nature human behaviour

SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41562-018-0488-z

In the format provided by the authors and unedited.

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

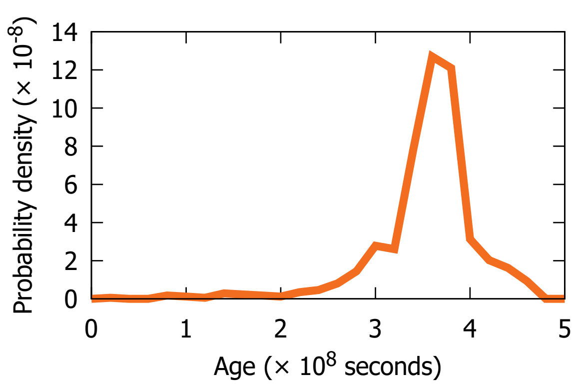
Jinhyuk Yun[1], Sang Hoon Lee[2*] and Hawoong Jeong[3,4*]

[1]Future Technology Analysis Center, Korea Institute of Science and Technology Information, Seoul, Korea. [2]Department of Liberal Arts, Gyeongnam National University of Science and Technology, Jinju, Korea. [3]Department of Physics, Korea Advanced Institute of Science and Technology, Daejeon, Korea. [4]Institute for the BioCentury, Korea Advanced Institute of Science and Technology, Daejeon, Korea. *e-mail: lshlj82@gntech.ac.kr; hjeong@kaist.edu

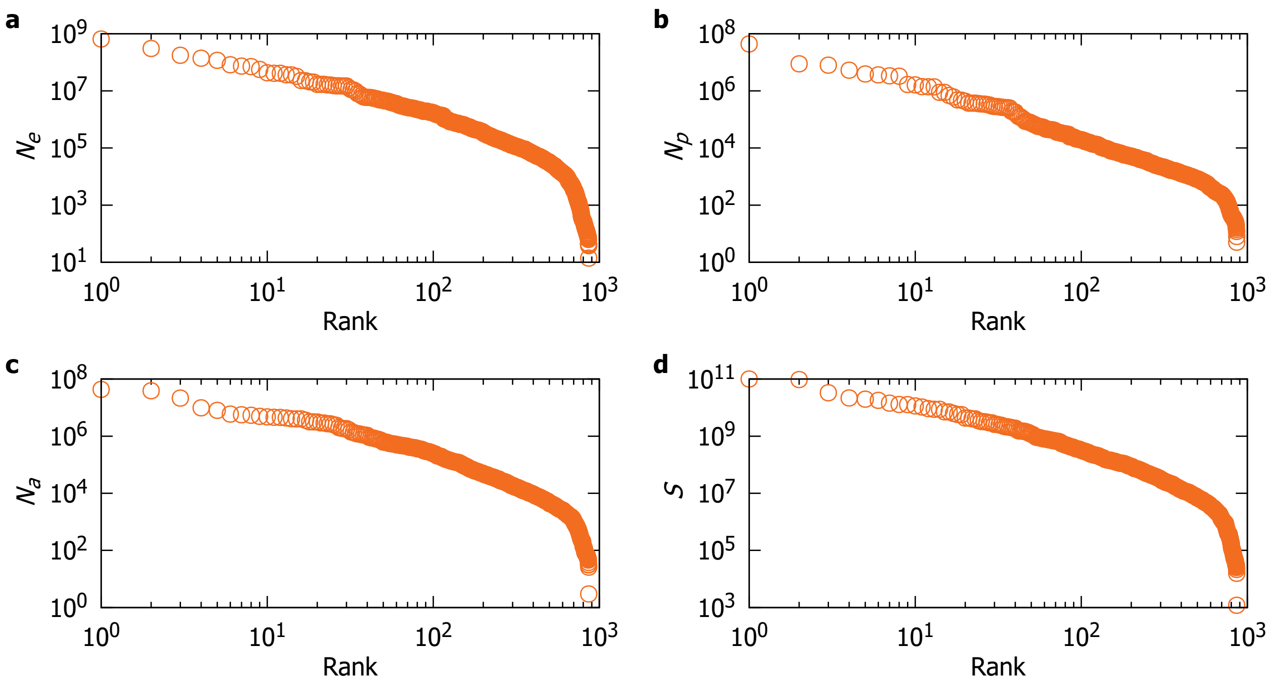
Supplementary Information

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

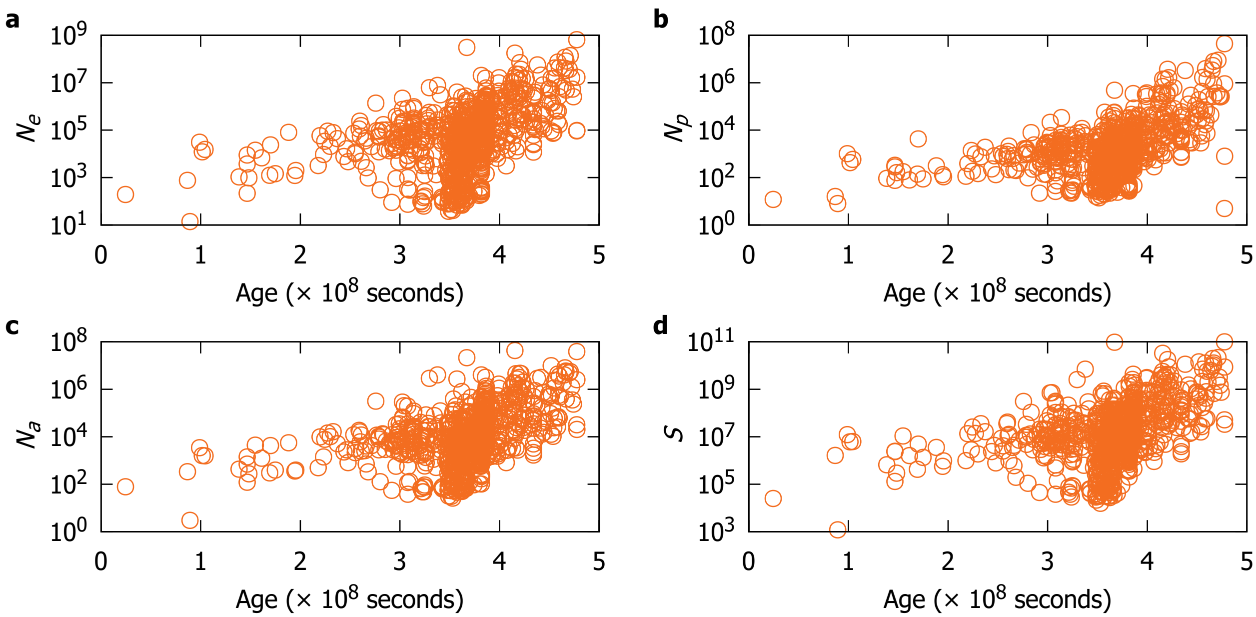
SUPPLEMENTARY FIGURES



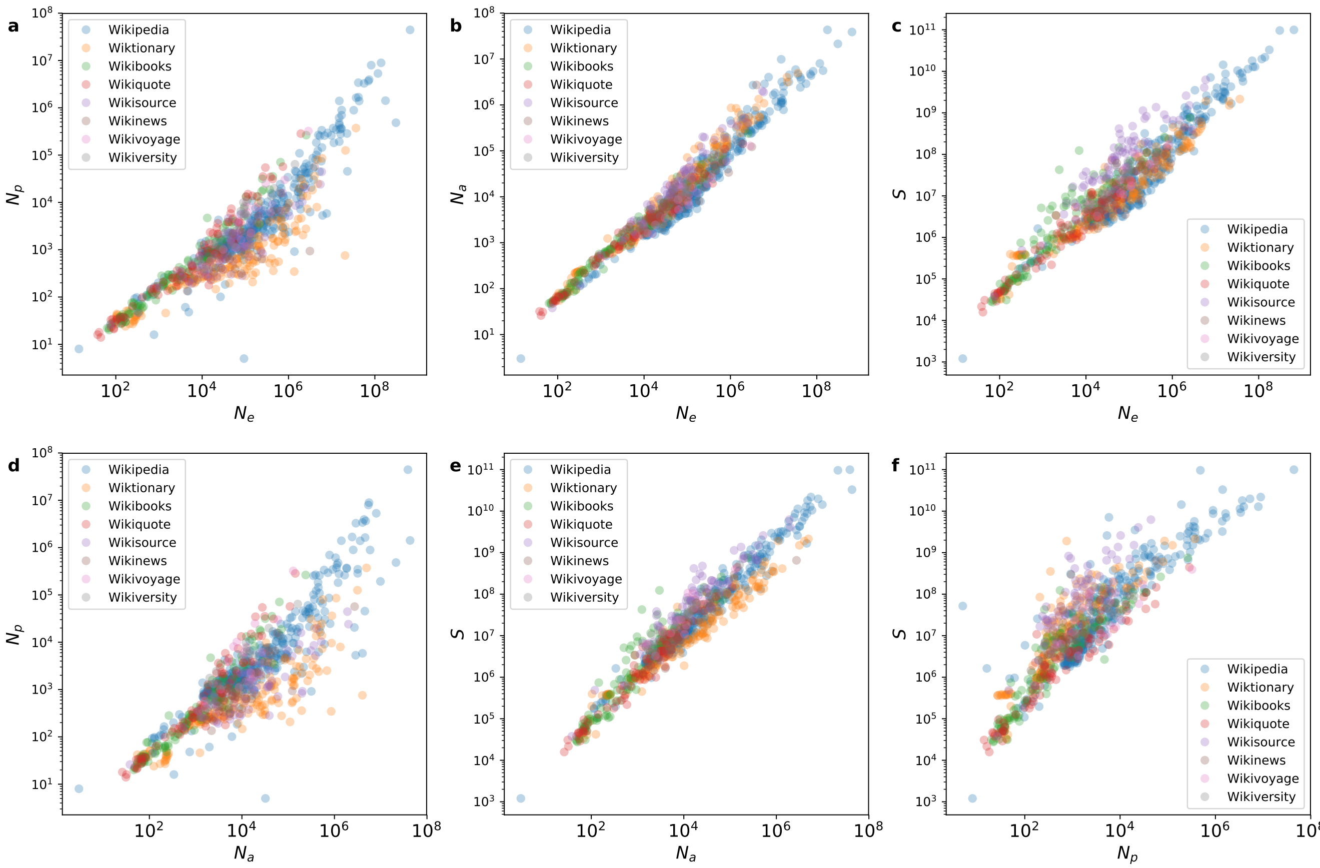
SUPPLEMENTARY FIGURE 1. The age distribution of Wikimedia projects, where we bin the data in the uniform length of 2×10^7 seconds (the resolution of the x axis).



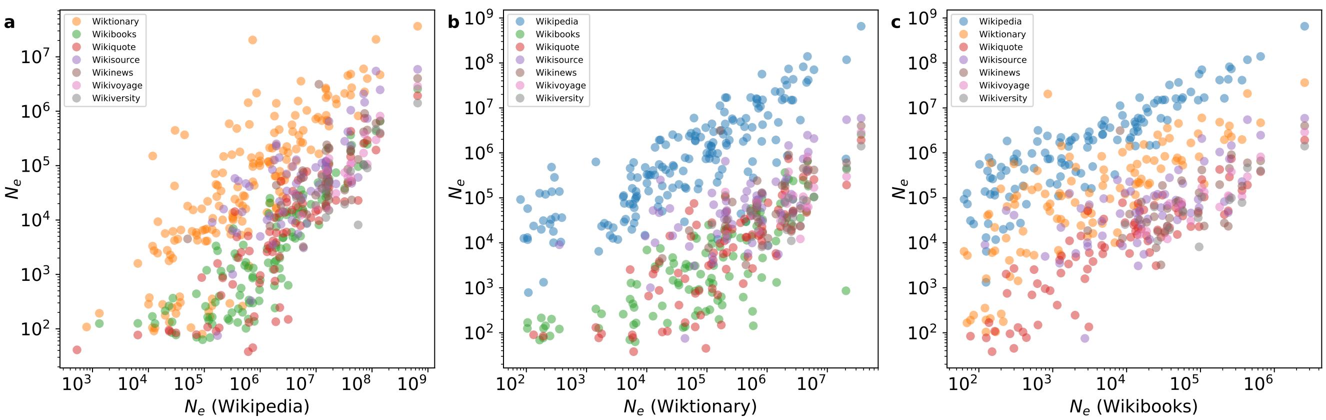
SUPPLEMENTARY FIGURE 2. The rank versus characteristic measures of Wikimedia projects: **a**, the number of edits N_e, **b**, the number of editors N_p, **c**, the number of articles N_a, and **d**, the total volume of texts S (in bytes).



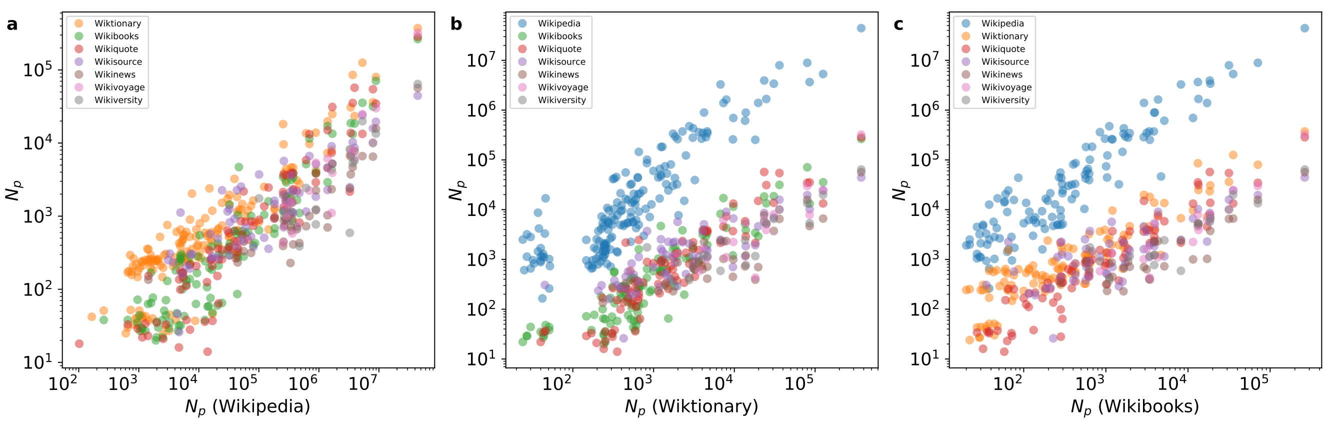
SUPPLEMENTARY FIGURE 3. Correlations for, **a**, the number of edits (Pearson correlation coefficient $\rho = 0.12$), **b**, the number of editors (Pearson correlation coefficient $\rho = 0.11$), **c**, the number of articles (Pearson correlation coefficient $\rho = 0.13$), and **d**, the total size of Wikimedia projects (Pearson correlation coefficient $\rho = 0.12$), with the age of Wikimedia projects.



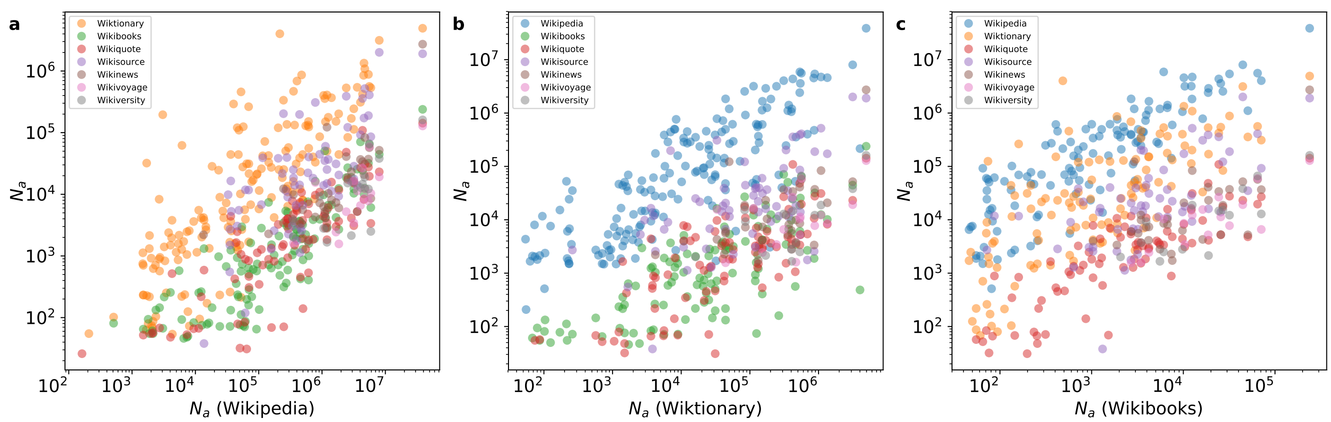
SUPPLEMENTARY FIGURE 4. Correlations between the number of edits N_e, the number of editors N_p, the number of articles N_a, and the total size of the data set S. The colours represent the types of Wikimedia projects.



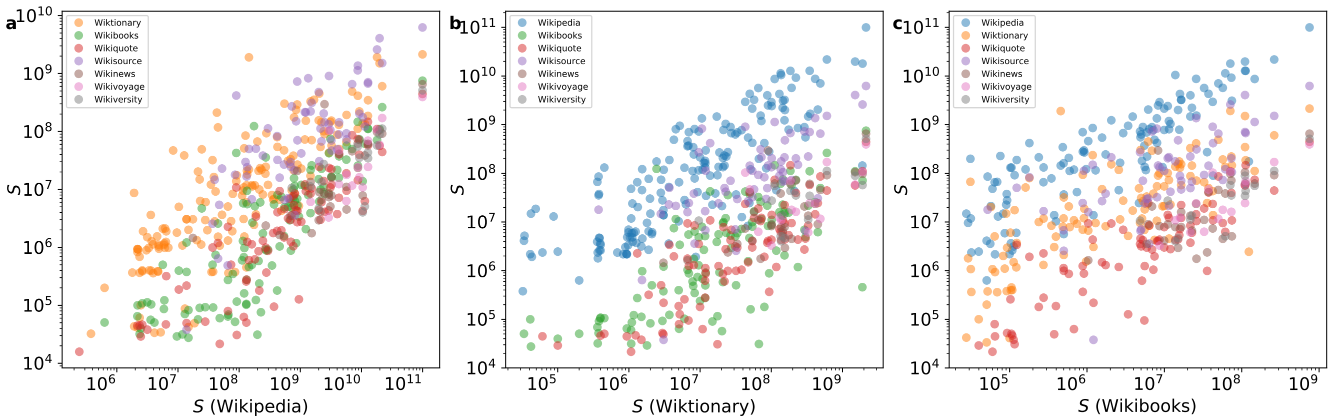
SUPPLEMENTARY FIGURE 5. Correlations between the number of edits N_e of each Wikimedia projects and that of the three largest Wikimedia projects for each language. **a**, The correlation between Wikipedia and the rest. **b**, The correlation between Wiktionary and the rest. **c**, The correlation between Wikibooks and the rest. The colours distinguish the types of corresponding Wikimedia projects.



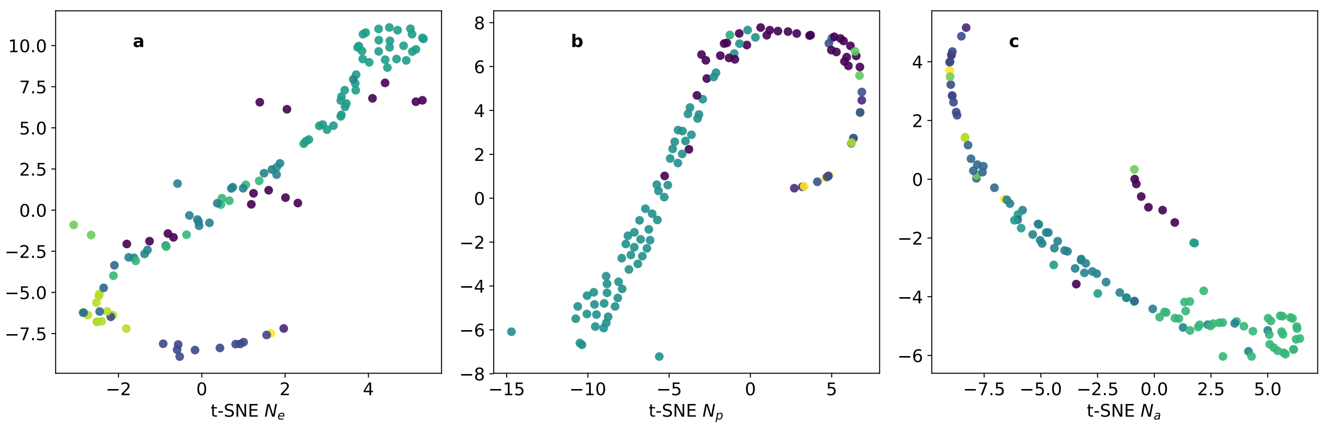
Supplementary Figure 6. Correlations between the number of editors N_p of each Wikimedia projects and that of the three largest Wikimedia projects for each language. **a**, The correlation between Wikipedia and the rest. **b**, The correlation between Wiktionary and the rest. **c**, The correlation between Wikibooks and the rest. The colours distinguish the types of corresponding Wikimedia projects.



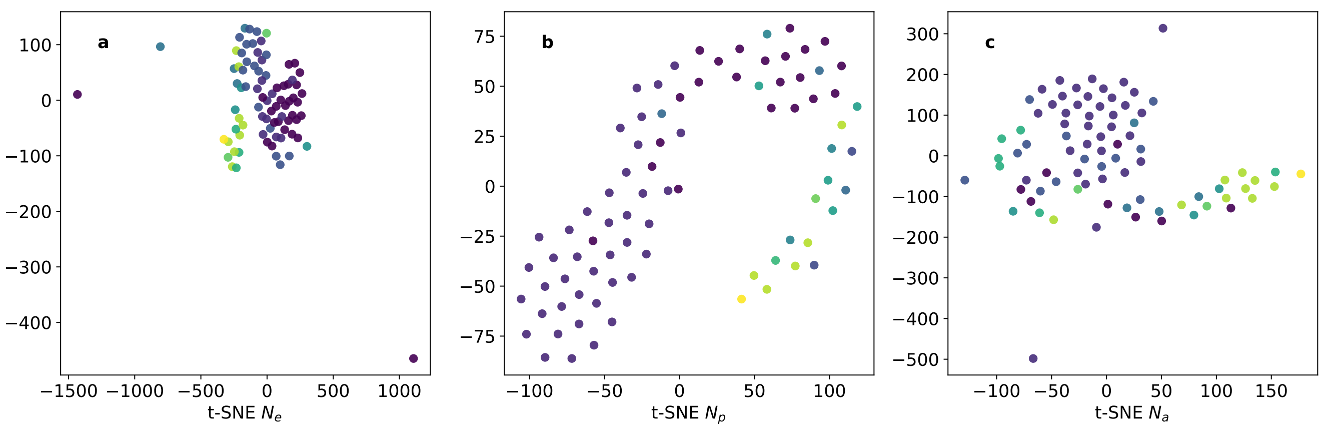
SUPPLEMENTARY FIGURE 7. Correlations between the number of articles N_a of each Wikimedia projects and that of the three largest Wikimedia projects for each language. **a**, The correlation between Wikipedia and the rest. **b**, The correlation between Wiktionary and the rest. **c**, The correlation between Wikibooks and the rest. The colours distinguish the types of corresponding Wikimedia projects.



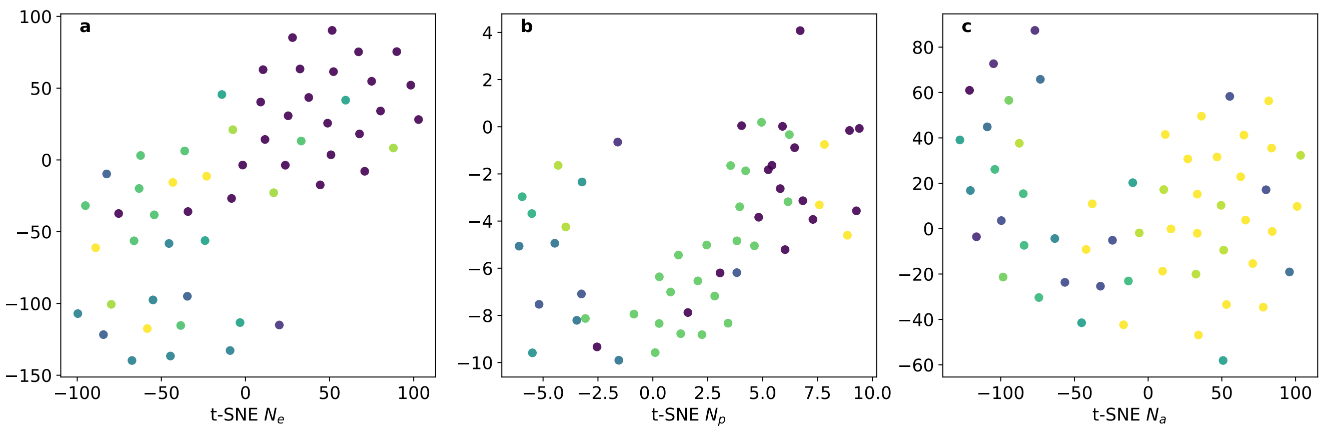
Supplementary Figure 8. Correlations between the total size of the data set S of each Wikimedia projects and that of the three largest Wikimedia projects for each language. **a**, The correlation between Wikipedia and the rest. **b**, The correlation between Wiktionary and the rest. **c**, The correlation between Wikibooks and the rest. The colours distinguish the types of corresponding Wikimedia rest.



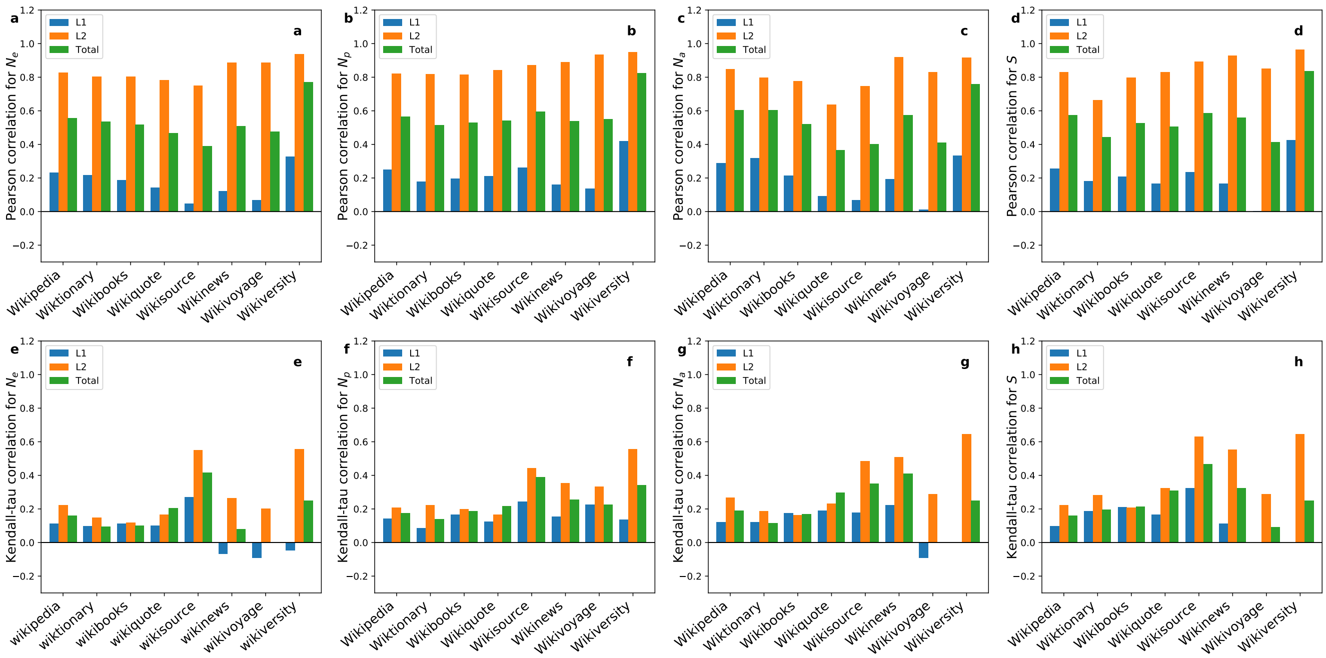
Supplementary Figure 9. The t-distributed stochastic neighbour embedding (t-SNE) [21] of 3-vector of Wikimedia projects onto 2-dimensional space (see Supplementary Methods for the details). **a**, embedded vector from 3-vector of number of edits N_e. **b**, embedded vector from 3-vector of number of editors N_p. **c**, embedded vector from 3-vector of number of articles N_a. Each point corresponds to a certain language in Wikimedia project, and its colour corresponds to a cluster to which the language belongs. Each language is clustered by Dirichlet Process Gaussian Mixture Model [20] with a parameter $\gamma_0 = 10^{-2}$, yet the result is stable within the range of $\gamma_0 \in [10^{-2}, 10^2]$.



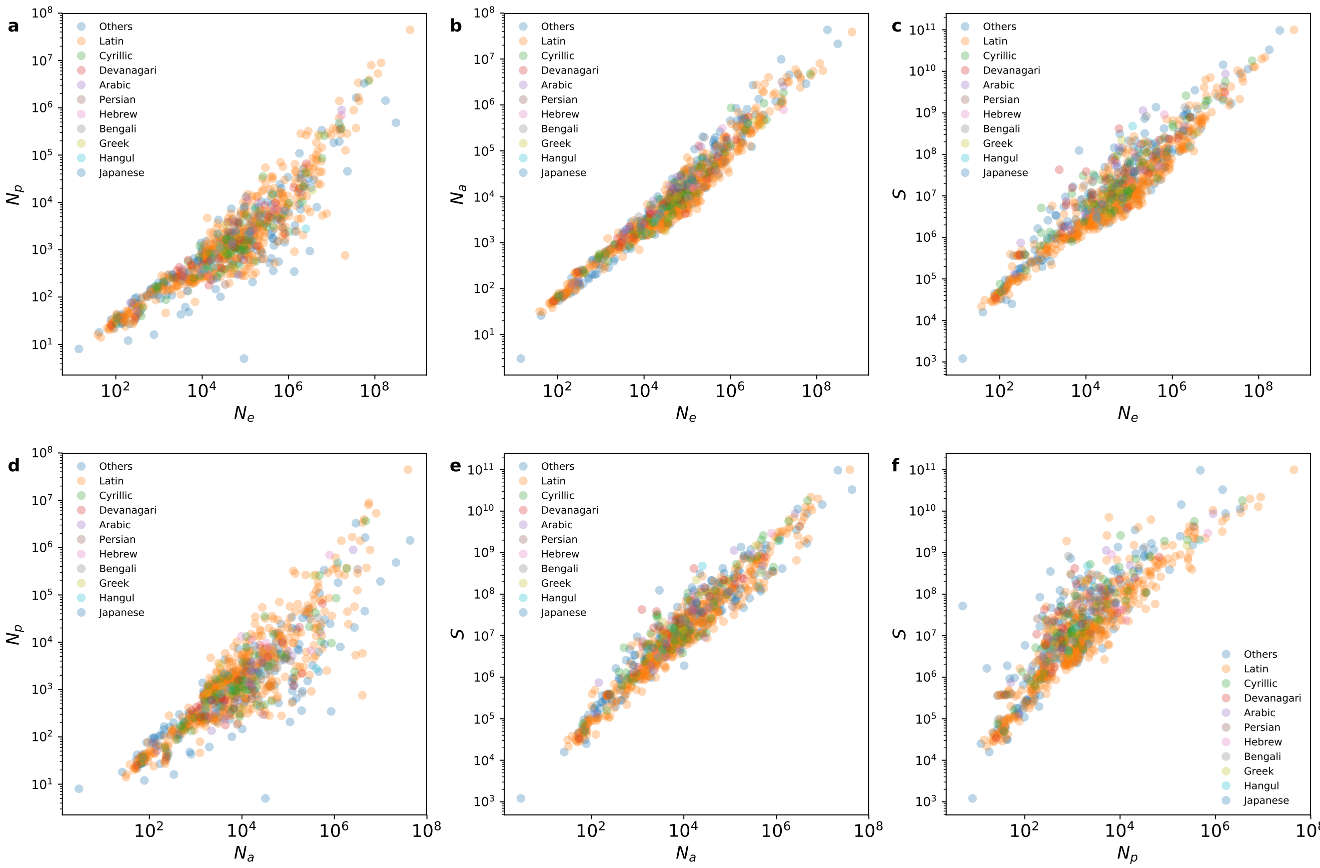
Supplementary Figure 10. The t-distributed stochastic neighbour embedding (t-SNE) [21] of 4-vector of Wikimedia projects onto 2-dimensional space (see Supplementary Methods for the details). **a**, embedded vector from 4-vector of number of edits N_e. **b**, embedded vector from 4-vector of number of editors N_p. **c**, embedded vector from 4-vector of number of articles N_a. Each point corresponds to a certain language in Wikimedia project, and its colour corresponds to a cluster to which the language belongs. Each language is clustered by Dirichlet Process Gaussian Mixture Model [20] with a parameter $\gamma_0 = 10^{-2}$, yet the result is stable within the range of $\gamma_0 \in [10^{-2}, 10^{2}]$.



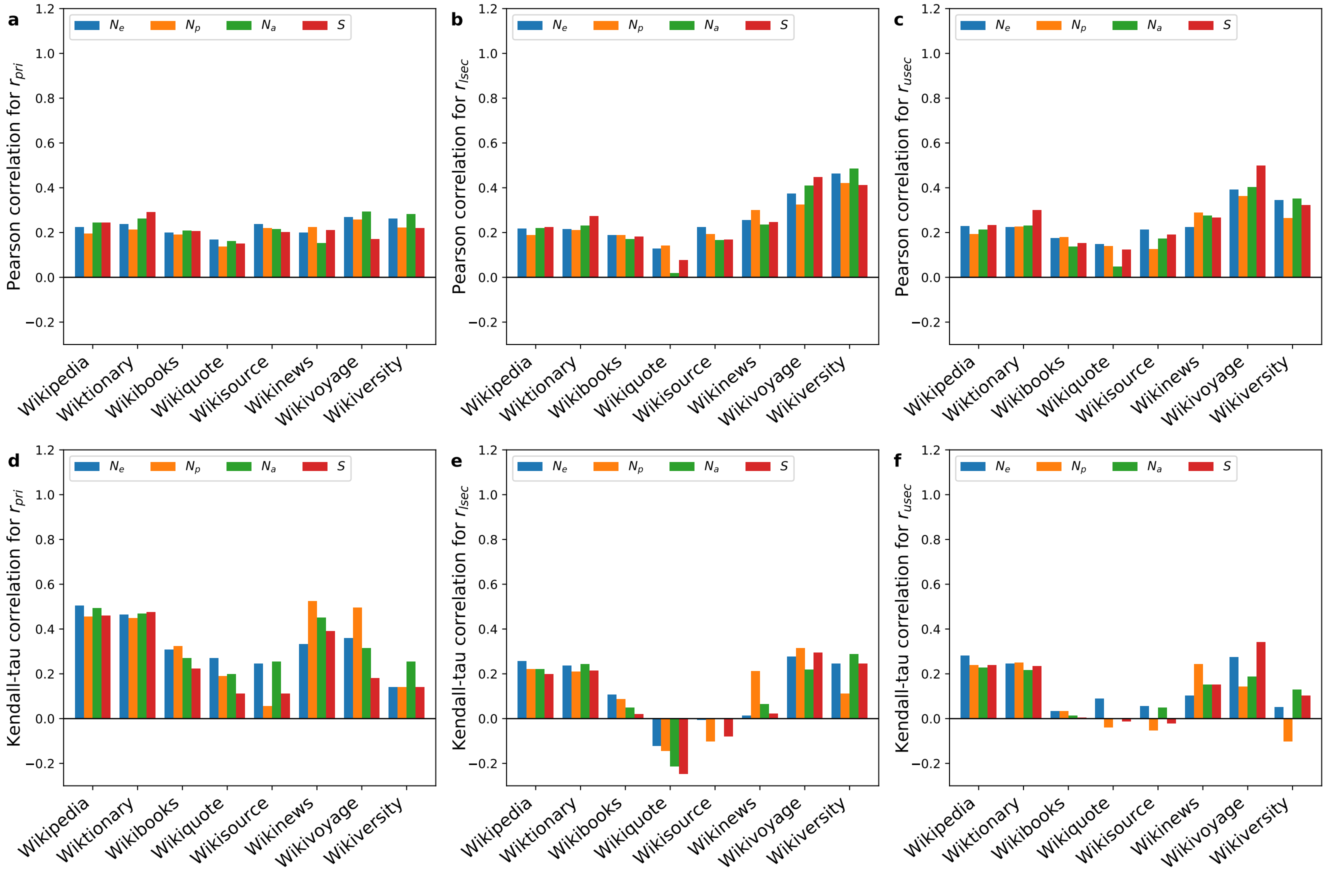
Supplementary Figure 11. The t-distributed stochastic neighbour embedding (t-SNE) [21] of 5-vector of Wikimedia projects onto 2-dimensional space (see Supplementary Methods for the details). **a**, embedded vector from 5-vector of number of edits N_e. **b**, embedded vector from 5-vector of number of editors N_p. **c**, embedded vector from 5-vector of number of articles N_a. Each point corresponds to a certain language in Wikimedia project, and its colour corresponds to a cluster to which the language belongs. Each language is clustered by Dirichlet Process Gaussian Mixture Model [20] with a parameter $\gamma_0 = 10^{-2}$, yet the result is stable within the range of $\gamma_0 \in [10^{-2}, 10^2]$.



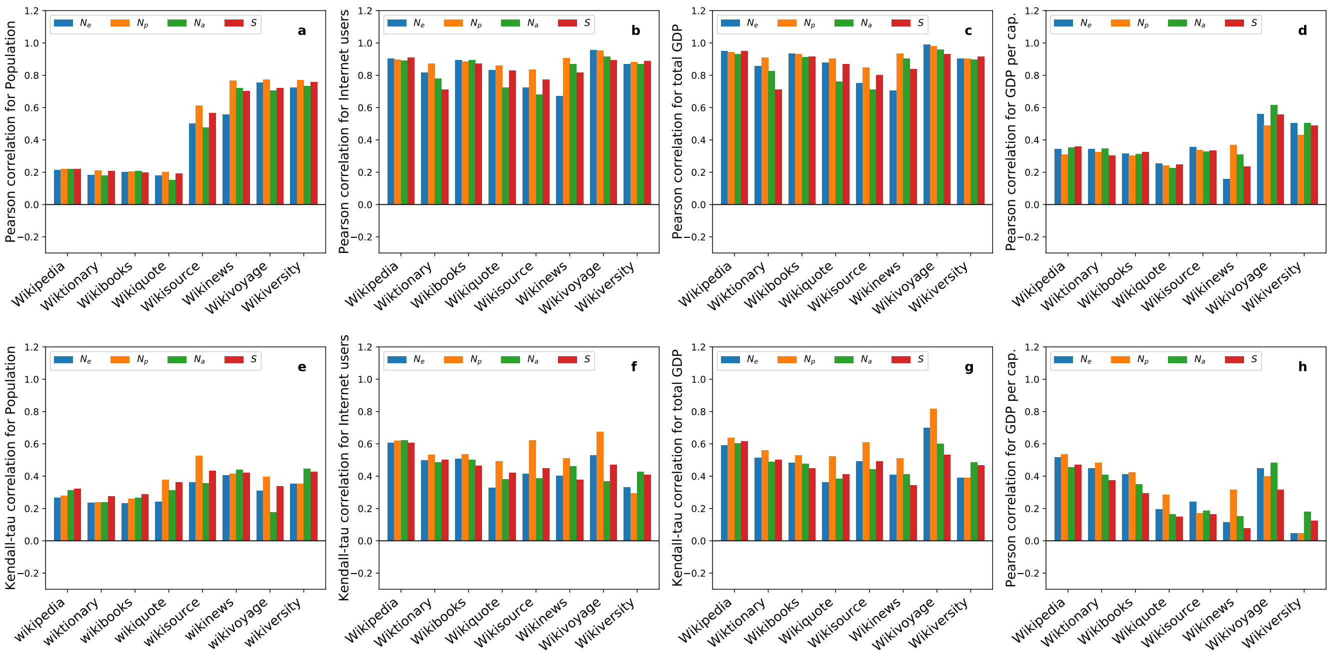
Supplementary Figure 12. The correlations of total number of speakers as functions of measures of Wikimedia project: (**a**,**e**), the number of edits N_e, (**b**,**f**), the number of editors N_p, **(c,g)**, the number of articles N_a, and (**d**,**h**), the total volume of texts S (in the unit of bytes). For (**a**–**h**) colours of the bar indicate number of the language's level 1 speakers (blue bars; as a native language), level 2 speakers (orange bars; as a secondary language), and total number of speakers (green bars). All of the plots are drawn under the conditions as follows: 1) the number of speakers are estimated by 2017 edition of Ethnologue [22], 2) at least 50 million speakers should be in total, and 3) languages with no information on level 2 speakers are assumed to not having level 2 speakers. As a result, we used 143 Wikimedia projects in 26 distinct languages in total. Please note that the estimation of language users is generally not reliable, because they were collected from different sources and dates [42]. Both the Pearson correlation coefficient (**a**–**d**) and the Kendall-tau rank correlation coefficient (**e**–**h**) are used to quantify interplays between measures, to exclude possible biases when we only observe a single measure.



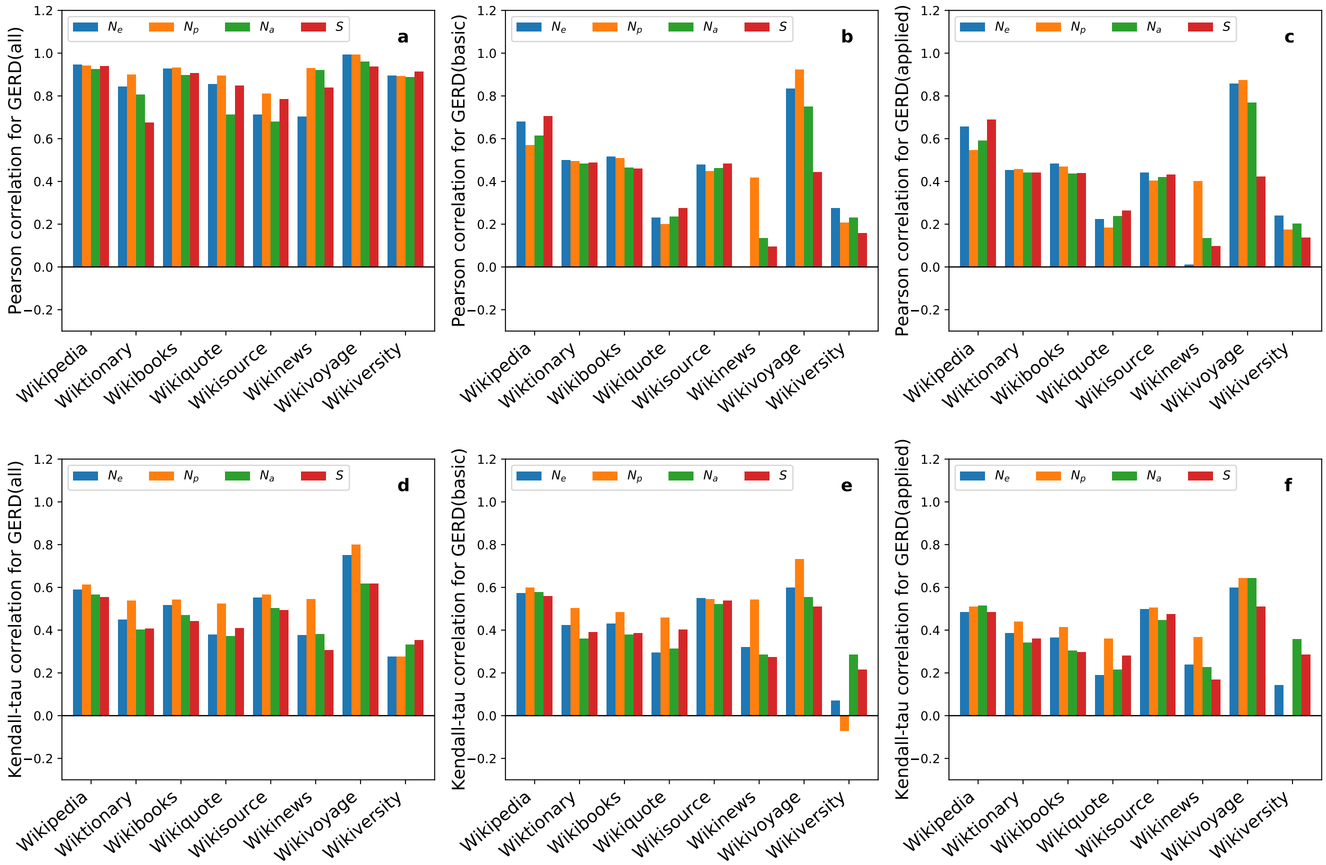
Supplementary Figure 13. The correlations between number of edits N_e, number of editors N_p, number of articles N_a, and total size of the data set S. The colours represent the script with which the corresponding Wikimedia projects are written.



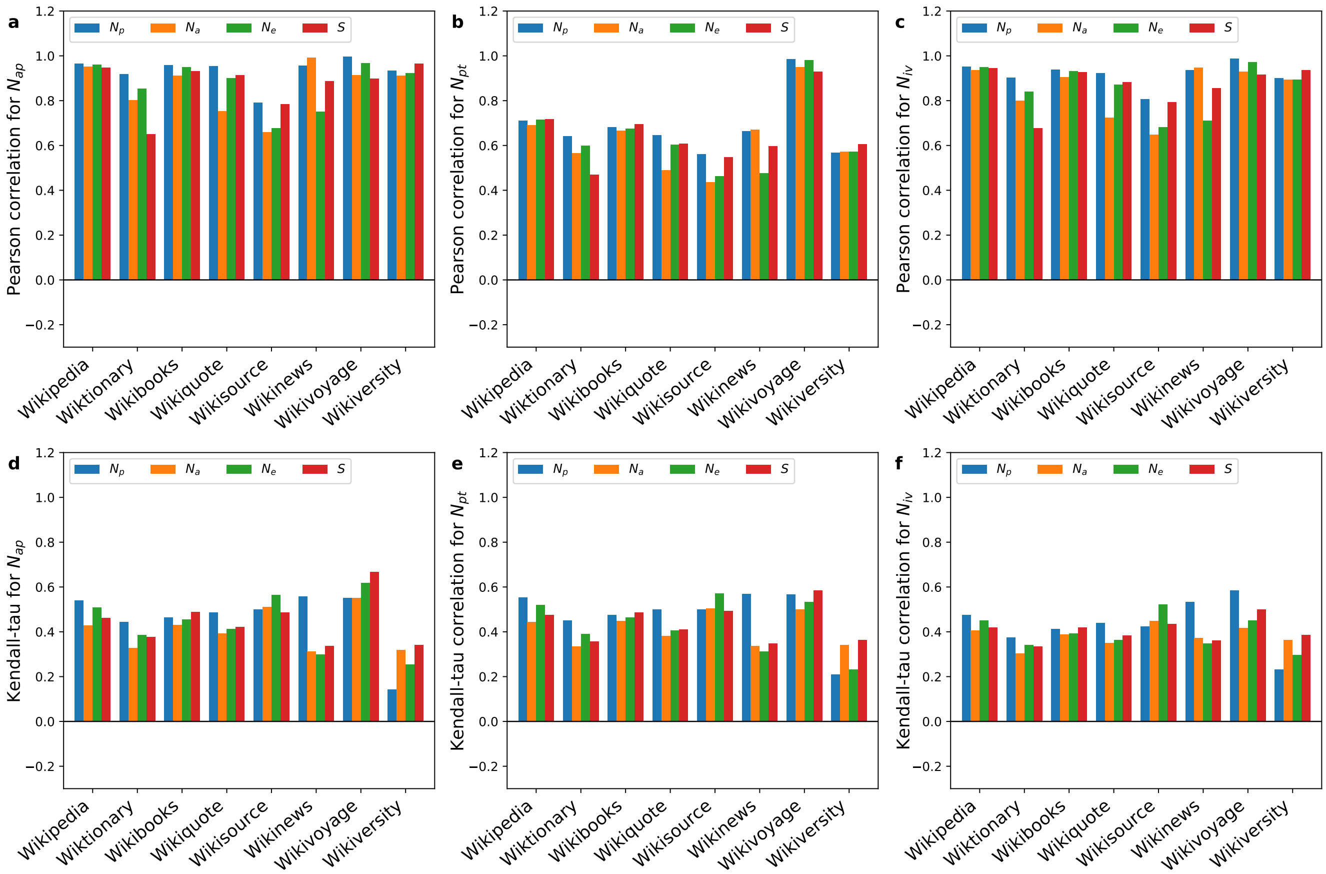
Supplementary Figure 14. The correlations between measures of different Wikimedia projects and education levels of the corresponding country (See Supplementary Methods for details). Education levels are collected from UNESCO Institute for Statistics [45] **a,d**, The correlation coefficients on measures of Wikimedia projects for attainment rate of a primary school of the corresponding country. **b,e**, Correlation coefficients on measures of Wikimedia projects for attainment rate of a lower secondary school of the corresponding country. **c,f**, The correlation coefficients on measures of Wikimedia projects for attainment rate of a upper secondary school of the corresponding country. For panels (**a**–**f**), the colours of the bars indicate the number of edits N_e (blue bars), the number of editors N_p (orange bars), the number of articles N_a (green bars), and the total volume of texts S (red bars). Both the Pearson correlation coefficient (**a**–**c**) and the Kendall-tau rank correlation coefficient (**d**–**f**) is used to quantify interplays between measures, to exclude possible biases when we only observe a single measure.



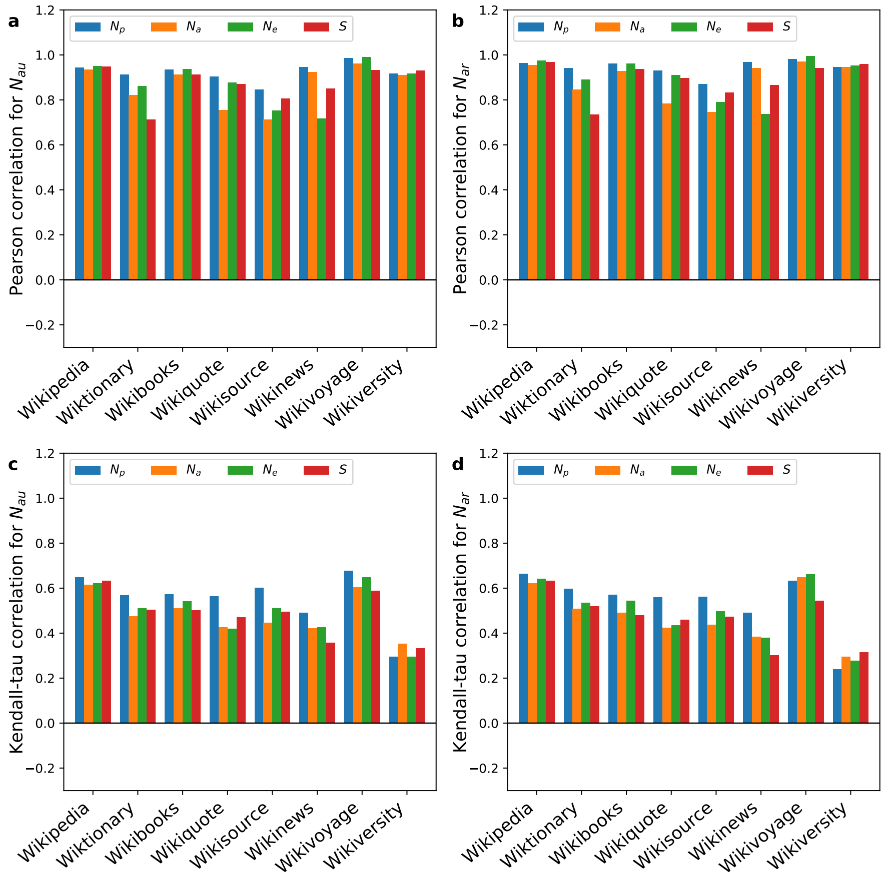
Supplementary Figure 15. The correlations between measures of different Wikimedia projects and scocio-economic indicators of the corresponding country (See Supplementary Methods for details). Population and numbers of Internet users are retrieved from Central Intelligence Agency (CIA) world FACT Book [43], and GDP measures are collected from UNESCO Institute for Statistics [45]. **a,e**, The correlation coefficients on measures of Wikimedia projects for the population of the corresponding country. **b,f**, The correlation coefficients on measures of Wikimedia projects for the number of Internet users of the corresponding country. **c,g**, Correlation coefficients on measures of Wikimedia projects for the GDP at market prices (in constant 2010 US$) of the corresponding country. **d,h**, The correlation coefficients on measures of Wikimedia projects for the GDP per capita, or PPP (in constant 2011 international $) of the corresponding country. For panels (**a–h**), the colours of the bars indicate the number of edits N_e (blue bars), the number of editors N_p (orange bars), the number of articles N_a (green bars), and the total volume of texts S (red bars). Both the Pearson correlation coefficient (**a–d**) and the Kendall-tau rank correlation coefficient (**e–h**) is used to quantify interplays between measures, to exclude possible biases when we only observe a single measure.



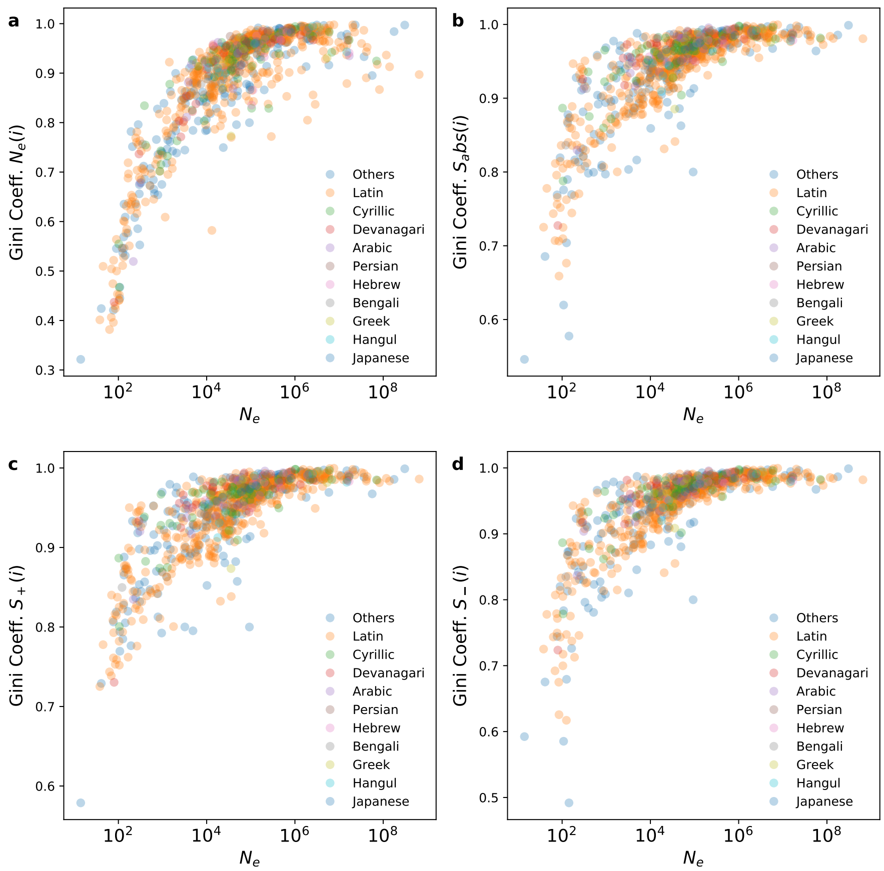
Supplementary Figure 16. The correlations between measures of different Wikimedia projects and gross domestic expenditure on R&D (GERD) of the corresponding country collected from UNESCO Institute for Statistics [45] (See Supplementary Methods for details). **a,d**, The correlation coefficients on measures of Wikimedia projects for the total gross domestic expenditure on R&D of the corresponding country. **b,e**, The correlation coefficients on measures of Wikimedia projects for the gross domestic expenditure on R&D for basic research of the corresponding country. **c,f**, The correlation coefficients on measures of Wikimedia projects for the gross domestic expenditure on R&D for applied research of the corresponding country. For (**a**–**f**) colours of the bar indicate the number of edits N_e (blue bars), the number of editors N_p (orange bars), the number of articles N_a (green bars), and the total volume of texts S (red bars). Both the Pearson correlation coefficient (**a**–**c**) and the Kendall-tau rank correlation coefficient (**d**–**f**) is used to quantify interplays between measures, to exclude possible biases when we only observe a single measure.



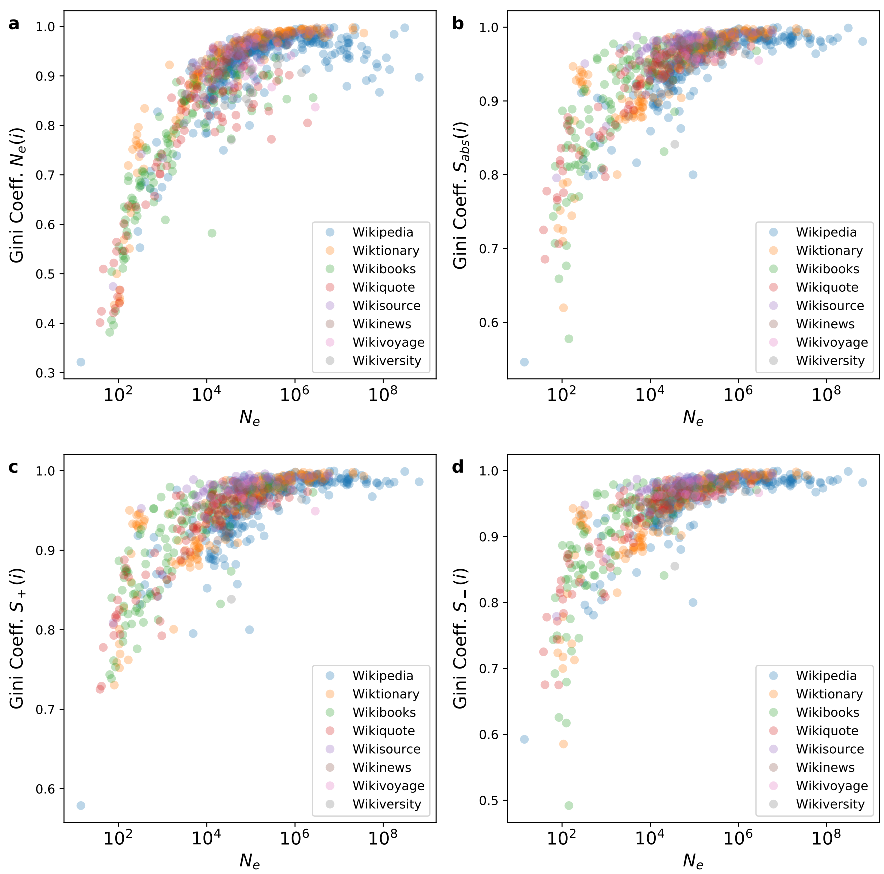
SUPPLEMENTARY FIGURE 17. The correlations between measures of different Wikimedia projects and measures of patent applications (see Supplementary Methods for details). **a,d**, The correlation coefficients on measures of Wikimedia projects for the number of patent applicants N_{ap} **b,e**, The correlation coefficients on measures of Wikimedia projects for number of patents N_{pt}. **c,f**, The correlation coefficients on measures of Wikimedia projects for the number of patent inventors N_{iv}. For panels (**a–f**), the colours of the bars indicate the number of edits N_e (blue bars), the number of editors N_p (orange bars), the number of articles N_a (green bars), and the total volume of texts S (red bars). Both the Pearson correlation coefficient (**a–c**) and the Kendall-tau rank correlation coefficient (**d–f**) is used to quantify interplays between measures, to exclude possible biases when we only observe a single measure.



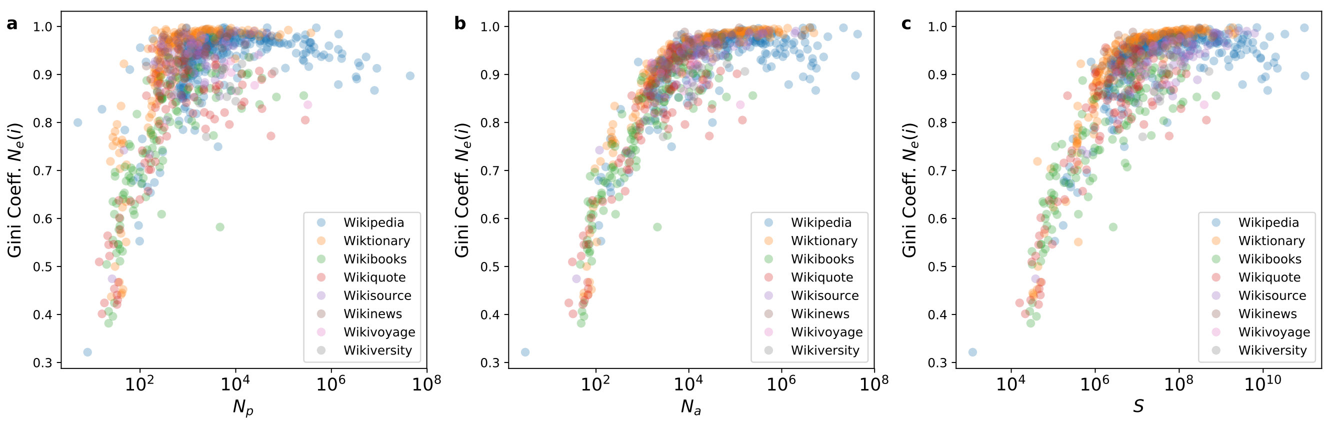
Supplementary Figure 18. The correlations between measures of different Wikimedia projects and measures of research articles (see Supplementary Methods for details). **a,c**, The correlation coefficients on measures of Wikimedia projects for the number of authors N_{au}. **b,d**, The correlation coefficients on measures of Wikimedia projects for number of research articles N_{ar}. For panels (**a–d**), the colours of the bar indicate the number of edits N_e (blue bars), the number of editors N_p (orange bars), the number of articles N_a (green bars), and the total volume of texts S (red bars). Both the Pearson correlation coefficient (**a,b**) and the Kendall-tau rank correlation coefficient (**c,d**) is used to quantify interplays between measures, to exclude possible biases when we only observe a single measure.



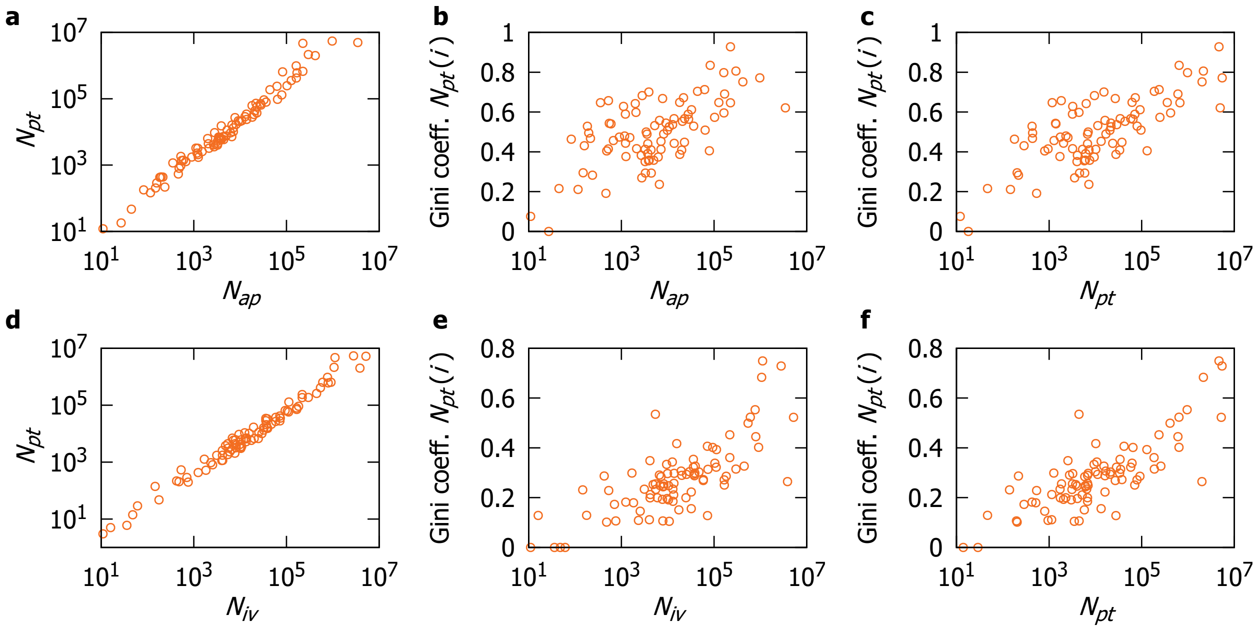
SUPPLEMENTARY FIGURE 19. The Gini coefficient of Wikimedia projects as a function of the number of edits. **a**, The Gini coefficient for numbers of edits performed by each editor. **b**, The average Gini coefficient for total sums of absolute amount of editing (in the unit of bytes) performed by each editor. **c**, The average Gini coefficient for the absolute number of sums of incremental change in edits performed by each editor. **d**, The average Gini coefficient for the absolute number of sums of decremental change in edits performed by each editor. The colours represent the script with which the corresponding Wikimedia projects are written.



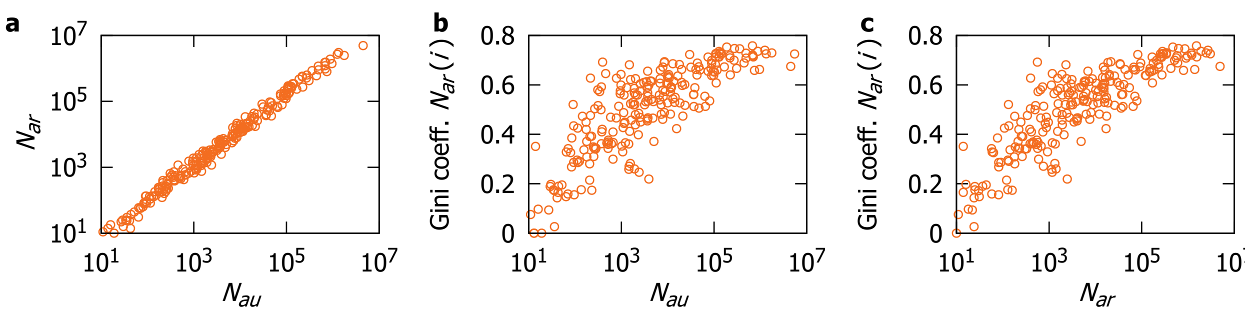
Supplementary Figure 20. The Gini coefficient of Wikimedia projects as a function of the number of edits. **a**, The Gini coefficient for numbers of edits performed by each editor. **b**, The average Gini coefficient for total sums of absolute amount of editing (in the unit of bytes) performed by each editor. **c**, The average Gini coefficient for the absolute number of sums of incremental change in edits performed by each editor. **d**, The average Gini coefficient for the absolute number of sums of decremental change in edits performed by each editor. The colours represent the types of Wikimedia projects.



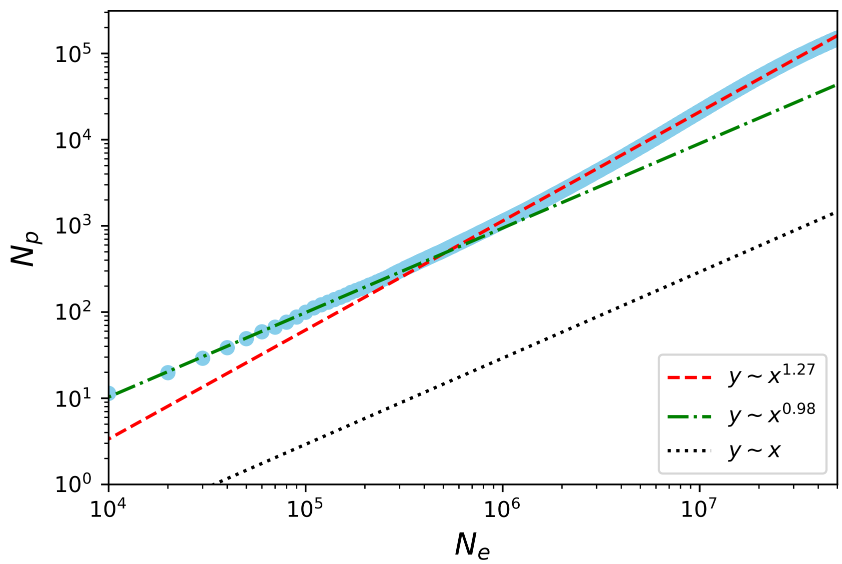
Supplementary Figure 21. The Gini coefficient of Wikimedia projects as functions of the characteristic measures of the projects. **a**, The Gini coefficient for numbers of edits performed by each editor as a function of the number of editors N_p. **b**, The Gini coefficient for numbers of edits performed by each editor as a function of the number of articles N_a. **c**, The Gini coefficient for numbers of edits performed by each editor as a function of the total size of the data set S. The colours represent the types of Wikimedia projects.



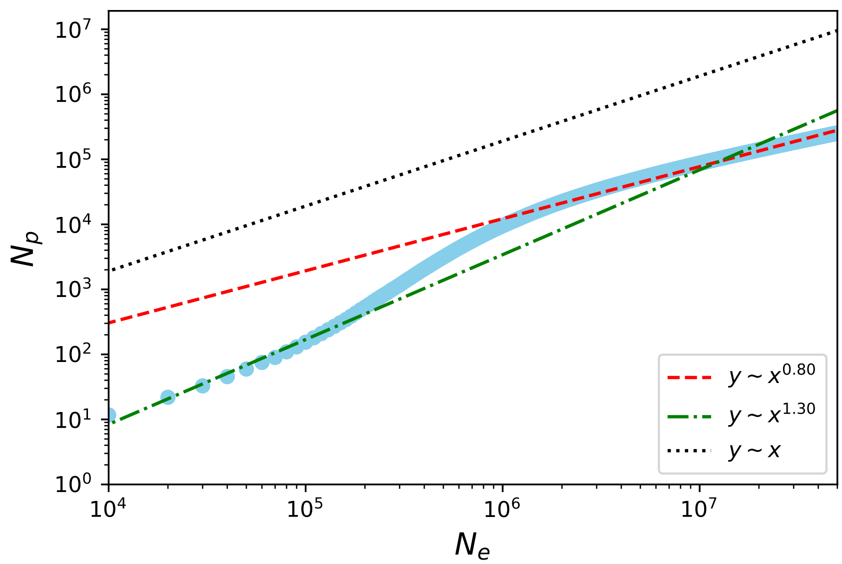
SUPPLEMENTARY FIGURE 22. The correlations and the Gini coefficients of patent applications as functions of number of patents N_{pt} for number of patent applicants N_{ap} (**a**–**c**) and patent inventors N_{iv} (**d**–**f**). **a**, The numbers of patent as functions of the number of patent applicants. **b**, The Gini coefficient for the numbers of patents filed by each applicant, as functions of the number of patent applicants. **c**, The Gini coefficient for numbers of patents filed by each applicant, as functions of the number of patents. **d**, The numbers of patent as functions of the number of patent inventors. **e**, The Gini coefficient for the numbers of patents filed by each inventor, as functions of the number of patent inventors. **c**, The Gini coefficient for numbers of patents filed by each inventor, as functions of the number of patents. For panels (**a**–**f**), each data point corresponds to a patent office that the patent initially filed. For the analysis, we use spring 2017 edition of European Patent Office (EPO) Worldwide Patent Statistical Database (PATSTAT), which contains metadata of patents from 91 national and international patent offices. Inventors and applicants are identified as a person IDs by EPO and we neglected the persons who are not identified by EPO (see Supplementary Methods for details).



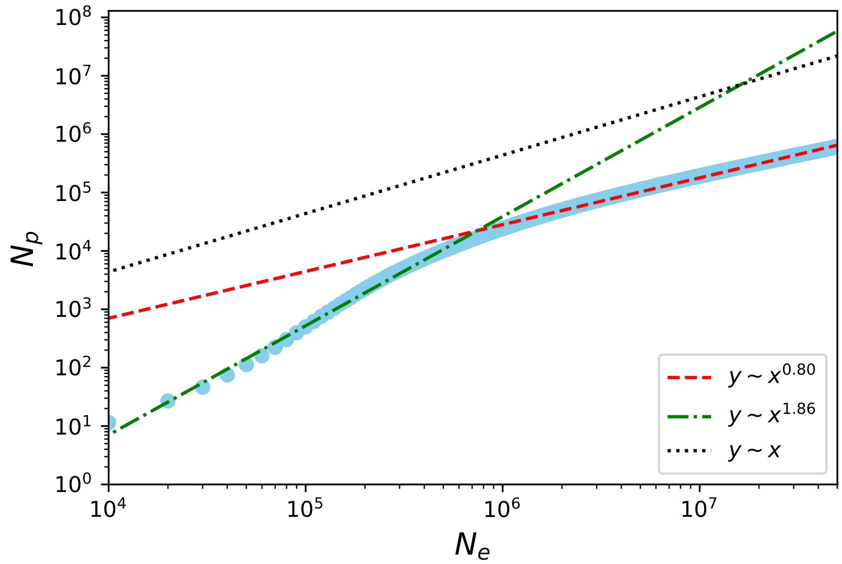
Supplementary Figure 23. The correlations and the Gini coefficients of number of authors as functions of number of research articles N_{ar} for number of authors N_{au}. **a**, Numbers of research articles as functions of the number of authors. **b**, The Gini coefficient for numbers of research articles published by each authors, as functions of the number of authors. **c**, The Gini coefficient for numbers of research articles published by each author, as functions of the number of research articles. Each data point corresponds to a nationality to which the affiliation of the author belongs. For the analysis, we use August 2017 edition of SCOPUS XML CUSTOM DATA, which contains meta data of worldwide research outputs. Authors and their affiliations are identified as an author IDs and affiliation IDs given by SCOPUS and we neglected the authors who are not identified by SCOPUS (see Supplementary Methods for details).



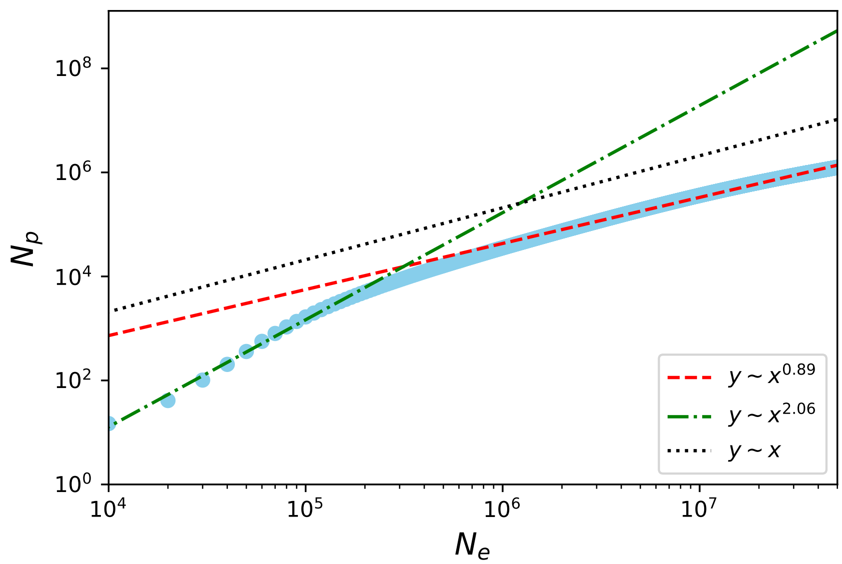
Supplementary Figure 24. The scaling relation between number of edits N_e and number of editors N_p of the model result. The relation is characterized by two simple power-law growth of the form $y \sim x^{\lambda}$. For the initial stage $N < 10^6$, the number of editors increases sublinearly with the exponent $\lambda \simeq 0.98$, whereas the number of editors increases superlinearly with the exponent $\lambda \simeq 1.27$ for later ($N > 10^6$). We use the following parameters with 10 realizations: $b = 0.0001$, $k = 0.4$, $\tau \to \infty$, and $r = 0.01$.



SUPPLEMENTARY FIGURE 25. The scaling relation between number of edits N_e and number of editors N_p of the model result. The relation is characterized by two simple power-law growth of the form $y \sim x^{\lambda}$. For the initial stage $N < 10^6$, the number of editors increases superlinearly with an exponent $\lambda \simeq 1.30$, whereas the number of editors increases sublinearly with an exponent $\lambda \simeq 0.80$ for the later $(N > 10^6)$. We use the following parameters with 10 realizations: $b = 0.0001$, $k = 0.6$, $\tau \to \infty$, and $r = 0.01$.



SUPPLEMENTARY FIGURE 26. The scaling relation between number of edits N_e and number of editors N_p of the model result. The relation is characterized by two simple power-law growth of the form $y \sim x^{\lambda}$. For the initial stage $N < 5 \times 10^5$, the number of editors increases superlinearly with an exponent $\lambda \simeq 1.86$, whereas the number of editors increases sublinearly with an exponent $\lambda \simeq 0.80$ for the later $(N > 5 \times 10^5)$. We use the following parameters with 10 realizations: $b = 0.0001$, $k = 0.7$, $\tau \to \infty$, and $r = 0.01$.



SUPPLEMENTARY FIGURE 27. The scaling relation between number of edits N_e and number of editors N_p of the model result. The relation is characterized by two simple power-law growth of the form $y \sim x^{\lambda}$. For the initial stage $N < 2 \times 10^5$, the number of editors increases superlinearly with an exponent $\lambda \simeq 2.06$, whereas the number of editors increases sublinearly with an exponent $\lambda \simeq 0.89$ for the later ($N > 2 \times 10^5$). We use the following parameters with 10 realizations: $b = 0.0001$, $k = 0.8$, $\tau \to \infty$, and $r = 0.01$.

SUPPLEMENTARY TABLES

SUPPLEMENTARY TABLE 1. The list of Wikimedia projects. In general, there are different language editions for each project.

Project	Editions	Description/Notes
Wikipedia	273	Encyclopedia articles
Wiktionary	172	Dictionary
Wikibooks	121	Educational textbooks and learning materials
Wikiquote	89	Collection of quotations
Wikisource	65	Library of source documents and translations
Wikinews	33	News source
Wikiversity	16	Educational and research materials and activities
Wikivoyage	17	Travel guide
etc.	77	Deactivated (not editable) ones are included
Total	863	

SUPPLEMENTARY TABLE 2. Statistical details about power-law fitting for the N_e–N_p relation. When we write that a P-value is 0, it means that this value is smaller than the minimum (approximately 2.23×10^{-308}) of the floating-point variables in Python.

Project	Power exponent	Proportionality coefficiant	P-value	Standard Error	Pearson ρ	Pearson ρ (log-log)
All	0.704249	0.738648	0	0.0110146	0.895056	0.908861
Wikipedia	0.842443	0.155088	8.82×10^{-138}	0.0192937	0.895667	0.924524
Wiktionary	0.539973	1.830600	4.72×10^{-63}	0.0201075	0.871989	0.899576
Wikibooks	0.800070	0.593431	2.52×10^{-85}	0.0148827	0.993453	0.980026
Wikiquote	0.800656	0.552054	9.37×10^{-49}	0.0258568	0.976789	0.957503
Wikisource	0.629638	1.278470	1.53×10^{-23}	0.0399435	0.926182	0.893163
Wikinews	0.622501	1.116610	2.68×10^{-8}	0.0844640	0.773823	0.797903
Wikivoyage	1.078590	0.017792	9.86×10^{-8}	0.1136140	0.979092	0.925912
Wikiversity	0.971437	0.068704	3.74×10^{-8}	0.0902719	0.989328	0.944534

Supplementary Table 3. Statistical details about power-law fitting for the N_e–N_a relation. When we write that a P-value is 0, it means that this value is smaller than the minimum (approximately 2.23×10^{-308}) of the floating-point variables in Python.

Project	Power exponent	Proportionality coefficiant	P-value	Standard Error	Pearson ρ	Pearson ρ (log-log)
All	0.851328	0.953075	0	0.0063968	0.851105	0.976546
Wikipedia	0.937205	0.256999	4.78×10^{-217}	0.0115751	0.849992	0.976168
Wiktionary	0.886418	0.876804	4.55×10^{-133}	0.0116775	0.971678	0.985567
Wikibooks	0.793356	1.773870	5.16×10^{-104}	0.0101733	0.979341	0.990358
Wikiquote	0.769828	1.939820	1.33×10^{-75}	0.0118208	0.949461	0.989899
Wikisource	0.971340	0.472484	4.70×10^{-51}	0.0204457	0.980636	0.986329
Wikinews	0.780295	1.999770	1.35×10^{-13}	0.0626789	0.789431	0.912861
Wikivoyage	0.809312	0.853837	1.04×10^{-10}	0.0516414	0.982997	0.970794
Wikiversity	0.928145	0.316529	3.56×10^{-9}	0.0717772	0.984704	0.960594

Supplementary Table 4. Statistical details about power-law fitting for the N_e–S relation. When we write that a P-value is 0, it means that this value is smaller than the minimum (approximately 2.23×10^{-308}) of the floating-point variables in Python.

Project	Power exponent	Proportionality coefficiant	P-value	Standard Error	Pearson ρ	Pearson ρ (log-log)
All	0.874691	1060.2	0	0.0102172	0.945518	0.945977
Wikipedia	0.987816	194.4	1.68×10^{-186}	0.0153652	0.946319	0.962968
Wiktionary	0.802352	1652.4	1.59×10^{-103}	0.0160625	0.892158	0.967582
Wikibooks	0.976815	728.3	6.50×10^{-61}	0.0301235	0.966832	0.947805
Wikiquote	0.891003	830.3	1.22×10^{-57}	0.0223718	0.977249	0.973655
Wikisource	0.989188	1610.7	7.95×10^{-34}	0.0405094	0.951270	0.951021
Wikinews	0.814366	1697.1	1.79×10^{-13}	0.0661310	0.955948	0.911193
Wikivoyage	0.846150	1612.5	3.33×10^{-7}	0.0980059	0.956239	0.912402
Wikiversity	0.894525	1310.5	1.21×10^{-8}	0.0760369	0.981032	0.952962

Supplementary Table 5. Statistical details about power-law fitting for the N_a–N_p relation. When we write that a P-value is 0, it means that this value is smaller than the minimum (approximately 2.23×10^{-308}) of the floating-point variables in Python.

Project	Power exponent	Proportionality coefficiant	P-value	Standard Error	Pearson ρ	Pearson ρ (log-log)
All	0.778094	1.199350	4.75×10^{-274}	0.0146429	0.649975	0.875402
Wikipedia	0.855545	0.817538	7.79×10^{-120}	0.0228277	0.651000	0.901426
Wiktionary	0.585252	2.473690	5.63×10^{-56}	0.0246021	0.773866	0.876922
Wikibooks	0.980838	0.403144	3.04×10^{-69}	0.0253558	0.956404	0.962462
Wikiquote	1.012750	0.338911	5.48×10^{-43}	0.0387238	0.878766	0.941891
Wikisource	0.616042	2.848200	4.00×10^{-20}	0.0459310	0.884531	0.860595
Wikinews	0.797727	0.642675	3.09×10^{-11}	0.0796528	0.955328	0.874015
Wikivoyage	1.188170	0.082902	1.53×10^{-5}	0.1898720	0.929965	0.850316
Wikiversity	0.940047	0.609241	5.83×10^{-6}	0.1334550	0.966194	0.883137

Supplementary Table 6. Statistical details about power-law fitting for the N_a–S relation. When we write that a P-value is 0, it means that this value is smaller than the minimum (approximately 2.23×10^{-308}) of the floating-point variables in Python.

Project	Power exponent	Proportionality coefficiant	P-value	Standard Error	Pearson ρ	Pearson ρ (log-log)
All	1.021710	1173.3	0	0.0097043	0.837651	0.963288
Wikipedia	1.050380	844.92	6.06×10^{-241}	0.0108699	0.838299	0.983090
Wiktionary	0.900004	1952.1	1.04×10^{-114}	0.0153493	0.916696	0.976157
Wikibooks	1.231870	358.1	4.13×10^{-66}	0.0340094	0.958276	0.957518
Wikiquote	1.154880	392.9	3.45×10^{-64}	0.0241913	0.938157	0.981443
Wikisource	0.999274	4167.2	1.49×10^{-32}	0.0430899	0.944978	0.946118
Wikinews	0.981187	1486.8	7.42×10^{-16}	0.0648819	0.910617	0.938418
Wikivoyage	1.022660	2346.8	1.83×10^{-7}	0.1130380	0.970173	0.919303
Wikiversity	0.944827	4725.6	2.93×10^{-10}	0.0604193	0.980323	0.972548

Supplementary Table 7. Statistical details about power-law fitting for the N_p–S relation. When we write that a P-value is 0, it means that this value is smaller than the minimum (approximately 2.23×10^{-308}) of the floating-point variables in Python.

Project	Power exponent	Proportionality coefficient	P-value	Standard Error	Pearson ρ	Pearson ρ (log-log)
All	1.059150	5791.2	4.34×10^{-292}	0.0187328	0.726656	0.887589
Wikipedia	1.022530	6194.5	1.11×10^{-124}	0.0261606	0.725107	0.908313
Wiktionary	1.223030	3538.5	2.02×10^{-58}	0.0492685	0.747880	0.885312
Wikibooks	1.160450	1966.0	5.32×10^{-50}	0.0455632	0.965096	0.919231
Wikiquote	1.029180	2743.8	1.57×10^{-42}	0.039893	0.967549	0.940423
Wikisource	1.304610	7218.7	1.68×10^{-22}	0.0868298	0.927855	0.884204
Wikinews	0.968762	16377.6	5.85×10^{-10}	0.1098120	0.896203	0.845663
Wikivoyage	0.593184	190739	5.99×10^{-4}	0.1370960	0.906938	0.745099
Wikiversity	0.809554	38139.5	4.66×10^{-6}	0.1126350	0.978002	0.887005

Supplementary Methods

Description of the Patent Data Set. For our analysis, we use the spring 2017 edition of the European Patent Office (EPO) Worldwide Patent Statistical Database (PATSTAT), which contains the metadata of patents from 91 national and international patent offices. Inventors and applicants are identified by the EPO using their person IDs. We excluded individuals who were not identified by the EPO. The database contains various types of intellectual property, but we only consider patent applications. The timestamps of patent applications are preferentially extracted from the application year field. Before 2000, the United States Patent and Trademark Office (USPTO) only published granted patents. We thus used patent applications after 2000, which was the year USPTO began to publish all patent applications. Furthermore, a certain publication was dated after 2017, which we excluded for data consistency. Please note that the person identification provided by the EPO is automatic; therefore, it is not perfect and may have unspecified errors.

Description of the Research Paper Data Set. For our analysis on paper metadata, we use the dump of the entire SCOPUS CUSTOM XML DATA for 22 August 2017. This custom data contains the complete copy of data from the SCOPUS website from the very beginning, i.e., January 1996 to August 2017, and includes the title, journal, abstract, author information, and citation records in XML format. Each type of document plays different roles for knowledge formation. For example, conference proceedings are a conventional method for presenting new research in the fields of computer science, whereas journal articles are the main method for many other disciplines. Some disciplines in the social sciences also consider books and reports as important archives of knowledge. Therefore, to prevent a possible bias towards specific disciplines, we use the entire metadata regardless of the citation type in SCOPUS.

The timestamps of the publications were preferentially extracted from the `publicationdate` element. This was replaced by the `xocs:sort-year` element only if the `publicationdate` element was missing or broken. If the timestamp of a certain publication was not between 1996 and 2017, we considered the data as erroneous and removed it. In addition, users were identified as `auid` elements, along with its affiliation element of `afid`. We only counted authors with a clear identification, and the nationality of the authors is assumed to be the country to which the authors affiliation belonged at the time of publication.

Matching Languages of Wikimedia Projects and Their Dominant Countries. One may ask how the current status of each Wikimedia project is related to the empirical socio-economic status of each society. In Supplementary Figs. 12–18, we present several examples in response to the above question. An essential prerequisite is the designation of a specific language in the Wikimedia project to its dominant country. We assume that a certain language is mainly spoken in countries wherein people use the language as their official or dominant language. Therefore, we obtained language usage statistics of the countries from the CIA World Factbook [43] and removed the subsidiary languages to retain the official or dominant languages alone. If a country has more than three official languages, we ignored the third and less used languages based on the magnitude of the populations that speak the languages. Subsequently, we collected the share of pageviews per Wikipedia language from a certain country based on the IP address of users [44]. Using these two data sets, we matched each language used in each Wikimedia project with a specific country as follows: first, if a language is used by only one country in the CIA database, we considered the country to be substantially in possession of the Wikimedia projects written in that language. Second, if there is more than one country that uses the language as the primary language, we considered the country with the largest share of Wikipedia page view traffic as substantially occupying the Wikimedia projects written in the language. Note that the *de facto* owner country of a Wikimedia project does not always match the country of origin of the language. For example, the Spanish Wikipedia is mainly used by Mexicans, and English Wikipedia is mainly used by people in the United States. We only use countries with ISO2 and ISO3 as the main country code. In total, 383 Wikimedia projects correspond to a language in the 80 distinct languages in the data set.

Calculating Representative Values of Socio-economic Indicators for Countries. To capture the socio-economic status of different countries, we used various socio-economic indicators provided by the Central Intelligence Agency (CIA) [43] and UNESCO Institute for Statistics (UIS) [45]. The CIA provides basic statistics such as population and the number of Internet users in different countries. UIS provides various statistics collected worldwide for four categories: 1) education and literacy, 2) science, technology, and innovation, 3) culture, and 4) communication and information. Some survey data are collected irregularly by different organisations; it is thus futile to choose a specific reference year to use. To circumvent this issue, we extracted all the existing data from 2000 to 2016, and averaged them over the entire duration for a certain country and its index value. Note that the economic indicators such as Gross Domestic Product (GDP) and Gross Domestic Expenditure on R&D (GERD) are usually collected yearly, whereas survey data such as attainment rate of school are collected irregularly.

Feature Vector Construction for Clustering of Languages in Wikimedia projects. To explore the relations between different languages in Wikimedia projects, we constructed an n-dimensional vector for each language. Each vector element represents a characteristic measure for a specific type of Wikimedia project based on the number of its language editions (see Table 1). Specifically, it follows the order 1) Wikipedia, 2) Wiktionary, 3) Wikibooks, 4) Wikiquote, and 5) Wikisource. As an illustrative example, the 3-vector representation of N_e for English Wikimedia Projects is $[N_e(\text{English Wikipedia}), N_e(\text{English Wiktionary}), N_e(\text{English Wikibooks})] = (654\,163\,757, 36\,453\,984, 2\,572\,276)$.

References

[42] Paolillo, J. C. and Anupam. D. Evaluating language statistics: The Ethnologue and beyond, *Contract report for UNESCO Institute for Statistics* (2006).

[43] CIA. *CIA World Factbook 2016-2017*. Available at: `https://www.cia.gov/library/publications/the-world-factbook/`.

[44] *Wikimedia Traffic Analysis Report - Page Views Per Wikipedia Language - Breakdown, Monthly requests or daily averages, for period: 1 Sep 2017 - 30 Sep 2017*, Available at: `https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm`

[45] *UNESCO Institute for Statistics*, Available at: `http://uis.unesco.org/`.