

RESEARCH ARTICLE

Upscaling human activity data: A statistical ecology approach

Anna Tovo^{1,2*}, Samuele Stivanello², Amos Maritan¹, Samir Suweis^{1,3}, Stefano Favaro⁴, Marco Formentin^{1,3*}

1 Dipartimento di Fisica e Astronomia “Galileo Galilei”, Istituto Nazionale di Fisica Nucleare, Università degli Studi di Padova, Padova, Italy, **2** Dipartimento di Matematica “Tullio Levi-Civita”, Università degli Studi di Padova, Padova, Italy, **3** Padova Neuroscience Center, Università degli Studi di Padova, Padova, Italy, **4** Dipartimento di Scienze Economico-Sociali e Matematico-Statistiche”, Università degli Studi di Torino, Torino, Italy

* anna.tovo@unipd.it (AT); marco.formentin@unipd.it (MF)



OPEN ACCESS

Citation: Tovo A, Stivanello S, Maritan A, Suweis S, Favaro S, Formentin M (2021) Upscaling human activity data: A statistical ecology approach. PLoS ONE 16(7): e0253461. <https://doi.org/10.1371/journal.pone.0253461>

Editor: Sandro Meloni, Instituto de Fisica Interdisciplinar y Sistemas Complejos, SPAIN

Received: January 9, 2021

Accepted: June 4, 2021

Published: July 1, 2021

Copyright: © 2021 Tovo et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data underlying the results presented in the study are available from - Formentin, M., Lovison, A., Maritan, A., and Zanzotto, G. (2014). Hidden scaling patterns and universality in written communication. *Physical Review E* 90, 012817 - Monechi, B., Ruiz-Serrano, A., Tria, F., and Loreto, V. (2017). Waves of novelties in the expansion into the adjacent possible. *PloS one* 12.

Funding: A. Tovo acknowledges financial support from neXt grant, Department of Mathematics “Tullio Levi-Civita” of University of Padova. S.

Abstract

Big data require new techniques to handle the information they come with. Here we consider four datasets (email communication, Twitter posts, Wikipedia articles and Gutenberg books) and propose a novel statistical framework to predict global statistics from random samples. More precisely, we infer the number of senders, hashtags and words of the whole dataset and how their abundances (i.e. the popularity of a hashtag) change through scales from a small sample of sent emails per sender, posts per hashtag and word occurrences. Our approach is grounded on statistical ecology as we map inference of human activities into the unseen species problem in biodiversity. Our findings may have applications to resource management in emails, collective attention monitoring in Twitter and language learning process in word databases.

Introduction

In ecology one of the most studied emerging patterns is the *Relative Species Abundance* (RSA), that gives the fraction of species with the same number of individuals. To determine large scale RSA features from the distribution of species abundances within a small random sample is a major challenge in ecology and through years plenty of methods have been proposed [1–7]. The success of such methods depends on the following notable fact: different ecosystems like tropical forests or coral reefs, despite their disparate locations and different evolutionary history, share a common shape of the empirical RSA [7–10]. Recently in [8, 11], it has been developed a rigorous statistical framework to predict global scale biodiversity from scattered local plots. The core of the method is that the empirical RSA patterns can be well described by negative binomials, a distribution which has a theoretical foundation within the framework of neutral theory of ecology [9, 12]. In [8, 11], authors show that the mathematical properties of the negative binomial allow to estimate the number of species populating an ecosystem from knowing species abundances in a small portion of that ecosystem. The framework in [8, 11] also links RSAs to another informative pattern studied in theoretical ecology, the *Species*

Suweis and A. Tovo acknowledge STARS grant 2019 from University of Padova. S. Stivanello acknowledges financial support from Progetto Dottorati - Fondazione Cassa di Risparmio di Padova e Rovigo. A. Tovo and A. Maritan acknowledge the support from University of Padova through "Excellence Project 2018" of the Cariparo foundation. S. Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. S. Favaro gratefully acknowledges the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Accumulation Curve (SAC), which describes how the number of different species increases with the population size. In general, a log-series like RSA implies that the SAC is a logarithmic function of the population size [7, 10].

In the present paper we adopted and extended the statistical framework presented in [8, 11] and firstly designed in ecology to get new insights into human activity databases (i.e. email communication, Twitter posts, Wikipedia articles and Gutenberg books) with the aim of inferring global statistics of a dataset from a random sample of it. Statistical regularities have been widely observed in many different contexts of human dynamics—for instance, Zipf's laws have been observed since decades in computational linguistic (see [13, 14] for a review)—and a variety of models have been proposed to explain such recurrent patterns [15–29]. In particular, the analogies between the systems here considered and ecological systems have been of interest for researchers and some modeling approaches move from an ecological perspective. For example, the analogy between the diffusion of online social media content and ecology has been investigated in [30–33]. The number of retweets of an hashtag may be interpreted as the offspring and one may think that the number of retweets depends on the attention a given hashtag is capable to gain. In this view, attention is the resource hashtags are competing for and lack of attention may bring to extinction. As far as the connection between words in books and ecological systems we can notice the following. Written language is (at least in first approximation) the account of spoken/natural language and like species, a language can split into several languages, it can mutate by modifying words/expressions over time, and it can face extinction. This connection has been explored intensively in the last decades. Many approaches to model language dynamics draw fully from ecology, for instance introducing the fitness of a language and competition between languages and words. A review of these research lines may be found in [34].

In the present work we focus on inference, not modeling. In particular, we consider four human activities with the following correspondence between species and individuals within each dataset (see Fig 1): (1) *Email communication* [35, 36]: here we set the sender identity to label a species and the number of sent emails to be the number of individuals pertaining to a species; (2) *Twitter posts* [37]: here hashtags play the role of species and the number of different tweets containing a certain hashtag represents its population size; (3) For *Wikipedia articles* [37] and (4) *Gutenberg books* [37] we use the following setting: each word is a different species while its abundance is given by the number of occurrences of the word in the dataset. Once defined what corresponds to species and individuals, the RSA of each dataset displays a negative binomial behavior (see S1.1 Section in S1 Appendix). This permits to apply a statistical ecology approach to upscale human activity data. Within this framework, we propose a statistical method:

1. that gives reliable estimates for the number of users, hashtags, and words from a random sample of mails, posts and word occurrences. We refer to the inference of global quantities of interest from random samples as *upscaling* (see Fig 2). Moreover, our framework predicts how the number of users/hashtags/words grows with the recorded activity (mails/posts/pages/books): the rate at which new elements appear (i.e. the analogous of the SAC for human-activity data) shows a sublinear power-law growth, signature of the Heap's law (see Fig 1).
2. We infer how the abundance of a species may change across scales (see Results section). This for example means that, observing a small portion of tweets and the popularity of a given hashtag among them, we can predict whether it will remain popular or not in the unseen part of the network.

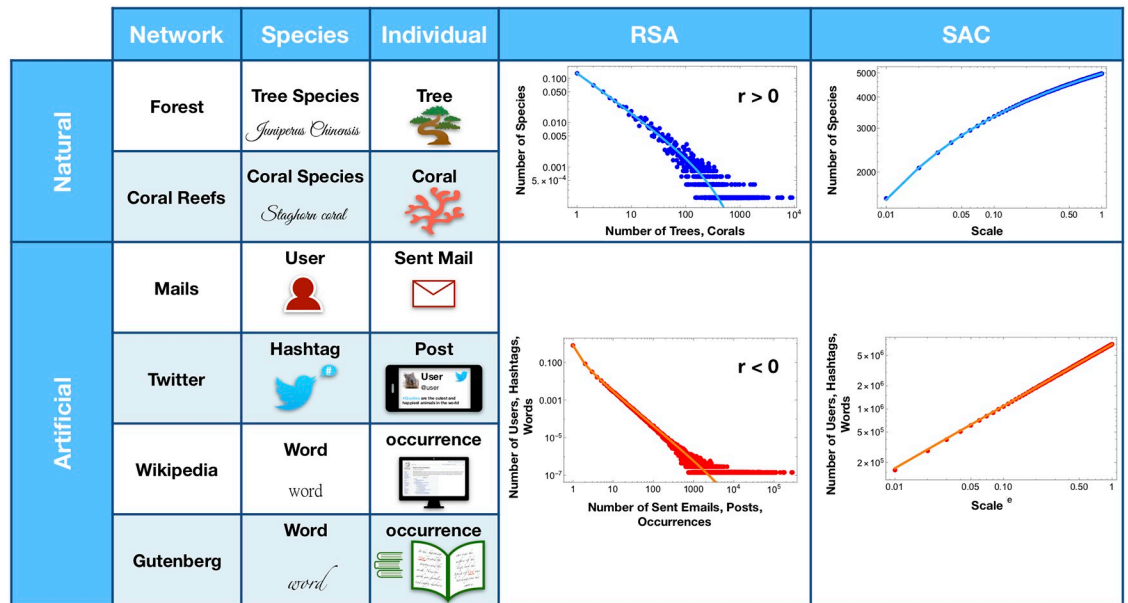


Fig 1. From ecology to human activities. The figure depicts the correspondence between species/individuals in a natural ecosystem and users/sent emails, hashtags/posts, words/occurrences in each one of the four datasets considered in the paper. Once the proper correspondence is established, it turns out that both natural and artificial RSAs can be well described by a negative binomial distribution. In the latest two columns, in order to show the typical shapes of the RSA (i.e. Relative Species Abundance) and SAC (i.e. Species Accumulation Curve) curves for natural versus human activity systems, we display the empirical patterns obtained for the Amazonia rainforest [8] and Twitter dataset. In general, all human activity RSA curves can be accommodated by with a negative value of clustering coefficient r in the interval $(-1, 0)$ (see Eq 1), whereas natural ecosystems prefer $r > 0$ (solid lines).

<https://doi.org/10.1371/journal.pone.0253461.g001>

In our statistical model we make the hypothesis the RSA distribution of the four human activity datasets can be described by a negative binomial with a clustering coefficient in the range $(-1, 0)$ (see the results of Kolmogorov-Smirnov tests in S1 Table and S1.1 Section in S1 Appendix). This choice is justified by the heavy tail of the observed RSAs and the fact that the tail exponents are invariant under random sampling. Indeed, the negative binomial is (to our knowledge) the simplest *form-invariant* distribution. Form-invariance should not be confused with scale-invariance, a property only satisfied by power-laws (see S1.3 Section in S1 Appendix). With form-invariance we mean that when a portion of individuals are randomly sampled, the resulting RSA is still negative-binomially distributed with a heavy tail showing the same exponent as of the whole dataset (see S1.3 Section in S1 Appendix). Form-invariance property of the RSA allows us to build reliable estimators for the number of new features (new email users, new hashtags, new words) at each scale of the dataset starting from random samples of the whole databases. Our approach brings two main novelties/advantages. First, the choice to model the distribution of the occurrence frequencies according to a negative binomial distribution, supported by Kolmogorov-Smirnov tests (see S1.1 Section in S1 Appendix) is new. In particular, its form-invariance property allows us to obtain, for the quantities of interest, effective yet simple estimators which explicitly depend on the scale. In linguistic different parametric and non parametric statistical models has been used to infer how the number of types grows as new samples are added [13]. Instead, to our knowledge, upscaling has never been investigated for email communication and Twitter datasets. Second, within our framework we also derive an estimator for how the type abundances change across scales. This problem, as far as we know, has not been previously investigated although it could be of interest when interpreting abundances of types as a measure of popularity in social network data.

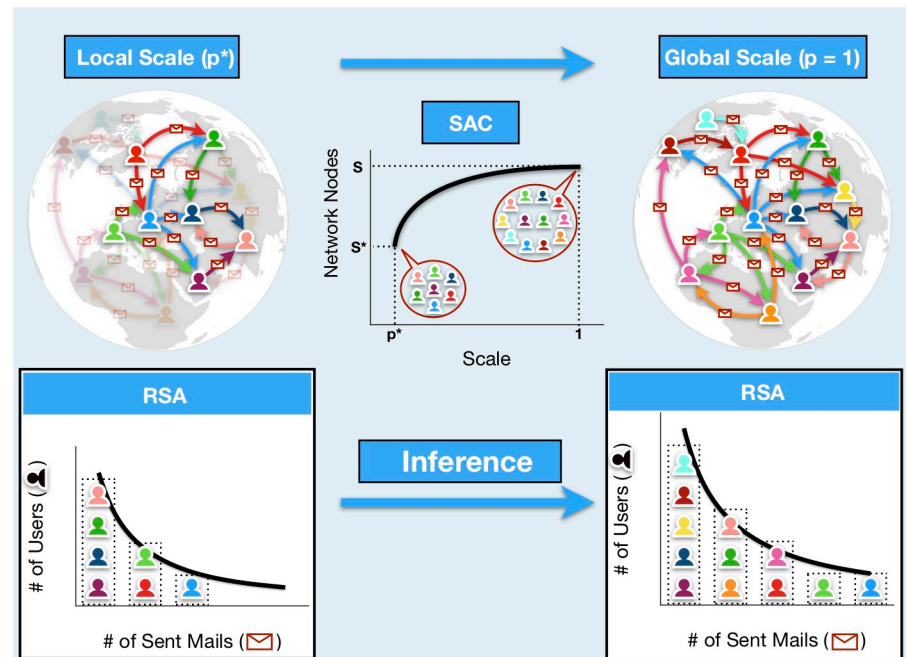


Fig 2. Sketch of our theoretical framework. Consider the email senders' network where each node is a sender and a directed link from node A to node B is an email issued from user A to user B . We set the identity of a sender to be the species and a sent email to be an individual of that species. For instance, if the user A has sent n emails, then the species A has n individuals. If an observer has access to a fraction p^* of the sent emails, s/he can partially recover the network (top-left) and the RSA curve at the local scale p (bottom-left). Within our framework, this information suffices to infer the number of species and the RSA curve at the global scale $p = 1$ (bottom-right). In terms of the network, the number of species corresponds to the number of users or nodes and the RSA gives the degree statistics. In this sense, our method reveals network features pertaining to the whole community activity initially unknown to the observer (top-right). Moreover, we can predict how the number of users increases with the number of links recorded, (i.e. the SAC curve in ecology), an information that may be used to optimize network design forecasting its growth.

<https://doi.org/10.1371/journal.pone.0253461.g002>

Results

To start with, we illustrate our approach, its potentiality and the kind of results it can provide as applied to e-mail communication. We consider the senders activity network where each node is a user and a directed link from node A to node B represents an email issued from user A to user B . We set the identity of a sender to label the species and the number of sent emails to be the individuals pertaining to a species. Thus, for instance, if user A has sent n emails we say that species A has n individuals. Suppose an observer have access to a small sample of sent emails, or, equivalently, to partial information on links and nodes of the email communication network. Our approach allows to infer the number of nodes (i.e. the number of users) and the link statistics of the whole network, thus revealing features previously unknown to the observer (see Fig 2).

Correspondence between species/individuals and human activities can be set similarly for the remaining datasets (see Fig 1). Our statistical ecology approach gives the following results:

- **RSA universality and form-invariance.** In each activity the RSA of the whole dataset turns out to be heavy tailed with an exponent between -1.8 and -1.4 (see Fig 3). Moreover, this exponent is maintained at different scales (see S1 and S2 Figs in S1 Appendix), supporting our choice of modeling the RSA by means of a negative binomial, which is form-invariant and keeps fixed the tail exponent through scales.

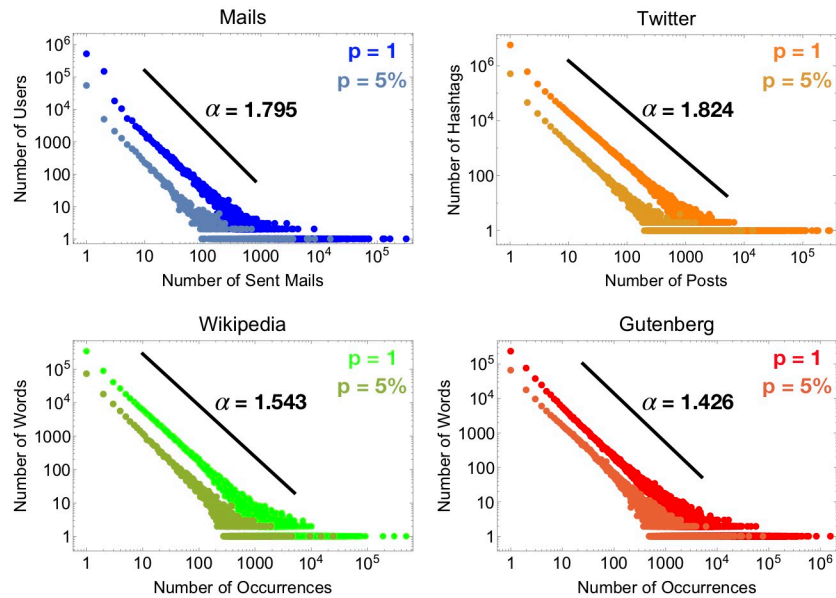


Fig 3. Universality and form-invariance of the empirical RSAs. Empirical RSA curves at the global scale ($p = 1$) and the local scale ($p = 5\%$) are shown. The patterns show a heavy-tailed form maintained through the different human activities and scales. This scale-invariance property of the RSAs allows for a successful implementation of our theoretical framework. In particular, our model predicts that the heavy-tail exponent α is related to the clustering parameter r of the RSA negative binomial distribution via the relation $\alpha = 1 - r$ (see [Materials and methods](#) and S1.4 Section in [S1 Appendix](#)). In each plot, for a visual inspection, we inserted a black line with slope $-\alpha = -1 + \hat{r}$, where \hat{r} have been obtained by fitting the local patterns at $p = 5\%$ through a negative binomial (see also [Table 1](#)). We can see that such lines describe very well the heavy-tail regime of the empirical RSAs at both local and global scale in all four cases. For the RSA fitting curves and predicted patterns, see S1 Fig in [S1 Appendix](#).

<https://doi.org/10.1371/journal.pone.0253461.g003>

- Inference of unseen human activities.** On the scale invariance property of the RSA we build a statistical framework which gives robust and accurate estimates for the number of email senders, Twitter hashtags, words of Wikipedia pages and Gutenberg books from a random sample of sent mails, posts and word occurrences (see [Table 1](#)). Moreover, our framework allows to reconstruct the growth of the number of users/hashtags/words with the recorded activity (mails/posts/pages/books), which represents another well-known pattern in ecological theory called the *Species-Accumulation Curve* (SAC) (see [Fig 2](#)).
- Popularity in social networks.** In Twitter and in social networks in general, popularity is known to be relevant, for instance, to manipulate mass opinion or to share information. One naive way to measure the popularity of a hashtag is to count the number of times it appears

Table 1. Predicted relative errors. Upscaling results for the number of species of the four analysed datasets from local samples covering a fraction $p^* = 5\%$ of the corresponding global dataset. For each human activity, we display the number of species (users, hashtags, words) and individuals (sent mails, posts, occurrences) at the global scale together with the average fitted RSA distribution parameters at the sampled scale and the relative percentage error (mean and standard deviation among 100 trials) between the true number of species and the one predicted by our framework. See S1 Fig in [S1 Appendix](#) for the corresponding fitting curves and predicted global RSA patterns.

	Emails	Twitter	Wikipedia	Gutenberg
Species	752, 299	6, 972, 453	673, 872	554, 193
Individuals	6, 914, 872	34, 696, 973	29, 606, 116	126, 289, 661
r	-0.796	-0.824	-0.543	-0.425
ξ_{p^*}	0.9999	0.9991	0.9985	0.9997
Relative Error	$0.45 \pm 0.45\%$	$3.38 \pm 0.16\%$	$6.09 \pm 0.38\%$	$-2.44 \pm 0.32\%$

<https://doi.org/10.1371/journal.pone.0253461.t001>

in other users' tweets. In our ecological interpretation, a hashtag represents a species, while the number of posts associated to it, gives the species' abundance. Within our framework, we can infer how the abundance of a species changes across scales (see Results section), thus allowing to monitor whether a locally popular hashtag will remain popular also in the undetected part of the network or not.

In the following we give the key steps of our upscaling framework. Denote with N the population size and with S the number of species (i.e. senders, hashtags, words) of the whole database. Given a scale $p^* \in (0, 1)$, consider a random sample of size $p^* N$ in which we recover $S_{p^*} \leq S$ species. In the sequel we denote by $P(n|p^*)$ the fraction of species with n individuals at scale p^* , i.e. the sample RSA. We assume that, at the global scale $p = 1$, $P(n|1)$ is proportional to a negative binomial distribution, $\mathcal{P}(n|r, \xi)$, with parameters $r \in (-1, +\infty) \setminus \{0\}$ and $\xi \in (0, 1)$:

$$P(n|1) = c(r, \xi)\mathcal{P}(n|r, \xi) \quad \text{for } n \geq 1 \tag{1}$$

where the normalizing factor $c(r, \xi) = 1/(1 - (1 - \xi)^r)$ takes into account that each of the S species consists of at least one individual at the global scale.

RSAs given in (1) have the following features: 1) values of $r \in (-1, 0)$ and ξ close to 1 reflect in a heavy-tailed behavior of the RSAs with an exponential cut-off. More precisely, the right tail of (1) has the form $n^{r-1} \exp(n \log \xi)$ (see S1.4 Section in S1 Appendix), where the exponential cut-off is "small" for ξ close to one and disappears in the limit $\xi \rightarrow 1$, where distribution (1) describes a pure power-law. Such heavy-tailed behavior with an exponential cut-off well describes the observed RSA patterns in human activities (see S1 Fig and S1.1 Section in S1 Appendix). Moreover, the exponent $\alpha = 1 - r$ matches very well with the empirical data (see also Fig 3).

2) Distribution (1) is *form-invariant*, meaning that the RSA $P(n|p)$ maintains the same functional form at different scales p (see S1.3 Section in S1 Appendix), a property observed in the empirical RSAs of all the four databases (see Fig 3). In mathematical terms, the RSA at any scale p is again proportional to a negative binomial distribution with the same r and a rescaled parameter

$$\xi_p = p\xi/(1 - \xi(1 - p)). \tag{2}$$

Properties 1) and 2) are the building blocks of our predictive statistical framework.

Our goal is to infer the total amount of species S (senders, hashtags, words) present in the entire database given the number of species S_{p^*} observed in a sample at the local scale p^* and their corresponding abundance (number of mails, posts, occurrences). From this limited information, we can construct the empirical values of the RSA, $P(n|p^*)$, and fit it to obtain the estimates \hat{r} and $\hat{\xi}_{p^*}$ of the parameters that best capture the behavior of the sampled data. Finally, thanks to the form-invariance property, one can obtain the value of the global parameter $\hat{\xi}$ via Eq (2) (henceforth we will denote with $\hat{\cdot}$ our estimation of any quantity).

Let us observe that the probability that a given species present at $p = 1$ is missing at $p < 1$ corresponds to the fraction of unobserved species $(S - S_{p^*})/S$. This value must be equal to $P(0|p) = 1 - c(r, \xi)/c(r, \xi_p)$, the probability for a species to have zero population in a sample of size pN (see S1.5 Section in S1 Appendix). Thus:

$$\hat{S} \simeq \frac{S_{p^*}}{1 - P(0|p^*)} \simeq \frac{1 - (1 - \hat{\xi})\hat{r}}{1 - (1 - \hat{\xi}_{p^*})\hat{r}} S_{p^*}, \tag{3}$$

where the last approximation is obtained by the definition of $c(r, \xi)$ and expressing $\hat{\xi}$ as a function of $\hat{\xi}_{p^*}$ by inverting Eq (2).

To test the reliability of estimator (3), we extracted, from each dataset, a hundred sub-samples each covering a fraction $p^* = 5\%$ of the databases' individuals (sent emails, posted hashtags, occurrences of words). We then inferred the total number of species (email senders, posted hashtags in Twitter data and different words in Wikipedia pages and Gutenberg books) from the empirical RSA constructed at $p^* = 5\%$. The average relative upscaling error is small in all four cases: about 0.4% for sent emails, 3% for Twitter hashtags, 6% for Wikipedia words and -2% for Gutenberg words. In Table 1 we report the average values of the fitted parameters together with the average relative percentage error between the predicted number of species, \hat{S} , and the true one, S (mean and standard deviation are displayed for all datasets). See S3 Table and S3, S4 Figs in S1 Appendix for the results obtained by considering different fractions p^* of the four datasets as starting information.

The second novelty that we introduce in our work is a method to estimate the variation of popularity, a fundamental concept arising naturally when investigating human dynamics [38–43]. Indeed, until now we exploited the information on the abundance of the observed species at the local scale only to estimate the number of unseen species, disregarding of their abundances. Instead, abundance information can be used to predict, for example, the most active users of the email network, the commonest words in a book or the popularity of a hashtag in Twitter database. In particular, focusing on Twitter, various sophisticated measures of popularity based on semantic analyses have been proposed (see for instance [44]). Here, by mean of the popularity of a hashtag we naively count the number of posts containing it that come to circulate within the network thanks to other users' tweets. This information is encompassed within the RSA pattern. Indeed, hashtags posted a low number of times are those positioned in the left side of the curve, whereas hashtags with high popularity are located in its right tail. Our goal now is to derive an estimator for the change in popularity of hashtags from a portion p^* of the observed tweets to the remaining $1 - p^*$ fraction of the unobserved tweets.

Let us thus denote by L a fixed threshold of posts above which we consider a hashtag popular at the sampled scale p^* and let us indicate with $S_{p^*}(\geq L)$ the number of different hashtags having abundance at least L in the surveyed collection of posts. We wish to check whether these (locally) popular species result to be popular also in the unseen fraction of the network, $1 - p^*$. Let us then denote by K the fixed popularity threshold at the unsurveyed scale. We are looking for an estimator of the number of species having popularity at least K in the $1 - p^*$ unseen part of the tweets, given that they have popularity at least L at scale p^* . These species, which we denote with $\hat{S}_{1-p^*}(\geq K | \geq L)$ are therefore globally popular within the network.

From our theoretical framework, we derive an estimator of such a quantity (see S1.6 Section in S1 Appendix). We define $S_{p^*}(l)$ to be the number of species having popularity exactly l at scale p^* and $S_{1-p^*}(k|l)$ to be the number of species having popularity exactly k at scale $1 - p^*$ given that they have popularity exactly l at scale p^* . Then we obtained the following estimator for $S_{1-p^*}(k|l)$ (see S1.6 Section in S1 Appendix for details):

$$\begin{aligned} \hat{S}_{1-p^*}(k|l) &= S_{p^*}(l) \cdot \frac{\binom{k+l}{l} p^{*l} (1-p^*)^k \binom{k+l+\hat{r}-1}{k+l} \hat{\xi}^{k+l} (1-\hat{\xi})^{\hat{r}}}{\binom{l+\hat{r}-1}{l} \hat{\xi}_{p^*}^l (1-\hat{\xi}_{p^*})^{\hat{r}}} \\ &= S_{p^*}(l) \binom{k+l+\hat{r}-1}{k} \cdot \frac{p^{*l} (1-p^*)^k \hat{\xi}^{k+l} (1-\hat{\xi})^{\hat{r}}}{\hat{\xi}_{p^*}^l (1-\hat{\xi}_{p^*})^{\hat{r}}} \end{aligned} \tag{4}$$

An estimator for $\hat{S}_{1-p^*}(\geq K | \geq L)$ can thus be obtained by summing up (4) for all $k \geq K$ and for all $l \geq L$. We tested the above estimator by fixing the (arbitrary) value of the threshold

Table 2. Percentage errors for popularity change predictions in Twitter database. For a fixed $L = 25$ and different values of K (first and second column), we estimated, from ten different Twitter sub-samples ($p^* = 5\%$), the number of species having abundance at least K at the unobserved scale $1 - p^* = 95\%$ given that they have abundance at least L at the sampled scale p^* via estimator (4). The average among the ten sub-samples of the true numbers of species, $S_{1-p^*}(\geq K | \geq L)$, and of the ones predicted by our method, $\hat{S}_{1-p^*}(\geq K | \geq L)$, among the ten sub-samples are displayed in the third and fourth columns, respectively. Finally, in the last two columns, we inserted the mean and the variance of the relative error obtained in the ten predictions. Similar results have been obtained for other values of L and K (see S4 Table in S1 Appendix).

L	K	$S_{1-p^*}(\geq K \geq L)$	$\hat{S}_{1-p^*}(\geq K \geq L)$	Relative Error	Variance
25	219	5,977	5,976.80	0.0018131	0.0000282
25	329	5,943	5,950.31	0.0448228	0.01097890
25	439	5,667	5,688.88	0.0896268	0.0609518
25	548	5,064	5,055.71	-0.1793290	0.0877951

<https://doi.org/10.1371/journal.pone.0253461.t002>

L equal to 25 and varying the value of K in the (arbitrary) range from 219 to 548 for ten sub-samples of Twitter database (for different choices of L and K see S4 Table in S1 Appendix). The average errors obtained in the predictions are displayed in Table 2. For all the considered cases, we achieved very good estimates, with an average relative percentage error below 0.2% in absolute value.

Discussions

To conclude, we show how our statistical ecology framework could be successfully applied to human activities. We tested our method in four databases: email sender activity, Twitter hash-tags, words in Wikipedia pages and Gutenberg books. Once set the correspondence to what we consider species and individuals of a species, our approach reveals that the RSA is scale-free in each mentioned dataset with a heavy-tailed form maintained at different scales—with roughly the same exponent—through the different human activities considered (see Fig 3). This form-invariant property allows for a successful implementation of our predictive statistical framework. The importance of the form-invariance has been already noticed in network science; in [45] authors state that “Only if the degree distributions of the network and randomly sampled subnets belong to the same family of probability distributions is it possible to extrapolate from subnet data to properties of the global network.” In our language and framework this translates into “upscaling is possible when the distribution is form invariant”. However, the heavy tail of the observed RSAs cannot be captured by a standard negative binomial distribution with $r \in \mathbb{R}^+$. Nevertheless, such behaviours can be accommodated when allowing the clustering parameter r to take negative values, $r \in (-1, 0)$ (see Materials and methods, S1 Fig and S1.1 Section in S1 Appendix). This allows us to exploit the form-invariance property of the negative binomial distribution to propose an estimator for the statistics of the unseen human activity from small random samples. In particular, from the activity (sent emails per senders, posts per hashtags, word occurrences) in a small random sample, we infer the number of species (senders, hashtags, words) at the global scale. Moreover, we predict how the popularity of species changes with the scale, an issue of evident importance when thinking of social networks like Twitter. Finally, we compare our estimates with the true known values and in all the considered databases the relative error is small (see Tables 1, 2 and S3 Section in S1 Appendix). This result confirms the ability of our theoretical method to capture hidden quantities of the human dynamics when only random samples are available. In this regard, we remark that within our approach what matters is the ratio between the size (in terms of the number of items) of the random sample and the size of the target one. This implies that our statistical model allows to upscale to two, three, n times the size of the given sample. This feature may be useful when the size of the whole dataset is unknown. Our results pave the way for new applications in upscaling problems beyond statistical ecology.

Indeed, our findings may have applications in different situations, spreading from resource management in emails to collective attention monitoring in Twitter and to language learning process in word databases. Let us see one example for each aforementioned context of how our framework could help in decision making processes related to different aspects of social activity networks. Let us start from the resource managing application. Suppose an internet/email provider starts a campaign to increase customers; for instance the provider wishes to double the number of subscribers. Now, in order to predict if more resources (e.g. number of servers in the email example) are necessary to supply the newly entered subscribers, the provider needs to infer the total amount of activity bursting thanks to these new users. Our method provides a possible solution to this inference problem. Indeed, by inverting Eq (3), which represents the well-known species-accumulation curve in theoretical ecology, one obtains an analytical link between the total amount of activity (e.g. number of sent emails) and the number of users. In particular, the activity does not grow linearly with the users, as one may naively guess. Thus, the information our framework provides on the species-accumulation curve may help the provider to decide how many further resources are needed for the expected number of new users. Clearly, this knowledge is useful either to avoid money waste in case no further resources are required, or to provide new structures/servers in advance in order to safely support the user activity and not to loose unsatisfied customers. Moreover, being aware of how many new structures are needed also helps balance their costs of installation, managing and maintenance with the profit coming from subscriptions.

A second application regards attention monitoring and information spreading. Nowadays social networks constitutes a fundamental source for spreading information and disinformation as well. They have being exploited to influence the mass opinion and attention in many different social contexts, from politics to economy [46]. It is enough to think about the influencer phenomenon arising in almost all social networks. In Twitter, popularity of a user may be read from the number of times a hashtag s/he initiated appears in other users' tweets. In our ecological interpretation, a hashtag represents a species, while the number of posts associated to it gives the species' abundance. Therefore, if the species s/he represents comes to be part of the right tail of the RSA distribution, it constitutes one of the community dominant species and thus we can say s/he is popular, whereas if it comes to fall at the left tail of the RSA, it is a hyper-rare species, thus not having received the desired attention. Therefore, in order to control someone's position within the global network, it is necessary to have access to the RSA at the whole community scale. However, this datum is usually not provided by the social network managing organization. Twitter, for example, only releases information on the total number of tweets posted across time. Nevertheless, there are other services as the Sample Tweets APIs or the Decahose stream service which provide the clients with real-time random samples covering small percentages (up to 10%) of the total tweets. With this information, our framework offers the possibility to fully reconstruct the global RSA as well as to monitor how the number of popular hashtags scales from the monitored sample up to the whole activity network. This latter information may also be useful for governments or public administrations in general to communicate important news (health information, emergency procedures, elections etc. . .) to the citizens. In particular, our method allows to know the number of further tweets one eventually needs to effectively spread the information, thus allowing to undertake the proper measures (a bigger publicity campaign to obtain more followers, the development of bot applications, etc.) to achieve the goal.

Finally, our theoretical framework may also be exploited in language learning process monitoring. For example, let us suppose one is learning a foreign speech. S/he may then be interested in the number of books that are needed s/he needs to read in order to be sure to expand her/his own vocabulary in order for it to cover a fixed percentage of all the speech words. The

species-accumulation curve emerging in this context thanks to our ecological correspondence between words/species and occurrences/abundances can thus be interpreted in a broader sense as a learning curve, with the total number of words encountered during the learning process (by dialogue experience, frontal lectures or personal readings) in the x-axis and the number of different words s/he manages to properly exploit in her/his speech in the y-axis.

Materials and methods

Datasets

In this study we considered four databases concerning human activities: emails, Twitter, Wikipedia and Gutenberg. Here we give a brief description of the data. For further details, see [35] for email dataset and [37] for Twitter, Wikipedia and Gutenberg data.

Emails. This dataset is a collection of almost 7 millions emails, that corresponds to the activity of a Department of the Università degli Studi di Padova during two years: 2012 and 2013. The collected data are in the form {sender, receiver, timestamp}. For our analysis, we selected the first column of the table [35].

Twitter. Our dataset consists of a table where each row is of the form {timestamp, hashtag, user}. For our purposes, we selected the second column of the table. Dataset can be found in <http://kreyon.net/waves-of-novelties/> [37].

Wikipedia and Gutenberg. Our data represents all words contained in a collection of Wikipedia pages and books. We label each different word with a different number. Note that the same word always maintain its correspondence to the same number, regardless of the Wikipedia page or book it belongs [37].

Power-law tails of the negative binomial with a negative clustering coefficient

A negative binomial density function with parameters ξ and $r > 0$ results to capture very well empirical RSA patterns in tropical forests [8, 11]. The observed RSAs in the analyzed human-activity databases, although displaying a similar universal character, do show a different behavior, characterized by heavy tails (see Fig 3). These heavy tails of the observed RSAs cannot be captured by a standard negative binomial distribution with $r \in \mathbb{R}^+$. Nevertheless, extending the clustering parameter region to take negative values, $r \in (-1, 0)$, reflects in a power-law behavior of the RSA distribution tail with an exponential cut-off. To show this, let us consider a truncated negative binomial distribution of parameters r and ξ at the global scale (henceforth we will write $P(n)$ for $P(n|1)$). The following theorem holds true [47, 48].

Theorem. Let $Y(z)$ be the generating function of a discrete random variable having probability mass function $P(\cdot)$ with dominant singularity R_Y . Let $\beta \in \mathbb{R} \setminus \{0, 1, 2, \dots\}$. If for $z \rightarrow R_Y$

$$Y(z) \sim c_Y (1 - z/R_Y)^\beta,$$

then the distribution $P(n)$ satisfies

$$P(n) \sim \frac{c_Y n^{-\beta-1} R_Y^{-n}}{\Gamma(-\beta)} \quad \text{for } n \rightarrow \infty,$$

where $\Gamma(\cdot)$ is the Gamma function.

Let us thus examine the probability generating function of our truncated negative binomial:

$$Y(z) = \sum_{n=0}^{\infty} P(n) z^n,$$

where $P(n)$ is given in (1). Now, since we are interested in the singularities of $Y(z)$, we can neglect the normalizing factor $c(r, \xi)$. Moreover, as the tail of a truncated negative binomial is exactly the same of a standard negative binomial, here we simply disregard of the truncation and conduct the analysis for a standard negative binomial. It then turns out (see S1.4 Section in S1 Appendix for details) that $Y(z)$ has a singularity at $z = 1/\xi$ and that it can be expressed as:

$$Y(z) = c_Y(1 - z\xi)^{-r} = c_Y(1 - z/R_Y)^\beta,$$

where we set $\beta = -r$ and $R_Y = \frac{1}{\xi}$. Thus, Theorem above provides a characterization of the tails of the (truncated) negative binomial:

$$P(n) \sim \frac{c_Y n^{r-1} \xi^n}{\Gamma(-\beta)} = \frac{c_Y n^{r-1} e^{n \ln(\xi)}}{\Gamma(-\beta)}, \quad n \gg 1.$$

The cut-off thus depends on ξ . In particular, the power-law range is greater for values of ξ close to 1.

Supporting information

S1 Appendix. Theoretical framework and additional results and figures.
(PDF)

Author Contributions

Conceptualization: Anna Tovo, Amos Maritan, Stefano Favaro, Marco Formentin.

Data curation: Anna Tovo, Samuele Stivanello.

Formal analysis: Anna Tovo, Samuele Stivanello, Stefano Favaro, Marco Formentin.

Funding acquisition: Amos Maritan, Stefano Favaro, Marco Formentin.

Investigation: Anna Tovo, Amos Maritan, Samir Suweis, Stefano Favaro, Marco Formentin.

Methodology: Anna Tovo, Samuele Stivanello, Amos Maritan, Samir Suweis, Stefano Favaro, Marco Formentin.

Project administration: Marco Formentin.

Resources: Marco Formentin.

Software: Anna Tovo, Samuele Stivanello.

Supervision: Amos Maritan, Samir Suweis, Stefano Favaro, Marco Formentin.

Validation: Amos Maritan, Samir Suweis, Stefano Favaro, Marco Formentin.

Visualization: Anna Tovo, Amos Maritan, Samir Suweis, Stefano Favaro, Marco Formentin.

Writing – original draft: Anna Tovo, Samuele Stivanello, Samir Suweis, Stefano Favaro, Marco Formentin.

Writing – review & editing: Anna Tovo, Marco Formentin.

References

1. Chao A. and Chiu C.-H. (2014). Species richness: estimation and comparison. *Wiley StatsRef: Statistics Reference Online*, 1–26.
2. Favretti M. (2018). Remarks on the maximum entropy principle with application to the maximum entropy theory of ecology. *Entropy* 20, 11. <https://doi.org/10.3390/e20010011>

3. Favretti M. (2018). Maximum entropy theory of ecology: a reply to Harte. *Entropy* 20, 308. <https://doi.org/10.3390/e20050308> PMID: 33265399
4. Good I. and Toulmin G. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* 43, 45–63. <https://doi.org/10.2307/2333577>
5. Harte J., Smith A. B., and Storch D. (2009). Biodiversity scales from plots to biomes with a universal species–area curve. *Ecology letters* 12, 789–797. <https://doi.org/10.1111/j.1461-0248.2009.01328.x> PMID: 19486123
6. Orłitsky A., Suresh A. T., and Wu Y. (2016). Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences* 113, 13283–13288. <https://doi.org/10.1073/pnas.1607774113> PMID: 27830649
7. Slik J. F., Arroyo-Rodríguez V., Aiba S.-I., Alvarez-Loayza P., Alves L. F., Ashton P., et al. (2015). An estimate of the number of tropical tree species. *Proceedings of the National Academy of Sciences* 112, 7472–7477. <https://doi.org/10.1073/pnas.1423147112> PMID: 26034279
8. Tovo A., Suweis S., Formentin M., Favretti M., Volkov I., Banavar J. R., et al. (2017). Upscaling species richness and abundances in tropical forests. *Science advances* 3, e1701438. <https://doi.org/10.1126/sciadv.1701438> PMID: 29057324
9. Volkov I., Banavar J. R., Hubbell S. P., and Maritan A. (2003). Neutral theory and relative species abundance in ecology. *Nature* 424, 1035. <https://doi.org/10.1038/nature01883> PMID: 12944964
10. Volkov I., Banavar J. R., Hubbell S. P., and Maritan A. (2007). Patterns of relative species abundance in rainforests and coral reefs. *Nature* 450, 45. <https://doi.org/10.1038/nature06197> PMID: 17972874
11. Tovo A., Formentin M., Suweis S., Stivanello S., Azae S., and Maritan A. (2019). Inferring macro-ecological patterns from local species' occurrences. *Oikos*. <https://doi.org/10.1111/oik.06754>
12. Hubbell S.P. (2001). *The unified neutral theory of biodiversity and biogeography* (Princeton University Press).
13. Baayen R. H. (2002). *Word frequency distributions*, vol. 18 (Springer Science & Business Media).
14. Kornai A. (2007). *Mathematical linguistics* (Springer Science & Business Media).
15. Deville P., Song C., Eagle N., Blondel V. D., Barabási A.-L., and Wang D. (2016). Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences* 113, 7047–7052. <https://doi.org/10.1073/pnas.1525443113> PMID: 27274050
16. Gao L., Song C., Gao Z., Barabási A.-L., Bagrow J. P., and Wang D. (2014). Quantifying information flow during emergencies. *Scientific reports* 4, 3997. <https://doi.org/10.1038/srep03997> PMID: 24499738
17. Karsai M., Jo H.-H., and Kaski K. (2018). *Bursty human dynamics* (Springer).
18. Loreto V., Baronchelli A., Mukherjee A., Puglisi A., and Tria F. (2011). Statistical physics of language dynamics. *Journal of Statistical Mechanics: Theory and Experiment* 2011, P04006. <https://doi.org/10.1088/1742-5468/2011/04/P04006>
19. Loreto V., Mukherjee A., and Tria F. (2012). On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences* 109, 6819–6824. <https://doi.org/10.1073/pnas.1113347109> PMID: 22509002
20. Malmgren R. D., Stouffer D. B., Campanharo A. S., and Amaral L. A. N. (2009). On universality in human correspondence activity. *science* 325, 1696–1700. <https://doi.org/10.1126/science.1174562> PMID: 19779200
21. Malmgren R. D., Stouffer D. B., Motter A. E., and Amaral L. A. (2008). A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* 105, 18153–18158. <https://doi.org/10.1073/pnas.0800332105> PMID: 19017788
22. Török J., Iniguez G., Yasseri T., San Miguel M., Kaski K., and Kertész J. (2013). Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment. *Physical review letters* 110, 088701. <https://doi.org/10.1103/PhysRevLett.110.088701> PMID: 23473207
23. Yasseri T., Hale S. A., and Margetts H. Z. (2017). Rapid rise and decay in petition signing. *EPJ Data Science* 6, 20. <https://doi.org/10.1140/epjds/s13688-017-0116-6>
24. Yasseri T., Sumi R., and Kertész J. (2012). Circadian patterns of wikipedia editorial activity: A demographic analysis. *PloS one* 7, e30091. <https://doi.org/10.1371/journal.pone.0030091> PMID: 22272279
25. Grauwin S., Szell M., Sobolevsky S., Hövel P., Simini F., Vanhoof M., et al. (2017). Identifying and modeling the structural discontinuities of human interactions. *Scientific reports* 7, 46677. <https://doi.org/10.1038/srep46677> PMID: 28443647

26. Alfi V., Parisi G., and Pietronero L. (2007). Conference registration: how people react to a deadline. *Nature Physics* 3, 746. <https://doi.org/10.1038/nphys761>
27. Bagrow J. P., Wang D., and Barabási A.-L. (2011). Collective response of human populations to large-scale emergencies. *PloS one* 6, e17680. <https://doi.org/10.1371/journal.pone.0017680> PMID: 21479206
28. Barabási A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* 435, 207. <https://doi.org/10.1038/nature03459> PMID: 15889093
29. Barabási A.-L. and Albert R. (1999). Emergence of scaling in random networks. *science* 286, 509–512. <https://doi.org/10.1126/science.286.5439.509> PMID: 10521342
30. Lorenz-Spreen P., Mønsted B.M., Hövel P., and Lehmann S. (2019). Accelerating dynamics of collective attention. *Nature communications* 10, 1–9. <https://doi.org/10.1038/s41467-019-09311-w> PMID: 30988286
31. Palazzi M., Solé-Ribalta A., Calleja-Solanas V., Plata C.A., Suweis S., Borge-Holthoefer J., et al. (2020). Resilience and elasticity of co-evolving information ecosystems. *arXiv preprint arXiv:2005.07005*.
32. Plata C.A., Pigani E., Azaele S., Calleja-Solanas V., Palazzi M., Solé-Ribalta A., et al. (2021). Neutral theory for competing attention in social networks. *Physical Review Research* 3, 013070. <https://doi.org/10.1103/PhysRevResearch.3.013070>
33. Borge-Holthoefer J., Baños R.A., Gracia-Lázaro C., and Moreno Y. (2017). Emergence of consensus as a modular-to-nested transition in communication dynamics. *Scientific reports* 7, 1–9. <https://doi.org/10.1038/srep41673> PMID: 28134358
34. Castellano C., Fortunato S., and Loreto V. (2009). Statistical physics of social dynamics. *Reviews of modern physics* 81, 591. <https://doi.org/10.1103/RevModPhys.81.591>
35. Formentin M., Lovison A., Maritan A., and Zanzotto G. (2014). Hidden scaling patterns and universality in written communication. *Physical Review E* 90, 012817. <https://doi.org/10.1103/PhysRevE.90.012817> PMID: 25122352
36. Formentin M., Lovison A., Maritan A., and Zanzotto G. (2015). New activity pattern in human interactive dynamics. *Journal of Statistical Mechanics: Theory and Experiment* 2015, P09006. <https://doi.org/10.1088/1742-5468/2015/09/P09006>
37. Monechi B., Ruiz-Serrano A., Tria F., and Loreto V. (2017). Waves of novelties in the expansion into the adjacent possible. *PloS one* 12. <https://doi.org/10.1371/journal.pone.0179303> PMID: 28594909
38. Jia T., Wang D., and Szymanski B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour* 1, 0078. <https://doi.org/10.1038/s41562-017-0078>
39. Mestyán M., Yasseri T., and Kertész J. (2013). Early prediction of movie box office success based on wikipedia activity big data. *PloS one* 8, e71226. <https://doi.org/10.1371/journal.pone.0071226> PMID: 23990938
40. Shen H., Wang D., Song C., and Barabási A.-L. (2014). Modeling and predicting popularity dynamics via reinforced poisson processes. In *Twenty-eighth AAAI conference on artificial intelligence*.
41. Sinatra R., Wang D., Deville P., Song C., and Barabási A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science* 354, aaf5239. <https://doi.org/10.1126/science.aaf5239> PMID: 27811240
42. Yucesoy B. and Barabási A.-L. (2016). Untangling performance from success. *EPJ Data Science* 5, 17. <https://doi.org/10.1140/epjds/s13688-016-0079-z>
43. Zhao Q., Erdogdu M. A., He H. Y., Rajaraman A., and Leskovec J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM)*, 1513–1522.
44. Colladon A. F. (2018). The semantic brand score. *Journal of Business Research* 88, 150–160. <https://doi.org/10.1016/j.jbusres.2018.03.026>
45. Stumpf M.P.H., Wiuf C., and May R.M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences* 102, 4221–4224. <https://doi.org/10.1073/pnas.0501179102> PMID: 15767579
46. Margetts H., John P., Hale S., and Yasseri T. (2015). *Political turbulence: How social media shape collective action* (Princeton University Press).
47. Flajolet P. and Sedgewick R. (2008). *Analytic Combinatorics* (Cambridge University Press).
48. Walraevens J., Demoor T., Maertens T., and Bruneel H. (2012). Stochastic queueing-theory approach to human dynamics. *Physical Review E* 85, 021139. <https://doi.org/10.1103/PhysRevE.85.021139> PMID: 22463184