Wikipedia: A quantitative analysis

Thesis · April 2009

1 author:

Felipe Ortega
King Juan Carlos University
35 PUBLICATIONS **298** CITATIONS

SEE PROFILE

Doctoral thesis summary.
Wikipedia: A quantitative analysis

Felipe Ortega (jfelipe@libresoft.es)
GSyC/Libresoft, Universidad Rey Juan Carlos, Madrid (Spain)

July 29, 2009

Abstract

In this doctoral thesis, we undertake a quantitative analysis of the top-ten language editions of Wikipedia, from different perspectives. Our main goal has been to trace the evolution in time of key descriptive and organizational parameters of Wikipedia and its community of authors. The analysis has focused on logged authors (those editors who created a personal account to participate in the project). Among the distinct metrics included, we can find the monthly evolution of general metrics (number of revisions, active editors, active pages); the distribution of pages and its length, the evolution of participation in discussion pages. We also present a detailed analysis of the inner social structure and stratification of the Wikipedia community of logged authors, fitting appropriate distributions to the most relevant metrics. We also examine the inequality level of contributions from logged authors, showing that there exists a core of very active authors who undertake most of the editorial work. Regarding articles, the inequality analysis also shows that there exists a reduced group of popular articles, though the distribution of revisions is not as skewed as in the previous case. The analysis continues with an in-depth demographic study of the community of authors, focusing on the evolution of the core of very active contributors (applying a statistical technique known as survival analysis). We also explore some basic metrics to analyze the quality of Wikipedia articles and the trustworthiness level of individual authors. This work concludes with an extended analysis of the evolution of the most influential parameters and metrics previously presented.

Based on these metrics, we infer important conclusions about the future sustainability of Wikipedia. According to these results, the Wikipedia community of authors has ceased to grow, remaining stable since Summer 2006 until the end of 2007. As a result, the monthly number of revisions has remained stable over the same period, restricting the number of articles that can be reviewed by the community. On the other side, whilst the number of revisions in talk pages has stabilized over the same period, as well, the number of active talk pages follows a steady growing rate, for all versions. This suggests that the community of authors is shifting its focus to broaden the coverage of discussion pages, which has a direct impact in the final quality of content, as previous research works has shown.

Regarding the inner social structure of the Wikipedia community of logged authors, we find Pareto-like distributions that fit all relevant metrics pertaining authors (number of revisions per author, number of different articles edited per author), while measurements on articles (number of revisions per article, number of different authors per article) follow lognormal shapes. The analysis of the inequality level of revisions performed by authors, and revisions received by articles shows highly unequal distributions. The results of our survival analysis on Wikipedia authors presents very high mortality percentages on young authors, revealing an endemic problem of Wikipedias to keep young editors on collaborating with the project for a long period of time. In the same way, from our survival analysis we obtain that the mean lifetime of Wikipedia authors in the core (until they abandon the group of top editors) is situated between 200 and 400 days, for

all versions, while the median value is lower than 120 days in all cases. Moreover the analysis of the monthly number of births and deaths in the community of logged authors reveals that the cause of the shift in the monthly trend of active authors is produced by a higher number of deaths from Summer 2006 in all versions, surpassing the monthly number of births from then on.

The analysis of the inequality level of contributions over time, and the evolution of additional key features identified in this thesis, reveals a worrying trend towards progressive increase of the effort spent by core authors, as time elapses. This trend may eventually cause that these authors will reach their upper limit in the number of revisions they can perform each month, thus starting a decreasing trend in the number of monthly revisions, and an overall recession of the content creation and reviewing process in Wikipedia. To prevent this probable future scenario, the number of monthly new editors should be improved again, perhaps through the adoption of specific policies and campaigns for attracting new editors to Wikipedia, and recover older top-contributors again.

Finally, another important contribution for the research community is WikiXRay, the software tool we have developed to perform the statistical analyses included in this thesis. This tool completely automates the process of retrieving the database dumps from the Wikimedia public repositories, process them to obtain key metrics and descriptive parameters, and load them in a local database, ready to be used in empirical analyses.

As far as we know, this is the first research work implementing a comparative analysis, from an quantitative point of view, of the top-ten language editions of Wikipedia, presenting results from many different scientific perspectives. Therefore, we expect that this contribution will help the scientific community to enhance their understanding of the rich, complex and fascinating working mechanisms and behavioral patterns of the Wikipedia project and its community of authors. Likewise, we hope that WikiXRay will facilitate the hard task of developing empirical analyses on any language version of the encyclopedia, boosting in this way the number of comparative studies like this one in many other scientific disciplines.

1 Highlights

1.1 Research questions tackled in this thesis

In this thesis work, we undertake a challenging objective: to build a quantitative, statistical model to explain the key factors affecting the evolution of the top-ten Wikipedias over the past years. We will concentrate on the content creation process in Wikipedia. Due to this, we will not consider in this research study any aspect concerning Wikipedia readers, that is, users that visit the website to consult information, but who do not contribute to the encyclopedia contents. Defining more concrete tasks, we want to answer the following research questions:

1. **How does the community of authors in the top-ten Wikipedias evolve over time?**: The size of the community and the large number of revisions performed on articles and other wiki pages, makes it difficult to analyze the whole history of changes registered in the database dump files. Our purpose will be to study the evolution in time of the number of contributions and number of active authors per month, searching for distinctive trends that we may find in these graphs. These are basic metrics, describing the activity level maintained by the community over time. We will also take into account the possible influence of anonymous authors and automatic programs that make contributions on articles, as well (the so-called *bots*).

2. **What is the distribution of content and pages in the top-ten Wikipedias?**: Different language versions may concentrate their collaboration efforts in distinct types of pages or content. Analyzing the percentage of each type of page (articles, redirects, user pages, discussion

pages...) produced in the top-ten Wikipedia will provide valuable insights about the different approaches selected by every community to develop their work. We will also obtain information about the importance of key organizational aspects for the community (like discussions and creation of user pages), as well as content categorization (category pages) and extension (through the definition of redirects). The analysis of the length of Wikipedia articles, and its evolution over time in different language versions will also reveal interesting features of the content creation process supported by each community under study.

3. **How does the coordination among authors in the top-ten Wikipedias evolve over time?**: The participation of Wikipedia authors in discussion pages associated to each article is critical to improve the quality of contents. At the same time, it is the natural forum to ensure the application of some important editing policies imposed by the Wikipedia community (we will describe them later on, in this chapter). The analysis of the evolution in time of active authors participating in talk pages, the evolution of the monthly number of active authors participating in discussions, and the evolution of the length of talk pages will contribute to complete the analysis of the inner behavioral patterns found in the Wikipedia community of authors.

4. **Which are the key parameters defining the social structure and stratification of Wikipedia authors?**: To address the specific problem of describing in detail the distribution of effort among community members, we develop an in-depth analysis of the distribution of revisions among authors and the number of different authors contributing to each article. We also examine the same picture from a different perspective, studying the sharing of revisions and authors among articles in each language version. Finally, we will apply several well-known metrics to study the inequality level of the distribution of revisions among authors and articles, thus characterizing the stratification of each community according to the effort spent by every member in the project.

5. **What is the average lifetime of Wikipedia volunteer authors in the project?**: One important aspect regarding the planning and sustainability of any collaborative project resides in identifying the average participation lifetime of individual volunteers in the community. If the project receives more new contributors than it loses, then we have a growing community that may confront more and more complex endeavors as time goes by. On the contrary, if the project is losing more members than it can attracts, then it may impose negative conditions for the future sustainability of the project in due course.

6. **Can we identify basic quantitative metrics to describe the reputation of Wikipedia authors and the quality of Wikipedia articles?**: Though analyzing the quality of Wikipedia content and the level of trustworthiness of each individual author is a quite complex task, we want to identify basic metrics that reveal common traits shared among all high quality articles in Wikipedia. We will built our measurements on the reviewing work performed by many community members, who has selected those articles deserving to be highlighted among all the rest, due to their high quality level. Previous metrics proposed by Stein and Hess [Stein and Hess, 2007] will be tested to check whether they can be applied to measure the quality of articles and the authors' reputation level, complementing other ongoing initiatives in the same research line [Adler and de Alfaro, 2007].

7. **Is it possible to infer, based on previous history data, any sustainability conditions affecting the top-ten Wikipedias in due course?**: To conclude this thesis, we will examine the evolution in time of some of the key parameters and metrics identified along the previous sections.

> The main objective of this analysis will be to infer relevant implications for the sustainability of Wikipedia in the future, specially regarding the number of authors needed to support its impressive growing rate and the broad range of terms and contents covered by the project.

As we can see, following a detailed quantitative analysis methodology we will study Wikipedia from many different points of view, such as the contribution level of Wikipedia authors, the frequency of these contributions, for how long they have been contributing so far and whether we can predict, or not, if a certain group of users will maintain, increase or cease their current workload in the project. We will also pay special attention to the evolution of Wikipedia contents over time, focusing on content authoring aspects that will let us know, for instance: how editors contributions are shared among the project contents; if there exist a high territoriality level in the work of Wikipedia editors (in the sense that they concentrate their contribution efforts in a reduced number of Wikipedia articles); and also, if we can identify common quantitative parameters shared among Wikipedia editors producing high quality contents.

To the best of our knowledge, this is the first comparative analysis of several Wikipedia communities of authors, and thus not focusing on specific language editions or individual communities of contributors. As a result, our model will be the first one to be applied to understand and compare some of the largest (if not the largest) communities of volunteers in the Internet, involved in an open content creation process.

2 Featured results

Table 1 presents some results illustrating the large amount of activity metadata that has been processed in this thesis. For instance, the English language version of Wikipedia comprises more than 11 million pages considering all namespaces (articles, talk pages, user pages, etc.). In addition, until December 2007 it has received contributions from more than 2 million registered users. The amount of talk pages in the top three language versions is also notable.

Lang.	Running time (days)	Tot.#pages	Pages in `main`	#Articles	#Redirects	#Talk pages
EN	2,546	11,405,052	4,623,811	2,183,496	2,440,315	1,764,252
DE	2,485	1,913,294	1,201,409	700,032	501,377	219,520
FR	2,470	2,374,156	1,398,441	629,927	768,514	366,512
PL	2,341	811,672	604,683	475,428	129,255	40,061
JA	2,008	1,161,559	747,834	476,457	271,377	92,712
NL	2,393	936,428	599,457	412,994	186,463	48,898
IT	2,365	1,171,826	593,582	416,694	176,888	83,707
PT	2,459	1,379,940	752,676	363,552	389,124	84,174
ES	2,478	969,864	568,779	338,792	229,987	73,562
SV	2,450	613,852	425,055	273,968	151,087	41,701

Table 1: Some general descriptive metrics about the top-ten Wikipedias. These results illustrate the huge size of the data repository that we have processed in this thesis, forcing us to search for performance-wise analyzing strategies to obtain the corresponding results in reasonable time.

Regarding the editorial activity, Table 2 shows that the English Wikipedia has received more than 167 million edits since its inception, back in 2001, until December 2007. This is a fairly impressive achievement, considering that the second largest version, the German Wikipedia, did not reached 1/4 of that total number of edits, yet. It is also remarkable that, despite the number of bots in some versions like Polish or Portuguese may look relatively low, further results in this thesis clearly show that these versions get a significant number of contributions from these automated programs. This is quite an important point for some communities of editors, like the Spanish Wikipedia, that directly try to minimize the amount of content retrieved or generated by automated means.

Lang.	#Logged authors	#Revisions	#Bots
EN	1,824,439	167,464,014	388
DE	226,912	37,367,801	169
FR	127,767	25,821,354	151
PL	51,796	10,465,003	100
JA	90,828	17,524,766	57
NL	60,749	10,691,679	103
IT	62,690	12,798,068	158
PT	64,994	8,904,662	69
ES	132,239	14,198,257	122
SV	26,972	5,583,020	93

Table 2: Additional overall descriptive parameters for the top-ten Wikipedias: total number of logged authors, number of revisions and number of bots.

The first important conclusion that we must highlight from this thesis is that the total number of revisions performed in Wikipedia articles by logged editors has broken its previous growing rate to enter in a steady-state phase. Figure 1 shows that the growing trend finished towards the end of 2006. From 2007 onwards, all top-ten language versions have registered a stable number of contributions. The main question to make at this point is: what caused this change in the trend of the total number of edits? The answer is shown in Figure 2, depicting the number of active editors that made at least one contribution in each month to a certain language version. From this graph, we can conclude that the number of editors performing revisions on Wikipedia articles has stabilized from 2007, as well. The same occurred with the number of active articles per month (those that registered at least one change in a certain month). These results show that the editorial effort of Wikipedia is not growing any more, because the group of active editors per month is not growing, either.

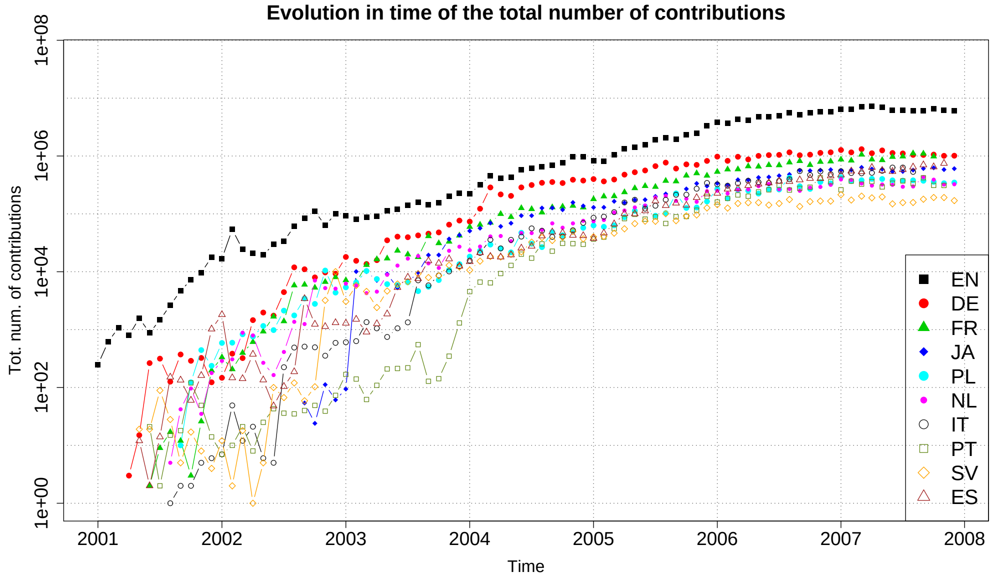

Figure 1: Evolution of the total number of revisions performed in all pages of the top-ten Wikipedias by logged authors. The vertical axis follows a logarithmic scale. The graph clearly shows that the number of contributions received from logged authors has stabilized over the last year, breaking the constant growing rate exhibited by all language editions since their creation.

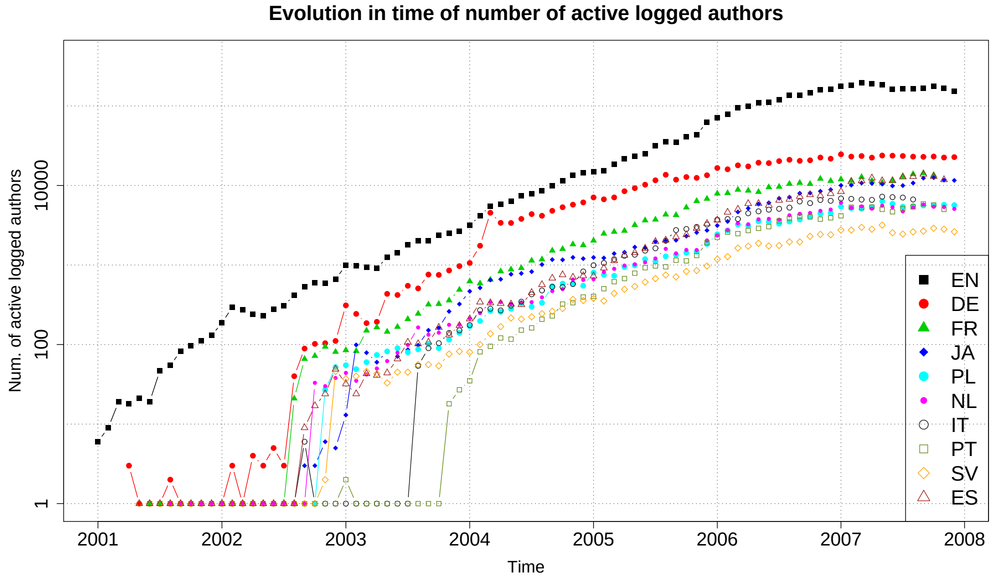

Figure 2: Evolution of the total number of active logged authors per month in the top-ten Wikipedias. The graphic exhibits the same leveraging effect already identified for the number of contributions over the last year, offering a possible cause for this effect

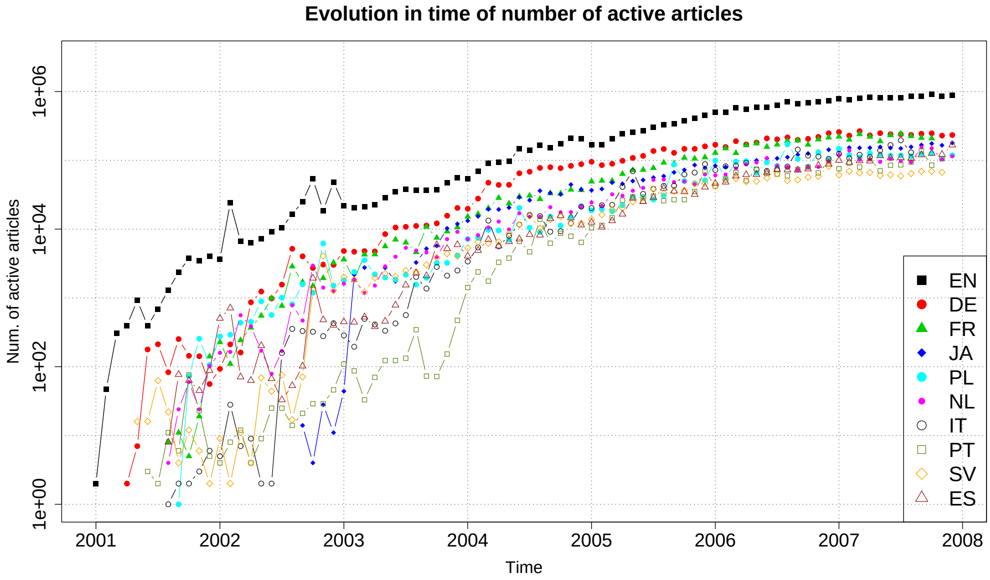

Figure 3: Evolution of active articles per month in the top-ten language versions of Wikipedia. We can appreciate the same leverage effect in 2007 already identified for the contributions from logged authors, thus demonstrating the influence of the leverage in the number of monthly active logged authors in this statistic, as well

Likewise, the number of monthly contributions to talk pages has remained stable from 2007, as we can see in Figure 4. On the contrary, the number of active talk pages per month has remain its growing slope, perhaps showing a shift in the trend of editorial patters of the Wikipedia community (from articles content to discussion about those contents and how to improve them) [1].

Figure 6 shows the evolution in time of the KDE of the length in bytes of Wikipedia articles, for all language versions under study. We clearly see that the shape of the density function is smoother as time goes by, with a trend towards increasing the amount of content included in articles (as we might have expected beforehand). Additionally, Figure 7 shows the distribution of the number of pages in each namespace for every language version. In this graph, we can also identify interesting patterns, like the very high proportion of user talk pages in some language versions (Portuguese, English, Italian and Spanish). On the opposite side, the disproportionate number of articles in the Polish Wikipedia, compared with other type of pages (in particular user and user talk pages), clearly shows the extremely high influence of bots in the content creation process of that version.

As for the social structure exhibited by the Wikipedia community of editors, Figure 8 shows that the number of different articles revised per author follows an upper truncated Pareto distribution. That is, the distribution is log-linear (following a power law shape) until the editors reach an upper limit in the maximum number of different articles that they are able to contribute to. This is also shown in the linear domain, in Figure 9. Finally, the number of authors that share the same number of different articles edited does follow a perfect power law shape (scale-free distribution).

[1] Recently, feedback from some community members in different language versions, most notably from editors in French Wikipedia, may partially rebate this conclusion. Apparently, around the end of 2007 several bots began to automatically open new talk pages for many Wikipedia articles. Though bots contributions have been explicitly filtered out of these results, chances are that some of those bots were not adequately identified in the activity metadata. Nevertheless, the growing rates clearly show up in all language versions, and thus, it seems that the bots issue may have a low influence in this effect

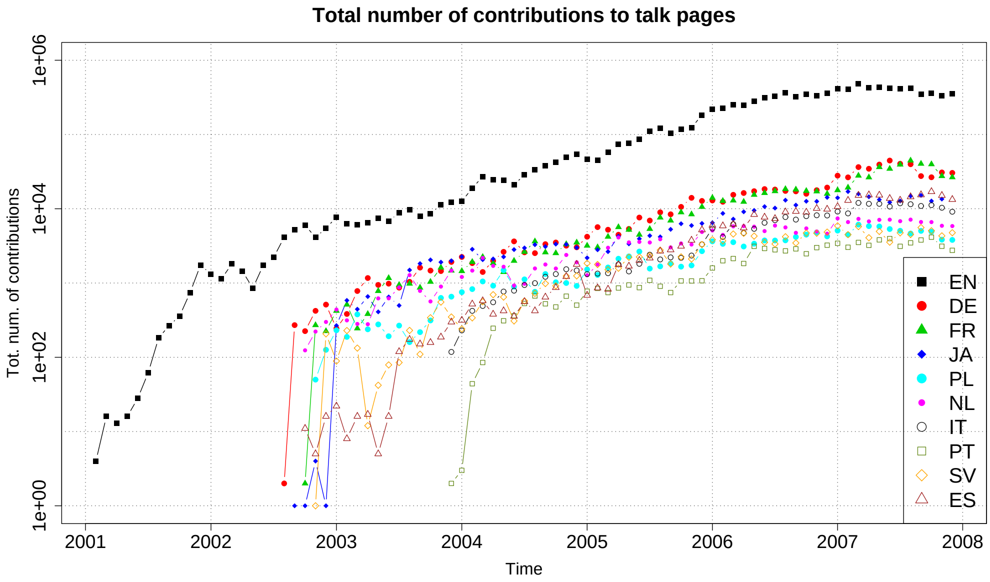

Figure 4: Evolution of the total number of revisions performed in talk pages of the top-ten Wikipedias by logged authors. The number of contributions to discussion pages has also suffered the same leveraging effect, as the number of contributions in articles becomes approximately constant over the last year in all language versions.

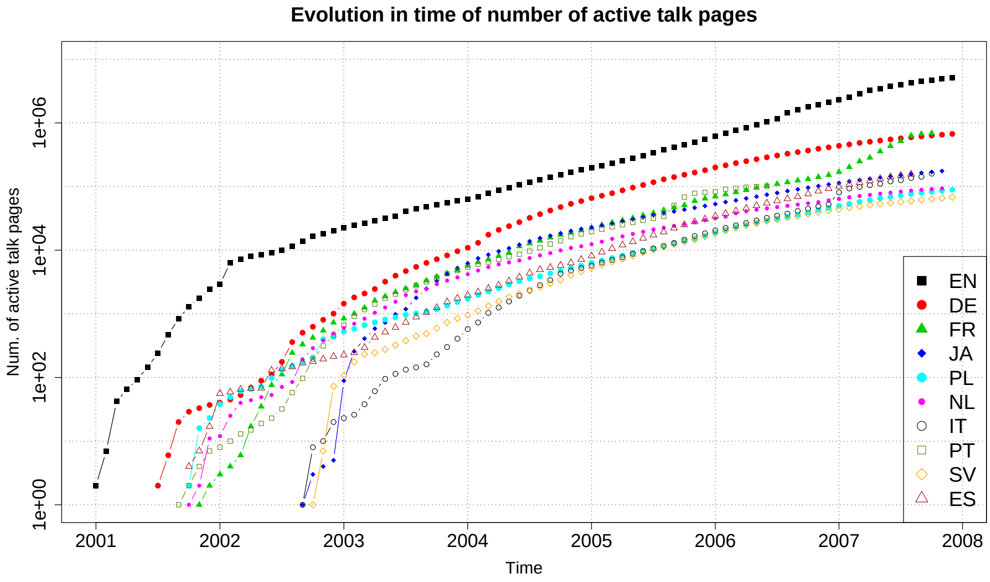

Figure 5: Evolution of active talk pages per month in the top-ten language versions of Wikipedia

The analysis of the inequality of the distribution of edits among authors also reveals interesting patterns in the Wikipedia community. The main finding here is that this distribution is highly unequal. The Gini coefficient (as well as other well-known inequality metrics) shows that less than 10% of the total number of logged editors in each language version is performing more than 90% of the total number of edits. Therefore, like in other collaborative communities in the Internet (for instance, Free, Libre, Open Source Software projects) in Wikipedia we can identify a core of very active editors in each version.

Language	**Gini**	**RS**	**Atkinson**	**Theil**	**Kolm**
EN	0.9306	0.8258	0.8077	3.5824	44.8299
DE	0.9394	0.8358	0.8188	3.3242	88.28795
FR	0.9479	0.8515	0.8391	3.4836	98.6112
PL	0.9468	0.8508	0.8355	3.3698	96.8605
JA	0.92571	0.8096	0.7851	2.9211	82.76989
NL	0.9562	0.8714	0.8628	3.8677	83.9347
IT	0.9418	0.8401	0.8236	3.3061	91.9483
PT	0.9328	0.8265	0.8130	3.6488	51.37997
ES	0.9331	0.8268	0.8086	3.4094	53.9749
SV	0.9515	0.8605	0.8477	3.5374	103.2733

Table 3: Gini coefficient and alternative inequality metrics found in the distribution of total number of revisions per logged author in the top ten Wikipedias

Another remarkable discovery of this thesis is the worrying trend in the number of editors that abandoned the project each month. From the beginning of 2007, the number of deaths (editors that left a certain version to never come back again) exceeds the number of births (editors that joined the project for the first time). In this scenario, we have a community of editors that gets smaller as time goes by. Thus, the editorial force of Wikipedia has less members, and this could be the primary course behind the stabilization of the number of editions and number of active editors per month in all versions under study. The survivability analysis performed on the community of logged authors reveals interesting insights about this situation. In particular, more than 50% of the total number of logged editors remain active in the project for half a year, approximately. Therefore, there exists a high mortality level among the group of young editors in Wikipedia. This fact questions the capacity of the project to retain these young editors, so that they are able to evolve into experienced Wikipedia community members.

This result also have important implications for the maintainability and improvement of quality content in Wikipedia. For example, in this thesis we have found that Featured Articles in Wikipedia usually need at least 1,000 days to reach their high quality level. These articles are usually edited by high experienced editors (with at least 2.5 years contributing to the project). This concordes with the intuitive notion that those editors with a deeper knowledge of Wikipedia guidelines and work flows have a high participation level in high quality content.

Finally, Figure 13 shows the evolution in time of the monthly Gini coefficient in all language version. An interesting pattern is shown up in this graph, namely the natural stabilization of the inequality coefficient in the monthly number of contributions from logged editors. For all language versions, since 2004 the monthly Gini coefficient is constrained within a very small range (80-85%). This means that the core of very active authors is making a higher number of contributions over time, since the inequality measure remains the same but the population of Wikipedia editors is decreasing.

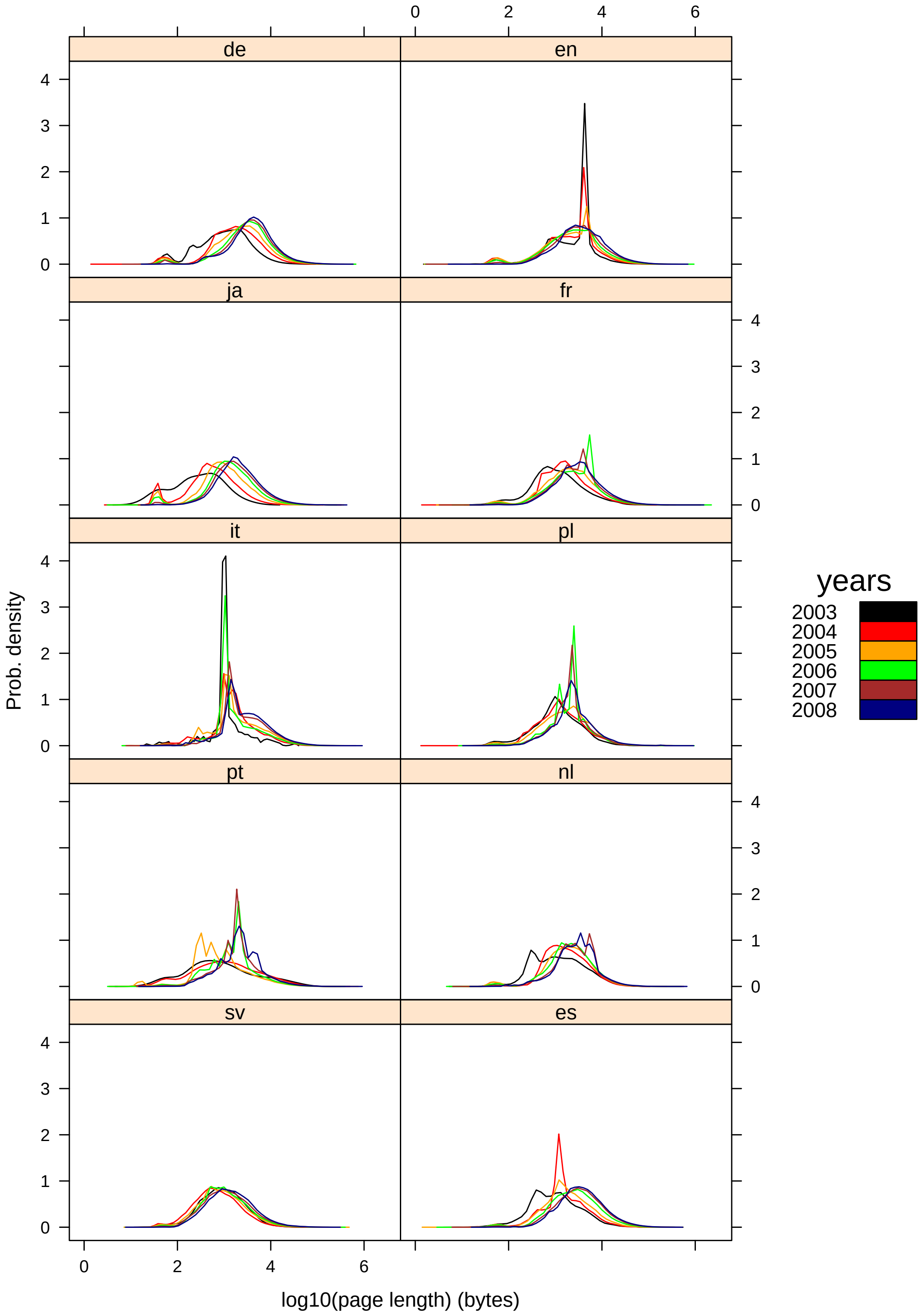

Figure 6: Evolution of KDE of the log10 of length in bytes of articles in the top-ten Wikipedias. Globally, we can see a clear pattern of the median of the length increasing as time goes by in most language versions, with the notorious exception of the English Wikipedia. Therefore, articles tend to become longer over time, as they receive a higher number of revisions

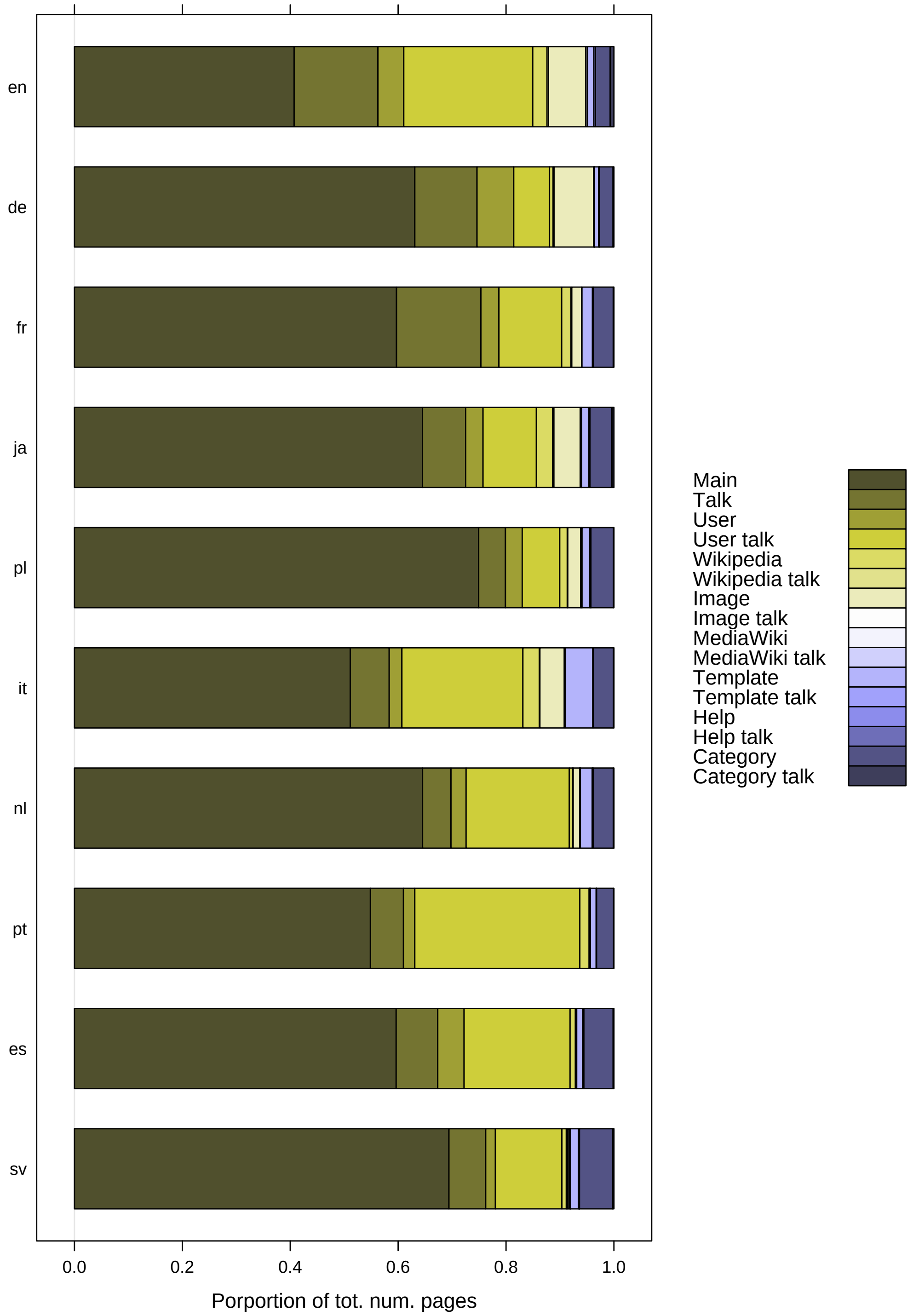

Figure 7: Proportion of total number of pages falling in every namespace for the top ten Wikipedias. In some language editions, the proportion is strongly biased towards the `main` namespace (which stores articles), while other versions present a strong bias towards discussion pages, like talk pages (in the case of the French Wikipedia) or the Portuguese and the English Wikipedias (with a significant proportion of `user_talk` pages. As well, we remark the significant proportion devoted to category pages in the smallest versions (Spanish and Swedish)

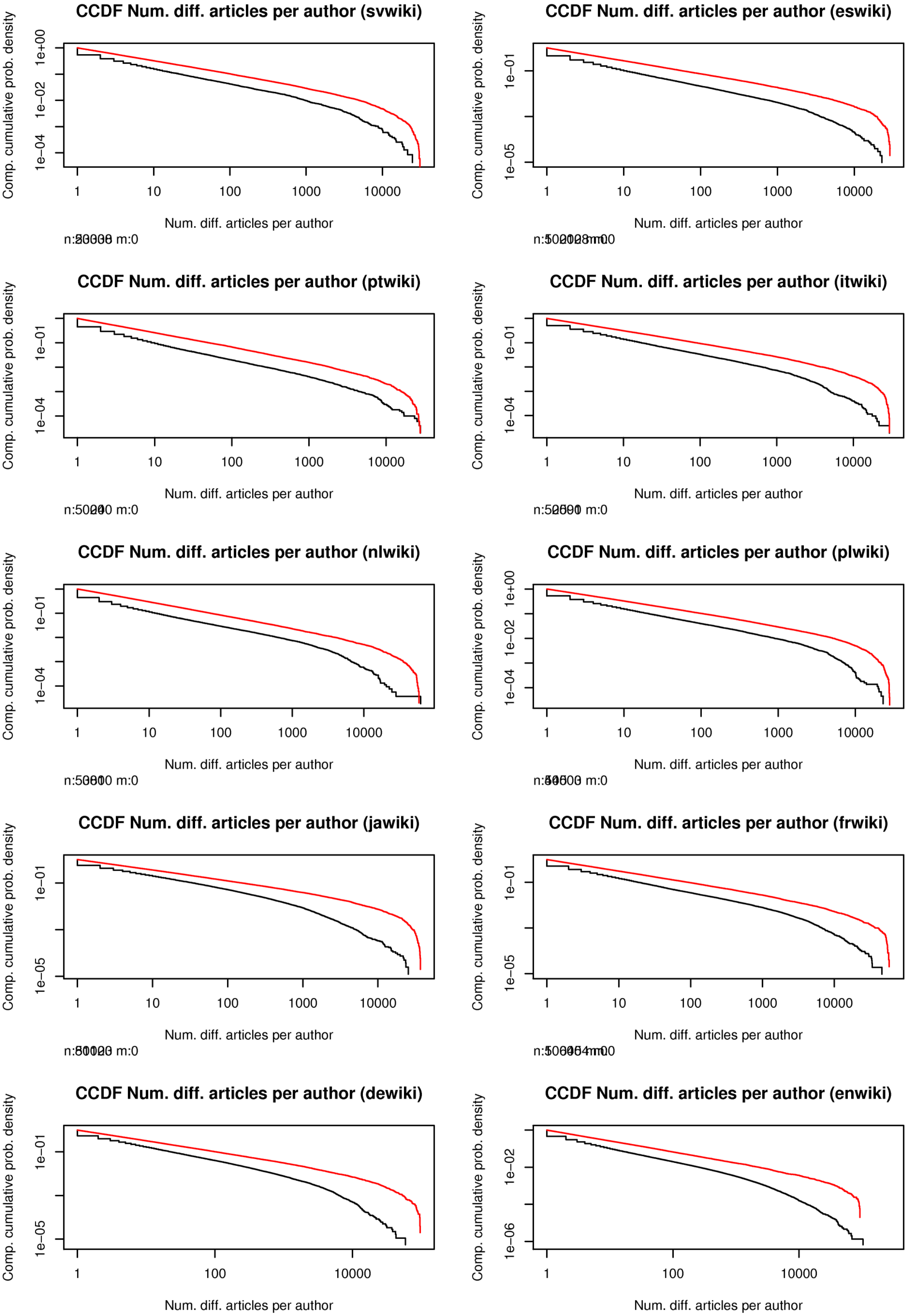

Figure 8: CCDF of number of different articles revised per author. All language versions seem to follow an upper truncated Pareto distribution. There exists a natural higher limit established by the maximum number of different articles that can be revised by a human author

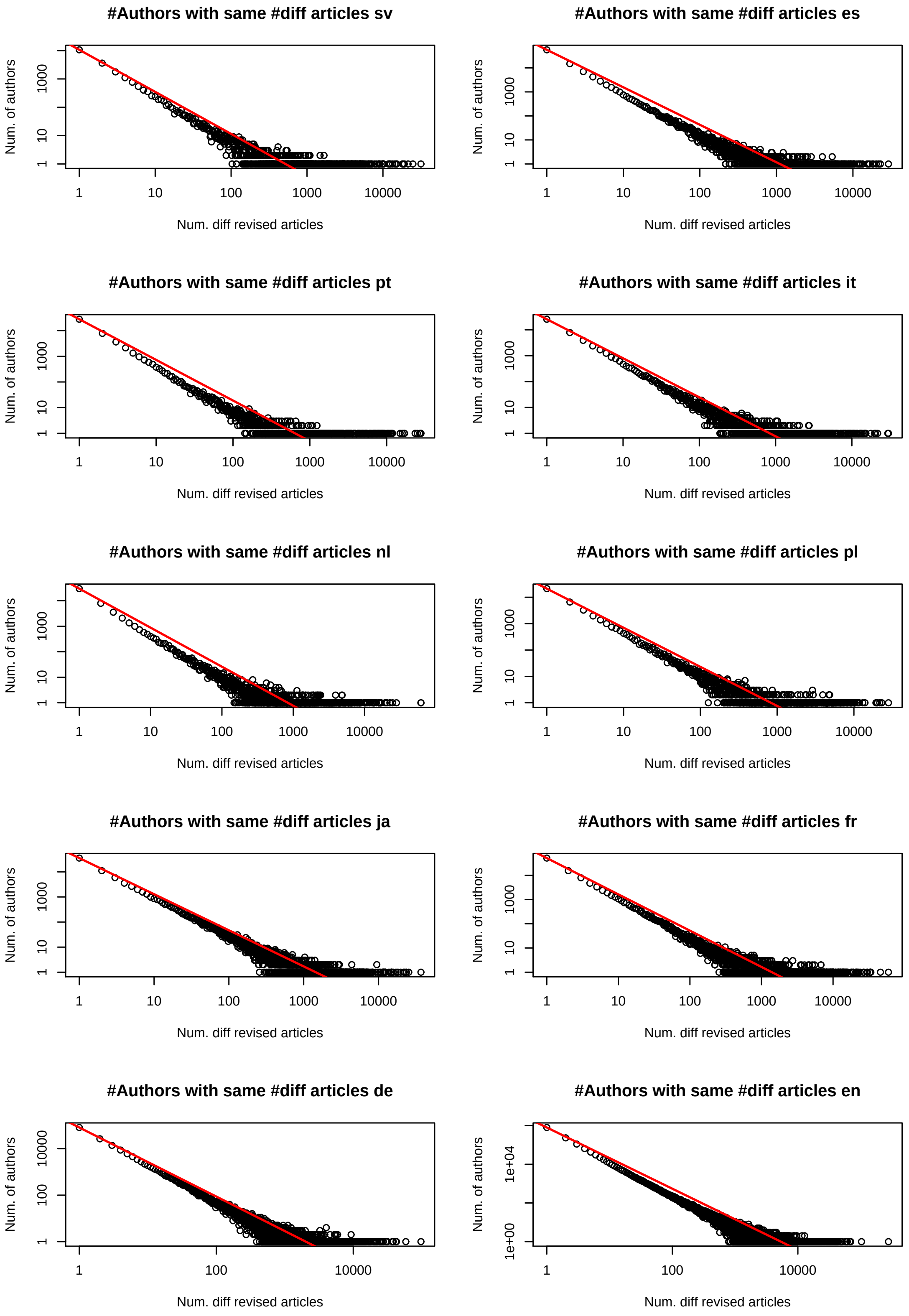

Figure 9: Scatterplot showing the number of users sharing the same number of different articles revised per author, in the top-ten Wikipedias. We also draw the best fit line, which follows a power-law in all language versions, for comparative purposes

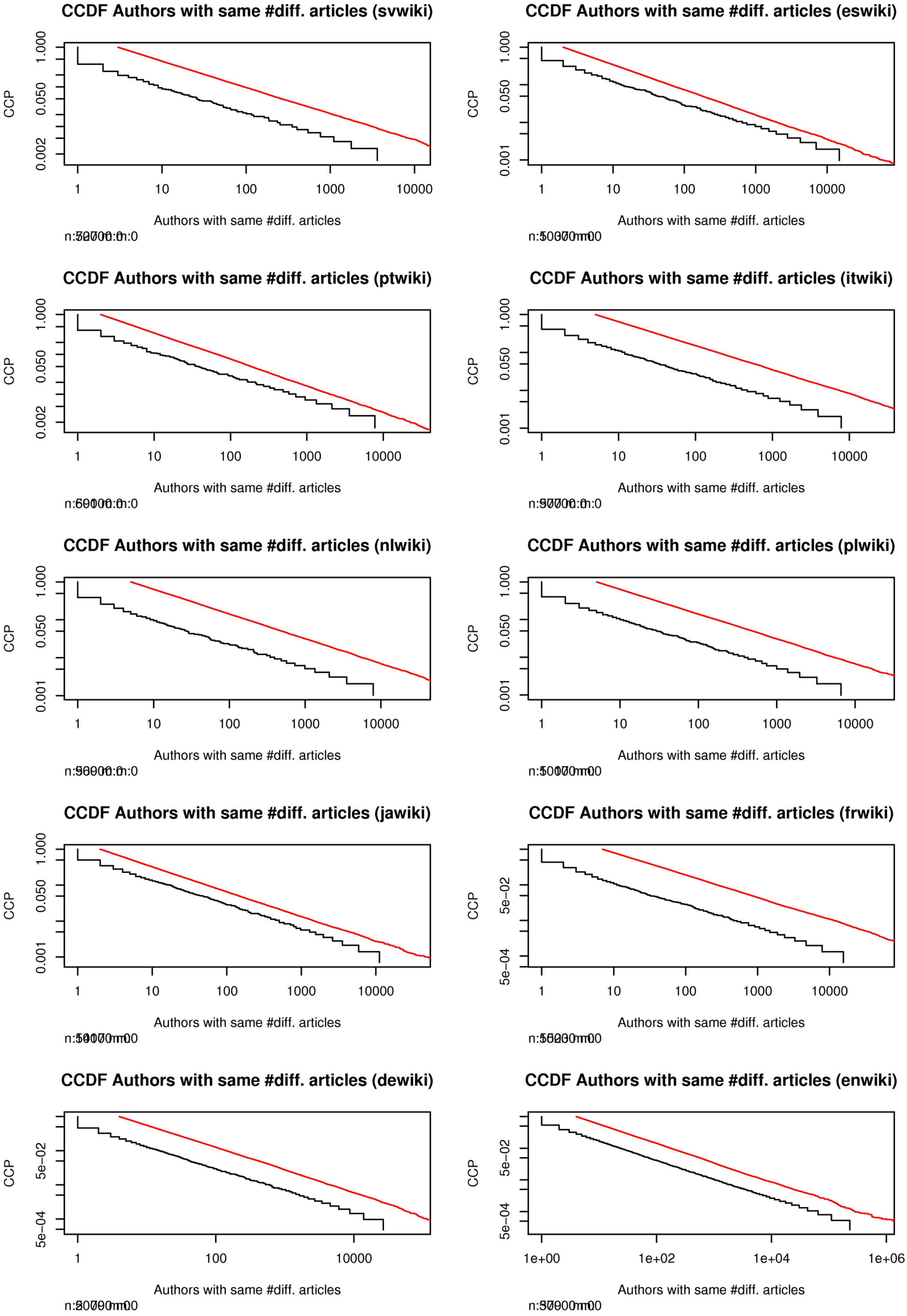

Figure 10: CCDF of number of authors sharing the same number of different articles revised in each language version. As we can see, in all versions the distribution perfectly follows a Pareto law. The best fitted Pareto line is also drawn in each graph for comparative purposes

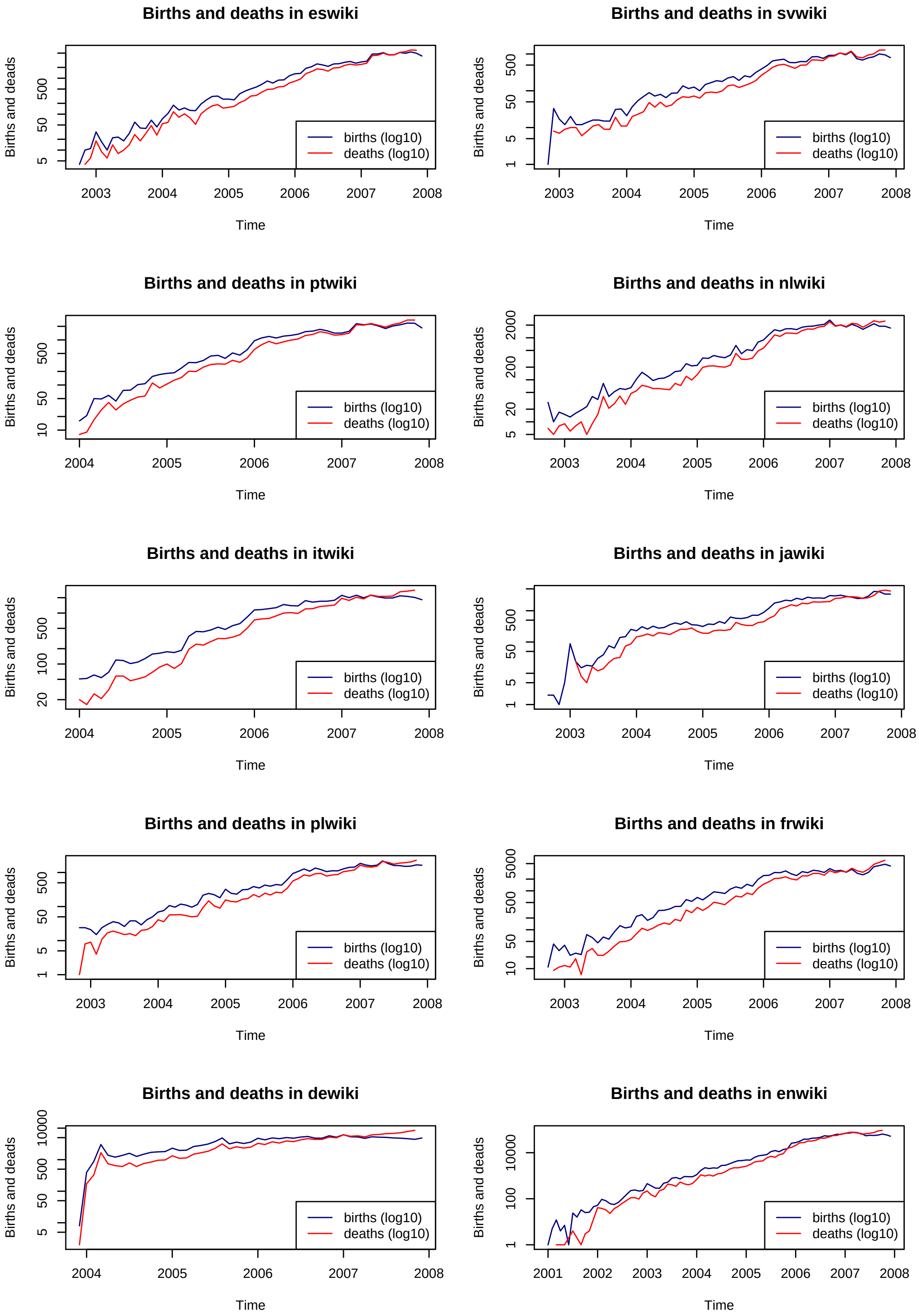

Figure 11: Monthly number of births an deaths of logged authors in the top ten Wikipedias. Both axes have a logarithmic scale. The graph shows that the number of deaths per month closely follows the number of births, suggesting a high mortality rate that prevents the population from growing at an exponential rate. We can also appreciate that in Summer-Fall 2006, there was a dramatic change in this tendency, in all language editions. The rate of deaths become higher than the number of births, and this trend has been followed consistently by all language versions over 2007

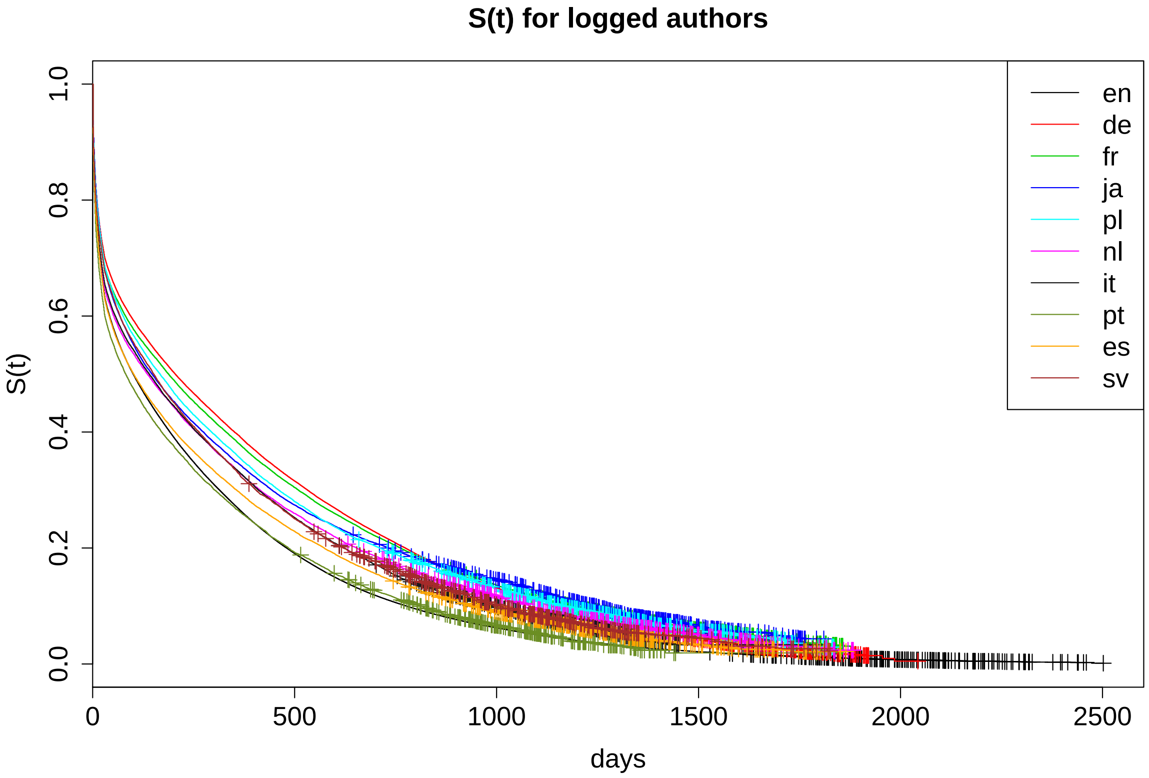

Figure 12: Survival functions of logged authors contributing in the top ten Wikipedias. The graph shows that the mortality level among young contributors (less than one year of participation in the project) is substantially high. It is also remarkable that less than 40% of authors in all language versions continue to participate in the project once they reached an age of more than 500 days

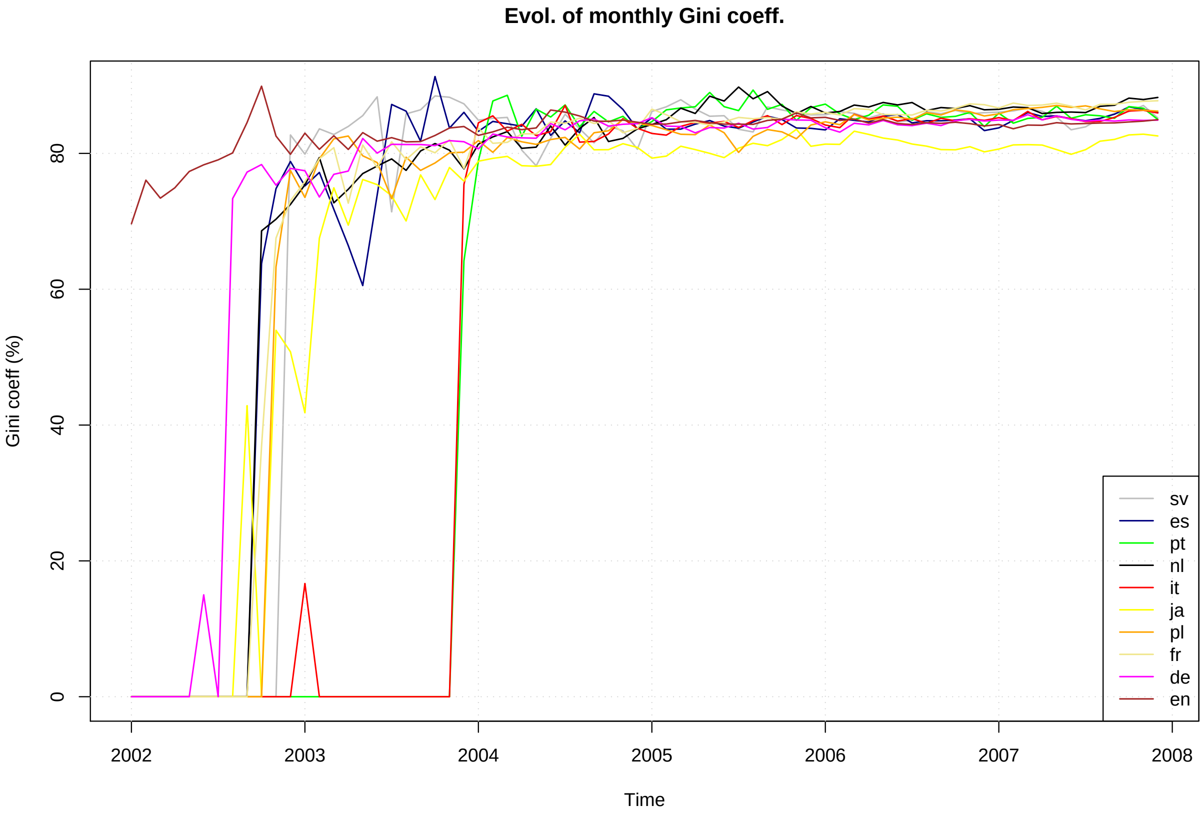

Figure 13: Evolution in time of the monthly Gini coefficient for logged authors in the top ten Wikipedias. The plot demonstrates that all language versions reach a self-regulated pattern after their first two years of history. The initial transitory state is debt to the initial influence of anonymous authors, which is subsequently overwhelmed by the number of contributions from logged authors in the following years. Interestingly, all language versions get stabilized within a small interval, with monthly Gini coefficients in the range 80-85%

3 Relevant conclusions

The main conclusion that we can infer from the overall results of our quantitative analysis is that there exists a severe risk on the capacity of the top-ten Wikipedias, to maintain their current activity level in due course. According to our graphs and numbers, the inequality level of the contributions from logged authors is becoming more and more biased towards the core of very active authors. At the same time, the monthly Gini coefficients show that the inequality level of contributions from logged authors has remained stable over time, at the cost of demanding more and more contributions from active authors to alleviate this deficit of monthly revisions.

Furthermore, we have seen that the distribution of the total number of revisions per author follows an upper truncated Pareto distribution. While more core authors begin to reach the upper limit of their human contribution capacity, we will see a point in the future of this language versions in which the steady-state of the monthly Gini coefficient will start to decrease. This situation would not pose a problem in itself, unless for the fact that we have demonstrated that the most significant part of the content creation effort in Wikipedia is not undertaken by casual, passing-by authors, but by members of the core of very active contributors. A recent study have shown that a preferential attachment process is responsible for the activity pattern that we have identified here [Spinellis and Louridas, 2008].

On top of that, the lack of new core members seriously threaten the scalability of the top-ten language versions regarding the quality of their content. We have demonstrated in the analysis previously presented that the eldest, top-active contributors are responsible for the majority of revisions in FAs, as well. Since the number of core authors has reached a steady-state (due to the leverage in the total number of active authors per month), the group of authors providing the primary source of effort in the revision of quality articles has stalled. Without new core members, the number of different articles who would potentially become FAs can not expand, since we do not have enough reviewers for that content. Since the total number of quality articles generated so far in the top-ten language editions is fairly low, we can conclude that this approach will not contribute to dynamize the creation of quality content in Wikipedia in due course. It is true that Wikipedia has succeeded to compete with other traditional encyclopedias, namely Britannica, but if we do not have a clear strategy for making the creation of quality content in Wikipedia more agile, the project will not ever evolve from its current character of “good starting point to look for a quick introduction of a new topic, from which we can jump to more serious information sources”.

To conclude this section, it would be disappointing to avoid offering some insights about possible solutions for the top-ten Wikipedias to improve their current trend. Nevertheless, some of the knowledge needed to formulate such recommendations could be perfectly a matter for a doctoral thesis on its own, namely the causes driving Wikipedia authors to eventually join the core of very active users. Since we have not answered such questions, we can simply settle for enumerating direct countermeasures to alleviate these findings.

In the first place, incrementing the number of core authors should become a priority for the project, and as a first step, Wikipedia should focus increasing the number of monthly active authors. Indeed, donations campaigns are necessary to aid in the financial support of the project, but attracting new contributors or recovering older ones should be an equally important goal, given the current situation. Apparently, a lot of work still has to be done, not only to create new articles, broadening Wikipedia coverage, but also revising current articles to let them reach the FAs distinction at some point. Whether the influence of featuring some of these quality articles in the main page may have a direct influence in the number of revisions received, it is undoubtedly that content featured in the main page of every language versions at least obtains superior visibility in the community. A good idea could then promote “candidate articles” on the main page, thus favoring the reception of new revisions. Many times, users do not know about the existence of articles until they are featured in the main page, or else, until they need to access them explicitly. In the same way, we recommend to display a "randomly selected" article (instead of the current approach of providing a simple link), to try and increase the number of revisions received in standard articles, as well.

Since the importance of the core of very active members has been demonstrated, thinking about possible tools to further automate their daily tasks, thus facilitating their normal activities, should also be taken into account. We know about current useful tools made with this goal in mind, but perhaps trying to recollect new ideas and suggestions from these users could be another option. Since Wikipedia is an open community, it would be quite difficult to further reduce vandalism, and the access of trolls and other undesirable contributors to articles and talk pages. Moreover, previous research works has demonstrated that these acts of vandalism against content or the community itself has been effectively controlled with the current approaches.

Finally, we can not ignore the potential benefits of large scale contributions coming from specific communities, specially from educational institutions at all levels. The potential applications of Wikipedia to learning environments has been also a matter of research, and some authors have shown that direct contribution approaches may have negative consequences for both the quality of content and the willingness of young authors to continue to contribute if the get strictly negative responses to their first revisions. All the same, semi-controlled strategies like providing a final version of the contribution, eventually created from an incremental local creative process may have better effects, for both the quality of content and maintaining the implication of young contributors. In this regard, providing special tools for highlighting these contributions could facilitate the work of experienced Wikipedia authors, who could then provide more focused comments.

References

[Adler and de Alfaro, 2007] Adler, T. B. and de Alfaro, L. (2007). A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, New York, NY, USA. ACM Press.

[Spinellis and Louridas, 2008] Spinellis, D. and Louridas, P. (2008). The collaborative organization of knowledge. *Commun. ACM*, 51(8):68–73.

[Stein and Hess, 2007] Stein, K. and Hess, C. (2007). Does it matter who contributes: a study on featured articles in the german wikipedia. In *HT '07: Proceedings of the 18th conference on Hypertext and hypermedia*, pages 171–174, New York, NY, USA. ACM.