frontiers

Upscaling human activity data: a statistical ecology approach

Anna Tovo [1,2,*], **Samuele Stivanello** [1], **Amos Maritan** [2], **Samir Suweis** [2,3], **Stefano Favaro** [4] **and Marco Formentin** [1,3,*]

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

Correspondence*:
Via Trieste, 63, 15121 Padova (PD), Italy
anna.tovo@unipd.it, marco.formentin@unipd.it

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.

Keywords: Upscaling of large datasets, Regular patterns in human activity data, Statistics of complex human dynamics, Computer-mediated social activities, Popularity of Twitter hashtags

1 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 Good and Toulmin (1956); Harte et al. (2009); Chao and Chiu (2014); Slik et al. (2015); Orlitsky et al. (2016). The success of such methods depends on the following notable fact: different ecosystems like tropical forests or coral reefs Volkov et al. (2003, 2007); Slik et al. (2015); Tovo et al. (2017), despite their disparate locations and different evolutionary history, share a common log-series shape of the empirical RSA which implies that the number of different species grows as the logarithm of the population size (see Figure 1). In the present paper we adopted and extended a statistical framework which was firstly designed in ecology Tovo et al. (2017, 2019a) to get new insights into human activity databases with the aim of inferring global statistics of a dataset from a random sample

of it. Indeed, we consider four human activities with the following correspondence between species and individuals within each dataset: (1) *Email communication* Formentin et al. (2014, 2015): 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* Monechi et al. (2017): 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* Monechi et al. (2017) and (4) *Gutenberg books* Monechi et al. (2017) 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. We remark that this latter dataset is somehow different from the other three as it consists of word occurrences in a corpus, representing thus a construction of natural languages. However, also natural language can be considered as the result of complex interactions among individuals over a long period of time and that is why we included this dataset in our study on human activities. Once defined, as we did above, what correspond to species and individuals, the RSA of each dataset displays a Zipf tail Monechi et al. (2017) and the rate at which new elements appear shows a sublinear power-law growth, signature of the Heap's law (see also Figure 1). Statistical regularities in human dynamics have been widely observed in many different contexts and a variety of models have been proposed to understand such recurrent patterns Barabási and Albert (1999); Barabási (2005); Alfi et al. (2007); Malmgren et al. (2008, 2009); Loreto et al. (2011); Bagrow et al. (2011); Loreto et al. (2012); Yasseri et al. (2012); Török et al. (2013); Gao et al. (2014); Deville et al. (2016); Grauwin et al. (2017); Yasseri et al. (2017); Karsai et al. (2018). In particular, Zipf's laws have been observed since decades in computational linguistic and many models generating such laws have been proposed (see Baayen (2002) and Kornai (2007) for a review). However, in the present work we focus on inference, not modeling. In particular, we propose a statistical framework: 1) that gives reliable estimates for the number of users, hashtags, and words from a random sample of mails, posts and word occurrences (see Table 1). We refer to the inference of global quantities of interest from random samples as *upscaling*. Moreover, our framework predicts how the number of users/hashtags/words grows with the recorded activity (mails/posts/pages/books) (see Figure 2); 2) We infer how the abundance of a species may change across scales (see Table 2). 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.

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 also Supplementary Section S1.1). This choice, justified by the heavy tail of the observed RSAs (see Figure 3), has the major consequence that the RSA is *form-invariant*. Form-invariance should not be confused with scale-invariance, a property only satisfied by power-laws (see Supplementary Section S1.2). 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 Figure 3 and Supplementary Section S1.2). 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. In particular, the idea of exploiting its form-invariance property to obtain an effective yet simple estimator which explicitly depends on the scale is new. Actually, to our knowledge, upscaling has never been investigated for email communication and Twitter datasets whereas 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 Baayen (2002). 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.

2 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 Figure 2).
Correspondence between species/individuals and human activities can be set similarly for the remaining datasets (see Figure 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 Figure 3). Moreover, this exponent is maintained at different scales (see Supplementary Figures S1 and S2), supporting our choice of modeling the RSA by means of a form-invariant distribution that keeps fixed the tail exponent through scales.
- **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 Figure 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 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 Table 2), 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 following 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 $\xi \in (0,1)$:

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

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 $r \in (-1,0)$ reflect in a heavy-tailed behavior of the RSAs. More precisely, the right tail of (1) has the form $n^{r-1}\exp(n\log\xi)$ (see Supplementary Section S1.3), where the exponential cut-off disappears in the limit $\xi \to 1$. In this latter case (1) describes a pure power-law tail behavior. Such heavy-tailed behavior well describes the observed RSA patterns in human activities (see Supplementary Figure S1). Moreover, the exponent $\alpha = 1-r$ matches very well with the empirical data (see also Figure 3). 2) Distribution (1) is *form-invariant*, meaning that the RSA $P(n|p)$ maintains the same functional form at different scales p (see Supplementary Section S1.2), a property observed in the empirical RSAs of all the four databases (see Figure 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 $\cdot$).

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 Supplementary Section S1.4). 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, ten 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 0.1% 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 Supplementary Table S1 and Figures S3-S4 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 Mestyán et al. (2013); Shen et al. (2014); Zhao et al. (2015); Yucesoy and Barabási (2016); Sinatra et al. (2016); Jia et al. (2017). 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 Colladon (2018)). 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 Supplementary Section S1.5). 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 Supplementary Section S1.5 for details):

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

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 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 Supplementary Table S2). 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.

3 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 hashtags, 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 Figure 3). This form-invariant property allows for a successful implementation of our predictive statistical framework. 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 and Supplementary Figure S1). 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 samples, 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 Table 1, Table 2 and Supplementary Section S2). This result confirms the ability of our theoretical method to capture hidden quantities of the human dynamics when only random samples are available. 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 Margetts et al. (2015). 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.

4 MATERIALS AND METHODS

4.1 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 Formentin et al. (2014) for email dataset and Monechi et al. (2017) 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 Formentin et al. (2014).

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/ Monechi et al. (2017).

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 Monechi et al. (2017).

4.2 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 Tovo et al. (2017, 2019a). 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 Figure 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 Walraevens et al. (2012); Flajolet and Sedgewick (2008).

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 $\beta \in \mathbb{R} \setminus \{0, 1, 2, ...\}$. If for $z \to R_Y$*

$$Y(z) \sim c_Y \left(1 - z/R_Y\right)^{\beta},$$

then the distribution $P(n)$ satisfies

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

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 Supplementary Section S1.3 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:

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

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

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

A.T, S.Stivanello, A.M., S.Suweis, S.F. and M.F. designed research, performed research, analysed data and wrote the paper.

FUNDING

A.T. acknowledges financial support from *neXt* grant, Department of Mathematics "Tullio Levi-Civita" of University of Padova. S. Suweis and A.T. 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.M. was supported by Excellence Project 2017 of the Cariparo Foundation. Stefano 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. Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

ACKNOWLEDGEMENTS

This manuscript has been released as a pre-print at arXiv, Tovo et al. (2019b).

SUPPLEMENTAL DATA

S1. Theoretical framework
S2. Additional results and figures

DATA AVAILABILITY STATEMENT

All the data and codes are available online (http://kreyon.net/waves-of-novelties/) or upon request to the corresponding authors.

REFERENCES

Alfi, V., Parisi, G., and Pietronero, L. (2007). Conference registration: how people react to a deadline. Nature Physics 3, 746

Baayen, R. H. (2002). Word frequency distributions, vol. 18 (Springer Science & Business Media)

Bagrow, J. P., Wang, D., and Barabási, A.-L. (2011). Collective response of human populations to large-scale emergencies. PloS one 6, e17680

Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. Nature 435, 207

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. science 286, 509–512

Chao, A. and Chiu, C.-H. (2014). Species richness: estimation and comparison. Wiley StatsRef: Statistics Reference Online , 1–26

Colladon, A. F. (2018). The semantic brand score. Journal of Business Research 88, 150–160

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

Flajolet, P. and Sedgewick, R. (2008). Analytic Combinatorics (Cambridge University Press)

Formentin, M., Lovison, A., Maritan, A., and Zanzotto, G. (2014). Hidden scaling patterns and universality in written communication. Physical Review E 90, 012817

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

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

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

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

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

Jia, T., Wang, D., and Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. Nature Human Behaviour 1, 0078

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

Kornai, A. (2007). Mathematical linguistics (Springer Science & Business Media)

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

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

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

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

Margetts, H., John, P., Hale, S., and Yasseri, T. (2015). Political turbulence: How social media shape collective action (Princeton University Press)

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

Monechi, B., Ruiz-Serrano, A., Tria, F., and Loreto, V. (2017). Waves of novelties in the expansion into the adjacent possible. PloS one 12

Orlitsky, 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

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

Sinatra, R., Wang, D., Deville, P., Song, C., and Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. Science 354, aaf5239

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

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

Tovo, A., Formentin, M., Suweis, S., Stivanello, S., Azaele, S., and Maritan, A. (2019a). Inferring macro-ecological patterns from local species' occurrences. Oikos doi:10.1111/oik.06754

Tovo, A., Stivanello, S., Maritan, A., Suweis, S., Favaro, S., and Formentin, M. (2019b). Upscaling human activity data: an ecological perspective. arXiv preprint arXiv:1912.03023

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

Volkov, I., Banavar, J. R., Hubbell, S. P., and Maritan, A. (2003). Neutral theory and relative species abundance in ecology. Nature 424, 1035

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

Walraevens, J., Demoor, T., Maertens, T., and Bruneel, H. (2012). Stochastic queueing-theory approach to human dynamics. Physical Review E 85, 021139

Yasseri, T., Hale, S. A., and Margetts, H. Z. (2017). Rapid rise and decay in petition signing. EPJ Data Science 6, 20

382 Yasseri, T., Sumi, R., and Kertész, J. (2012). Circadian patterns of wikipedia editorial activity: A
383 demographic analysis. PloS one 7, e30091
384 Yucesoy, B. and Barabási, A.-L. (2016). Untangling performance from success. EPJ Data Science 5, 17
385 Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., and Leskovec, J. (2015). Seismic: A self-exciting point
386 process model for predicting tweet popularity. In Proceedings of the 21th ACM SIGKDD International
387 Conference on Knowledge Discovery and Data Mining (ACM), 1513–1522

FIGURE CAPTIONS

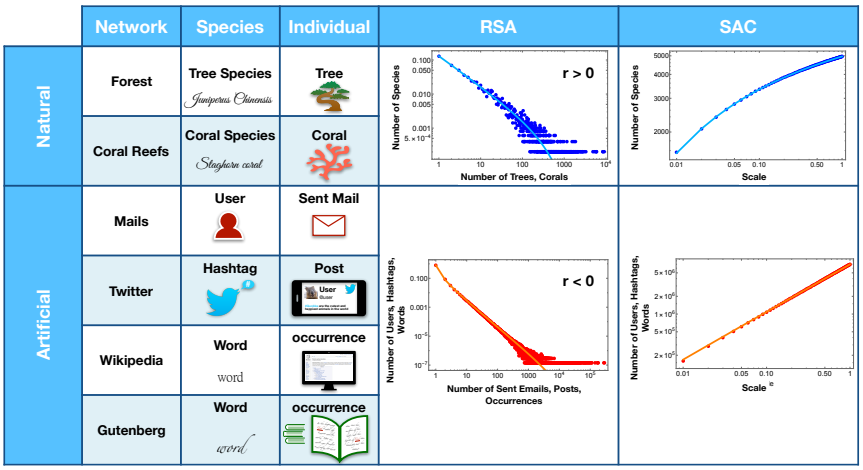


Figure 1. From Ecology to Human Activities. The figure depicts the correspondence between species/individuals in a natural ecosystem and users/sent emails, hashtags/posts, words/occurrences in each one of the four datasets considered in the paper. Once the proper correspondence is established, it turns 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 and SAC curves for natural versus human activity systems, we display the empirical patterns obtained for the Amazonia rainforest Tovo et al. (2017) and Twitter dataset. In general, all human activity RSA curves can be accommodated by with a negative value of r in the interval $(-1, 0)$, whereas natural ecosystems prefer $r > 0$ (solid lines).

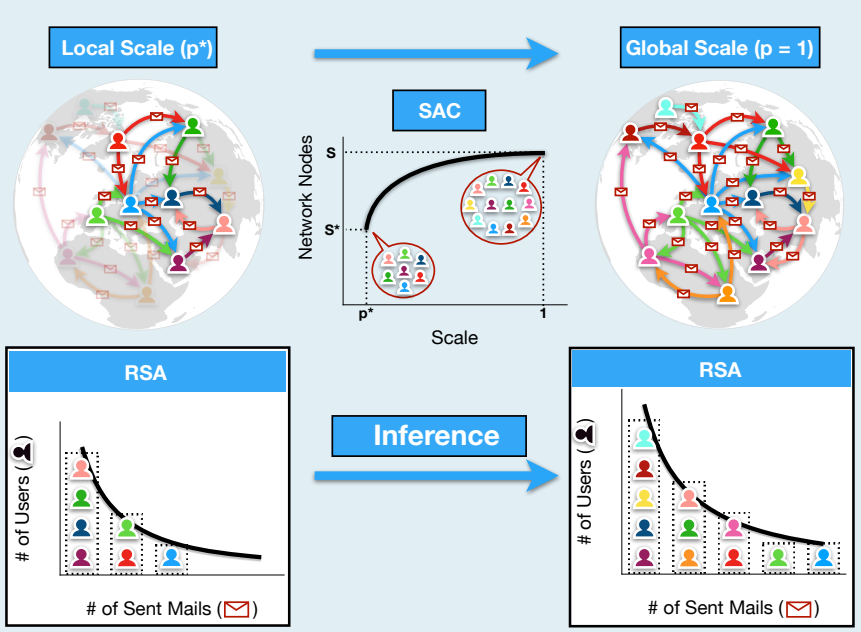


Figure 2. Sketch of our theoretical framework. Consider the email senders' network where each node is a sender and a directed link from node A to node B is an email issued from user A to user B. We set the identity of a sender to be the species and a sent email to be an individual of that species. For 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.

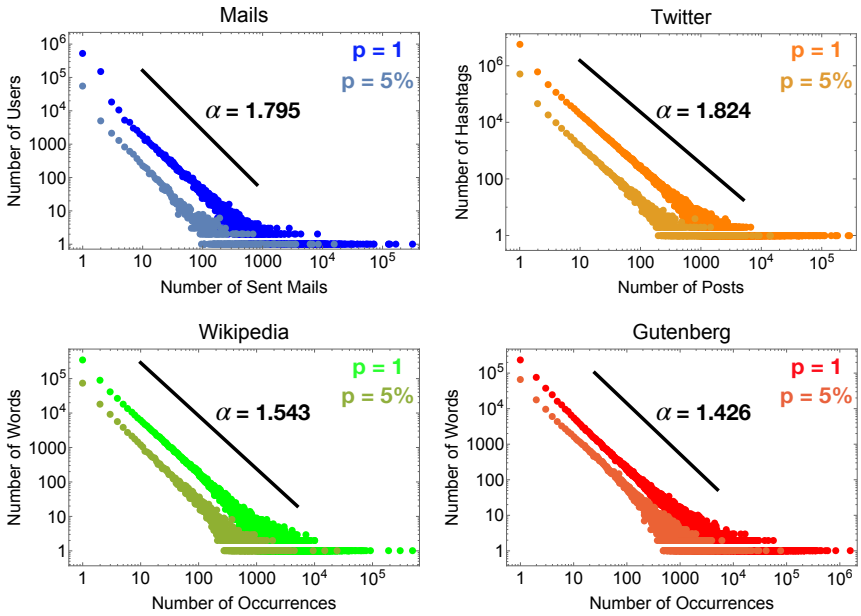


Figure 3. Universality and form-invariance of the empirical RSAs. Empirical RSA curves at the global scale ($p = 1$) and the local scale ($p = 5\%$) are shown. The patterns result scale-free in all the analysed datasets, with a heavy-tailed form maintained through the different human activities and scales. 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 Supplementary Section S1.3). 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 Supplementary Figure S1.

TABLES

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 fitted RSA distribution parameters at the sampled scale and the relative percentage error (mean and standard deviation) between the true number of species and the one predicted by our framework. See Supplementary Figure S1 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	$6,914,872$	$34,696,973$	$29,606,116$	$126,289,661$
$\mathbf{r}$	-0.795	-0.824	-0.543	-0.426
$\xi_{\mathbf{p}^*}$	0.9999	0.9991	0.9985	0.9997
Relative Error	$0.112 \pm 0.385\%$	$3.33 \pm 0.17\%$	$6.11 \pm 0.118\%$	$-2.30 \pm 0.23\%$

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 Supplementary Table 2).

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

Supplementary Material

S1 THEORETICAL FRAMEWORK

S1.1 Statistical model

Once it has been defined what are species and individuals of a species in each of the four human activities considered, we can proceed in the explanation of our statistical model from an ecological perspective.
We denote with N the total population size and with S the number of different species populating an ecosystem.
The *species abundance distribution* (SAD) at a scale p depicts the number of species in a subpopulation of size pN having exactly n individuals. In the following we will quote as RSA the corresponding probability distribution, denoted by $P(n|p)$.
Let us now consider the whole system, i.e. the entire population. We assume that, at the global scale $p = 1$, the RSA distribution is proportional to a negative binomial with parameters r and ξ. It reads:

$$P(n|1) = c(r, \xi) \cdot \mathcal{P}(n|r, \xi) \qquad \text{for } n \geq 1 \tag{S1}$$

where $\mathcal{P}(n|r, \xi)$ is the well known negative binomial density function with parameters r and ξ, i.e.

$$\mathcal{P}(n|r, \xi) = \binom{n + r - 1}{n} \xi^n (1 - \xi)^r$$

and where the normalizing factor $c(r, \xi)$ takes into account the fact that each of the existing S species at the global scale consists of at least one individual:

$$c(r, \xi) = \left[\sum_{n=1}^{\infty} \binom{n + r - 1}{n} \xi^n (1 - \xi)^r \right]^{-1} = \frac{1}{1 - (1 - \xi)^r}.$$

Through the paper we always consider the generalized negative binomial distribution where the binomial coefficient is expressed by means of Gamma functions, i.e. $\binom{n+r-1}{n} = \frac{\Gamma(n+r)}{\Gamma(n+1)\Gamma(r)}$.
The reason why we chose to model the RSA distribution with a negative binomial will be clear in few lines. For the moment, let us anticipate that the negative binomial has two properties that are essential for the development of our estimators: it is form-invariant (see Section S1.2) and, varying the values of ξ and r, it can describe very well different tail behaviors, from exponential to power-law (see Section S1.3).

S1.2 Form-invariance of the RSA distribution

Zooming at a sub-scale p, i.e. considering a subpopulation of size pN, we will recover $S_p \leq S$ species. Note that S_p may depend on which pN individuals we select. In other words, different samples of the same size may lead to different values of S_p. We wish to derive the distribution of the local RSA $P(k|p)$ under the hypothesis of random sampling.
Under random sampling, it can be proven that, if the RSA at the global scale is distributed according to (S1), then the local RSA at a local scale p is again proportional to a negative binomial, with rescaled

parameter ξ_p and same r:

$$P(k|p) = \begin{cases} c(r,\xi)\cdot\mathcal{P}(k|r,\xi_p) & k \geq 1 \\ 1 - c(r,\xi)/c(r,\xi_p) & k = 0 \end{cases} \tag{S2}$$

with

$$\xi_p = \frac{p\xi}{1-\xi(1-p)}. \tag{S3}$$

The fact that the RSA maintains the same functional form at different scales will be central in our framework. We will refer to this property as *form-invariance*. We remark that form-invariance should not be confused with *scale-invariance*. In fact, this latter is defined as the following property: a distribution f is said to be scale-invariant if $f(px) = g(p)f(x)$ where $g(p)$ is a multiplicative scale-dependent constant. It can be proven that power-laws are the only distributions satisfying this property. In contrast, with form-invariant we mean a distribution which maintains the same functional form under random sampling.
We wish now to prove that relation (S2) holds.
Suppose that a species consists of n individuals among the whole population. Under random sampling, the conditional probability that the species has k individuals at the sub-scale p, given that it has total abundance n at the global scale, is given by a binomial distribution of parameters n and p:

$$\mathcal{P}_{binom}(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, ..., n$$

and $\mathcal{P}_{binom}(k|n,p) = 0$ if $k > n$. Let us now prove that the RSA at the local scale $P(k|p)$ is indeed distributed according to (S2).
We start by noticing that, in order to compute the probability that a species has abundance $k \geq 1$ at a local scale p, we need to condition on the fact that the species has abundance n at the whole scale $p = 1$, and then to sum over n, i.e.

$$\begin{aligned} P(k|p) &= \sum_{n\geq k} \mathcal{P}_{binom}(k|n,p) P(n|1) \\ &= \sum_{n\geq k} \binom{n}{k} p^k (1-p)^{n-k} \cdot c(\xi,r) \binom{n+r-1}{n} \xi^n (1-\xi)^r \\ &= c(\xi,r) \binom{k+r-1}{k} \left(\frac{p\xi}{1-\xi(1-p)}\right)^k \left(\frac{1-\xi}{1-\xi(1-p)}\right)^r \\ &= c(\xi,r) \binom{k+r-1}{k} \xi_p^k (1-\xi_p)^r \\ &= c(\xi,r)\cdot\mathcal{P}(k|r,\xi_p), \end{aligned}$$

with ξ_p given in (S3). For $k = 0$ we have

$$P(0|p) = 1 - \sum_{k\geq 1} \mathcal{P}_{sub}(k|p) = 1 - c(\xi,r) \sum_{k\geq 1} \mathcal{P}(k|r,\xi_p) = 1 - \frac{c(\xi,r)}{c(\xi_p,r)}.$$

Our method proceeds as follows: after fitting the parameters $\hat{\xi}_{p^*}$ and $\hat{r}$ from the empirical RSA observed at a local scale p^*, we upscale them so to obtain an estimation of the global parameter $\hat{\xi}$ at $p = 1$ by inverting (S3). The formula reads explicitly:

$$\xi = \frac{\xi_{p^*}}{p^* + \xi_{p^*}(1 - p^*)}. \tag{S4}$$

Note that this form-invariance holds between any two scales $q \leq p$. Indeed, from

$$\xi_p = \frac{p\xi}{1 - \xi(1-p)} \qquad \text{and} \qquad \xi_q = \frac{q\xi}{1 - \xi(1-q)}$$

we obtain

$$\begin{aligned} \xi_q &= \frac{q\xi}{1 - \xi(1-q)} = \frac{q\frac{\xi_p}{p+\xi_p(1-p)}}{1 - \frac{\xi_p}{p+\xi_p(1-p)}(1-q)} = \frac{q\xi_p}{p + \xi_p(1-p) - \xi_p(1-q)} \\ &= \frac{q\xi_p}{p - \xi_p(p-q)} = \frac{\frac{q}{p}\xi_p}{1 - \xi_p(1 - \frac{q}{p})}. \end{aligned}$$

With the same argument, for any $q \geq p$ it holds

$$\xi_q = \frac{\xi_p}{\frac{p}{q} + \xi_p(1 - \frac{p}{q})}. \tag{S5}$$

Hence what really matters is the relative ratio of the two scales.

S1.3 Power-law tails of $\mathcal{P}(n|r, \xi)$ with $r \in (-1, 0)$

A negative binomial density function with parameters ξ and $r > 0$ results to capture very well empirical RSA patterns in tropical forests Tovo et al. (2017, 2019). 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 Figure S2 and Figure 3 of the main text). These heavy tails of the observed RSAs cannot be captured by a standard negative binomial distribution with $r \in \mathbb{R}^+$. Nevertheless, they can be accommodated when allowing the clustering parameter r to take negative values, $r \in (-1, 0)$, thus enabling us to adapt and generalize the theoretical work of Tovo et al. (2017) to portray regular statistics of human activities and to use information on local scales to predict hidden features of the human dynamics at the global scale.

We wish now to show that the extension of the parameter region reflects in a power-law behavior of the RSA distribution tail with an exponential cut-off, which well describes the observed patterns in human activities. We point our that both the parameters intervene in the shape of the RSA patterns, being r responsible for the power-law tail with exponent $\alpha = 1 - r$ and ξ for the position of the exponential truncation of the distribution. Note that, although this section is purely theoretical, the predicted exponent $\alpha = 1 - r$ matches very well our findings when we empirically fit the data (see also Figure 3 of the main text).

We start by considering our truncated negative binomial distribution of parameters r and ξ at the global scale (henceforth we will write $P(n)$ for $P(n|1)$, thus omitting the explicit dependence on the scale $p = 1$):

$$P(n) = c(r, \xi)\binom{n + r - 1}{n}\xi^n(1 - \xi)^r, \tag{S6}$$

The following theorem holds true Walraevens et al. (2012); Flajolet and Sedgewick (2008).

THEOREM S1.1. *Let $Y(z)$ be the generating function of a discrete random variable having probability mass function $P(\cdot)$ with dominant singularity R_Y. Let $\beta \in \mathbb{R} \setminus \{0, 1, 2, ...\}$. If for $z \to R_Y$*

$$Y(z) \sim c_Y \left(1 - z/R_Y\right)^{\beta}, \tag{S7}$$

then the distribution $P(n)$ satisfies

$$P(n) \sim \frac{c_Y n^{-\beta-1} R_Y^{-n}}{\Gamma(-\beta)} \qquad \text{for } n \to \infty, \tag{S8}$$

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

We wish to apply this theorem to our truncated negative binomial distribution. Let us first recall that a singularity of a complex function is a point in the complex plane where the function is not analytic. Examples are poles, square-root branch points and branch cuts.
We now start by examining the probability generating function:

$$Y(z) = \sum_{n=0}^{\infty} P(n) z^n, \tag{S9}$$

where $P(n)$ is given in (S6). Observe that, since we wish to investigate the singularities of $Y(z)$, the normalizing factor $c(r, \xi)$ does not play any significant role. Moreover, the tail of a truncated negative binomial is exactly the same of a standard negative binomial, hence we simply disregard of the truncation and conduct the analysis for a standard negative binomial.
Since we aim at finding the lowest-norm singularity of the probability generating function $Y(z)$, we proceed with the computation by replacing the term $P(n)$ in (S9) with its definition (S6):

$$\begin{aligned} Y(z) &= \sum_{n=0}^{\infty} \binom{n+r-1}{n} \xi^n (1-\xi)^r z^n \\ &= \sum_{n=0}^{\infty} \binom{n+r-1}{n} (z\xi)^n (1-z\xi)^r \cdot \frac{(1-\xi)^r}{(1-z\xi)^r} \\ &= \frac{(1-\xi)^r}{(1-z\xi)^r} \cdot \sum_{n=0}^{\infty} \binom{n+r-1}{n} (z\xi)^n (1-z\xi)^r. \end{aligned}$$

For $z\xi < 1$, i.e. for $z < \frac{1}{\xi}$, the sum converges to 1 as we are summing over $\mathbb{N}$ the marginals of a standard negative binomial of parameters r and $z\xi$.
Thus we are left with

$$Y(z) = \frac{(1-\xi)^r}{(1-z\xi)^r} = c_Y (1-z\xi)^{-r}.$$

It turns out that $Y(z)$ has a singularity at $z = 1/\xi$.
We now wish to express $Y(z)$ as in (S7) to apply the theorem. In our case:

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

where we set $\beta = -r$ and $R_Y = \frac{1}{\xi}$. Thus, Theorem (S1.1) provides a characterization of the tails of the (truncated) negative binomial:

$$P(n) \sim \frac{c_Y n^{r-1}\xi^n}{\Gamma(-\beta)} = \frac{c_Y n^{r-1} e^{n\ln(\xi)}}{\Gamma(-\beta)}, \qquad n >> 1. \tag{S10}$$

S1.4 Estimator for the total number of species and SAC

We proceed now in the description of our procedure. Recall that our method only uses the information available at a sub-sample covering a fraction p^* of the entire system. Therefore, we only have information on the abundances of the S_{p^*} species present within the surveyed area. We now wish to determine the relationship between the total number of species S in the entire population, i.e. at $p = 1$, and the number of observed species at the sub-scale p^*.
Note that the probability that a species among the existing S has null abundance at scale p^* corresponds to the fraction of unsurveyed species. Hence we obtain

$$P(k = 0|p^*) \simeq \frac{S - S_{p^*}}{S}. \tag{S11}$$

Arranging the latter equation, we get a formula to predict the total number of species:

$$\begin{aligned}
\hat{S} &\overset{\text{eq (S11)}}{=} \frac{S_{p^*}}{1 - P(k = 0|p^*)} \\
&\overset{\text{eq (S2)}}{=} S_{p^*}\frac{1 - (1 - \hat{\xi})^{\hat{r}}}{1 - (1 - \hat{\xi}_{p^*})^{\hat{r}}} \\
&\overset{\text{eq (S4)}}{=} S_{p^*}\frac{1 - \left(1 - \dfrac{\hat{\xi}_{p^*}}{p^* + \hat{\xi}_{p^*}(1 - p^*)}\right)^{\hat{r}}}{1 - (1 - \hat{\xi}_{p^*})^{\hat{r}}}.
\end{aligned} \tag{S12}$$

Thus we derived a formula to estimate the total number of species of a community given a sample at scale p^*.
Let us note that we can do more. Indeed, for any $q \in (p^*, 1)$ we can apply the same chain of equations as above with some slight modifications to estimate $\hat{S}_q$:

$$\hat{S}_q = S_{p^*}\frac{1 - \left(1 - \dfrac{\hat{\xi}_{p^*}}{\frac{p^*}{q} + \hat{\xi}_{p^*}(1 - \frac{p^*}{q})}\right)^{\hat{r}}}{1 - (1 - \hat{\xi}_{p^*})^{\hat{r}}} = S_{p^*}\frac{1 - \left(\dfrac{p^*\left(1 - \hat{\xi}_{p^*}\right)}{p^* + \hat{\xi}_{p^*}\,(q - p^*)}\right)^{\hat{r}}}{1 - (1 - \hat{\xi}_{p^*})^{\hat{r}}}. \tag{S13}$$

Hence we obtained an explicit formula for the species-accumulation curve for every $q \leq 1$ from the local up to the global scale.
Moreover we can express the RSA distribution at the global scale by plugging the estimated parameters $\hat{\xi}$ and $\hat{r}$ into (S1).

S1.5 Popularity and abundance variation through scales

Note that until now we studied the abundance distribution of the observed species at the local scale, but only to estimate the number of unseen species, disregarding of their abundances. However, abundance information may of relevance in some contexts. For example, if one is interested in measuring the popularity of hashtags in Twitter, one naive way to do that is to count the number of times it has been posted. A second novelty we introduced in our work is indeed a method to estimate the variation of popularity in social networks. Let us first recall our previous findings using a more detailed notation which turns out to be essential in the following.

DEFINITION S1.2. *For every $s = 1, ..., S$, we indicate with $n_s^{p^*}$, $n_s^{1-p^*}$ the abundance of species s in the observed (resp. unobserved) fraction p^* (resp. $1 - p^*$) of the population.*

- First, let us introduce the statistics:

$$S_{p^*} = \sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*} > 0\}}$$

  whose expected value can be computed as follows:

$$\mathbb{E}\left[S_{p^*}\right] = \mathbb{E}\left[\sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*} > 0\}}\right] = \sum_{s=1}^{S} \mathbb{E}\left[\mathbb{1}_{\{n_s^{p^*} > 0\}}\right] = \sum_{s=1}^{S} \mathbb{P}\left(n_s^{p^*} > 0\right)$$

$$= S \cdot P\left(k > 0 | p^*\right) = S \cdot \left[1 - P\left(k = 0 | p^*\right)\right].$$

- Arranging the latter equation, we can isolate the quantity we are interested to estimate:

$$S = \frac{\mathbb{E}\left[S_{p^*}\right]}{1 - P\left(k = 0 | p^*\right)}. \tag{S14}$$

- An estimator of S can be thus obtained by replacing the mean $\mathbb{E}\left[S_{p^*}\right]$ by the observable S_{p^*}:

$$\hat{S} = \frac{S_{p^*}}{1 - P\left(k = 0 | p^*\right)} \tag{S15}$$

  With no surprise, we recover the same result as in (S12). We wish to stress that this new formulation allows us to push further our investigation, as we are going to show.

We wish now to apply the same procedure to different statistics.
Recall that we are sampling S_{p^*} species at scale p^* from a pool consisting of N individuals belonging to S different species. If a species s is not observed in the sample at scale p^*, we say that s is a "new" species. The meaning of this definition can be easily explained. If you imagine to further sample your population, you can either pick individuals belonging to species already observed or you can discover indeed "new"

species.
Let us then consider the following statistics for the new species:

$$S^{\text{new}}_{1-p^*} = \sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=0, n_s^{1-p^*}>0\}}. \tag{S16}$$

The following chain of equality turns out to be meaningful in the following:

$$\begin{aligned} S^{\text{new}}_{1-p^*} &= \sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=0, n_s^{1-p^*}>0\}} = \sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=0, n_s^{1}>0\}} \\ &= \sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=0\}} = \sum_{s=1}^{S} \left(1 - \mathbb{1}_{\{n_s^{p^*}>0\}}\right) = S - S_{p^*}. \end{aligned}$$

We can recover an estimator for the "new" species from estimator (S15) for S.
This remark seems trivial, and the chain of equation above appears redundant. Nevertheless, it is crucial for the development of our work. We stress that the statistics $S^{\text{new}}_{1-p^*}$ uses both the information at the sample scale p^* and the information contained in the unseen fraction of the population $1-p^*$. In contrast, the statistics for S_{p^*} only consider the observed individuals.
Given now the statistics (S16) representing the number of species unobserved in the sample of size p^*N but present in the remaining population of size $(1-p^*)N$. We wish to recover an estimator for the new species $S^{\text{new}}_{1-p^*}$. We thus compute the expected value of the corresponding statistics:

$$\begin{aligned} \mathbb{E}\left[S^{\text{new}}_{1-p^*}\right] &= \mathbb{E}\sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=0, n_s^{1-p^*}>0\}} = S \cdot \mathbb{P}\left(n_s^{p^*}=0, n_s^{1-p^*}>0\right) \\ &= S \cdot \mathbb{P}\left(n_s^{p^*}=0, n_s^{1}>0\right) = S \cdot \underbrace{\mathbb{P}\left(n_s^{p^*}=0\right)}_{P\left(k=0|p^*\right)}. \end{aligned}$$

The expected value turns out to be a product of two factors: $P\left(k=0|p^*\right) = \mathbb{P}(n_s^{p^*}=0)$, which can be computed via (S2), and S, a quantity we can estimate via (S15). Hence we derive the following estimator:

$$\hat{S}^{\text{new}}_{1-p^*} = \frac{S_{p^*}}{1-P\left(k=0|p^*\right)} \cdot P\left(k=0|p^*\right).$$

This procedure captures the techniques which allows us to derive other useful estimators.
In particular, this turning point leads us to new statistics that consider also the popularity.
Let us start from the statistics:

$$S^{\text{new}}_{1-p^*}(l) = \sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=0, n_s^{1-p^*}=l\}}. \tag{S17}$$

Note that if we get an expression for $S^{\text{new}}_{1-p^*}(l)$, than we can easily extend the result to

$$S^{\text{new}}_{1-p^*}(\geq L) = \sum_{l=L}^{\cdot} S^{\text{new}}_{1-p^*}(l).$$

Moreover, results from the previous section can be included here, simply noticing that:

$$S^{\text{new}}_{1-p^*} = S^{\text{new}}_{1-p}(\geq 1) = \dot{\sum_{l=1}} S^{\text{new}}_{1-p^*}(l).$$

We proceed as before by computing the expected value:

$$\begin{aligned}
\mathbb{E}\left[S^{\text{new}}_{1-p^*}(l)\right] &= \mathbb{E}\left[\sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=0, n_s^{1-p^*}=l\}}\right] \\
&= S \cdot \mathbb{P}\left(n_s^{p^*} = 0, n_s^{1-p^*} = l\right) \\
&= S \cdot \mathbb{P}\left(n_s^{p^*} = 0, n_s^{1} = l\right) \\
&= S \cdot \underbrace{\mathbb{P}\left(n_s^{p^*} = 0 | n_s^{1} = l\right)}_{Binomial(n_s^1, p^*)} \underbrace{\mathbb{P}\left(n_s^{1} = l\right)}_{P(l|1)},
\end{aligned}$$

where in the third equality we used the following relation:

$$\mathbb{P}\left(n_s^{p^*} = x, n_s^{1-p^*} = y\right) = \mathbb{P}\left(n_s^{p^*} = x, n_s^{1} = x + y\right).$$

Let us note now the following facts:

- From the sampling binomial distribution, it holds that $\mathbb{P}\left(n_s^{p^*} = 0 | n_s^{1} = l\right) = (1-p^*)^l$;
- $\mathbb{P}\left(n_s^{1} = l\right) = P(l|1)$ is given by (S1);
- S is unknown and we thus need an estimator for it.

Again, we can use the results of the previous subsection to define $\hat{S} = \dfrac{S_{p^*}}{1 - P\left(k = 0 | p^*\right)}$ and hence to obtain

$$\hat{S}^{\text{new}}_{1-p^*}(l) = \hat{S} \cdot (1-p^*)^l \cdot P(l|1) = \frac{S_{p^*}}{1 - P\left(k = 0 | p^*\right)} \cdot (1-p^*)^l \cdot P(l|1), \tag{S18}$$

which is the estimator for the new species with abundance l.
Thus, as a first partial result, we obtained an estimator for the popularity of the new species.
Let us now consider the statistics:

$$S_{1-p^*}(l \to k) = \sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=l, n_s^{1-p^*}=k\}}, \tag{S19}$$

which represents the number of species having contemporarily abundance l at the observed scale p^* and abundance k at the unobserved scale $1-p^*$. Note that we can compute the number of species having an abundance that lies within a population interval by summing up on different values of l and k. We proceed

by computing the expected value of the statistics (S19):

$$\begin{aligned}\mathbb{E}\left[S_{1-p^*}(l \to k)\right] &= \mathbb{E}\left[\sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=l, n_s^{1-p^*}=k\}}\right] \\ &= S \cdot \mathbb{P}\left(n_s^{p^*} = l, n_s^{1-p^*} = k\right) \\ &= S \cdot \mathbb{P}\left(n_s^{p^*} = l, n_s^1 = k+l\right) \\ &= S \cdot \underbrace{\mathbb{P}\left(n_s^{p^*} = l | n_s^1 = k+l\right)}_{Binomial(n_s^1, p^*)} \underbrace{\mathbb{P}\left(n_s^1 = k+l\right)}_{P(k+l|1)}.\end{aligned}$$

Now we have the following:

- From the sampling binomial distribution, it holds that $\mathbb{P}\left(n_s^{p^*} = l | n_s^1 = k+l\right) = \binom{k+l}{l} p^{*l}(1-p^*)^k$;
- $\mathbb{P}\left(n_s^1 = k+l\right) = P(k+l|1) = c(r,\xi)\binom{k+l+r-1}{k+l}\xi^{k+l}(1-\xi)^r$;
- S is unknown. However, we can estimate it via $\hat{S} = \dfrac{S_{p^*}}{1 - P\left(k=0|p^*\right)}$.

Hence we obtained

$$\begin{aligned}\hat{S}_{1-p^*}(l \to k) &= \hat{S} \cdot \mathbb{P}\left(n_s^{p^*} = l | n_s^1 = k+l\right) \cdot P(k+l|1) \\ &= \frac{S_{p^*}}{1 - P(0|p^*)} \cdot \binom{k+l}{l} p^{*l}(1-p^*)^k \cdot c(r,\hat{\xi})\binom{k+l+\hat{r}-1}{k+l}\hat{\xi}^{k+l}(1-\hat{\xi})^{\hat{r}}.\end{aligned}$$

Estimator $\hat{S}_{1-p^*}(l \to k)$ above gives the number of species with abundance l at the observed scale p^* and abundance k at the unobserved scale $1-p^*$. Note that this estimator is independent of the number of species with abundance l at scale p^*; indeed, we are using the sample at scale p^* only to estimate the parameters ξ_{p^*} and r, which we need to predict $\hat{S}$. Hence we are only using partial information at the local scale.

We wish now to take into account the information about the number of species with abundance l at the surveyed scale, $S_{p^*}(l)$. In particular, we are looking for an estimator of the species with abundance k in the unobserved fraction $1-p^*$ of the population, given that they have abundance l in the sample at the observed scale p^*.

We thus define $S_{p^*}(l) := \sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=l\}}$.

In the following we will need to use quantities of the type $\mathbb{P}(n_s^{1-p^*} = k | n_s^{p^*} = l)$.

Using Bayes' theorem, we obtain

$$\begin{aligned}\mathbb{P}(n_s^{1-p^*} = k | n_s^{p^*} = l) &= \mathbb{P}(n_s^1 - n_s^{p^*} = k | n_s^{p^*} = l) \\ &= \mathbb{P}(n_s^1 - l = k | n_s^{p^*} = l) \\ &= \mathbb{P}(n_s^1 = k + l | n_s^{p^*} = l) \\ &= \frac{\mathbb{P}(n_s^{p^*} = l | n_s^1 = k+l)\mathbb{P}(n_s^1 = k+l)}{\mathbb{P}(n_s^{p^*} = l)}.\end{aligned}$$

Note that we all the probabilities appearing in the latter formula are known, since:

- $\mathbb{P}\left(n_s^{p^*} = l | n_s^1 = k+l\right) = \binom{k+l}{l} p^{*l}(1-p^*)^k$ is the sampling binomial distribution;
- $\mathbb{P}\left(n_s^1 = k+l\right) = P(k+l|1) = c(r,\xi)\binom{k+l+r-1}{k+l}\xi^{k+l}(1-\xi)^r$ is the global truncated negative binomial distribution of parameters r and ξ as in (S1);
- $\mathbb{P}\left(n_s^{p^*} = l\right) = P(l|p^*) = c(r,\xi)\binom{l+r-1}{l}\xi_p^{*l}(1-\xi_{p^*})^r$ is again a truncated negative binomial with rescaled parameter ξ_p as in (S2).

Let us now retrace the same steps as for $\hat{S}_{1-p^*}(l \to k)$ for the conditional estimator $\hat{S}_{1-p^*}(k|l)$. We start from the statistics

$$S_{1-p^*}(k|l) = \sum_{s=1}^{S} \mathbb{1}_{\{n_s^{p^*}=l\}} \mathbb{1}_{\{n_s^{1-p^*}=k, n_s^{p^*}=l\}} = \sum_{s=1}^{S_{p^*}(l)} \mathbb{1}_{\{n_s^{1-p^*}=k | n_s^{p^*}=l\}}.$$

We proceed by computing the expected value

$$\mathbb{E}\left[S_{1-p^*}(k|l)\right] = S_{p^*}(l) \cdot \mathbb{P}\left(n_s^{1-p^*} = k | n_s^{p^*} = l\right) = S_{p^*}(l) \cdot \frac{\mathbb{P}\left(n_s^{p^*} = l | n_s^1 = k+l\right)\mathbb{P}\left(n_s^1 = k+l\right)}{\mathbb{P}\left(n_s^{p^*} = l\right)}.$$

Note that empirically $\mathbb{P}\left(n_s^{p^*} = l\right) = S_{p^*}(l)/S$, so that we can recover $\mathbb{E}\left[S_{1-p^*}(l \to k)\right]$.
Let us now insert into the above formula the probabilities computed by using the fitted parameters:

$$\hat{S}_{1-p^*}(k|l) = S_{p^*}(l) \cdot \frac{\binom{k+l}{l} p^{*l}(1-p^*)^k \cdot \binom{k+l+\hat{r}-1}{k+l}\hat{\xi}^{k+l}(1-\hat{\xi})^{\hat{r}}}{\binom{l+\hat{r}-1}{l}\hat{\xi}_{p^*}^l(1-\hat{\xi}_{p^*})^{\hat{r}}},$$

where the terms $c(r,\hat{\xi})$ in the numerator has cancelled out with the one at the denominator.
Estimator $\hat{S}_{1-p}(k|l)$ is theoretically unbiased.
Note that, again, we can pass from punctual estimation to cumulative ones, by summing up over all l and k values above some fixed thresholds L and K, respectively:

$$\hat{S}_{1-p^*}(\geq K | \geq L) = \sum_{l \geq L} \sum_{k \geq K} \hat{S}_{1-p^*}(k|l) \tag{S20}$$

Estimator (S20) is the one we are going to test in our databases.

S2 ADDITIONAL RESULTS AND FIGURES

In this section we collect some additional results not presented in the main text.

S2.1 Upscaling results from sample scale $p^* = 3\%$

In the main text we showed the results we obtained with our upscaling method when sampling a fraction $p^* = 5\%$ of the four databases. We performed the same tests also for a local scale $p^* = 3\%$, with similar

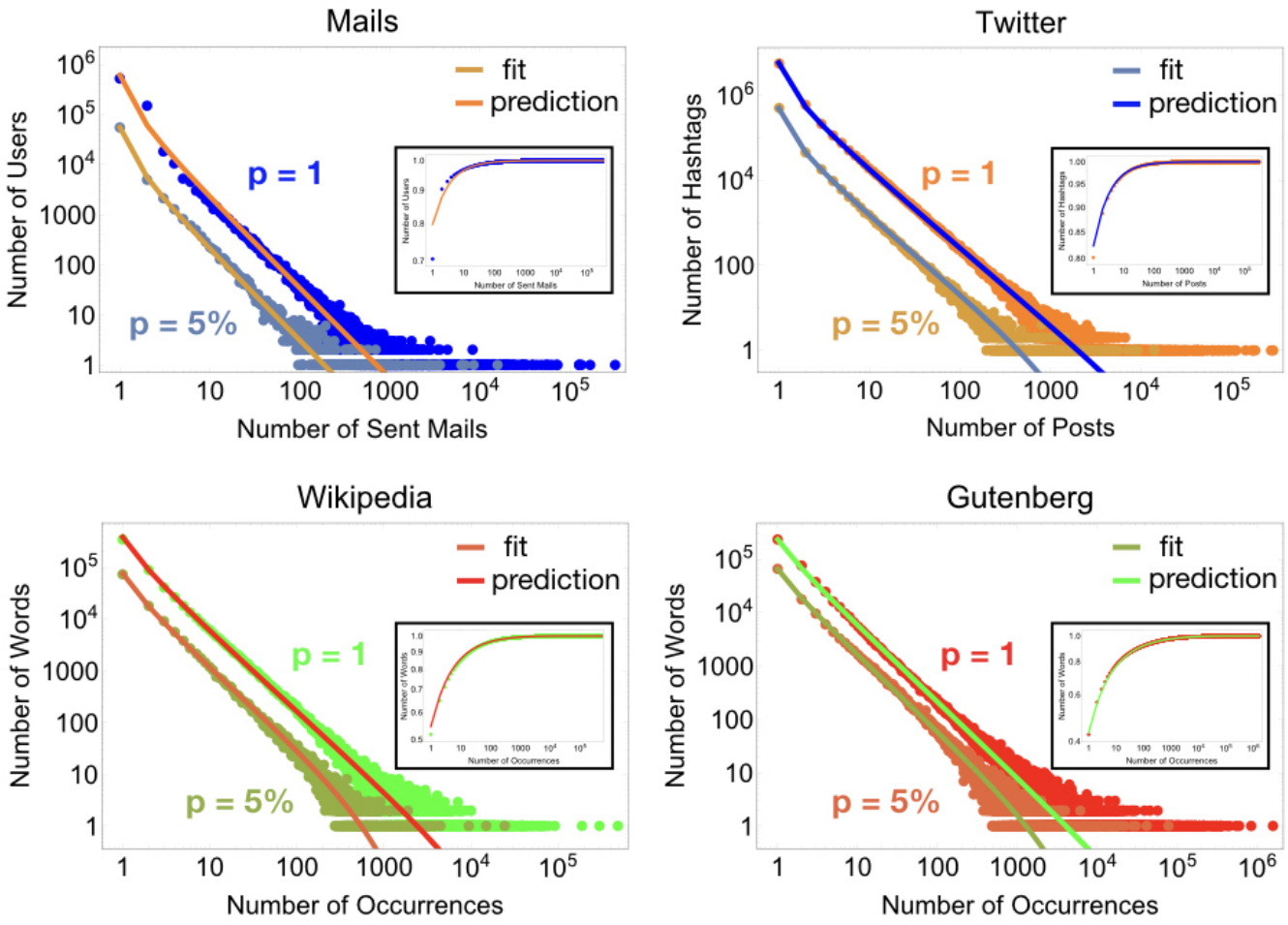


Figure S1: **Best-fit and predicted patterns from a local sample scale $p^* = 5\%$.** Empirical RSA curves at global scale ($p = 1$) and local scale ($p^* = 5\%$) are shown. In each panel, coloured lines over the local RSAs represent the distribution obtained via a best-fit of the empirical pattern with a negative binomial having $r \in (-1, 0)$. Lines over the global RSA distributions represent our prediction for the RSAs at the global scales obtained via our upscaling equations for both the parameters and the biodiversity. In each panel, insets showing the corresponding global cumulative RSA (both empirical and predicted) are added.

results.

First of all, as shown in Figure S2, also for the case $p^* = 3\%$ we observe the form-invariance property of the empirical RSAs for all the considered human activity datasets.

Moreover, as for $p^* = 5\%$, we tested the reliability of estimator (S12) in predicting the total number of species in the different networks when only a random portion of them is extracted. Table S1 displays the relative percentage error we obtained for the different databases together with the total dataset composition and the values of the parameters fitted from the empirical RSAs at $p^* = 3\%$.

S2.2 Upscaling results for popularity change

In the main text we exhibited in Table 2 the results for the predictions of popularity (via the conditional estimator S20) in the unsurveyed fraction $1 - p^* = 0.95$ of the population for a fixed value of the local popularity threshold $L = 10$. In Table S2 we show the results obtained for different values of L and K.

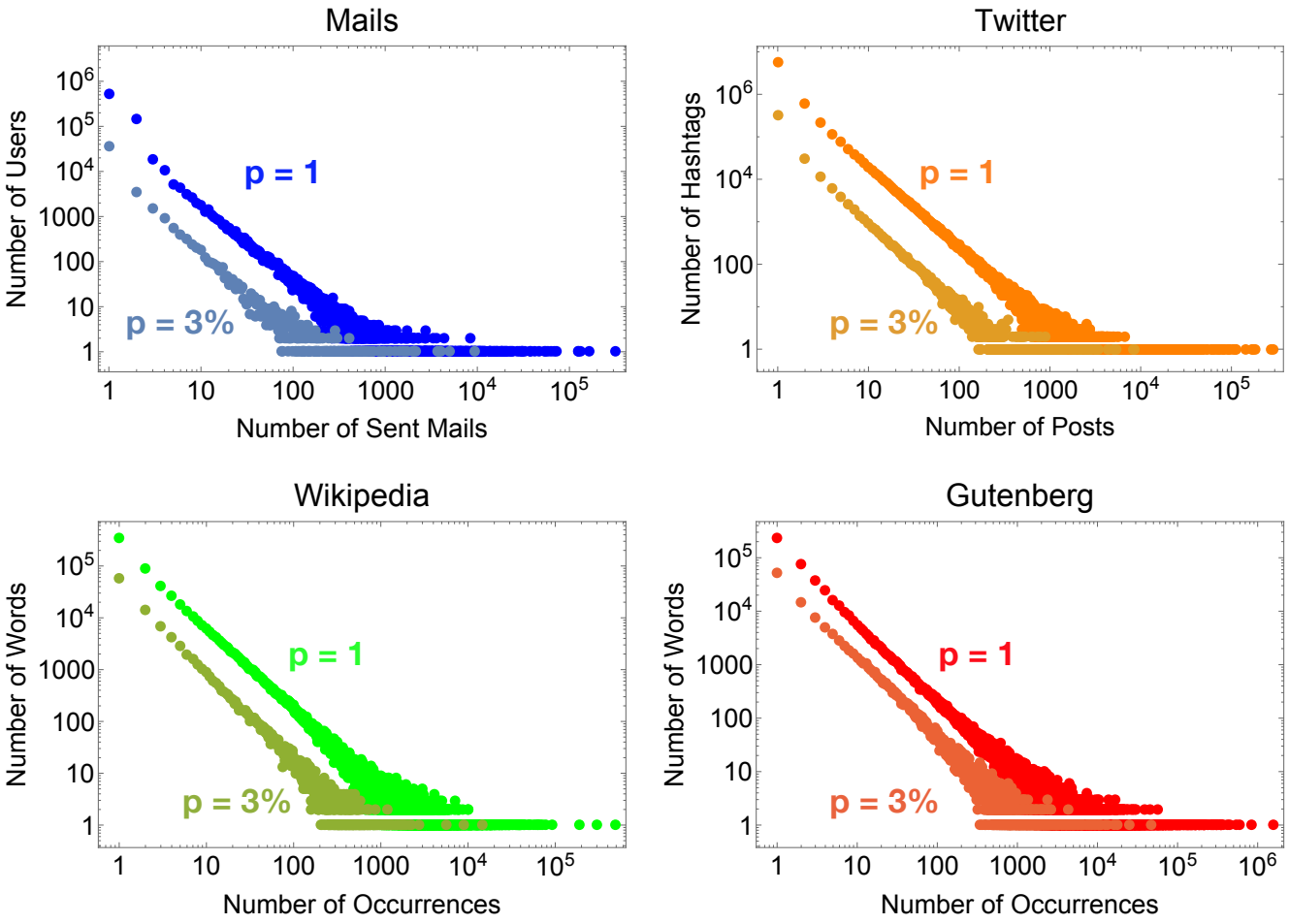


Figure S2: **Universality and form-invariance of the empirical RSAs.** Empirical RSA curves at the global scale ($p = 1$) and the local scale ($p^* = 3\%$) are shown. RSA is scale-free in all the four datasets analyzed, with a power-law form maintained through the different human activities and scales. RSA form-invariance property is at the core of our theoretical framework.

	Emails	Twitter	Wikipedia	Gutenberg
Species	$752,299$	$6,972,453$	$673,872$	$554,193$
Individuals	$6,914,872$	$34,696,973$	$29,606,116$	$126,289,661$
$\mathbf{r}$	-0.788	-0.828	-0.549	-0.422
$\xi_{\mathbf{p}^*}$	0.9997	0.9976	0.9987	0.9994
Relative Error	-2.74%	4.41%	8.22%	-3.52%

Table S1. Predicted relative errors. Upscaling results for the number of species of the four analysed datasets from a local sample covering a fraction $p^* = 3\%$ of the global database. For each database, we display the number of species (users, hashtags, words) and individuals (sent mails, posts, occurrences) at the global scale, together with the fitted RSA distribution parameters at the sampled scale and the relative percentage error between the true number of species and the one predicted by our framework.

S2.3 Local Analysis

We also tested how estimator (S12) performs on different spatial sub-scales. In this case, due to the huge amount of data, we chose to work with a smaller datasets for a systematic analysis. In particular, we considered as global four samples of the original datasets each covering a fraction $p^* = 5\%$ of the total

Table S2. Percentage errors for popularity change predictions in Twitter database. For $L = 10,\ 40,\ 55$ (first column) and different values of K(second column), we estimated, from ten different Twitter samples (at $p = 5\%$), 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^* (see estimator 4 of the main text). The average true number of species $S_{1-p^*}(\geq K| \geq L)$ and the average one predicted by our method 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 among the ten predictions.

L	K	$S_{1-p^*}(\geq K\| \geq L)$	$\hat{S}_{1-p^*}(\geq K\| \geq L)$	Relative Error	Variance
10	77	14,266	14,274.38	−0.0029	0.0012
10	115	14,113	14,105.65	0.0534	0.0151
10	154	13,551	13,544.76	0.2457	0.0428
10	192	12,509	12,584.32	0.4679	0.0731
10	231	11,305	11,366.66	0.5372	0.0965
40	362	3,749	3,748.99	−0.0001	≈ 0
40	543	3,742	3,741.96	0.0393	0.0058
40	724	3,591	3,578.83	−0.0715	0.0668
40	905	3,096	3,091.45	0.0368	0.0660
40	1,086	2,600	2,582.75	−0.5634	0.0370
55	504	2,673	2,673.00	≈ 0	≈ 0
55	756	2,672	2,670.96	−0.0141	0.0013
55	1,008	2,569	2,567.71	−0.0978	0.0565
55	1,260	2,195	2,199.11	0.0023	0.0557
55	1,512	1,806	1,820.01	0.1286	0.2070

amount of data (see Figure S3).

We then randomly sub-sampled the reduced 5% databases at different sub-scales p^{**} ranging from 10% to 90% and applied our framework to predict the number of species observed at p^* (here considered as $p = 1$). In Figure S3, bottom panels, we displayed the relative percentage error graphs between the true number of species, S^*, and the one predicted from the local information at the different sub-scales p^{**}. We see that, for all datasets and sub-scales, our method always led to an error below 5%. Moreover, it displays an intuitive decreasing behavior as the available information increases, a desirable property for an estimator. We performed the same analysis also starting from a sample at the scale $p^* = 3\%$, obtaining comparable results (see Figure S4).

REFERENCES

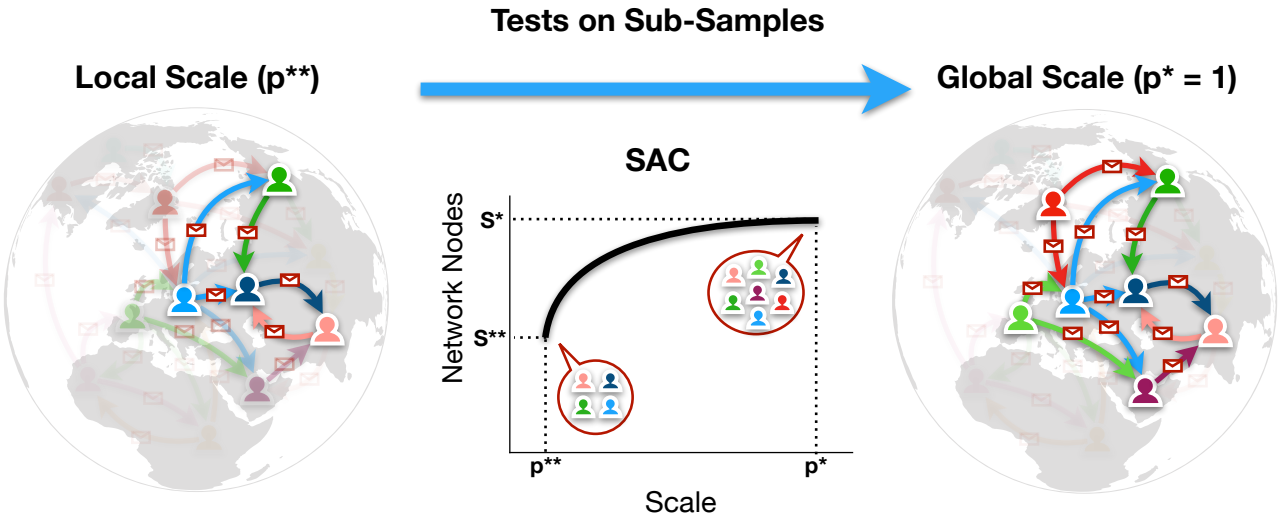
Flajolet, P. and Sedgewick, R. (2008). *Analytic Combinatorics* (Cambridge University Press)

Tovo, A., Formentin, M., Suweis, S., Stivanello, S., Azaele, S., and Maritan, A. (2019). Inferring macro-ecological patterns from local species' occurrences. *Oikos* doi:10.1111/oik.06754

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

Walraevens, J., Demoor, T., Maertens, T., and Bruneel, H. (2012). Stochastic queueing-theory approach to human dynamics. *Physical Review E* 85, 021139

Figure S3: **Relative percentage errors at different sub-scales from $p^* = 5\%$.** Starting from a sample at $p^* = 5\%$ of each human activity database, we sub-sampled it at different spatial sub-scales $p^{**} \in \{10\%, \ldots, 90\%\}$ of p^* and computed the relative percentage error between the number of predicted species, $\hat{S}^*$, and the true number of species, S^*, observed in the sample at p^*, here considered as the global scale ($p^* = 1$).



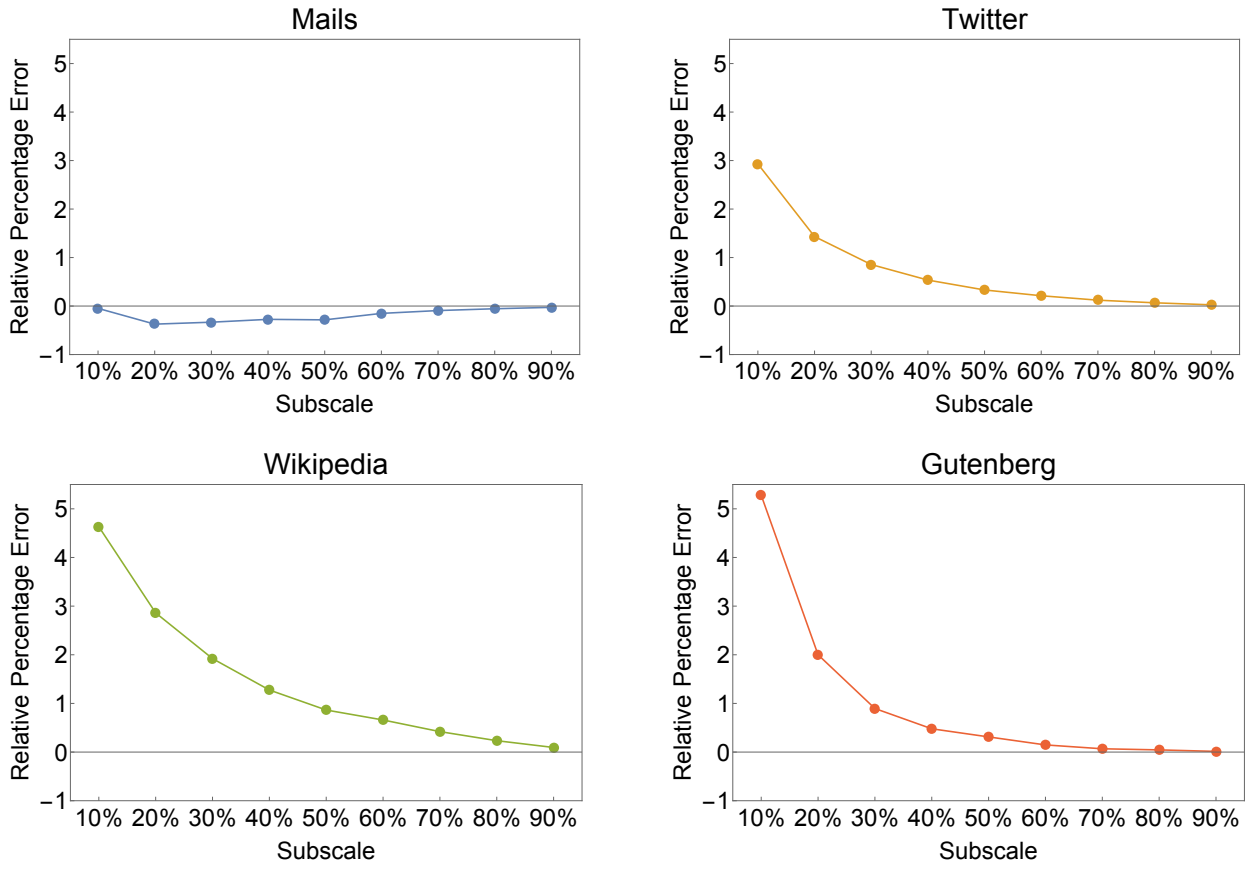


Figure S4: **Relative percentage errors at different sub-scales from $\mathbf{p^* = 3\%}$.** Starting from a sample at $p^* = 3\%$ of each human activity database, we sub-sampled it at different spatial sub-scales $p^{**} \in \{10\%, \ldots, 90\%\}$ of p^* and computed the relative percentage error between the number of predicted species, $\hat{S}^*$, and the true number of species, S^*, observed in the sample at p^*, here considered as the global scale ($p^* = 1$).