Early onset of structural inequality in the formation of collaborative knowledge in all Wikimedia projects

Jinhyuk Yun1, Sang Hoon Lee2,3* and Hawoong Jeong2,3,4*

The Wikimedia project, including Wikipedia, is one of the largest communal data sets and has served as a representative medium to convey collective knowledge in the twenty-first century. Researchers have believed that the analysis of these collaborative digital data sets provides a unique window into the processes of collaborative knowledge formation; yet, in reality, most previous studies have usually focused on its narrow subsets. Here, by analysing all 863 Wikimedia projects (various types and in different languages), we find evidence for a universal growth pattern in communal data formation. We observe that inequality arises early in the development of Wikimedia projects and stabilizes at high levels. To understand the mechanism behind the observed structural inequality, we develop an agent-based model that considers the characteristics of the editors and successfully reproduces the empirical results. Our findings from the Wikimedia projects data, along with other types of collaboration data, such as patents and academic papers, show that a small number of editors have a disproportionately large influence on the formation of collective knowledge. This analysis offers insights into how various collaboration environments can be sustained in the future.

The contemporary world is the result of an enormous amount of accumulated human knowledge. To maintain and enhance societal progress, it is essential that we understand the process of knowledge creation and, perhaps more importantly, the collaborative human behaviours behind it. However, quantitative data that analyse human history have been mostly far from satisfactory because of the lack of systematically preserved records describing the details of human knowledge development. Thus, the investigation of human knowledge creation and collaboration has long been anecdotal, with historians and anthropologists reconstructing the past based on fragmentary evidence. However, this situation radically changed at the turn of the twenty-first century. The exponential growth of information technology in the twenty-first century has created online environments to share up-to-date information generated by everyone. The segregating lines between information producers and consumers have been blurred, a phenomenon called produsage1. It has been argued that these new online environments enable the ‘democratization’ of knowledge2. At the same time, accumulation of an inconceivable amount of information produced by everyone with online access every minute of the day has generated an unprecedented amount of exhaustive records of digital footprints, which processes of human knowledge creation. Wikipedia (https://www.wikipedia.org/), a representative openly edited encyclopaedia, represents a paradigmatic example of an online environment for contributors who generate information. Wikipedia’s credibility has been questioned: it is sometimes considered unstable, imprecise and even misleading3,4. On the other hand, studies have shown that the current accuracy of Wikipedia is remarkable, even surpassing traditional encyclopaedias5,6. Nevertheless, researchers have identified substantial heterogeneities in the editing process, due to the monopoly or oligopoly of a few ‘supereditors’ who govern Wikipedia’s content7–11. Thus, Wikipedia is still far from being the ideal of communal knowledge that we desire it to be.

The majority of studies, including our own previous work, focused on just a few language editions, mostly the English edition, of Wikipedia to examine the dynamics and properties of the communal data set, an editable data set that is shared within a community to build collective knowledge12–14. Although these studies warned of the potential risks behind the current social structure in the English edition, cultural background affects the of individuals; therefore, the results of those studies may not be generalizable. Wikipedia users may be affected by their social norms or cultures. For instance, people belonging to different cultural backgrounds tend to use different symbols on the Web15; cultural background also affects Web page design16,17. The editors of different language editions edit Wikipedia using distinctive patterns18. Moreover, there are differences in linguistic complexity. For example, the linguistic complexity of the English edition of Wikipedia differs from that of the German and Spanish editions19. Finally, previous studies have been based on small samples obtained from non-identically sized data sets, comparing, for instance, the same entries across a number of languages. However, the number of articles in the English edition of Wikipedia and the number of articles in the other language editions are of different orders of magnitude; the English edition has at least five times more edits. Thus, it is unclear whether the heterogeneity in editing processes is a general phenomenon or applies to English users only.

To investigate the general features of editing processes and dynamics, we have extended our previous analysis of the English edition of Wikipedia20 to all 863 Wikimedia projects (https://wikimediafoundation.org/our-work/wikimedia-projects/); these consist of various types of communal data sets hosted by the Wikimedia Foundation (https://wikimediafoundation.org/). For this purpose,
we investigated heterogeneity in contributions and supereditors’ share in Wikimedia projects to understand the psychosocial processes and dynamics behind openly edited communal data sets. In particular, we examined the complete editing history of every Wikimedia project to assess their growth over time. We mainly focused on the number of edits, editors and articles, article size and the level of heterogeneity captured by an inequality coefficient that measures the disparity of contribution in the editing activities (edits and number of contributions) of editors. We demonstrate the existence of typical growth patterns of openly edited communal data sets that eventually establish strong heterogeneity among the contributions of editors. In addition, to understand the mechanism behind such disparity, we introduce an agent-based model that replicates the interactions between communal data sets and editors. Our model takes into account the competition between editors’ natural reduction in motivation over time, their stronger memory for more recent editing activities and their attachment to the articles they edit. The model reproduces the actual universal growth patterns in a manner that is consistent with the data.

Results
Our main goal was to identify the underlying principles in the development of communal data sets. For our analysis, we used the March 2016 dump of all Wikimedia projects (https://dumps.wikimedia.org/backup-index.html). We considered a single Wikimedia project as a sample of such communal data sets. To proceed with the in-depth analysis of the evolution of communal data sets, we need to stress that most data sets were aged approximately 3.5 × 10^8 s (about 11 years; see Supplementary Fig. 1). Thus, most Wikimedia projects are of similar age, suggesting that raw measures without time rescaling are appropriate.

Universality and disparity in communal data sets. We present evidence of a universal growth pattern shared by all Wikimedia projects, as displayed by characteristic measures that are based on the current snapshot of the communal data set, such as the total number of edits (N_e), editors (N_e) and articles (N_a), and the current article size (S, in bytes). Our primary interest was to identify the generality in growth patterns of the communal data set, not individual articles. Thus, we used the total sum of these values for all articles in a Wikimedia project, without considering the individual properties of its constituent articles. For example, N_a is the total number of edits for a given edition of a Wikimedia project, or the sum of the number of edits of individual articles for that edition. Our first analysis of the interplay between such measures indicates their regularity across projects, regardless of age, language and data set type.

Growth scale of communal data sets. We begin our analysis with the study of the intercorrelations between N_a, N_e, N_e and S in the current Wikimedia projects. One may speculate the absence of a general rule between measures because of the excessive heterogeneity of current Wikimedia projects (see Supplementary Fig. 2) compared to their age distribution (see Supplementary Fig. 1). As an example of the differences between different language editions of Wikipedia projects, the degree of language proficiency among editors of the English edition of Wikipedia is qualitatively different from that of other language editions19. Despite such a difference, we found common positive correlations between the measures. First, the number of editors and articles, and the size of the data sets, gradually vary as functions of the number of edits. The growing patterns are characterized by the simple sublinear growth y ∝ x^λ, where x is the number of edits, the exponent λ ≃ 0.70 is the number of editors (Fig. 1a), λ ≃ 0.85 is the number of articles (Fig. 1b) and λ ≃ 0.87 is the total size of the data sets in bytes (Fig. 1c). One may argue that this sublinear growth of λ < 1 does not directly indicate a slowdown in growth. For instance, exponent λ = 1/2 is required to retain the mean editing rate for a single article if the editing rate is proportional to the current number of articles because dN/dt ∝ N, so λ < 1/2 indicates stagnation under this condition. However, every event in the Wikimedia projects can be considered as a different type of editing that leads to an increment of 1 in the number of edits N_e. In other words, the participation of a new editor and the creation of a new article are not independent editing actions; they are instead special cases of editing. Consequently, each parameter N_e, N_a and S can be written as functions of N_e, for example, as αN_e^c. Thus, λ indicates the temporal trend of the ratio between regular editing and special events, that is, the participation of a new editor and the creation of a new article. In summary, λ < 1 implies an overall slowdown of growth with regard to the frequency of appearance of new editors and brand-new articles, as more edits take place; from the perspective of editability, larger data sets are more inefficient than smaller ones.

To identify the reason for this stagnation with regard to the number of edits, we also examined the interrelationships between the other measures. The number of editors increases with the number of articles (λ ≃ 0.78; Fig. 1d). Meanwhile, the size of each article roughly linearly increases with the number of articles (λ ≃ 1.02; Fig. 1e) and the number of editors (λ ≃ 1.06; Fig. 1f), respectively. Briefly, the rate of text accumulation remains almost constant regardless of the number of articles and editors. Our previous study on the English edition of Wikipedia suggested that (1) the inter-event period between two consecutive edits in an individual article follows a universal distribution regardless of its age in real time and (2) the size difference between two consecutive edits also follows a universally right-skewed distribution regardless of size2. In light of these findings, the current results suggest that stagnation is caused by the decreased appearance of new editors and not by the decreased productivity of existing editors. In addition, N_e, N_a, N_e and S are not correlated with the age of the data sets (Supplementary Fig. 3), indicating that the raw number of edits is the appropriate measure to compare the various data sets, rather than actual time. As shown in Supplementary Fig. 1, most of the Wikimedia projects are of similar age; therefore, our analysis suggests that the rate of growth per unit time decreases as the size of Wikimedia projects increases, as we showed in our previous study8. All Wikimedia projects display this growth scale, regardless of their institutional aim, suggesting that the common growth pattern is caused by the common characteristics of communal data sets (see Supplementary Fig. 4).

Along with the common scaling patterns observed in the different measures of a Wikimedia project, that is, N_a, N_e, N_e and S, it is also important to identify the possible scaling patterns between different types of Wikimedia projects in the same language. As we show in Supplementary Figs. 5–8, the majority of different types of Wikimedia projects show almost identical tendencies with regard to the size ratio. Each of the four measures (N_a, N_e, N_e and S) for Wikipedia is considerably larger than that of any other types; Wikitongue is the second largest. The differences between other types of Wikimedia projects are not as great.

The next question is whether languages in Wikimedia projects can be categorized into distinct clusters that share growth patterns. To answer this question, we performed a cluster analysis by constructing a simple feature vector for language, which consists of characteristic measures from different types of Wikimedia projects (see Supplementary Methods). We present results from two machine learning techniques. In particular, we used the Dirichlet process Gaussian mixture model14, which can efficiently partition vectors in the case of unknown numbers of groups. Then, we used the t-SNE algorithm to visualize the higher dimensional feature vectors in two-dimensional space, while preserving their original degrees of separation12. We did not observe any clear-cut clustering for the different dimensions of feature vectors and different clustering...
parameters (Supplementary Figs. 9–11). Therefore, our cluster analysis results support the existence of universal rules governing the growth of a communal data set, regardless of its language.

**Wikimedia projects and their corresponding socio-economic indicators.** Next, we examined if there are any possible external factors that affect the current status of different Wikimedia projects with regard to the number of editors, articles, and the total size of the data set. The number of Wikimedia projects is not simply determined by the total number of users. As an illustrative example, the Spanish edition of Wikipedia is approximately ten times larger than the Hindi edition, despite the fact that both Spanish and Hindi have around half a billion speakers each. Bearing in mind that the age of both editions is comparable (16 years for the Spanish edition and 14 years for the Hindi edition), the growth of the Hindi edition has been much slower up to this point. To elucidate the reason behind such a big difference, we inspected the factors affecting the current status of Wikimedia projects. The simplest factor is the ratio of the number of people using the language as a native language to those using it as a second language (see Supplementary Methods). One may assume that the people using a particular language as a second language have less of an impact on the formation of a communal data set written in that specific language, compared to its native language users. However, we found that the number of Wikimedia projects is better correlated with the number of second language users than the total number of language users or the number of users who are native speakers (Supplementary Fig. 12). Other linguistic properties may also influence the growth rate of Wikimedia projects, so we tried categorizing Wikimedia projects according to their written scripts. Rather surprisingly, there are no notable differences between the scripts (Supplementary Fig. 13) because, as mentioned earlier, a single character in each script takes a different size in bytes depending on the language.

In addition to simple linguistic factors, we tried to consider more complex ones by cross-correlating the language editions of Wikimedia projects with the socio-economic status of countries the Wikimedia projects belong to. We assigned the dominant country of a certain language edition according to the following criteria: (1) a country using the language as a primary or official language; and (2) the first country in terms of the page view share of the Wikipedia edition in that language (see Supplementary Methods). First, the educational level of a country shows a positive correlation with the number of corresponding Wikimedia projects, although not in a statistically significant manner (Supplementary Fig. 14). In addition, the total population weakly impacts the status of Wikimedia projects, whereas the total population of Internet users shows a strong positive correlation with the number of Wikimedia projects (Supplementary Fig. 15). Gross domestic product is also well correlated with the growth of Wikimedia projects (Supplementary Fig. 15). However, the gross domestic product per capita is not correlated with the number of Wikimedia projects (Supplementary Fig. 15). In summary, the scale of the economy, which is partly reflected in the number of Internet users, affects the growth of Wikipedia projects.

We also observe that national expenditure and products for research and development show a significant correlation with the current number of individual Wikimedia projects. Larger investors in research also tend to have larger Wikimedia projects compared to their smaller counterparts (Supplementary Fig. 16). Consequently, the number of patents and academic papers are also strongly correlated with the size of individual Wikimedia projects (Supplementary Figs. 17 and 18; see Supplementary Methods for further details). Such a research and development scale is determined by the scale of the national economy. Taken together with the results described earlier, the size of individual Wikimedia projects is closely tied to the overall size of the economy of a country, yet per-capita levels do not impact greatly on the current size of individual Wikimedia projects. In other words, the Wikimedia projects of richer countries grow faster and larger.

**Disparity in contributions.** The general growth patterns of $N_e$, $N_a$, $N_p$ and $S$ suggest an interesting hypothesis: could there be a universal rule underpinning the genesis of structural heterogeneity in the

---

**Fig. 1** | Correlation between the number of edits ($N_e$), editors ($N_e$) and articles ($N_a$), and the total size of the data set ($S$). Every correlation is characterized by the simple power-law growth of $y \sim x^\lambda$. a, With regard to the number of edits, the other measures grow sublinearly with $\lambda \approx 0.70$ for the number of editors (Pearson’s correlation coefficient $\rho \approx 0.90$). b, $\lambda \approx 0.85$ for the number of articles ($\rho \approx 0.85$). c, $\lambda \approx 0.87$ for the total size of the data set in bytes ($\rho \approx 0.95$). d, The number of editors also sublinearly increases by the number of articles with $\lambda \approx 0.78$ ($\rho \approx 0.65$). e, The size of the data set increases almost linearly by the number of articles with $\lambda \approx 1.02$ ($\rho \approx 0.83$). f, Nearly linear relationship ($\lambda \approx 1.06$) between the number of editors and the size of the data set, which in turn indicates that the average productivity of a single editor is maintained ($\rho \approx 0.73$). a–f, We estimated the power-law exponent using simple linear regression methods in a double logarithmic scale (see Methods). The statistical details for the regressions are shown in Supplementary Tables 2–7.
formation of communal data sets²? To examine the validity of the hypothesis, we employed a variant of the Gini coefficient, a conventional measure of inequality of income³₁. In our analysis, a variant of the Gini coefficient quantifies how the number of edits are distributed among different editors involved in a specific Wikimedia project, that is, those editors who have edited an article in the project at least once. This variant of the Gini coefficient ranges from 0 for minimal heterogeneity (or maximal homogeneity, when every editor contributes equally) to 1 for maximal heterogeneity (when only a single editor contributes everything). We considered the number of edits and the data size for individual editors as the variables of interest; these variables are referred to as ‘total contributions’ (they parallel the quantification of inequality in economic wealth) unless specified otherwise.

The trend of the variant of the Gini coefficient as an increasing function of \( N_t \) displayed in Fig. 2 suggests that the disparity in the level of total contributions among editors is intensified as the communal data set grows. Larger values of \( N_t \) produce more intense disparity not only for the number of edits performed by the editors (Fig. 2a), but also for the total data changes (in bytes) made by the editors (Fig. 2b). This increasing trend is still valid when addition and subtraction are performed separately (Fig. 2c,d). In addition, because the age of an article does not severely affect heterogeneity (Fig. 2e,f), the results of the variant of the Gini coefficient are consistent with our observation that the age of an article does not affect the current state of the communal data sets. We predict that heterogeneity becomes more severe if a given data set is edited more frequently. No notable differences among different written scripts of Wikimedia projects (see Supplementary Fig. 19) and the institutional objectives of different Wikimedia projects (see Supplementary Figs. 20 and 21) were observed. To summarize, based on the current snapshot of the communal data set, we observed that the universal pattern of heterogeneity increases with the number of edits.

Evidence for the establishment of supereditor dominance. We have shown that current Wikimedia projects exhibit a high level of heterogeneity and that a variant of the Gini coefficient increases with the number of edits (Fig. 2). Although the current status of all Wikimedia projects appears to follow a specific function of \( N_t \), this could be coincidental. Thus, we further tracked the actual history of individual Wikimedia projects to confirm or reject the possibility of such a coincidence, so that we could determine if the trend for increasing inequality in contributions is inherent to the formation of communal data sets. We set the initial number of edits of all 863 data sets to the same value (\( N(0) = 0 \)) and recorded the trajectories of the variant of the Gini coefficient as functions of \( N_t \) (see Fig. 2g for the curve averaged over the data sets, which also shows the standard deviation). Following the conventional use of the Gini coefficient (used to measure inequality of income or wealth), we used the accumulated number of edits up to \( N_t \) (note that the unit of time in this case is \( N_t \)) for each editor, which captures the total contributions made by an editor. Technically, a variant of the Gini coefficient is undefined when a single editor has edited a data set, since we defined the set of editors as the editors who have contributed at least once; however, we take a variant of the Gini coefficient as 1 because it describes the completely monopolized state well. Our result shows that the average of the variant of the Gini coefficient coincides with the current status of Wikimedia projects (Fig. 2g); thus, the current status of a specific data set can be taken as the midpoint of a single master curve described as a function of \( N_t \). For example, a history of the Cebuano edition of Wikipedia clearly follows the typical growth pattern for \( N_t > 10^4 \) (Fig. 2h), except for the initial fluctuations due to small values of \( N_t \).

Although we employed a variant of the Gini coefficient as a measure of inequality in the distribution of accumulated total contributions, which parallels the conventional use of the coefficient to characterize inequality in wealth, an alternative approach of the coefficient is widely used to capture inequality due to time-bound contributions. In economics, income is defined as the value gained within a specific time frame⁴. Similarly, we considered the number of edits for individual editors per unit time frame as the ‘income’ variable in the Gini coefficient; this is called ‘time-bound contributions’, unless otherwise specified. In other words, the ‘total contributions’ analysed previously are the accumulated contributions from the onset of an individual editor’s first activity. In this analysis, we used the time window of 10⁴ edits, but different time frame values do not meaningfully affect the result. With regard to the distribution of time-bound contributions for the communal data set as a function of \( N_t \), the variant of the Gini coefficient indicates that the larger \( N_t \) values suggest less marked heterogeneity in contributions (Fig. 2i). Indeed, it suggests that the distribution of time-bound contributions becomes more homogeneous with time (Fig. 2i), whereas disparity in the total distribution of contributions is maintained (Fig. 2g).

Therefore, for all editors, the heterogeneity in the total distribution of contributions intensifies over time, whereas disparity in contributions per time frame becomes less severe with time. To consolidate the two results, we examined in detail how the ‘rich get richer’ concept affects the communal data set. Fig. 3a,b suggest that editors not only tend to keep their short-term social positions, but also maintain their long-term social positions. For instance, 58.1% of editors remain in the [0%, 10%] rank range (10% from the top) for the next 10⁴ edits if editors are ranked in the [0%, 10%] range (10% from the top) in the 990,000 ≤ \( N_t < 1,000,000 \) time window; meanwhile, only 32.6% of editors are ranked in the [0%, 90%] range, that is, except for the bottom 10%, editors retain their positions (Fig. 3a). In other words, editors who edit more often within a specific time window tend to edit more often in other time windows also. Although the exact proportion and number of edits for each percentile vary over time, the distinction among different groups of editors classified according to the level of contributions is preserved. As a result, a hierarchical structure between editors is gradually established.

The trend is even clearer for the accumulated number of edits (Fig. 3b). Early on, only highly ranked editors, whose number of contributions are much larger than the median, maintain the positions represented by their cumulative number of edits up to that time. Meanwhile, editors whose number of contributions are much smaller than the median change their positions in the editing hierarchy more frequently. For every percentile, over time the percentile of revisiting editors becomes more associated with their previous class; eventually, this groups most editors under a stratified percentile. Therefore, solid classes are established at a very early stage and remain thus for a long time. The oligopoly of supereditors is thus clear⁴. We have not only revisited the existence of such an oligopoly; we have also observed how its degree of influence changes as more edits are performed. The territories of these editing conglomerates extend beyond single articles and span the entire Wikimedia project community; their leverage on Wikipedia is still growing (Fig. 3b).

To understand the formation of the considerable share of such supereditors, we further examined the interrelationship between the number of editors in two consecutive editing sequences in various time windows from the onset of the data set (Fig. 3c). We calculated the Pearson’s correlation coefficient between the lists of number of edits in two successive time frame windows for an editor. Initially, the two consecutive sequences of number of edits are highly correlated across various lengths of time-frame windows; however, short-term correlations gradually decrease as more edits are performed. In addition, a boundary between the high- (correlation \( > 0.7 \)) and low-correlated (correlation \( \leq 0.7 \)) domains builds over time; consequently, only long-term correlation is maintained.

In light of this information, the results shown in Fig. 3 explain the results shown in Fig. 2g,i; the disparity in the total distribution of...
of contributions is preserved by long-term correlations, whereas the
disparity in the time-bound distribution of contributions is steadily
resolved due to the decline of short-term correlations. Although the
short-term activities of editors may vary, the dominance of a few
editors is not resolved in the long-term because such dominance is
established at a very early stage of the communal data set.

Other collaborative knowledge creation. One clear advantage
of investigating online data such as Wikimedia projects is that we
can identify individual contributions in the formation of collective
knowledge. On the other hand, observing ubiquitous growth pat-
terns and the formation of strong heterogeneity also prompts a key
question: is the early onset of heterogeneity specific to the creation
of online communal data sets? In other words, is it also possible to
find similar growth patterns in conventional knowledge formation
processes? Although the Internet revolution has meant that online
media play an important role in constructing and spreading knowl-
edge in the twenty-first century, conventional platforms remain a
major route for disseminating expertise. To explore the wider land-
scape of collective knowledge formation, we extended our analysis
to academic papers and patents, two pivotal media of traditional
knowledge formation and dissemination.

For our analysis, we used patent data from the spring 2017 edi-
tion of the European Patent Office Worldwide Patent Statistical
Database and academic paper data from a 22 August 2017 dump of
the entire Scopus Custom Data in XML format (see Supplementary
Methods for details). For the patent data set, we assumed that 91 dis-
tinct patent offices play roles analogous to the different editions of
the Wikimedia projects. Similarly, we also considered each author’s
affiliated country as a unit of knowledge formation, analogous to a
language edition of the Wikimedia projects. In theory, a single pat-
et or a single academic article can be considered as equivalent to
an article in the Wikimedia projects. Unfortunately, it is impossible
to trace the entire editing process involved in the composition of
a single patent or academic paper. Accordingly, we only used the
information of the number of patents/articles and the number of

participants for each country. In addition, unlike the Wikimedia projects, the time frame of our patent/academic paper data set does not cover the very beginning of this knowledge platform. Considering the long history of patents and academic papers, we could examine only a small contemporary subset, specifically from 2000 for the patents and 1996 for the academic papers.

Based on our analysis of the Wikimedia projects, one may expect the existence of a general rule between the number of participants and the number of patents/academic papers. We found strong positive correlations between measures for both patents and academic papers (Supplementary Figs. 22 and 23). Specifically, the Pearson’s correlation coefficient for the number of patents and the number of inventors (who originally designed the technology) is \( \rho = 0.85 \), whereas it is \( \rho = 0.74 \) for the number of applicants who originally filed the patent to obtain intellectual property rights. The statistics of academic papers show a larger Pearson’s correlation coefficient of \( \rho = 0.97 \) for the number of articles and authors.

Our finding of general growth patterns across conventional and online knowledge platforms prompted us to seek possible inequalities among participants of conventional knowledge formation. Once more, we employed a variant of the Gini coefficient to measure the degree of inequality between players in conventional knowledge formation. As with the Wikimedia projects, the heterogeneity of both patents and academic papers grows due to the increasing function of the number of participants and research outputs. In contrast to the steep increase observed in the Wikimedia projects, patent and academic paper data sets show a more gradual increase (compare Supplementary Figs. 21–23), yet the variant of the Gini coefficient as a function of \( N_e = 10^4 \) is high (\( \rho = 0.97 \) for the number of articles and authors). The model begins with a single agent. Each agent represents a single editor who participates in the editing process. A single medium represents the communal data set or a single language edition of a Wikimedia project. In our model, we considered the actions of editors to be driven by a set of motivations, and introduced parameters for editors to describe their activities. First, for an editor \( i \), we denoted the accumulated number of edits as \( N_i(t) \) at time \( t \). Next, we specified the time of the first edit by an editor as \( t_{b;i} \) and the time when the last edit occurred as \( t_{e;i} \). The dynamics rules are as follows: for each simulation step, the debut of a new agent and the return (or re-edit) by an already existing agent occur in turns; for every simulation step, a new agent appears with a constant probability \( b \) and begins to participate in the editing process; once a new agent appears in the data set at time \( N_e \), the agent edits the data set at the time of its entry into the system so that \( t_{e;i} \) and the time unit \( t \) is increased by 1 (the unit of the edit number). Note that the time scale of \( N_e \) for the model and data is not identical because the time scale of the model can vary with system size and differing parameters.

In the second step, an agent chosen uniformly at random attempts to edit the data set. Many factors affect the motivation for the editing, but we assumed three: (1) the decrease in motivation to edit over the long term; (2) the inertia of continuing the action; and (3) the attachment editors have for an article, which is increased by the past contributions they have made. In general, editors are highly motivated at the beginning of their participation, but their motivation fades steadily.\(^{28,30,31}\) Thus, participants diminish their engagement as time goes by; this is modelled by the power-law decay as factor \( (t - t_{b;i})^{-x} \), where \( k \) represents the decay in motivation that is observed in many temporally varying systems.\(^{28,29}\) In addition, a fat-tailed distribution is observed for the time between consecutive edits;\(^{32}\) this suggests that the editing time scale of Wikipedia shows a bursty. The inertia of continuing the (editing) action means that it is harder to edit when the interval between an editor’s latest editing attempt and the current time \( t \) increases.\(^{28,30,31}\) This inertia is modelled as \( [1 + e^{-(x(t-t_{b;i}))}]^{-1} \), where \( x \) represents the characteristic sustaining time of the inertia in this stimulation. Finally, editors tend to be more engaged when they have already participated more frequently in the past.\(^{32,33}\) The number of edits is assigned 1 at the time of first participation of the editor; this value increases by unity every time an agent participates in the editing process, so that it is equivalent to the number of edits \( N_e(t) \) up to the time point \( t \).
Taking these factors together, in our model, when an agent $i$ is chosen for editing, they participate in the editing with the probability

$$P[t; N_i(t), t_{b,e}, t_e] = \min \{1, N_i(t)(t-t_{b,e})^{-k}[1 + e^{-(t-t_e)/\tau}]\}$$

Once an agent decides to participate, $t_{e}$ is newly set as $t + 1$ and $N_i(t + 1) = N_i(t) + 1$. In addition, we also included the possibility for an agent to abandon the editing process permanently, this departure being based on the loss of motivation to edit. Therefore, in our model, agents leave the system when they choose not to edit and $P[t; N_i(t), t_{b,e}, t_e] < r$, where $r$ is a pre-assigned cutoff parameter common to all editors. In the next section, we provide some evidence that inequality is created by these factors, regardless of the personal characteristics of an individual editor.

**Model results.** Previously, we have shown that the variant of the Gini coefficient increases as the number of edits increases. The variant of the Gini coefficient increases rapidly at the early stage of data set formation and stabilizes at a high value (approximately $0.8$ for $N_e \gtrsim 10^4$, see Fig. 2). The results of our model result are consistent with the empirical observations. With regard to the model data set, the variant of the Gini coefficient rapidly increases until a high value is reached at $N_e \approx 10^4$ for $k = 0.8$ (compare Fig. 4a with Fig. 2g). Smaller $k$ values yield a slower increase in the variant of the Gini coefficient, whereas $\tau$ does not affect it significantly. The coefficient does not reach a high value ($\geq 1$) if we assign $k \geq 1$; this suggests that a moderate decrease in motivation is essential to reproduce the current state of the communal data sets. With regard to the time-bound distribution of contributions, the variant of the Gini coefficient displays results from our model that are similar to those from the data. For $k = 0.8$, the variant of the Gini coefficient for time-bound contributions steadily decreases from $N_e \approx 10^4$ (Fig. 4b), which is observed in the data for $N_e \gtrsim 10^4$ (Fig. 2i).

It may be suggested that, with regard to time-bound contributions, the early tendency of our model ($N_e \lesssim 10^4$) for the variant of the Gini coefficient disagrees with the data. In addition, undulation points that are absent in the data are observed at $N_e \approx 2 \times 10^4$ for the variant of the Gini coefficient with regard to the total contributions in cases other than $k = 0.8$. Although our model reproduces the patterns in the data at a later stage ($N_e \gtrsim 10^4$ for $k = 0.8$), this seemingly different growth pattern at an early stage should be stressed. First, as a minimalist model, we do not intend to reproduce the inherent disparity due to the characteristics of an agent, for example, social class, educational level and language fluency. Furthermore, we do not intend to explain the data all the way from the complex early procedure, when people launch a new project, that is available to a limited number of users, such as (1) language proposal and (2) incubator. The dynamics of this early stage is very different from the public launch of a project. Despite such discrepancies, our model starts with a regular dynamics from the very first agent and shows an early stabilizing period. The relationship between the number of editors and the number of edits in the model also displays two different stages (Supplementary Figs. 24–27). Although the transition point between two stages varies according to the specific values of $k$, we assume that this point may correspond to the undulation points for the variant of the Gini coefficient for total and time-bound contributions (compare Fig. 4 with Supplementary Figs. 24–27).

In addition to the variant of the Gini coefficient, our model also reproduces the trend of decreased short-term correlations for the number of edits between time windows reported in Fig. 3c. As shown in Fig. 4c, the interrelationship between the number of edits in two consecutive sequences in various time frames, from the onset of the data set, produces a similar result. In the model and empirical data, we observed a significant correlation between two consecutive sequences regardless of sequence length. With time, the short-term correlation is steadily reduced, whereas the long-term correlation is sustained. Like the empirical data, the boundary delimiting the size of time windows between large- ($\gtrsim 0.7$) and small-correlated ($\lesssim 0.7$) domains increases as more edits are performed (Fig. 4). The slope of a boundary is different for different $k$ values, but $\tau$ does not affect the slope.

Briefly, the parameter $k$ mainly governs the overall dynamics despite the fact that the rapid increase in inequality of total contributions happens at an early stage and the gradual decrease in inequality of time-bound contributions always occurs. In other words, the loss of long-term motivation induces the inequality, whereas the inertia of continuing the (editing) action does not affect the system notably. Therefore, the ‘rich get richer’ effect is mainly driven by the accumulated engagement induced by previous edits. Such a long-term engagement has led to the formation of the supereditors’ oligopoly that is lasting to date. In addition, our model indicates that the supereditors’ oligopoly can be formed without direct communication between editors.

**Discussion**

In this study, we have examined the common formation patterns displayed in all language editions of different types of Wikimedia...
projects. Although previous studies have reported these general patterns, they were usually based on partial observations of specific types of data sets or specific languages, leaving many unanswered questions. The extensive data set represented by Wikimedia projects, which record large-scale collaboration for the online creation of collective knowledge, provided us with an unprecedented opportunity to explore human collaborative behaviours quantitatively. In this data set, we observed a universal interplay between the number of editors, articles and edits, and article size, which are characterized by power-law scaling with a single set of exponents. The existence of universal growth patterns in all 273 languages and 12 types of Wikimedia projects suggests a pan-human behaviour with regard to collaboration.

This universal pattern is seen not only in the four key features of the data sets (number of edits, editors and articles, and article size), but also in their heterogeneity as quantified by the variant of the Gini coefficient; the disparity in who contributes the most is formed at a very early stage of the communal data sets and continues thereafter. It is widely hoped that communal data sets will bring about democratization of knowledge, yet studies reveal that current Wikimedia projects are hampered by strong heterogeneity in editing. We have demonstrated that heterogeneity between editors is more deep-rooted than expected. The existence of the supereditors’ very substantial share is a universal phenomenon across all Wikimedia communal data sets, regardless of size and/or activity. We have also observed a universal trend of intensified disparity for all types of data sets, which suggests that the vast influence exerted by a few dedicated editors will be intensified further. The value of such dedicated editors must be acknowledged because their voluntary dedication has enabled the current level of accuracy in Wikipedia. However, biased coverage of some topics has been reported; lost diversity may intensify systematic biases, notwithstanding Wikipedia’s continuing efforts towards neutrality. In addition, we have shown that a hierarchy of contributors in such communal data sets can be formed at a very early stage and that the polarization of editors is already underway.

Our study is not limited to a description of the current state of Wikimedia projects, but arguably provides a general insight into the future direction of communal data sets. For instance, our simulation suggests that inequality can arise without direct interaction between editors. Indeed, editors tend to obey pre-established authorities created by supereditors. Again, we acknowledge the dedication of supereditors, for instance, in maintaining the high quality of Wikipedia content by their (by definition) large number of contributions to it. However, the total productivity of each editor decreases as the number of edits increases, which may result in reduced productivity and accuracy in the future. The growth of Wikipedia has slowed down, and our analysis warns that inequality in contributions cannot be easily resolved without active efforts.

Since the turn of this century, Wikimedia projects have spearheaded the international open knowledge market. However, when considering the nature of such a social structure, strategic actions are required to sustain its openness to worldwide collaborations. Giving incentives to new editors may help, but a suitable tutorial system that prevents vandalism and encourages productive editing activities is also needed. Fork-and-merge systems commonly used in open-source communities also improve the editability of Wikimedia projects by serving as a secondary talk pages method where editors can debate. With the fork system, new editors fork their own versions of articles to share their ideas, which can be merged with the original article following discussion and debate.

Our findings provide evidence of imbalances in the formation of a particular set of communal data from the outset, but the results and implications of our study are relevant beyond Wikimedia projects. With the Internet being such a central aspect of daily life, online environments have become a mainstream platform for collaborative knowledge creation. Therefore, researchers have started to study the contribution patterns in various communal data sets, such as open-source and free software. Compared to traditional (offline) collaboration systems, the products of online collaboration are immediately available in a collaborative fashion. Future analyses of GitHub, Apache, GNU software, free software and copyleft will contribute to a more detailed landscape of collaborative knowledge creation. These analyses can further our understanding of human collective behaviours, and, we hope, provide clues to address social inequalities at an even larger scale.

Methods
Data description. We used the March 2016 dump of all Wikimedia projects (https://dumps.wikimedia.org/back-up-index.html), which contains the complete copy of Wikimedia articles from the 15 January 2001 to 5 March 2016, including the raw text and metadata in XML format. This data set not only includes Wikipedia, but also its sibling collaborative projects, such as Wiktionary, Wikibooks, Wikiquote, Wikisource, Wikinews, Wikiversity and Wikivoyage in different languages (see Supplementary Table 1 for further details). Each of these openly edited projects has a distinct subject and object. For example, each language edition of Wiktionary aims to describe all words in all languages. For example, the English edition of Wiktionary aims to describe all words of all languages using definitions and descriptions in English. Differences between objects may yield gaps in the editing behaviours of editors belonging to each project, which is caused by a difference in demographic pools, accessibility, degree of interest, and so on.

This data contains a total of 267,304,095 articles across all Wikimedia projects with their complete editing histories. Each article documents either the Wikipedia account identification or the Internet protocol address of the editor for each edit, the article size, the timestamp for each edit, and so on. A single character takes one byte, except for a few languages such as Korean (two bytes) and Chinese (two or three bytes). Thus, article size in bytes is a direct measure of article length. Each data set contains a number of articles ranging from 43,124,816 (Wikimedia Commons, a database of freely usable audiovisual media) to 3 (Wikipedia Login, a database used for administrative purposes); the number of editors ranges from 44,349,908 (English Wikipedia) to 5 (Wikipedia Nostalgia, a read-only copy of the old English language Wikipedia); the number of edits ranges from 654,163,757 (English Wikipedia) to 5 (Wikipedia Login); and the total article size ranges from 99,519,138,751 bytes (English Wikipedia) to 1,206 bytes (Wikipedia Login) (see Supplementary Fig. 2 for the distributions of various measures).

Estimation of power exponent for the correlation between N, Np, Ns and S. To estimate the power-law scaling relationship between measures, we applied a simple linear regression method to the logarithm of the values of interest, assuming a simple power-law scaling of $y = Cx^\alpha$. Inevitably, various types of noise and fluctuations affected the empirical observations, so the distribution should in fact be written as $y = \kappa(x/x_{\text{min}}) + \eta + \epsilon$, where $x_{\text{min}}$ is the minimum value of $x$ from which the power-law is observed, $\kappa$ is the constant background offset and $\eta$ represents random fluctuations. We neglected the noises and fluctuations to obtain the overall collective trends for all Wikimedia projects. This simple method has clear advantages over complex multivariate regression; it is less sensitive to the heterogeneous disparity in empirical distribution. The aforementioned power-law scaling can be transformed as $\ln(y) = \ln(\kappa) + \ln(x) + \ln(\lambda)$. We performed simple linear regression on the logarithmic values of $\ln(x)$ and $\ln(y)$ to yield the exponent $\alpha$ and the proportionality constant $C$ (see Supplementary Tables 2–7 for the statistical details).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The codes used for data processing, agent-based modelling and figure drawing are available from GitHub (https://github.com/bluekura/wikimedia-inequality).

Data availability. Wikipedia dumps for the main analysis are available from the Wikimedia Downloads (https://dumps.wikimedia.org/). Additional public data sets are also available from OECD (http://data.oecd.org), UNESCO (http://data.uis.unesco.org/); and the CIA (https://www.cia.gov/library/publications/the-world-factbook/). The data set of the total number of speakers for each language is owned by SIL. International and can be accessed from their website by means of a subscription (https://www.ethnologue.com/). Bibliographic metadata of academic papers and patents were retrieved from the in-house system of the Knowledge Science and Technology Information and were licensed from Scopus (https://www.scopus.com/) and the European Patent Office (https://www.epo.org/searching-for-patents/business/psstat.html); distribution is prohibited.
The pre-processed data used to create the figures are available from GitHub (https://github.com/bluekura/wikimedia-inequality), along with codes.

Received: 31 May 2017; Accepted: 31 October 2018; Published online: 17 December 2018

References
1. Bruns, A. Blogs, Wikipedia, Second Life, and Beyond: From Production to Produsage (Peter Lang, New York, 2008).
2. Lemke, C. & Coughlin, E. The change agents. Educ. Leadersh. 67, 54–59 (2009).
3. Walker, L. Spreading knowledge, the wiki way. Washington Post 9 (September 2004); http://www.washingtonpost.com/wp-dyn/articles/A5430-2004Sep8.html
4. Seeley, K. Q. Snared in the web of a Wikipedia liar. The New York Times (4 December 2005); https://www.nytimes.com/2005/12/04/weekinreview/snared-in-the-web-of-a-wikipedia-liar.html
5. Chesney, T. An empirical examination of Wikipedia’s credibility. First Monday 11, https://doi.org/10.5210/fm.v11i11.1413 (2006).
6. Giles, J. Internet encyclopaedias go head to head. Nature 438, 900–901 (2005).
7. Gandica, Y., Carvalho, J. & Sampaio dos Aidos, F. Wikipedia editing dynamics. Phys. Rev. E 91, 012824 (2015).
8. Yun, J., Lee, S. H. & Jeong, H. Intellectual interchanges in the history of the massive open online editing encyclopedia, Wikipedia. Phys. Rev. E 93, 012307 (2016).
9. Heaberlin, B. & DeDeo, S. The evolution of Wikipedia’s norm network. Future Internet 8, 14 (2016).
10. Zha, Y., Zhoua, T. & Zhou, C. Unfolding large-scale online collaborative human dynamics. Proc. Natl Acad. Sci. USA 113, 14627–14632 (2016).
11. Ortega, F., Gonzalez-Barahona, J. M. & Robles, G. On the inequality of contributions to Wikipedia. In Proc. 41st Annual Hawaii International Conference on System Sciences (IEEE, 2008).
12. Kittur, A., Suh, B. & Chi E. H. Can you ever trust a wiki? Impacting perceived trustworthiness in Wikipedia. In Proc. 2008 ACM Conference on Computer Supported Cooperative Work 477–480 (ACM, 2008).
13. Adler, B. T. et al. Assigning trust to Wikipedia content. In Proc. 4th International Symposium on Wikis 26 (ACM, 2008).
14. Yasseri, T., Sumi, R., Rung, A., Kornai, A. & Kertész, J. Dynamics of conflicts in Wikipedia. PLoS ONE 7, e38869 (2012).
15. Barber, W. & Badre, A. Culturalitity: the merging of culture and usability. In Proc. 4th Conference on Human Factors and the Web http://www.ncl.ac.uk/government/hfweb/att4/proceedings/badre/ (1998).
16. Marcus, A. & Gould, E. W. Crosscurrents: cultural dimensions and global Web-user interface design. Interactions 7, 32–46 (2000).
17. Schmid-Isler, S. The language of digital genres: a semiotic investigation of style and iconology on the World Wide Web. In Proc. 33rd Hawaii International Conference on System Sciences https://doi.org/10.1109/HICSS.2000.926695 (IEEE, 2000).
18. Pfeil, U., Zaphiris, P. & Ang, C. Cultural differences in collaborative authoring of Wikipedia. In J. Comput. Mediat. Commun. 12, 88–113 (2006).
19. Kim, S. et al. Understanding editing behaviors in multilingual Wikipedia. PLoS ONE 11, e0153305 (2016).
20. Blei, D. M. & Jordan, M. I. Variational inference for Dirichlet process mixtures. Bayesian Anal. 1, 121–143 (2006).
21. van der Maaten, L. J. P. & Hinton, G. E. Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605 (2008).
22. Simon, G. F. & Fennig, C. D. (eds) Summary by language size. In Ethnologue: Languages of the World, Twenty-first Edition Online Version (SIL International, Dallas, 2018); https://www.ethnologue.com/statistics/size
23. Gini, C. Variabilità e Mutabilità: Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche (Tipog. di P. Cappini, Bologna, 1912).
24. Mankiw, N. G. Principles of Economics (Dryden, Forth Worth, 1998).
25. George, B. P. & George, B. P. Past visits and the intention to revisit a destination: place attachment as the mediator and novelty seeking as the moderator. J. Tour. Stud. 15, 37–50 (2004).
26. Crane, R. & Sornette, D. Robust dynamic classes revealed by measuring the response function of a social system. Proc. Natl Acad. Sci. USA 105, 15649–15653 (2008).
27. Wu, F. & Huberman, B. A. Novelty and collective attention. Proc. Natl Acad. Sci. USA 104, 17599–17601 (2007).
28. Karsai, M., Kaski, K., Barabási, A.-L. & Kertész, J. Universal features of correlated bursty behaviour. Sci. Rep. 2, 397 (2012).
29. Karsai, M., Perra, N. & Vespignani, A. Time varying networks and the weakness of strong ties. Sci. Rep. 4, 4001 (2014).
30. Io, H.-H., Perotti, J. I., Kaski, K. & Kertész, J. Correlated bursts and the role of memory range. Phys. Rev. E 92, 022814 (2015).
31. Gandica, Y. et al. On the origin of burstiness in human behavior: the Wikipedia edits case. Preprint at https://arxiv.org/abs/1601.00864 (2016).
32. Hube, C. Bias in Wikipedia. In Proc. 26th International Conference on World Wide Web Companion 717–721 (IW3C2, 2017).
33. Callahan, E. S. & Herring, S. C. Cultural bias in Wikipedia content on famous persons. J. Assoc. Inf. Sci. Technol. 62, 1899–1915 (2011).
34. Reagle, J. & Rhue, L. Gender bias in Wikipedia and Britannica. Int. J. Commun. 5, 1138–1158 (2011).
35. Kutt, A. & Kraut, R. E. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In Proc. 2008 ACM Conference on Computer Supported Cooperative Work 37–46 (ACM, 2008).
36. Suh, B., Convertino, G., Chi, E. H. & Pirolli, P. The singularity is not near: slowing growth of Wikipedia. In Proc. 5th International Symposium on Wikis and Open Collaboration 8 (ACM, 2009).
37. Gherardi, M., Mandrà, S., Bassetti, B. & Cosentino Lagomarsino, M. Evidence for soft bounds in Ubuntu package sizes and mammalian body masses. Proc. Natl Acad. Sci. USA 110, 21054–21058 (2013).
38. Tsay, J., Dabbish, L. & Herbsleb, J. Influence of social and technical factors for evaluating contribution in GitHub. In Proc. 36th International Conference on Software Engineering 356–366 (ACM, 2014).
39. Padhye, R., Mani, S. & Sinha, V. S. A study of external community contribution to open-source projects on GitHub. In Proc. 11th Working Conference on Mining Software Repositories 332–335 (ACM, 2014).
40. Benkler, Y. Wealth of Networks: How Social Production Transforms Markets and Freedom (Yale Univ. Press, New Haven, 2006).
41. Vergeau, E. UTF-8, A Transformation Format of ISO 10646 https://tools.ietf. org/html/rfc3629 (2003).

Acknowledgements
This work received institutional supports from the Korea Institute of Science and Technology Information. The National Research Foundation (NRF) of Korea grant funded the Korean Government also supported this work through grant nos. NRF-2017R1A1A103070975 (EY), NRF-2018R1C1B5083863 (S.H.L.) and NRF- 2017R1A2B3060930 (H.I.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions
All three authors designed the experiment and wrote the manuscript. J.Y. contributed to the analysis, decision to publish or preparation of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0488-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.H.L. or H.I.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2018
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed |
|-----|-----------|

☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐ The statistical test(s) used AND whether they are one- or two-sided
☐ A description of all covariates tested
☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
☐ Give P values as exact values whenever suitable.
☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated
☐ Clearly defined error bars
☐ State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

Data access was conducted using GNU wget to download Wikimedia xml dumps.

Data analysis

Data was analyzed using Python version 2.7 with the packages: numpy, scipy, scikit-learn, and pandas.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

- Public data sets are available in following Webpages: Wikimedia (https://dumps.wikimedia.org/), OECD (https://data.oecd.org/), UNESCO (http://data.uis.unesco.org/), and CIA (https://www.cia.gov/library/publications/the-world-factbook/).
- The data set containing languages by total number of speakers is propriety SIL International and accessible at their website with subscription (https://
Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences  ☒ Behavioural & social sciences  ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Study description | We analysed the activities in the entire Wikimedia projects with four characteristic measures (number of articles, number of edits, number of editors, and total size in bytes) and an inequality index (a variant of the Gini index). |
|-------------------|--------------------------------------------------------------------------------------------------|
| Research sample   | We used the complete editing history of 267,304,095 Wikimedia items until 2016. |
| Sampling strategy | No sample size calculation was executed due to the design of the research. |
| Data collection   | All data was independently collected by the Wikimedia foundations, OECD, UNESCO, CIA, SIL International, SCOPUS, and European Patent Office. |
| Timing            | All data was collected between 2000 and 2016. |
| Data exclusions   | No data were excluded. |
| Non-participation | No participant dropped out of participation. |
| Randomization     | In the model study, we assigned random seed of the random number generators. |

Reporting for specific materials, systems and methods

Materials & experimental systems

- Unique biological materials
- Antibodies
- Eukaryotic cell lines
- Palaeontology
- Animals and other organisms
- Human research participants

Methods

- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging