a document d thus read:

$$P(w^{N_d}, \lambda^{N_d}) = \prod_{n=1}^{N_d} p(w_{i,d}|\lambda_{1,d}) p(\lambda_{i,d}) \tag{6}$$

where $p(\lambda_{i,d}) \equiv \theta_{i,d}$. To compute the posterior probability of the observable sequence w^{N_d} we must integrate out the latent variables. This is an exponentially hard problem and is solved with methods for numerical approximation by using, for instance, Markov Chain Monte Carlo algorithms. A more flexible approach is to use the Dirichlet or PD processes instead of the Dirichlet distributions over topics. This allows the number of topics k to remain unspecified a priori.

The probabilities $\theta_{i,d}$ are the elements of a sequence generated by a Dirichlet or PD process, for each document d. The processes characterising each document share the same discrete base distribution, which is, in turn, generated by a Dirichlet or PD process with a non-atomic P_0. Again, efficient sampling algorithms for computing the posterior distributions[33,36,37] have been developed in this framework.

In the framework of authorship attribution, methods relying on LDA are more widely adopted than those based on the Dirichlet or PD processes, primarily due to their simplicity and comparable accuracy[34].

The procedure followed by the LDA-H algorithm to address the author attribution task is described in the Supplementary Methods.

Data availability

The corpora used to validate our approach, with the exception of the Italian literature, are available online at the following addresses: Blog corpus (https://u.cs.biu.ac.il/~koppel/BlogCorpus.htm), PAN'11 email corpus (https://doi.org/10.5281/zenodo.3713246), Polish literature (https://github.com/computationalstylistics/100_polish_novels), English literature (https://github.com/GiulioTani/InnovationProcessesInference/tree/main/sample_data/English_literature). The Italian literature corpus is currently covered by copyright, we make accessible the list of included titles (https://github.com/GiulioTani/InnovationProcessesInference/tree/main/sample_data/Italian_literature). Please refer to the Supplementary Note 1 for details about which parts of the public datasets where used in this study.

Code availability

All the code we used to compute all the attributions with the CP2D approach is publicly available under the GNU GPL v3.0 license at https://github.com/GiulioTani/InnovationProcessesInference[68].

Received: 27 June 2023; Accepted: 24 June 2024;
Published online: 06 September 2024

References

1. Tria, F., Loreto, V., Servedio, V. & Strogatz, S. The dynamics of correlated novelties. *Sci. Rep.* **4**, 1–8 (2014).
2. Heaps, H. S. *Information Retrieval, Computational And Theoretical Aspects* (Academic Press, 1978).
3. Taylor, L. Aggregation, variance and the mean. *Nature* **189**, 732 (1961).
4. Gerlach, M. & Altmann, E. G. Scaling laws and fluctuations in the statistics of word frequencies. *N. J. Phys.* **16**, 113010 (2014).
5. Tria, F., Loreto, V. & Servedio, V. Zipf's, heaps' and taylor's laws are determined by the expansion into the adjacent possible. *Entropy* **20**, 752 (2018).
6. Tria, F., Crimaldi, I., Aletti, G. & Servedio, V. D. P. Taylor's law in innovation processes. *Entropy* **22**, 573 (2020).
7. Zipf, G. K. *The Psychobiology of Language* (Houghton-Mifflin, 1935).
8. Moreno-Sánchez, I., Font-Clos, F. & Corral, Á. Large-scale analysis of zipf's law in english texts. *PLoS ONE* **11**, e0147073 (2016).
9. Kauffman, S. A. *Investigations* (Oxford University Press, 2000).
10. Pólya, G. Sur quelques points de la théorie des probabilités. *Ann. de. l'I. H. P.* **1**, 117–161 (1930).
11. Ubaldi, E., Burioni, R., Loreto, V. & Tria, F. Emergence and evolution of social networks through exploration of the adjacent possible space. *Commun. Phys.* **4**, 28 (2021).
12. Monechi, B., Ruiz-Serrano, Ã., Tria, F. & Loreto, V. Waves of novelties in the expansion into the adjacent possible. *PLoS ONE* **12**, e0179303 (2017).
13. Aletti, G., Crimaldi, I. & Ghiglietti, A. Interacting innovation processes. *Sci. Rep.* **13**, 17187 (2023).
14. Iacopini, I., Milojević, Scv & Latora, V. Network dynamics of innovation processes. *Phys. Rev. Lett.* **120**, 048301 (2018).
15. Iacopini, I., Di Bona, G., Ubaldi, E., Loreto, V. & Latora, V. Interacting discovery processes on complex networks. *Phys. Rev. Lett.* **125**, 248301 (2020).
16. Di Bona, G. et al. Social interactions affect discovery processes. *arXiv* preprint arXiv:2202.05099 (2022).
17. Di Bona, G. et al. The dynamics of higher-order novelties. *arXiv* preprint arXiv:2307.06147 (2023).
18. Lijoi, A., Mena, R. H. & Prünster, I. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786 (2007).
19. Favaro, S., Lijoi, A., Mena, R. H. & Prünster, I. Bayesian non-parametric inference for species variety with a two-parameter poisson–dirichlet process prior. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **71**, 993–1008 (2009).
20. Chakraborty, S., Arora, A., Begg, C. B. & Shen, R. Using somatic variant richness to mine signals from rare variants in the cancer genome. *Nat. Commun.* **10**, 5506 (2019).
21. Holec, P. V., Berleant, J., Bathe, M. & Birnbaum, M. E. A bayesian framework for high-throughput t cell receptor pairing. *Bioinformatics* **35**, 1318–1325 (2019).
22. Masoero, L., Camerlenghi, F., Favaro, S. & Broderick, T. More for less: predicting and maximizing genomic variant discovery via bayesian nonparametrics. *Biometrika* **109**, 17–32 (2022).
23. Gershman, S. J. & Blei, D. M. A tutorial on bayesian nonparametric models. *J. Math. Psychol.* **56**, 1–12 (2012).
24. Ni, Y. et al. Scalable bayesian nonparametric clustering and classification. *J. Comput. Graph. Stat.* **29**, 53–65 (2020).
25. Schmidt, M. N. & Morup, M. Nonparametric bayesian modeling of complex networks: an introduction. *IEEE Signal Process. Mag.* **30**, 110–128 (2013).

26. Hu, L., Chan, K. C., Yuan, X. & Xiong, S. A variational bayesian framework for cluster analysis in a complex network. *IEEE Trans. Knowl. Data Eng.* **32**, 2115–2128 (2019).
27. Teh, Y. W. & Jordan, M. I. Hierarchical bayesian nonparametric models with applications. *Bayesian Nonparametric* **1**, 158–207 (2010).
28. Blei, D. M., Griffiths, T. L. & Jordan, M. I. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM* **57**, 1–30 (2010).
29. Pitman, J. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900 (1997).
30. Ferguson, T. S. A bayesian analysis of some non-parametric problems. *Ann. Stat.* **1**, 353–355 (1973).
31. De Blasi, P. et al. Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 212–229 (2013).
32. Fadel, A. et al. *Overview of the PAN@FIRE 2020 Task on the Authorship Identification of SOurce COde*. 4–8 (ACM, 2020).
33. Blei, D. M. Probabilistic topic models. *IEEE Signal Process. Mag.* **27**, 55–65 (2010).
34. Seroussi, Y., Zukerman, I. & Bohnert, F. Authorship attribution with topic models. *Comput. Linguist.* **40**, 269–310 (2014).
35. Yang, M., Zhu, D., Tang, Y. & Wang, J. Authorship attribution with topic drift model. In *Thirty-First AAAI Conference on Artificial Intelligence* (2017).
36. Teh, Y., Newman, D. & Welling, M. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *Advances in neural information processing systems.* Vol. 19 (2006).
37. Porteous, I. et al. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 569–577 (2008).
38. de Finetti, B. *Annales de l'institut Henri Poincaré*. Vol. **7**, p. 1–68 (1937).
39. Kingman, J. F. C. Random partitions in population genetics. *Proc. R. Soc.* **361**, 1–18 (1978).
40. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
41. Cover, T. M. & Thomas, J. A. *Elements of Information Theory.* 2nd edn (Wiley-Interscience, 2006).
42. Stamatatos, E. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.* **60**, 538–556 (2009).
43. Koppel, M., Schler, J. & Argamon, S. Authorship attribution in the wild. *Lang. Resour. Eval.* **45**, 83–94 (2011).
44. Ziv, J. & Lempel, A. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory* **23**, 337–343 (1977).
45. Lalli, M., Tria, F. & Loreto, V. *Drawing Elena Ferrante's Profile: Workshop Proceedings, Padova, 7 September 2017* (eds. Tuzzi, A. & Cortelazzo, M. A.) (Padova UP, 2018).
46. Pitman, J. *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002* (Springer, 2006).
47. Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. Hierarchical dirichlet processes. *J. Am. Stat. Assoc.* **101**, 1566–1581 (2006).
48. Teh, Y. W. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, 985?992 (Association for Computational Linguistics, 2006). https://doi.org/10.3115/1220175.1220299.
49. Rybicki, J. & Eder, M. Deeper Delta across genres and languages: do we really need the most frequent words? *Lit. Linguist. Comput.* **26**, 315–321 (2011).
50. Argamon, S. & Juola, P. *Overview of the International Authorship Identification Competition at PAN-2011* (2011).
51. Saedi, C. & Dras, M. Siamese networks for large-scale author identification. *Comput. Speech Lang.* **70**, 101241 (2021).
52. Schler, J., Koppel, M., Argamon, S. & Pennebaker, J. W. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, vol. 6, 199–205 (2006).
53. Seroussi, Y., Zukerman, I. & Bohnert, F. Authorship attribution with latent Dirichlet allocation. In *Proceedings of the fifteenth conference on computational natural language learning*, 181–189 (2011).
54. Seroussi, Y., Bohnert, F. & Zukerman, I. Authorship attribution with author-aware topic models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 264–269 (2012).
55. Yang, M., Mei, J., Xu, F., Tu, W. & Lu, Z. Discovering author interest evolution in topic modeling. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 801–804 (2016).
56. Mnih, A. & Kavukcuoglu, K. Learning word embeddings efficiently with noise-contrastive estimation. In *Proc. Advances In Neural Information Processing Systems.* Vol. 26 (2013).
57. Benedetto, D., Caglioti, E. & Loreto, V. Language trees and zipping. *Phys. Rev. Lett.* **88**, 048702 (2002).
58. Baronchelli, A., Caglioti, E. & Loreto, V. Artificial sequences and complexity measures. *J. Stat. Mech.: Theory Exp.* **2005**, P04002 (2005).
59. Neal, T. et al. Surveying stylometry techniques and applications. *ACM Comput. Surv.* **50**, 1–36 (2017).
60. Fabien, M., Villatoro-Tello, E., Motlicek, P. & Parida, S. *BertAA : BERT fine-tuning for Authorship Attribution*. p. 127–137 (2020).
61. Bauersfeld, L., Romero, A., Muglikar, M. & Scaramuzza, D. Cracking double-blind review: authorship attribution with deep learning. *PLoS ONE* **18**, e0287611 (2023).
62. Romanov, A., Kurtukova, A., Shelupanov, A., Fedotova, A. & Goncharov, V. Authorship identification of a russian-language text using support vector machine and deep neural networks. *Future Internet* https://www.mdpi.com/1999-5903/13/1/3 (2021).
63. Kestemont, M. et al. *Working Notes Papers of the CLEF 2019 Evaluation Labs*, vol. 2380 of *CEUR Workshop Proceedings* (eds. Cappellato, L., Ferro, N., Losada, D. & Müller, H.) https://ceur-ws.org/Vol-2380/paper_264.pdf (2019).
64. Stamatatos, E. et al. *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, vol. 3497 of *CEUR Workshop Proceedings*, 2476–2491 (eds. Aliannejadi, M., Faggioli, G., Ferro, N. & Vlachos, M.) https://ceur-ws.org/Vol-3497/paper-199.pdf (2023).
65. Hoppe, F. M. Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.* **20**, 91–94 (1984).
66. Buntine, W. & Hutter, M. A. *Bayesian view of the Poisson-Dirichlet Process*. Tech. Rep. arXiv:1007.0296, NICTA and ANU http://arxiv.org/abs/1007.0296 (2010).
67. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
68. Tani Raffaelli, G., Lalli, M. & Tria, F. GiulioTani/InnovationProcessesInference: Accepted (Zenodo, 2024), https://doi.org/10.5281/zenodo.12163218.

Acknowledgements

M.L. acknowledges support by the European Project: XAI: Science and technology for the explanation of AI decision making (https://xai-project.eu/), ERC-2018-ADG G.A. 834756. G.T.R. was supported by the Czech Science Foundation project No. 21-17211S. This work has been realised in the framework of the agreement between Sapienza University of Rome and the Sony Computer Science Laboratories.

Author contributions

F.T. conceived and designed research; F.T. and M.L. performed a preliminary investigation, collected in M.L. master thesis. M.L. and G.T.R. collected the datasets. G.T.R. wrote the code and performed the numerical calculations. F.T. and G.T.R. analysed the data, refined the model and discussed the results. All authors contributed to writing the paper and revising the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-024-01714-6.

Correspondence and requests for materials should be addressed to Francesca Tria.

Peer review information *Communications Physics* thanks Gabriele Di Bona, Liubov Tubikina and Tao Jia for their contribution to the peer review of this work. A peer review file is available.