Law of corresponding states for open collaborations

Marco Gherardi, Federico Bassetti, Marco Cosentino Lagomarsino

Law of corresponding states for open collaborations

Marco Gherardi,[1, 2, 3, *] Federico Bassetti,[4] and Marco Cosentino Lagomarsino[1, 5]

[1] *Sorbonne Universités, UPMC Univ Paris 06, UMR 7238,*
Computational and Quantitative Biology, 15 rue de l'École de Médecine Paris, France
[2] *Dipartimento di Fisica, Università degli Studi di Milano, via Celoria 16, 20133 Milano, Italy*
[3] *I.N.F.N. Milano*
[4] *Dipartimento di Matematica, Università di Pavia, Pavia, Italy*
[5] *CNRS, UMR 7238, Paris, France*

We study the relation between number of contributors and product size in Wikipedia and GitHub. In contrast to traditional production, this is strongly probabilistic, but is characterized by two quantitative nonlinear laws: a power-law bound to product size for increasing number of contributors, and the universal collapse of rescaled distributions. A variant of the random-energy model shows that both laws are due to the heterogeneity of contributors, and displays an intriguing finite-size scaling property with no equivalent in standard systems. The analysis uncovers the right intensive densities, enabling the comparison of projects with different numbers of contributors on equal grounds. We use this property to expose the detrimental effects of conflicting interactions in Wikipedia.

I. INTRODUCTION

Achieving a quantitative understanding of collective human activities is both a challenge and an opportunity for contemporary statistical physics [1–3]. In our society, new important forms of production are promoted by information-communication technology and the internet. *Crowdsourcing* is the process of obtaining contributions to a project (services, ideas, or content) by soliciting input from a large online community. This new way of collaborating is changing the scale and efficiency of social endeavors [4–7].The success of open collaborations as participative self-organised projects has catalysed new ways of thinking about innovation and sustainability [8]. Some of the main open questions concern the efficiency and the predictability of such social collaboration processes [9–15]. However, despite the large amount of available data, they are still largely unexplored quantitatively, making it difficult to interpret empirical data and make useful predictions [16–19].

A long-standing question in software engineering, relevant for other productive processes, concerns the relations between different variables characterising a project, such as size, number of developers, effort, and duration. Classic empirical studies have found polynomial scaling laws relating these quantities [20, 21]. The existence of clear relationships among these variables is crucial for estimating the costs for a project of a given size and complexity [22, 23]. In particular, the question of how the size of a piece of software is related to the total effort (the number of contributors times the time spent) was first addressed in the pioneering book "The mythical man month" [24], where a superlinear scaling was conjectured.

In this Letter, we focus on the relationship between the number of contributors and the size of open collaborative projects [25], and consider two paradigmatic examples: Wikipedia, the free on-line encyclopaedia, and GitHub, the web-based platform for collaborative software development. Taking a statistical-mechanics approach to these data, we define a novel variant of the random-energy model (REM), an exactly tractable model used to describe disordered systems in statistical physics. This "neutral" model captures all the salient aspects relating

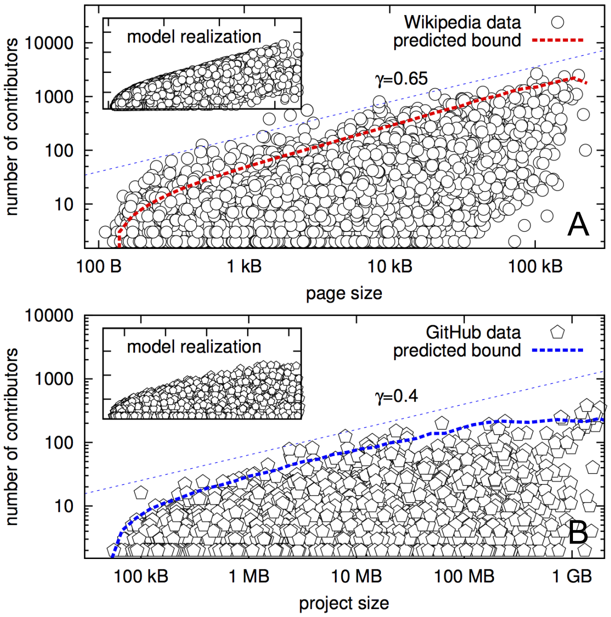

FIG. 1. The relation between effort and project size is probabilistic and shows the presence of sublinear bounds. Both features are demonstrated by scatterplots of the number of contributors versus the project size in Wikipedia pages (A) and GitHub software (B). The thick dashed lines show the average bounds from several realisations of the model described in the text (examples are shown in the insets): dotted lines are reference power laws fitted from the data (we used the procedure developed in [30] for smoothing out the roughness and fitting the power-law exponent of the bound).

[*] Corresponding author: marco.gherardi@mi.infn.it

size to contributors without the need to explicitly include agent interactions [26]. The crucial ingredient is the highly heterogeneous activity of contributors, whose commitment, measured as the number of edits, is power-law distributed [26–29] (possibly also due to interactions).

II. RESULTS

A. Number of contributors and project size are not related deterministically

We collected data on size (in bytes) and number of contributors for 20 000 GitHub projects and 400 000 Wikipedia pages in English. Precisely, the GitHub set contains 20 000 projects chosen at random among the 100 000 projects with the largest number of followers. The Wikipedia set contains 200 000 alphabetically-ordered pages starting with letter A, and the same number starting with M, comprising ∼10% of the English Wikipedia. Data were retrieved with the APIs of Wikipedia (English Wikipedia; en.wikipedia.org/w/api.php, accessed 8 May 2014) and GitHub (developer.github.com, accessed 21 March 2014).

Figure 1 shows that the variability in the size of projects with a given number of authors is very high. Clearly, no simple functional dependency can give a satisfactory description of the trends. Rather, they are better expressed by the joint probability distribution of the two variables, which we investigate in detail below.

B. An anomalous upper bound limits the number of contributors for a project of a given size

Albeit highly dispersed, the number of developers shows a clear size-dependent bound. This is visible in Fig. 1, where it is compared with (markedly sublinear) power laws. While the relation is probabilistic, a scaling law appears to describe the minimum size of a project with a given number of contributors, or, equivalently, the maximum number of contributors to a project of a given size. Remarkably, both Wikipedia and GitHub display the same non-extensive feature. Such a constraint is unusual, but a similar phenomenology (a “soft bound”) was found in other empirical systems [30, 31]. In brief, the scenario for open collaborations is one where no deterministic law exists between product size and “man-cost” (measured here as number of contributors, not man-months), but a sublinear scaling relates the *maximum cost* with the size. These results give a fresh look to the question of the “Mythical man month” [24] for the case of open collaborations. Central to this debate is the impossibility of measuring progress as number of men times number of months. Since complex tasks cannot be partitioned, because of hierarchical constraints (e.g., efforts in communication, coordination, etc.), cost is expected to scale with project size in a poorer than linear way.

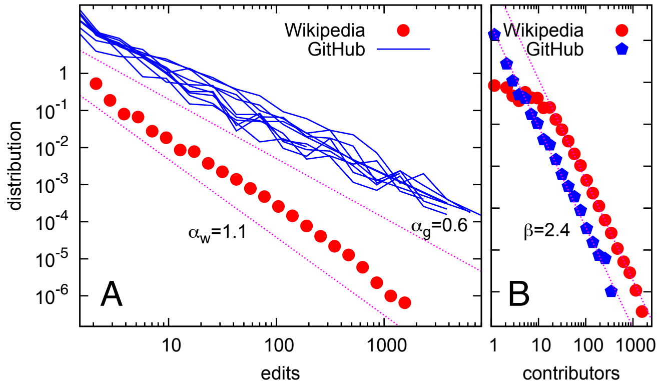

FIG. 2. The number of edits per contributor (A) and the total number of contributors per project (B) follow fat-tailed distributions. Wikipedia edits (circles) are aggregated for all pages, while GitHub edits (lines) are shown separately for the 10 projects with most contributors (shifted upwards by a factor 100 to increase readability). Dotted lines show reference power laws with exponents $\alpha+1$ (left panel) and β (right panel) fitted from data: $\alpha_{\mathrm{w}} = 1.1(1)$, $\alpha_{\mathrm{g}} = 0.6(1)$, and $\beta_{\mathrm{w}} = \beta_{\mathrm{g}} = 2.4(1)$.

C. Effort and number of contributors are widely distributed

The marginal distribution of the number of contributors, i.e., the number $N(n)$ of projects with a given number n of contributors, is well described by a power law, $N(n) = N_1 n^{-\beta}$, for both Wikipedia and Github (Fig. 2), where N_1 is the number of one-man projects. Such a wide distribution has been already noted, and may reflect preferential-attachment dynamics [32], or the variable intrinsic appeal of projects [33]. A relevant additional observation regards the distribution of contributor activity, estimated by the total number of edits per contributor: it has a large-activity tail that follows a power law $P(A) \sim A^{-(\alpha+1)}$ (Fig. 2), with $\alpha \approx 1.1$ in Wikipedia and $\alpha \approx 0.6$ in GitHub. This confirms previous results on human activity in these and other systems (see e.g. [26, 32, 34]).

D. Size distributions collapse onto a single characteristic curve

Conversely, conditional size distributions at fixed number of contributors show a striking regularity. Rescaling size with an appropriate power of n exposes a universal curve common to all marginal distributions (separately for the two model systems, Fig. 3). The best-collapse exponent [35] is approximately 0.8 for Wikipedia and 1.7 for Github. We note that the inverse of these values (1.25 and 0.59 respectively) coincide quite precisely with the exponents α obtained above. This leads to the definition of the *reduced size* as the empirical size (in bytes) multiplied by $n^{-1/\alpha}$. This nonlinear rescaling realises the correct intensive quantity (the “specific size”) in these

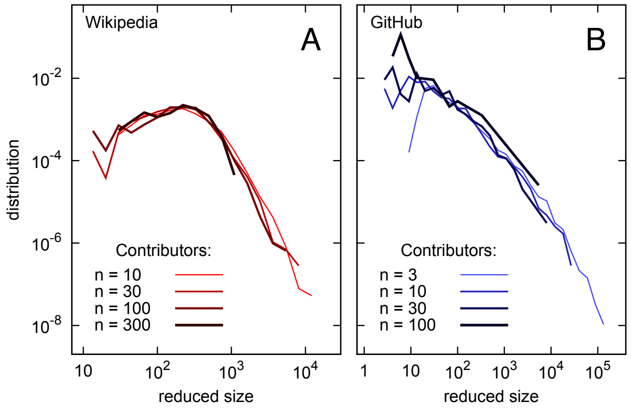

FIG. 3. The marginal size distributions at fixed number of contributors collapse onto universal curves, characterising Wikipedia (A) and GitHub (B). The power-law rescaling yielding the best collapse is given by n raised to the exponent 0.8 for Wikipedia and 1.7 for Github. These exponents (obtained by a a minimization procedure [35]) are equal to the inverse of the activity exponent α, as predicted by the model.

non-extensive systems and defines a "law of corresponding states", in analogy with the eponymous thermodynamic law, asserting that all van der Waals gases behave alike at the same reduced conditions. Importantly, the universal collapse allows to compare projects with widely different numbers of contributors through their reduced size.

E. A random-energy model explains the scaling laws in a non-interacting scenario

Capturing the regularities within the strong stochasticity of these data demands an approach based on the methods of statistical mechanics. We find that the basic mechanisms and observations are fully elucidated by the following analytically solvable stochastic model. Each of the n users working on a project (piece of software or page) adds a number of contributions proportional to her/his intrinsic activity, independently of the actions of other users. The model assumes that each edit contributes a fixed amount to the project's size; this choice is justified by the compact edit-size distribution (see [13]). The size of a project with n active users is modelled as the sum X of n independent and identically-distributed random variables $\{A_i\}_{i=1\dots n}$, extracted from the activity distribution $P(A)$. For economy of parameters, subtractive contributions are neglected; this ingredient enhances noise near the bound and is likely responsible for the points lying outside the predicted bounds in Fig. 1 (more details are given in Sec. II G). (We note in passing the similarity of this framework with a model of bursty human dynamics [36]; see also [37].) One key result can be obtained via the generalised central-limit theorem (CLT) [38], which constrains the sum of a large number of random variables to obey the *α-stable* distribution, where the parameter α is related to the tail exponent of $P(A)$. Such variables satisfy a notable scaling relation, namely $X \sim n^{1/\alpha} A$ [39]. Therefore, the model predicts the collapse of the distributions at different n, when size is rescaled by $n^{1/\alpha}$, as observed empirically.

We now turn to the bound, i.e., the minimum $x_{\rm m}$ of the sum of the contributions at fixed n. This quantity shows an intriguing finite-size scaling property. One may expect an "extensive" scaling, where $x_{\rm m}$ is linearly proportional to the number of contributors n, since the convolution of n distributions with a fixed lower cutoff in a_0 has support in (na_0, ∞): taking an infinite number of samples N at fixed n gives a linear scaling independently of the tail exponent. In our case, instead, the reverse order of taking this limit (physically meaningful observations are performed on a finite system), together with the sub-exponential scaling of the sampling $N(n)$, may lead to a non-trivial scaling law for the bound, $x_{\rm m}^\gamma \sim n$, with $\gamma < 1$. We have computed the exponent γ considering in particular $N(n) = N_1 n^{-\beta}$ (for positive β, one has the empirical case; note that $N(n) \geq 1$). Our calculation is based on the asymptotic form of the cumulative distribution function of a stable law L_α, which has the following behavior [39] for x close to x_0 (with $x_0 \equiv 0$ if $\alpha < 1$, and $x_0 = -\infty$ if $\alpha > 1$):

$$\int_{x_0}^{x} L_\alpha(y)\,\mathrm{d}y \sim |x|^{\frac{\alpha}{2(1-\alpha)}} \exp\left[-|1-\alpha|\left|\frac{x}{\alpha}\right|^{-\frac{\alpha}{1-\alpha}}\right]. \quad (1)$$

The typical minimum of X is estimated as the value $x_{\rm m}$ such that $P\{X \leq x_{\rm m}\} = \frac{1}{N(n)}$. When $x_{\rm m}$ is small this condition becomes

$$\tilde{x}^{1/2}\exp\left[-\left(\frac{1-\alpha}{\tilde{x}}\right)\alpha^{\frac{\alpha}{1-\alpha}}\right] = \frac{n^\beta}{N_1}\ , \quad (2)$$

where $x_{\rm m}$ appears only through the scaling variable $\tilde{x} = x_{\rm m}^{\alpha/(1-\alpha)} n^{-1/(1-\alpha)}$. For constant $N(n)$ ($\beta = 0$) the equation only depends on $\tilde{x}$, thus giving the scaling $n \sim x_{\rm m}^\alpha$ (i.e., $\gamma = \alpha$). For $\beta > 0$, deviations set in, such that the effective exponent γ observed is a slowly varying function of $x_{\rm m}$ (see the Appendix). Specifically, for small n the effective exponent takes the value $\gamma = \alpha$ independently of β, while for large n (i.e, when $N(n) \approx 1$), the bound is well described by a power law of exponent

$$\gamma = \alpha - \frac{2(1-\alpha)\alpha\beta}{1 + 2(1-\alpha)\beta + 2(1-\alpha)/e}.$$

These considerations apply to the case $\alpha < 1$; when $\alpha > 1$, one simply obtains $\gamma = 1$ independently of α and β, due to an additional translation required by the CLT (see the Appendix; we do not expect $\alpha = 1$ to be pathologic). We have studied also the asymptotics of all possible diverging $N(n)$ (including the case $\beta < 0$), and proved that for $\alpha < 1$ the bound becomes linear as soon as the sampling $N(n)$ grows at least exponentially fast with n. It is possible to further characterize the fluctuations of the minimum of X, and prove that they follow the extreme-value Gumbel distribution; however, X does not belong to any min-stable basin of attraction,

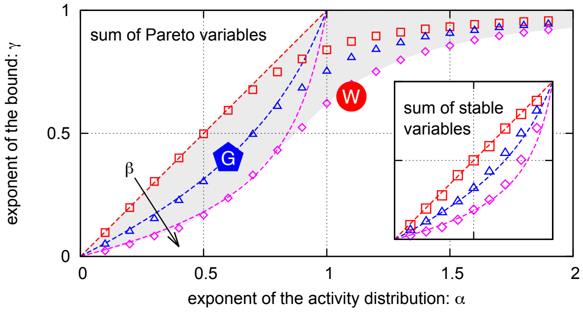

FIG. 4. The exponent γ of the bound depends on the exponent α of the activity distribution. Open symbols are fits from simulated data (Pareto variables in the main plot, α-stable in the inset), for $\beta = 0$ (red squares), $\beta = 0.8$ (blue triangles), and $\beta = 2.4$ (magenta rhombi). The dashed lines are the theoretical estimates for stable variables. The solid red circle and blue pentagon are the empirical values in Wikipedia and GitHub: $\alpha_{\rm w}$ and $\alpha_{\rm g}$ as in Fig. 2, and $\gamma_{\rm w} = 0.65(5)$, $\gamma_{\rm g} = 0.4(1)$. The shaded region displays the allowed values of γ for the empirical case $\beta = 2.4$ (i.e., those lying between the case of Pareto variables and the asymptotic result, $\gamma = \alpha$ for $\alpha < 1$, and $\gamma = 1$ for $\alpha > 1$).

thus a non-linear transformation is needed to obtain this result (details are reported in the Appendix).

The analytical predictions for the bounds are fully in line with both computer simulations and empirical data (Fig. 1 and 4). The results strictly apply only to α-stable variables. Wikipedia and GitHub data are better described by Pareto distributions, whose left tails drop abruptly at the cutoff $a_0 \approx 1$. The convergence to the stable distribution in the small-activity regime close to a_0 corrects the measured exponents γ. Numerical analysis (Fig. 4) shows that γ deviates from the α-stable prediction only for $\alpha \gtrsim 0.8$. The corrections have negative sign, and are largest for $\alpha \approx 1$.

We note the similarity of this approach with the random-energy model (REM), widely used to investigate disordered systems [40–42]. The REM considers an ideal system of n binary variables ("spins"), with $M = 2^n$ states, and assumes that the energy of each state is the sum of $O(n)$ IID random variables with a fixed distribution. The minimum of X then corresponds to the ground-state energy (minimum over the states) in the REM. At variance with the standard REM, however, the number $N(n)$ of projects built by n contributors (the analogue of M above) scales as a negative power law for both Wikipedia and GitHub (Fig. 2). This quantity has an impact on the observed bounds, which can be understood as a complex finite-size effect. Note that for the standard REM the linear scaling is a requirement, as the ground-state energy of a magnetic system has to be extensive [43]. One can imagine that our results may also apply to disordered systems of physical nature where, due to constraints, few of the $O\left(e^n\right)$ possible configurations are accessible. The calculations reported in the Appendix show that, in order to exhibit this anomalous finite-size scaling of the bound, it is sufficient that the accessible configurations grow less than exponentially with the number of degrees of freedom.

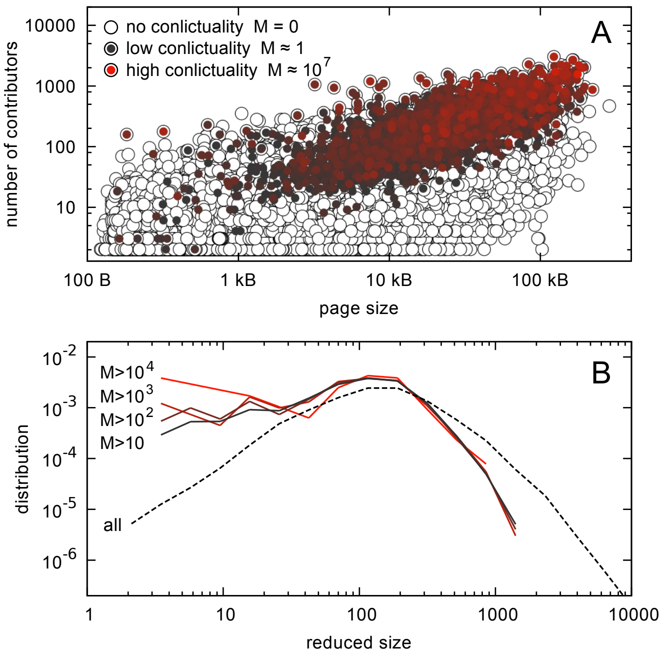

FIG. 5. Controversial Wikipedia pages are biased towards small sizes. Circles in the contributors/size scatterplot (A) are shaded according to the conflictuality M (lighter red corresponds to larger M). Panel B compares the reduced-size distributions (see Fig. 3) of all pages (dashed line) with those of conflictual pages in four ranges of the controversiality index M (solid red lines). [Wikipedia data (accessed 12 February 2015) and conflictuality values (obtained from wwm.phy.bme.hu, accessed 12 February 2015) were matched by page title.]

F. Conflicts decrease project size

Our model assumes that contributions are independent and always additive. In reality, contributors interact, and crowdsourced collaborative projects are not fully consensual. Controversies (*edit wars* in Wikipedia jargon) can arise during the creation and maintenance of projects [44, 45]. A proposed measure M of the conflictuality of each Wikipedia page [46] is based on revert counts, and comprehensive lists of M values are available publicly (at wwm.phy.bme.hu). By means of this quantity, we asked whether controversial topics in Wikipedia fall in notable areas in the contributors/size plots. The upper panel of Fig. 5 shows joint data for 200 000 pages of the English Wikipedia starting with the letter A. While, as might be expected, the general trend shows some correlation between conflictuality and page size (Fig. 5A), the upper bound is populated mostly by conflictual pages, while medium to large pages with few contributors are always consensual ($M = 0$). The same phenomenology is captured by a simple variant of our model that includes

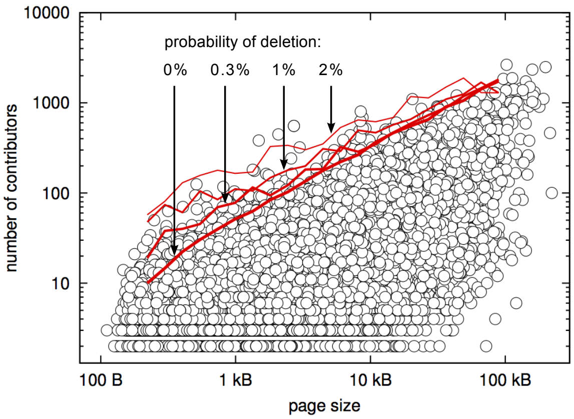

FIG. 6. Subtractive contributions affect the bound mildly. Their main effect is the blurring of the bound (red lines) because of the larger variance, which increases with increasing deletion probability; a secondary effect is a small readjustment of the slope. The points are the Wikipedia data set.

subtractive contributions, without relaxing the hypothesis of independent agents (Sec. II G). Most importantly, it is possible to gain a full quantitative grasp on the effect of conflicts by making use of the reduced size, whose distributions fall into two distinct classes, depending on the conflictuality (Fig. 5B).

Altogether, these considerations suggest the existence of at least two different forms of user interactions: edit wars, biasing the size of projects, and other interactions, affecting the individuals' activity patterns.

G. Deletions affect the bound mildly

We briefly report here the numerical results obtained for a model variant taking into account deletions as well as additions. Contributions in the pure-growth model described above are always additive, so the definition of project size is $X = \sum_{i=1}^{n} A_i$. Deletions can be incorporated — in a simplified description — by allowing the terms in the sum to be negative. Let σ_i be independent Bernoulli random variables, taking the value 1 with probability p and 0 with probability $1-p$. Then project size is defined as $X = \sum_{i=1}^{n}(-1)^{\sigma_i} A_i$. This is a rough approximation of the real production processes; for instance, the partial sums in the computation of X should be constrained to be positive, as size can never become negative during the life of a project. However, this simple description already helps to clarify the role of deletions in shaping the bound.

Figure 6 shows the bounds computed by simulation of this model for small values of the deletion probability p. The main effect of p is that of increasing the roughness of the bound, making it more probable for outliers to appear. This suggests that the discrepancies between model and data in Fig. 1 are due to code or text removal, reverts, deletions, etc. The prevalence of conflictual pages among those populating the bound in Wikipedia confirms this observation, as the measure M of conflictuality is based on the dynamics of reverts. Moreover, the exponent of the bound slightly changes as a consequence of deletions, giving a possible explanation for the small deviations between the data and the simulations of the model in Fig. 4. However, we expect these effects to be less marked in a more refined model of deletions, taking into account positivity of size, as this would decrease the fluctuations for small projects.

III. DISCUSSION

Scaling laws involving size (or "allometry") are a common feature of complex structures as diverse as cities, languages, living matter, and software [47–52]. Such a behavior is usually explained in terms of cooperative mechanisms, correlations, feedback, or non-linear system response [53]. Contrary to physical intuition, the power law in the bound emerges here in the neutral hypothesis of independent agents. Similarly to what happens in the standard REM, non-trivial behaviour stems from the simple assumption of random non-cooperative elements, which in turn can be interpreted as an effective description of the cooperative interactions observed e.g. in mean-field spin-glass models. Activity patterns may well be related to interactions between the agents, but the joint distribution and bounds in size are consequences solely of the activity patterns themselves. We should add that, as in the case of human correspondence, some features may emerge from non-stationarity, such as circadian patterns [54, 55] (which can be detected in Wikipedia as well [56]).

A significant finding of this study is the fact that the marginal size distributions, when properly rescaled by a power of the number of contributors, collapse onto a universal master curve. This curve can be interpreted as a quantitative signature of the interactions present in the system. Similar considerations apply when scaling behavior emerges in other systems [57–60]. On the practical side, this feature is essential for revealing the anomalous size distribution of projects affected by conflict. More in general, the reduced size realises a size metric independent of manpower, a significant goal in software sizing [61]. For instance, it allows to define an intensive measure of project growth, which can be used for quantifying the influence of extrinsic changes on the developmental process, without being sensitive to the fluctuations in the number of contributors.

ACKNOWLEDGMENTS

We are grateful to M. Barthelemy, P. Rotondo, B. Bassetti, A. Maritan, G. Tiana, and P. Cicuta for useful discussions and advice

Appendix A: Analytical calculations

1. Details of the model and relation with the REM

In our model of open-collaborative production, the distribution of contributor activity — to be meant as the total number of edits per contributor — has a power law tail, $P(A) \sim A^{-(\alpha+1)}$. Each user working on a project (piece of software or page) adds a number of contributions proportional to her/his activity, independently of the other users' actions. Thus, the size X of a project with n active users is defined as the sum $X = \sum_{i=1}^{n} A_i$ of n independent random variables $\{A_i\}_{i=1\dots n}$ sampled from the activity distribution $P(A)$. The minimum size of a project with n users among $N(n)$ projects corresponds to the sampling of the minimum of X among $N(n)$ independent realizations $X_1, \dots, X_{N(n)}$, in symbols $X_\mathrm{m} = \min\left\{X_1, \dots, X_{N(n)}\right\}$.

In the REM, a disordered system of n spins, with $M = 2^n$ possible states, is modelled by assigning a random energy to each of its configurations. Each state is assumed to be the sum of $O(n)$ IID random variables; the interplay between this scaling and that of M is crucial for the emergence of non-trivial behaviour. In our model $N(n)$ (the number of projects with n contributors) is the equivalent of M in the REM. At variance with the standard REM, however, $N(n)$ can scale differently from 2^n. Empirically, it scales as a power law, $N(n) = N_1 n^{-\beta}$ for both Wikipedia and Github (Fig. 2).

The lowercase letter x_m will be used to denote the typical value of the random variable X_m. Note that n as a function of x_m, gives information on how the *maximum* number of contributors at a fixed size depends on size. One may expect the extensive scaling $x_\mathrm{m} \sim n$ to occur, since the sum of n random variables with support in (a_0, ∞) has support in (na_0, ∞). However, the sampling of this distribution (i.e., the number of samples) affects the actual value of the minimum observed. In a standard REM, the linear scaling corresponds to the physical requirement that the ground-state energy of a magnetic system be extensive. On the contrary, a slow scaling of the sampling $N(n)$ with the number of summands n, may lead to a nontrivial bound $x_\mathrm{m}^\gamma \sim n$, with $\gamma < 1$. A sufficient condition for the breaking of the linear scaling is derived below. In the empirically relevant case $N(n) = N_1 n^{-\beta}$, if $\beta = 0$ the number of summands scales as a power law with the minimum, $n = x_\mathrm{m}^\gamma$, while for $\beta > 0$ deviations set in, such that the effective exponent γ observed is a slowly varying function of x_m.

2. Generalized central limit theorem and stable distributions

The starting point for studying the scaling of X_m is the analysis of the asymptotic distribution of the size of a single project. The sum X of n variables whose distribution has power law tails of exponent $\alpha + 1$ and infinite variance converges (when scaled and shifted appropriately) to an α-stable variable, defined by its characteristic function

$$\Psi_{\alpha,c,\sigma}(t) = \exp\left\{-\left|ct\right|^\alpha \left[1 - \mathrm{i}\sigma\,\mathrm{sgn}(t)\tan(\pi\alpha/2)\right]\right\}$$

(with $\alpha \neq 1$). The probability density function $L_\alpha(x)$ of an α-stable variable has support on $[0, \infty)$ if $\alpha < 1$ and $\sigma = 1$, it has support on $(-\infty, 0]$ if $\alpha < 1$ and $\sigma = -1$, while in all other cases the support is the whole real line. Since in our case $A_i > 0$, the corresponding stable distributions attracting X are the ones with the largest allowed skewness parameter, namely $\sigma = 1$. We will stick to this value in the following and also fix $c = 1$.

In order to examine the minimum of the size X_m, it is important to control the (asymptotic) distribution of small sizes. The cumulative distribution function of a stable law has the following behavior for x close to x_0, with $x_0 \equiv 0$ for $\alpha < 1$ and $x_0 \equiv -\infty$ for $\alpha > 1$ [39]

$$\int_{x_0}^{x} L_\alpha(y)\,\mathrm{d}y \sim |x|^{\frac{\alpha}{2(1-\alpha)}} \exp\left[-\left|1-\alpha\right|\left|\frac{x}{\alpha}\right|^{-\frac{\alpha}{1-\alpha}}\right]. \quad \text{(A1)}$$

[This expression neglects an α-dependent prefactor, which only slightly corrects Eq. (A7).]

3. Asymptotic analysis for diverging $N(n)$

When both n and $N(n)$ diverge one can give a complete characterisation of the asymptotic behaviour of X_m.

a. The case $\alpha < 1$. Since the distribution of $n^{-1/\alpha}X$ converges for $n \to \infty$ to an α-stable distribution, we can assume that X has the same distribution of $n^{1/\alpha}A$. It is worth noticing that this is true only asymptotically, but it holds for any finite n whenever the A_i's are stable random variables. Under this hypothesis, when $N(n)$ diverges, X_m is *self-averaging.* Specifically, for large n,

$$X_\mathrm{m} \simeq x_m := \frac{n^{\frac{1}{\alpha}}\alpha(1-\alpha)^{\frac{1-\alpha}{\alpha}}}{\log(N(n))^{\frac{1-\alpha}{\alpha}}}. \quad \text{(A2)}$$

We prove this by showing that $X_\mathrm{m}/x_\mathrm{m}$ converges in probability to 1, that is $P\{X_\mathrm{m}/x_\mathrm{m} > y\}$ goes to 1 if $y \leq 1$ and to 0 if $y > 1$. The fact the distribution for each X is the same as $n^{1/\alpha}A$ implies that X_m has the same law as $n^{1/\alpha}\min\left\{A_1, \dots, A_{N(n)}\right\}$. Hence, the cumulative distribution of $X_\mathrm{m}/x_\mathrm{m}$ is

$$\begin{aligned} P\{X_\mathrm{m}/x_\mathrm{m} > y\} &= P\{A_1 > yx_\mathrm{m}/n^{1/\alpha}, \\ &\qquad \dots, A_{N(n)} > yx_\mathrm{m}/n^{1/\alpha}\} \\ &= (1 - P\{A_1 \leq yx_\mathrm{m}/n^{1/\alpha}\})^{N(n)}. \end{aligned}$$

This can be written as $P\{X_\mathrm{m}/x_\mathrm{m} > y\} = \left[1 - g_n(y)\right]^{N(n)} \sim \exp\left[-N(n)g_n(y)\right]$, where $g_n(y)$ is given by the right-hand side of Eq. (A1) evaluated at $x = yx_\mathrm{m}/n^{\frac{1}{\alpha}}$. A brief calculation then shows that $N(n)g_n(y) \to 0$ if $y \leq 1$ and $N(n)g_n(y) \to \infty$ if $y > 1$, which proves the statement.

Note that, in particular, a polynomially diverging sampling $N(n) = N_1 n^\beta$ contributes to x_{m} only via logarithmic corrections, where β is a prefactor. Importantly Eq. (A2) gives the rate of divergence of $N(n)$ needed to recover the scaling $x_{\mathrm{m}} \sim n$: the sample size must grow exponentially in n. From the point of view of a REM with power-law-distributed energies, this is the condition for which the ground-state energy remains extensive.

The fluctuations of $X_{\mathrm{m}}/x_{\mathrm{m}}$ around 1 can be characterized by a similar calculation, which shows that

$$\log(N(n))\Big[\Big(\frac{x_{\mathrm{m}}}{X_{\mathrm{m}}}\Big)^{\frac{\alpha}{1-\alpha}} - 1 - \frac{\log(\log^{1/2}(N(n)))}{\log(N(n))}\Big]$$

obeys, for large $N(n)$, the extreme-value Gumbel distribution. Note that the random variables X do not belong to any min-stable basin of attraction of extreme value distributions. This explains why the nonlinear transformation $X_{\mathrm{m}}^{-\alpha/(1-\alpha)}$ is needed (in the case of the standard REM this problem does not arise since the energies summed are not bounded from below [43]).

b. The case $\alpha > 1$. The variables A_i have finite mean μ, and the generalised central limit theorem states that $n^{-1/\alpha}(X - n\mu)$ is asymptotically an α-stable random variable (notice the additional translation of the mean). Hence we may assume that X_{m} has asymptotically the same law as $n^{1/\alpha} \min\{A_1, \ldots, A_{N(n)}\} + n\mu$ where A_i are α-stable random variables. Again, this is true for finite n, and not only asymptotically, whenever the distribution of contributor activity is an α-stable law with mean μ. Arguing as for $\alpha < 1$, recalling that in this case $x_0 = -\infty$ in (A1), one can prove that $(X_{\mathrm{m}} - n\mu)/b_n$ converges in probability to -1 for

$$b_n = \frac{n^{\frac{1}{\alpha}} \log(N(n))^{\frac{\alpha-1}{\alpha}}}{\alpha(\alpha-1)^{\frac{\alpha-1}{\alpha}}}.$$

To do this one shows that $P\{(X_{\mathrm{m}} - n\mu)/b_n > y\}$ converges to 0 for $y > -1$ and to 1 for $y \le -1$. Combining these facts, one gets

$$X_{\mathrm{m}} \simeq x_m := \begin{cases} b_n(c-1) & \text{if } n/\log(N(n)) \to c \neq \infty, \\ n\mu & \text{if } n/\log(N(n)) \to +\infty. \end{cases} \tag{A3}$$

In particular, a polynomially diverging sampling $N(n) = N_1 n^\beta$ gives the linear scaling $n \sim x_{\mathrm{m}}$.

4. Nonlinear bounds for finite $N(n)$

We now show how nonlinear bounds emerge for finite $N(n)$. In this case it is difficult to obtain a convergence result akin to the one above. However, an analytical estimate for the typical value x_{m} of the variable X_{m} can be attained by looking for the value x_{m} such that

$$P\{X \le x_{\mathrm{m}}\} = \frac{1}{N(n)}, \tag{A4}$$

which states that the average number of samples lying beyond x_{m} is 1. We consider again the case $\alpha < 1$. When x_{m} is small and (A1) is applicable, (A4) becomes

$$\Big(\frac{x_{\mathrm{m}}}{n^{1/\alpha}}\Big)^{\frac{\alpha}{2(1-\alpha)}} \exp\Big[-(1-\alpha)\Big(\frac{x_{\mathrm{m}}}{\alpha n^{1/\alpha}}\Big)^{-\frac{\alpha}{1-\alpha}}\Big] = \frac{n^\beta}{N_1}. \tag{A5}$$

The left-hand side is a function only of the scaling variable $\tilde{x} = x_{\mathrm{m}}^{\alpha/(1-\alpha)} n^{-1/(1-\alpha)}$.

If $\beta = 0$ (i.e., if the number of samples is independent of n), the whole equation can be expressed in terms of $\tilde{x}$ only, giving the scaling $n \sim x_{\mathrm{m}}^\alpha$, i.e., $\gamma = \alpha$. Otherwise, by taking the logarithm of Eq. (A5) and differentiating with respect to $\log x_{\mathrm{m}}$ one obtains an effective exponent $\gamma = \mathrm{d}\log n/\mathrm{d}\log x_{\mathrm{m}}$ as a function of $\tilde{x}$:

$$\gamma = \alpha \frac{1 + f(\alpha)/\tilde{x}}{1 + f(\alpha)/\tilde{x} + 2(1-\alpha)\beta}\ , \tag{A6}$$

where $f(\alpha) = 2(1-\alpha)\alpha^{\alpha/(1-\alpha)}$. The prefactor of α in (A6) is less than unity, so $\gamma < \alpha$ for $\beta > 0$. Note the asymptotic values $\gamma_0 \equiv \gamma(\tilde{x} \to 0) = \alpha$ and $\gamma_\infty \equiv \gamma(\tilde{x} \to \infty) = \alpha/[1 + 2\beta(1-\alpha)]$. For finite $\tilde{x}$, the effective power-law exponent depends on how far into the tail of the distribution the minimum lies, and this depends on n and N_1. As $n \ge 1$, the small-n regime is realized at fixed n in Eq. (A5) by taking the limit $N_1 \to \infty$. Since x_{m} and $\tilde{x}$ go to zero, $\gamma = \gamma_0$ — in accordance with Eq. (A2) — thus recovering the same exponent as for $\beta = 0$. For large n instead, at fixed N_1, $\tilde{x}$ diverges (since $\tilde{x} \sim n^{(\alpha-\gamma)/\gamma(1-\alpha)}$ with $\gamma < \alpha$). However, n cannot become larger than $N_1^{1/\beta}$, since $N(n) \ge 1$. Therefore the full asymptotic scaling is never attained, and the exponent γ_1 observed around the largest available value of n (i.e., when $N(n)$ is of order 1) will be different from γ_∞. This effect can be quantified by considering the solution to Eq. (A5), $\tilde{x} = f(\alpha)^{-1} W\big(f(\alpha) N(n)^2\big)$, where W is the Lambert function, defined by $W(y) = \log y - \log W(y)$, and $f(\alpha)$ is defined as above. By setting $N(n) = 1$, noting that $W(f(\alpha))$ is well approximated by its expansion $2(1-\alpha)/e$ around $\alpha = 1$, and substituting the expression into (A6) one obtains

$$\gamma_1 = \alpha - \frac{2(1-\alpha)\alpha\beta}{1 + 2(1-\alpha)\beta + 2(1-\alpha)/e}\ , \tag{A7}$$

which gives a remarkably good approximation of simulated data at finite $N(n)$. Figures 1 and 4 show the accord between model and data.

We conclude by discussing the case $\alpha > 1$. Recalling that in this case X_{m} has the same law as $n^{1/\alpha} \min\{A_1, \ldots, A_{N(n)}\} + n\mu$, condition (A4) now holds for $n\mu - x_{\mathrm{m}}$, and takes a similar form to Eq. (A5), with $\tilde{x} = (n\mu - x_{\mathrm{m}})^{\alpha/(\alpha-1)} n^{-1/(\alpha-1)}$. Following the same reasoning as for the infinite-mean case, one obtains estimates of $\tilde{\gamma} = \mathrm{d}\log n/\mathrm{d}\log(n\mu - x_{\mathrm{m}})$. However, in this case $x_{\mathrm{m}} \sim n\mu - kn^{1/\tilde{\gamma}}$, where k is a constant. Hence, for large enough n, the effective behavior will always be $n \sim x_{\mathrm{m}}$, i.e., $\gamma = 1$, in agreement with the case in which $N(n)$

diverges polynomially. We did not treat the case $\alpha = 1$, but it is not expected to be pathologic. Notice that the values of γ for $\alpha = 1^+$ and $\alpha = 1^-$ agree.

[1] A. Vespignani, Nat. Phys. **8**, 32 (2012).

[2] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, Science **323**, 721 (2009).

[3] D. Brockmann, L. Hufnagel, and T. Geisel, Nature **439**, 462 (2006).

[4] T. Gowers and M. Nielsen, Nature **461**, 879 (2009).

[5] S. Weber, *The success of Open Source* (Harvard University Press, 2005).

[6] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. Players, Nature **466**, 756 (2010).

[7] S. E. Minson, B. A. Brooks, C. L. Glennie, J. R. Murray, J. O. Langbein, S. E. Owen, T. H. Heaton, R. A. Iannucci, and D. L. Hauser, Science Advances **1** (2015), 10.1126/sciadv.1500036.

[8] J. Howe, *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*, 1st ed. (Crown Publishing Group, New York, NY, USA, 2008).

[9] H. Sauermann and C. Franzoni, Proceedings of the National Academy of Sciences **112**, 679 (2015), http://www.pnas.org/content/112/3/679.full.pdf.

[10] S. S and van Krogh G, Information Systems Research **26**, 224 (2014).

[11] S. S. Levine and M. J. Prietula, Organization Science **0** (2013), 10.1287/orsc.2013.0872.

[12] A. Forte and C. Lampe, American Behavioral Scientist **57**, 535 (2013).

[13] T. Yasseri and J. Kertész, Journal of Statistical Physics **151**, 414 (2013).

[14] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing, Proceedings of the National Academy of Sciences of the United States of America **108**, 9020 (2011).

[15] B. A. Huberman, D. M. Romero, and F. Wu, J. Inf. Sci. **35**, 758 (2009).

[16] C. R. Myers, Phys. Rev. E **68**, 046116 (2003).

[17] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, Phys. Rev. E **74**, 036116 (2006).

[18] C. Bird, D. Pattison, R. D'Souza, V. Filkov, and P. Devanbu, in *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, SIGSOFT '08/FSE-16 (ACM, New York, NY, USA, 2008) pp. 24–35.

[19] V. Naroditskiy, N. R. Jennings, P. V. Hentenryck, and M. Cebrián, Preprint arXiv:1304.3548v3 [cs.SI] (2013).

[20] C. Jones, *Software Assessments, Benchmarks, and Best Practices*, Addison-Wesley information technology series (Addison Wesley, 2000).

[21] V. Basili and K. Freburger, The Journal of Systems and Software **2**, 47 (1981).

[22] C. F. Kemerer, Commun. ACM **30**, 416 (1987).

[23] B. Boehm, *Software engineering economics* (Prentice Hall, Englewood Cliffs, NJ, 1981).

[24] F. P. Brooks, Jr., *The Mythical Man-month (Anniversary Ed.)* (Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995).

[25] W. Scacchi, J. Feller, B. Fitzgerald, S. Hissam, and K. Lakhani, Software Process–Improvement and Practice **11**, 95 (2006).

[26] L. Muchnik, S. Pei, L. C. Parra, S. D. S. Reis, J. S. Andrade Jr, S. Havlin, and H. A. Makse, Sci. Rep. **3**, 1783 (2013).

[27] J. Voss, in *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics* (Stockholm, Sweden, 2005).

[28] F. Ortega, J. M. Gonzalez-Barahona, and G. Robles, 2014 41st Hawaii International Conference on System Sciences **0**, 304 (2008).

[29] N. Perra, B. Gonçalves, R. Pastor-Satorras, and A. Vespignani, Sci. Rep. **2**, 469 (2012).

[30] M. Gherardi, S. Mandrà, B. Bassetti, and M. C. Lagomarsino, Proceedings of the National Academy of Sciences **110**, 21054 (2013).

[31] S. Mandrà, M. C. Lagomarsino, and M. Gherardi, Phys. Rev. E **90**, 032805 (2014).

[32] F. Schweitzer, V. Nanumyan, C. J. Tessone, and X. Xia, Advances in Complex Systems **17**, 1550008 (2014).

[33] A. S. Chakrabarti, B. K. Chakrabarti, A. Chatterjee, and M. Mitra, Physica A: Statistical Mechanics and its Applications **388**, 2420 (2009).

[34] A.-L. Barabasi, Nature **435**, 207 (2005).

[35] S. M. Bhattacharjee and F. Seno, Journal of Physics A: Mathematical and General **34**, 6375 (2001).

[36] H.-H. Jo, R. K. Pan, J. I. Perotti, and K. Kaski, Phys. Rev. E **87**, 062131 (2013).

[37] S. N. Majumdar, M. R. Evans, and R. K. P. Zia, Phys. Rev. Lett. **94**, 180601 (2005).

[38] B. V. Gnedenko and A. N. Kolmogorov, *Limit distributions for sums of independent random variables* (Addison-Wesley, Cambridge, MA, 1954).

[39] V. M. Zolotarev, *One-dimensional stable distributions* (American Mathematica Society, 1986).

[40] B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[41] J. D. Bryngelson and P. G. Wolynes, *The Journal of Physical Chemistry*, J. Phys. Chem. **93**, 6902 (1989).

[42] G. B. Arous, A. Bovier, and V. Gayrard, Phys. Rev. Lett. **88**, 087201 (2002).

[43] J.-P. Bouchaud and M. Mézard, J. Phys. A: Math. Gen. **30**, 7997 (1997).

[44] J. Török, G. Iñiguez, T. Yasseri, M. San Miguel, K. Kaski, and J. Kertész, Phys. Rev. Lett. **110**, 088701 (2013).

[45] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész, PLoS ONE **7**, e38869 (2012).

[46] R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertész, in *SocialCom/PASSAT* (IEEE, 2011) pp. 724–727.

[47] L. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, and G. West, Proceedings of the National Academy of Sciences **104**, 7301 (2007).

[48] R. Louf and M. Barthelemy, Phys. Rev. Lett. **111**, 198702 (2013).

[49] R. Louf and M. Barthelemy, Sci. Rep. **4**, 5561 (2014).

[50] T. Louail, M. Lenormand, O. G. Cantú, M. Picornell,

R. Herranz, E. Frias-Martinez, J. J. Ramasco, and M. Barthelemy, Sci. Rep. **4**, 5276 (2014).

[51] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc, Sci. Rep. **2**, 943 (2012).

[52] G. B. West, J. H. Brown, and B. J. Enquist, Science **276**, 122 (1997).

[53] D. Sornette, *Critical phenomena in natural sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools* (Springer-Verlag, Berlin Heidelberg, 2006).

[54] R. D. Malmgren, D. B. Stouffer, A. S. L. O. Campanharo, L. A. N. Amaral, and S. Paulo, Science, 1696 (2009).

[55] A. Vázquez, J. a. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási, Phys. Rev. E **73**, 036127 (2006).

[56] T. Yasseri, R. Sumi, and J. Kertész, PLoS ONE **7**, e30091 (2012).

[57] A. Giometto, F. Altermatt, F. Carrara, A. Maritan, and A. Rinaldo, Proceedings of the National Academy of Sciences **110**, 4646 (2013).

[58] H. Stanley, *Introduction to Phase Transitions and Critical Phenomena*, International series of monographs on physics (Oxford University Press, 1971).

[59] J. R. Banavar, J. Damuth, A. Maritan, and A. Rinaldo, Phys. Rev. Lett. **98**, 068104 (2007).

[60] M. Gherardi and M. C. Lagomarsino, Sci. Rep. **5**, 10226 (2015).

[61] F. Wilkie, I. McChesney, P. Morrow, C. Tuxworth, and N. Lester, Information and Software Technology **53**, 1236 (2011).