UNIVERSIDAD REY JUAN CARLOS

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

Wikipedia: A quantitative analysis

Doctoral Thesis

José Felipe Ortega Soto

Ingeniero de Telecomunicación

Madrid, 2009

Thesis submitted to the Departamento de Sistemas Telemáticos y Computación in partial fulfillment of the requirements for the degree of
Doctor

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universidad Rey Juan Carlos
Madrid, Spain

DOCTORAL THESIS

Wikipedia:
A quantitative analysis

Author:
José Felipe Ortega Soto
Ingeniero de Telecomunicación

Director:
Jesús M. González Barahona
Doctor Ingeniero de Telecomunicación

Madrid, Spain, 2009

March , 2009

WE HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER OUR SUPERVISION BY *José Felipe Ortega Soto* ENTITLED *Wikipedia: A quantitative analysis* BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF *Doctor of Philosophy in Computer Science*.

Jesús M. González Barahona, Ph.D.
Thesis Director

The committee named to evaluate the Thesis above indicated, made up of the following doctors

Carlos Delgado Kloos, Ph.D.
Universidad Carlos III de Madrid
President

Daniel German, Ph.D.
University of Victoria
Member

Nicolas Jullien, Ph.D.
TELECOM Bretagne
Member

Juan Julián Merelo Guervós, Ph.D.
Universidad de Granada
Member

Gregorio Robles Martínez, Ph.D.
Universidad Rey Juan Carlos
Secretary

has decided to grant the qualification of

Móstoles, Madrid (Spain), March , 2009.

The secretary of the committee.

We are at the very beginning of time for the human race. It is not unreasonable that we grapple with problems. But there are tens of thousands of years in the future. Our responsibility is to do what we can, learn what we can, improve the solutions, and pass them on.

Richard Feynman
US Educator & Physicist (1918-1988)

The world we have made as a result of the level of thinking we have done thus far creates problems that we cannot solve at the same level at which we created them.

Albert Einstein
Theoretical Physicist (1879-1955)

Acknowledgements

This thesis is the final outcome of a long, hard and very difficult journey that I began 3 years ago, when I firstly talked to my advisor Jesús M. González Barahona about starting a research line on Wikipedia in our group. Along these years, the invaluable help and support from a number of people has made it possible to eventually succeed in this challenging endeavor.

In the first place, I would like to mention the incredible support I received from Chus. She has been always at my side, encouraging me to continue to work, to solve all problems I encountered during this time. She has also been generous enough to sacrifice some of our spare time to let me undertake pending tasks related to this thesis, until it was eventually finished. I would also like to thank my parents for their love and support, specially since I decided to leave my home town 10 years ago to come to Madrid and start my studies in Telecommunications Engineering. You know that they will be there, at any time, for anything. You made me stronger, more confident and sensible. This thesis is now a reality in part because of your unparalleled support over these years. My grandmother and my other relatives there (with a special mention to Antonio) also share this dedication from my heart.

My thesis advisor, Jesus, has played a major role during this process, not only on the professional aspect but also on the personal side. He always trusted in what we did, since the very beginning of this journey, in spite of the numerous obstacles that we had to overcome until we started to define a coherent line of work. It is because of his unequalable vision of future, and his invaluable intuition to predict the next steps that you should follow, even before the path appears in front of you, that we are now one of the very few research groups in the world capable of performing a major role in this field. Without a doubt, the remaining part of my research career will always be influenced by what I have learned from Jesus.

It is now the turn to thank all my workmates at Libresoft, my research group and, in some sense, my home during these years. It is very difficult to try to mention here all of them. All the same, Gregorio (the restless researcher), Miquel (the "Spanish librarian", fountain of wisdom), Luis (the oriental "veggie" master), Isra (the geek researcher), Pedro, Alicia, Isa, Lili, Paco, Rober, Jose, Álvaro, Dani, Javier, Santi, Roca, Fran, Teo, Gato, Miguel, and all remaining current members (and past members like Juanjo) of the team must receive my gratitude for all the great moments we enjoyed during this time. My friends at UAX also deserved a special mention at this point (Antonio, Ricardo, Basil, Juan Carlos, Goyo, Fran, Juan Antonio, Jesús...). This thesis is also dedicated to all of you.

I would also like to specially thank all external reviewers of this thesis work, including my friends from the WikiSym and the folks from the Wikipedia Research mailing list for their pertinent comments, corrections and suggestions that made this thesis even better and more useful for the research community worldwide.

To conclude, like many other times, I must remember the person who (along with my parents) has probably influenced most on my personality, my view of life and how to face everyday challenges, as well as all personal values that will come with me for the rest of my life: my grandfather. My grandparents and him would probably have been proud of this. Surely, they are.

Abstract

Ortega Soto, José Felipe. M.S. in Telecommunications Engineering, Departamento de Sistemas Telemáticos y Computación, Universidad Rey Juan Carlos, Móstoles, Madrid, 2009. *Wikipedia: A quantitative analysis*.

Presently, the Wikipedia project lodges the largest collaborative community ever known in the history of mankind. Due to the large number of contributors, along with the amazing popularity level of Wikipedia in the Web, it has soon become a topic of interest for researchers of many academic disciplines. However, in spite of the increasing significance of Wikipedia in scholar publications over the past years, we oftenly find studies concentrating either on very specific aspects of the project, or else, on a specific language version.

As a result, there is a need of broadening the scope of previous research works to present a more complete picture of the Wikipedia project, its community of contributors and the evolution of this project over time. This doctoral thesis offers a quantitative analysis of the top ten language editions of Wikipedia, from different perspectives. The 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 is focused on logged authors (those editors who created a personal account to participate in the project). The comparative study encompasses general evolution parameters, a detailed analysis of the inner social structure and stratification of the Wikipedia community of logged authors, a study of the inequality level of contributions (among authors and articles), a demographic study of the Wikipedia community and some basic metrics to analyze the quality of Wikipedia articles and the trustworthiness level of individual authors. This work concludes with the study of the influence of the main findings presented in this thesis for the future sustainability of Wikipedia in the following years.

The analysis of the inequality level of contributions over time, and the evolution of additional key features identified in this thesis, reveals an untenable trend towards progressive increase of the effort spent by the most active authors, as time passes by. 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. 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, massaging it to obtain key metrics and descriptive parameters, and loading 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 complementary results from different research 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 encyclopaedia, boosting in this way the number of comparative studies like this one in many other scientific disciplines.

Contents

List of Figures

List of Tables

Chapter 1

Motivation

> " The problem with Wikipedia is that it only works in practice. In theory, it can never work." *Mikka Ryokas, CS student quoted on the NY Times* [1]

1.1 Introduction

Wikipedia is one of the most dynamic, ambitious and largest collaborative projects in the Internet. Building a universal compendium of the human knowledge has been a long wanted goal pursued by some of the most brilliant minds worldwide, since the 18th century. Nowadays, the Internet of the 21st century makes it possible to enrol literally millions of volunteer authors in this challenging goal. They conform what is, most probably, the largest fine-grain-traceable human group ever established, defying all previous limits related to collective intelligence, the so-called *wisdom of crowds* [114], and collaborative information systems. Without a doubt, the totally open approach followed by the Wikipedia has been determining to reach their current popularity level, and the immense number of contributions received. Nevertheless, it is precisely this totally open policy which intuitively raises many questions about the effective ability of this project to articulate a coherent corpus of knowledge. Wikipedia must find its way to merge this large number of contributions, from users with disparate backgrounds, in an worthy effort to coordinate them accordingly.

In fact, Wikipedia has revealed itself as a useful resource for many users around the world, who visit the website in a daily basis, and who made this project the 8th most visited website in the Internet, according to Alexa's [2] top-ranked websites list. Besides that, Wikipedia is not only open regarding its editing and accessing policy, but also about the access to its log files, registering every single edit performed by Wikipedia authors in any language version. As a result, the database dumps containing these log entries provide detailed information about the largest open virtual collaborative project we can find today. This is an unparalleled opportunity for researchers in many different areas like computer science, sociology, education, behavioral sciences, linguistics, semantics and so forth, to analyze this singular project and gain knowledge about its unique features from many different perspectives.

[1] `http://www.nytimes.com/2007/04/23/technology/23link.html?_r=1&ex=1178510400&en=c0eb1b23e5c579f7&ei=5070`. A fairly similar, and earlier comment in the same line, was performed by David Gerard in response to an entry on Mark Glaser's blog, on April 14, 2006 `http://www.pbs.org/mediashift/2006/04/how-much-do-you-trust-wikipedia104.html`

[2] `http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none`

All the same, most previous research works on Wikipedia have not been able to profit from this totally open access to its activity log files. Numerous studies only focus on the English version of the encyclopaedia, or even on smaller subsets of pages or authors from this language version. Empirical analyses comparing different language versions of Wikipedia are still scarce, and frequently focused on very specific areas. The cause of this lack of general comparative studies among different language versions is that it is extremely difficult for scholars and researchers to process the huge database dumps files in an efficient way. This initial step frequently frustrates their original plans, reducing the coverage of ambitious research initiatives and limiting our overall understanding of this very important project.

Given this initial situation, along with the successful background of our research group on the analysis of very large open collaborative projects and communities in the Open Source Software arena, I decided to undertake the ambitious enterprise of automating this process. As a result in the context of this doctoral thesis I created a software tool, *WikiXRay*, that efficiently retrieve available information from Wikipedia log files in any language version, summarizes descriptive metrics from those files and creates a convenient, ready-to-use database that can be directly accessible by researchers from all disciplines to develop their own studies. At the same time, due to the absence of broad comparative analyses of different language versions, I applied our home-grown software tool to perform an empirical analysis of the top ten language versions of Wikipedia, according to the official count of their respective number of articles. This analysis encompasses many different aspects, such as describing the Wikipedia community of authors, the organization of Wikipedia content, the distribution of reviewing effort among community members, or a complete demographic analysis of the community of authors in these language versions. The underlying objective of this quantitative study is to identify key evolution patterns in the Wikipedia project (at least, valid for the top ten language versions), as well as to infer relevant trends that may affect the sustainability and feasibility of this ambitious project in the future.

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, it is expected 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 encyclopaedia, boosting in this way the number of comparative studies like this one in many other scientific disciplines.

1.2 Wikipedia and the *open movements* phenomenon

Wikipedia may seem to be a one-of-a-kind project, thanks to its high popularity level, its working philosophy based on totally open access, and its quite active community of contributors and readers. Nevertheless, Wikipedia is just another example (though, admittedly, a very important one) of the general trend towards *open movements*, that we have witnessed in the Internet over the past years [81]. *Open movements* is a general term, describing a broad range of projects, initiatives and working strategies put in practice in the virtual world of Internet. All of these initiatives share 3 common features:

- They are built around a community of (frequently) volunteer contributors, who participate in the project without economic rewards, and who follow a self-organizing policy to organize the

community and share the different responsibilities among their members.

- The main goal of these initiatives is to produce knowledge outcomes, supported in digital formats, which may come in multiple flavors: multimedia content, software, documentation, news, comments and editorials, and so forth.

- They must provide complete access to the outcomes of their production process, without imposing any limits to non-members of their respective communities to retrieve, display and learn from those outcomes. However, there may be some previous requirements to participate in the content creation process (though some of these projects, for instance Wikipedia, do accept contributions from any interested individual, disregarding her previous background). Usually, this benefit comes by means of licensing knowledge outcomes under any kind of open license (like GFDL, CC, etc). Most of times, these licenses preserve the right of original authors of receiving attribution when derived works are produced based on their previous content.

In this context, it is remarkable the case of *Free, Libre, Open Source Software* (FLOSS), because of the many potential similarities that we may find in common with Wikipedia. FLOSS development projects are focused on the production of software solutions, along with appropriate documentation and sometimes other additional products (like development tools, integration environments and so on). We can find an extensive collection of prior research works which revolve around the analysis of FLOSS projects and their associated development communities. This is an effort to understand such initiatives, that have proven to be capable of creating quality, durable products even valid for industrial environments. The challenge for many experts is to discovered how this quality outcomes may raise from an apparently uncontrolled, self-organized community of volunteer contributors, without the traditional tight conditions and organizational policies imposed in industrial production environments.

From a scientific perspective, FLOSS projects offer and unparalleled opportunity for researchers due to the huge amount of publicly accessible data that they offer, as a result of their open access working strategy. We can find an extensive analysis of the multiple research strategies and applications of these publicly accessible repositories in [97]. Given this previous research experience on the analysis of FLOSS development projects, it would have been a nonsense to ignore all previous lessons learned and strategies already applied in the past to study this field. Wikipedia also provide a large amount of logging information and additional documentation, that can be used to study the organizational structure and working patterns exhibited by its community of authors in the same way we have conducted similar analyses for FLOSS projects.

Nevertheless, Wikipedia also presents unique features that make it difficult to apply some of the same tools and strategies used for FLOSS projects:

1. The archives containing all information and activity records from the daily work of the Wikipedia authors is several orders of magnitude larger than the data repositories we must face in FLOSS projects. This imposes the adoption of drastic changes in the implementation of our analyses, since the size of the data files to be processed prevents us of performing any manual analyses. As a result, we are forced to create appropriate tools to automate the data analysis stage of our research process.

2. Wikipedia is focused on the production of multimedia encyclopaedic articles, while FLOSS projects are focused on the creation of software (either complete solutions, or modules/libraries). The different nature of their respective outcomes creates very different communities. Each FLOSS project attracts those developers who hold specific knowledge of the

problem it addresses, with reasonable practice in the programming languages and development methodologies utilized in the project and with a particular willingness of contributing to that community. In other words, only certain contributors have the adequate skills to participate in the software development process. In Wikipedia, only the last condition remains valid, since it accepts contributions from virtually any volunteer, disregarding her prior background. The main consequence is a much broader production environment, both in number of contributors and in number of pages (and consequently in the amount of content) produced. The analytical techniques applied on Wikipedia must deal with the size of the target community and the set of outcomes, again raising the need of creating automated procedures to successfully undertake these tasks.

3. Finally, as a consequence of the two aforementioned conditions, when we undertake any empirical analysis of Wikipedia, we are forced to find a valid, yet efficient and optimized methodology if we want to obtain results in a reasonable time period. Therefore, one of the main contributions of this thesis work is to define efficient procedures to develop the quantitative analysis of empirical data obtained from Wikipedia, and provide other researchers with adequate tools automating such procedures, allowing them to reproduce these results and build on these tools to further extend the scope of the empirical analyses performed on the Wikipedia project.

However, despite the numerous and important differences between Wikipedia and FLOSS development projects, the same open access and open contribution paradigms are applicable on both types of initiatives. Therefore, it is worth to keep in mind the main research guidelines and results obtained in previous research works on FLOSS projects. We may find some underlaying characteristics and behavioral patterns indicating interesting common points between both fields.

1.3 Motivation of this doctoral thesis

As we have already stated in the introduction of this chapter, the popularity of Wikipedia, its singular features, the size of its community of authors and the publicly available repositories, with detailed information about all individual contributions, make it a insuperable research topic among the open virtual communities. Wikipedia presents peculiar organizational features, a broad coverage of millions of entries in many distinct languages, as well as a high popularity level and other unparalleled accomplishments, to justify attracting the research interest of the scientific community. More specifically, Wikipedia stands out of similar projects pertaining to the *open movements* phenomenon, from 3 different points of view:

- The size of the Wikipedia community: As we will see in chapter 3, if we consider the 250 different language editions of Wikipedia, this project has received contributions from more than 5,000,000 registered users worldwide, and an undetermined number of anonymous users. Most probably, this is the largest community of users in history focusing on a collaborative project at a global scale on the Internet. This fact offers us a unique opportunity to identify possible behavioral patterns driving the production process of this group of users. At the same time, if we build a statistical model showing the key parameters that affect the workflow of this community, we would be in a great position to offer valuable assessment regarding other content creation processes involving free open communities.

- The growth rate of Wikipedia: So far, Wikipedia has always improve its position in Alexa's website traffic ranking year after year, until it became the 8th most popular website. Its growing

popularity makes this project one of the most dynamic and active initiatives of the Internet, both by its number of users and the size of its contents. Thus, it presents us the most difficult challenge up to date to model behavioral patterns of a large community of users, as we will face: a huge number of contributions, continuous changes in its group of users (mainly corresponding to new contributors coming to participate) and a large number of individual communities of users focused on their own language editions, that let us compare different behavioral patterns influenced by customs and habits from different locations.

- Totally open project: Every piece in the Wikipedia engine is totally open to public scrutiny. Its server side infrastructure is completely based on free open source software. Database transactions and contents are publicly available on a special website supported by Wikimedia Foundation [3]. This encyclopaedia is open to receive contributions from any user without any restriction other than respecting the project contribution policy, and specially, disregarding the user's previous knowledge or experience. So we have unlimited access to one of the largest information repositories of the Internet, storing valuable data about users' contributions, contents evolution over time and, in summary, every conceivable aspect we can think of related to this open contents creation process.

In the past years, two main controversies have raised around the Wikipedia project and its capacity to achieve its goals in the long term, following a sustainable approach. The first one concerns the distribution of effort among Wikipedia authors. On September 4, 2006, Aaron Swartz wrote a frequently cited blog entry about this topic [115]. This author argued that, if we count the number of characters introduced in each individual contribution by a Wikipedia author, and then we carefully follow the number of text blocks that remained unmodified until the final (current) version of the article, it can be shown that casual contributors are the main force behind the creation of most of the contents in Wikipedia articles. This findings contradicted the previous statements by Wikipedia Founder, Jimmy Wales, assuring that to the best of his knowledge, the majority of the content creation effort in Wikipedia is performed by a small group of very active authors, many of which have even meet other colleagues in person, once in a while. If we reflect for a moment about this debate, we can conclude that both authors are talking about slightly different aspects. While Swartz focuses on the final result of the content creation process (that is, how many text block were produced by each author in the final version), Wales was concerned about the *reviewing* process, that may or may not translate into actual changes in the article (perhaps the author just checked a fraction of the whole text and ignored the rest). Wales approach seems more reasonable to us, since we want to characterize the behavioral patterns found in the community of authors from the top ten Wikipedias, and thus, we must concentrate on reviewing activity rather than on actual products.

In fact, the content of Wikipedia articles is quite involved in the second controversial debate around Wikipedia, which has been also the most active and recurrent both in academic forums and mass media: the quality of Wikipedia articles, and the debate around the trustworthiness level of Wikipedia content. Despite some articles published in scientific publications with undoubtedly reputation, showing that the quality of some Wikipedia articles is perfectly comparable to the trust level of other serious encyclopaedias like Britannica [44], an overwhelming majority of articles, editorial columns and presentations has been devoted to highlight the poor quality of Wikipedia articles. The article by Neil Waters [130], published in the CACM magazine in 2007, is the perfect example summarizing the natural concerns of many educational experts and professional about the emerging role acquired by Wikipedia, acting as a natural source of information for their

[3] `http://download.wikimedia.org`

students. However, Waters establishes the difference between rigorous scientific information sources and encyclopaedias, of any kind. As third level citation sources, encyclopaedias can not be cited in academic works, in a regular basis. On the contrary, Wikipedia has found its place as a good point for Internet users to get a quick look at a new term, acting as the initial step of a longer search path to find the information they are looking for. Of course, the quality of Wikipedia articles has become a matter of concern also for Wikipedia management, starting with Jimmy Wales. This is an ambitious, yet very complex goal, as we will describe in the following chapter of this thesis. Our purpose has been to spot some light in this field, looking for simple metrics that may help us to identify the common profile of Wikipedia authors who produce quality content, shared traits among Featured Articles in Wikipedia, and initial metrics of the quality of Wikipedia pages and the reputation of Wikipedia authors.

Thus, we undertake in this thesis work 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 encyclopaedia 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 articles contributed by each author. 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 organization 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 [108] 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 [6].

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.

Among the most relevant implications that we pursue to clarify in this thesis work, we can mention:

- Obtaining a deep understanding of the inner behavioral patterns driving the creativity effort and organization of large size communities of users in a collaborative project.
- Developing automated tools to extract valuable information to constantly track the current status of such projects.
- Offering users real time information about the progress of the contents evolution, so that they can easily identify featured contents providing high quality, trustable information.
- Evaluating eventual updates of system requirements on the sever side to accommodate the users workload.
- Making educated assessments about the project future trend, focusing on possible further enhancements that could be applied to improve the users experience.

We can not forget to talk about the software tool developed as a result of this doctoral thesis. So far, we have remarked several times the importance role of available automation tools to facilitate the complete process of retrieving information from Wikipedia repositories, process the database dump archives and store the resulting data in the adequate format in our local database. To fulfill this objective, we have developed a completely new tool, named *WikiXRay*, which we use throughout this thesis to undertake our analyses, creating both graphical and numeric results and summaries. In Chapter 3, the reader can find and in-depth description of our tool, its current capabilities and the most relevant extensions we have thought of to expand its current features in due course. This is an important contribution for the whole Wikipedia research community. As far as we know, this is the first software suite providing complete automation of these tasks. We have already receive positive feedback and comments from external researchers, who have utilized some of the modules of our tool to undertake their own empirical analyses, thus making easier for them to concentrate in their research goals, rather than struggling against technical implementations tricks and difficulties in their daily work. Although the development of WikiXRay has been triggered by our own necessities to undertake these empirical analyses, we have tried our best to ensure that WikiXRay is stable and useful to perform a wide range of studies, and it can be easily extendible in case other researchers needs further features in their own research works.

1.4 An overview of the Wikipedia project

Before going further into the presentation of this research work, the reader may profit from an introductory description of the Wikipedia project, covering both historical and technical aspects. It seems a sensible approach to acquire some knowledge about the project we are going to analyze. These details may help us to contextualize some of the results and conclusions that we will present in subsequent chapters of this document. Following this approach, we proceed in the first place with a basic historical presentation of the inception of Wikipedia and its evolution during its early years of history. Then, we continue with a description of MediaWiki, the core wiki engine adopted by Wikipedia, and we introduce an overall description of the hardware infrastructure and information

systems that make it possible to run this project. Finally, we conclude with an overview of the top-level aspects that we should have in mind when we explore the content creation process in the Wikipedia community, including the main policies and guidelines established within the community to regulate this process. As we will see, this apparently random and disorganized collaborative process is controlled, in fact, by a very active community of volunteer who try their best to conserve and enhance the quality of the most valuable active in Wikipedia: its set of encyclopaedic articles.

1.4.1 The inception of Wikipedia

Many people do not know that shortly after the creation of the World Wide Web by Tim Berners-Lee in 1989 [4], and the Hypertext Transfer Protocol in 1991 [5], one of the central technologies for collaborative projects was about to birth. In 1994, Ward Cunningham started to develop the first known wiki, called the *WikiWikiWeb*. The purpose of this project was to help software programmers share their knowledge about Design Patterns on Object-Oriented Programming, acting as a complement to the Portland Pattern Repository [6]. The site got a high popularity level almost immediately, but only within the group of programmers interested in this topic.

It was much later, in 2001, when wikis became broadly known, and reached the popularity status they hold today. On January 15th, 2001, Jimmy Wales set up the first version of Wikipedia, and put it online, according to the information provided by the Wikipedia project on `http://en.wikipedia.org/wiki/History_of_Wikipedia`. The initial purpose of Wikipedia was to serve as a feeder for the Nupedia project, a peer-review free encyclopedia, written by highly qualified contributors in each field, founded by Bomis.com. Larry Sanger, graduated student of Philosophy hired by Wales to act as full time editor in chief of Nupedia, introduced the wiki concept to the project. The main objective was to speed up the contents creation process, since only 12 articles were finished in Nupedia during its first year of existence.

Many editors from Nupedia were reluctant to associate the name of the peer-reviewed project to a wiki-based, totally open site, so Sanger suggested to give it its own name, Wikipedia, and run it on a separate domain, `wikipedia.org`. The project was a complete success, after receiving the attention of mass media (the first coverage was in *The New York Times* on September 20th, 2001) and other news sites like Slashdot. Soon after its inception, the internationalization of Wikipedia began with the first non-english version, the German Wikipedia, on March 16th, 2001. It was followed only several minutes later by the Catalan version. The French version started on March 23rd, and later, in May 2001, several new language editions were created (Chinese, Dutch, Esperanto, Hebrew, Italian, Japanese, Portuguese, Russian, Spanish and Swedish). The goal of creating a completely free, open encyclopaedia that could eventually concentrate all the human knowledge in one site was about to start.

As we will see in this thesis work, Wikipedia has experimented an exponential growth rate over the past years. Its openness, as well as the self-management approach to work with contents, are often two recurrent arguments given to explain the huge success of the project. Despite that, this policy also created some major controversies among Wikipedia initial pioneers, specially between Larry Sanger and Jimmy Wales. When Bomis.com ceased to fund both projects, Larry Sanger give them up, and later began a crude campaign centered in criticizing Wikipedia working philosophy, the role of Wales in Wikipedia's success, and his ideas about how Wikipedia should be manage to face

[4] `http://www.w3.org/People/Berners-Lee/`

[5] `http://www.w3.org/Protocols/HTTP/AsImplemented.html`

[6] `http://c2.com/cgi/wiki?WikiHistory`

future challenges. Then, Sanger began Citizendium, an alternative free encyclopedia with specific acknowledgement of expert's contributions and direct supervision by an editors committee. Several alternative projects have also raised in the past few years, including Interpedia, Veropedia and the Spanish *Enciclopedia Libre*, but none of them has reach the level of success of Wikipedia.

Nowadays, Wikipedia continues its own battle towards the maximum popularity level on the Internet, struggling against other giant projects and companies with overwhelming financial resources: Google, Yahoo!, Youtube, MSN... [7] . All the same, Wikipedia, with its auto-financed policy, has manage to reach the 8th position among the most popular web pages on the Internet, according to Alexa's website traffic ranking. Its English language edition has currently surpassed the 2,300,000 articles mark, and the top 20 language editions (ordered by their total number of articles) store more than 8,260,000 articles in total. All these versions store more than 100,000 articles each. Among other possible reasons, there is no doubt that one of the key aspects that has contributed to the very big success of Wikipedia is its simple, intuitive and easy-to-use editing interface, and powerful capabilities to merge multimedia contents and display information in an organized way. The following section focuses on MediaWiki, the software package providing all these excellent features in Wikipedia.

1.4.2 MediaWiki: The core engine of Wikipedia

The first version of the Wikipedia utilized a simple wiki engine named UseModWiki, which was developed in Perl. It was later migrated to a PHP based user interface developed by Magnus Manske, a student and developer from the German University of Cologne. However, the great level of success reached by the project forced him to rewrite the code, mainly to provide a more scalable database backend, based on MySQL. Lee Daniel Crocker was the first responsible of the current software, MediaWiki [8]. Later, Brion Vibber assumed the role of most active developer and release manager. The name "MediaWiki" was coined by Wikipedia contributor Daniel Mayer, playing with the name of The Wikimedia Foundation, established on June 20th, 2003, to manage the Wikipedia and other related projects.

Currently, MediaWiki supports all projects launched by the Wikimedia Foundation, including of course Wikipedia, plus all wikis hosted by the Wikia project [9] and many other external wikis, including some of the most popular ones like Wikitravel [10]. Figure 1.1 shows the aspect of the main page of the English Wikipedia, running on MediaWiki. MediaWiki is libre software, released under the GNU General Public License (GNU GPL).

MediaWiki currently offers an impressive list of automations and tools that facilitates the work of wiki readers, editors and managers. Figure 1.2 presents a close-up of a random page found in Wikipedia.

As we can see, the user interface on every page presents several tabs, offering a convenient, intuitive way for accessing contents and functionalities. Actual contents of the article are displayed on the tab `article`. Users can navigate through those contents just like on any other typical web page, and follow *internal links* to other articles in that language edition of Wikipedia, as well as *external links* to web pages hosted in other websites different from Wikipedia. Newer versions of MediaWiki incorporate specific support to separate external links from internal links, presenting the former ones as footnote bibliographic entries.

[7] `http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none`

[8] `http://en.wikipedia.org/wiki/Mediawiki`

[9] http://www.wikia.com/wiki/Main_Page

[10] `http://wikitravel.org/`

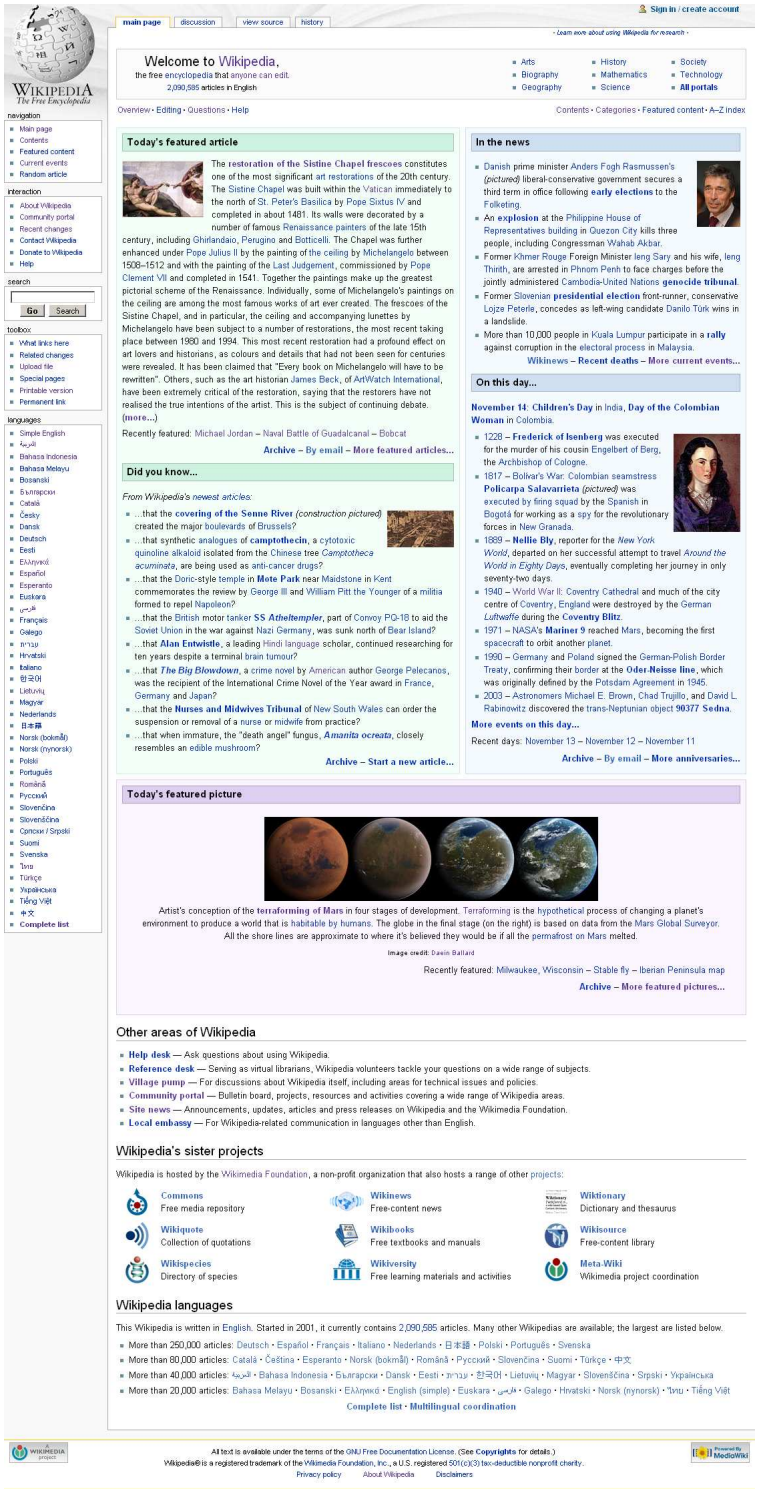

Figure 1.1: Screenshot showing the main page of the English edition of Wikipedia

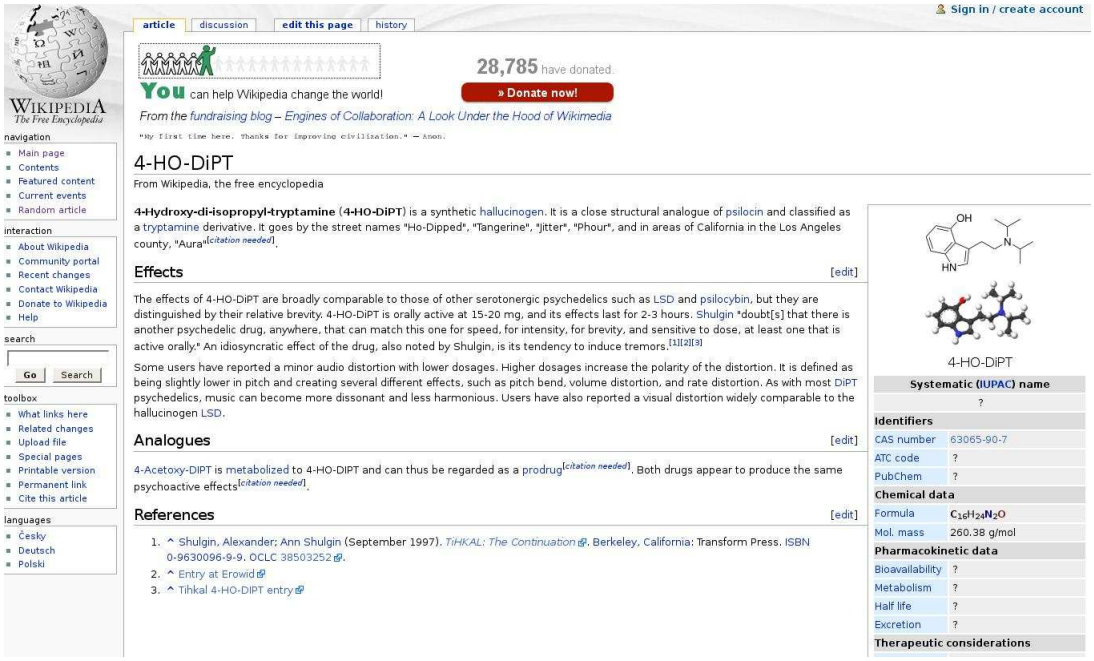

Figure 1.2: Close-up snapshot of a random article from the English edition of Wikipedia

If the reader feels like editing the article's contents, to add new information, images, correct any mistake she may have found, or whatever other reason, she only needs to push on the `edit` tab, to switch to the edit interface on MediaWiki. This interface is presented on Figure 1.3.

As we can see, we have plenty of options to choose from. Above the contents box, we have the main menu to access editing tools, automating the content creation process. Aside from the typical buttons to stress words using bold and italic fonts, inserting internal and external links, specially formatted headers and images and multimedia files, we have specialized tools to include mathematical formulae, hidden comments, super and subindexes, picture galleries, blocks of quoted text, tables, etc. Below the contents box, we find a set of special characters, symbols and magic tags, very useful to speed up the editing process. We can also check how the aspect of the wiki page will be before actually committing our changes to the database, using the *Show Preview* button, and we can look for changes on contents clicking on the *Show Changes* button. MediaWiki also include some additional features, like automated table of contents for pages with more than 4 second level headlines, summary tables for special types of articles, categorization (currently implemented manually, simply by adding the corresponding tags to the end of the article) as well as *page transcluding*, which allows the editor to embed contents from other wiki pages in the current page.

Thus, MediaWiki has supposed a determinant factor to eliminate possible barriers stopping average, non technically skilled users, from contributing to Wikipedia. So, the client side is powerful but, how about the server side of Wikipedia infrastructure? It has been stated previously that the Wikimedia Foundation is a non-profit organization, and that Wikipedia is totally financed through donations. How can a self-financed project reach the top 8th position in the list of most visited websites? The following section presents some details about this infrastructure, and presents some answers to this challenging question.

Figure 1.3: Close-up snapshot of the edit interface in MediaWiki

1.4.3 Wikipedia Server Side Infrastructure

It may seem obvious that a totally centralized infrastructure, implementing the server side of the Wikipedia project would have failed very soon, when the project began to acquire its impressive level of success and popularity among Internet users. However, that was the situation of the server infrastructure in Wikipedia, at least until June, 2005.

The initial set of Wikipedia servers was located at Tampa, Florida (USA). Nowadays, the vast majority of Wikipedia servers has been moved to the new headquaters of Wikimedia Foundation at San Francisco, California (USA). At the Florida Power Medium data center. Bomis.com paid for the bandwith costs until it leaved the project on 2004. The Tampa cluster comprises 12 nodes for MySQL database servers, 50 nodes for Squid proxy and web caching servers, 158 nodes executing Apache web servers and 14 additional nodes for miscellaneous purposes. The acquisition cost of this hardware exceeds 800,000$ [11]. Figure 1.4 shows a picture of the cluster of servers at Tampa, Florida (USA) [12].

On June 2005, Kennisnet (Netherlands), a public Dutch foundation providing ICT support to educational organizations in the Netherlands, reached and agreement with Wikimedia Foundation to start providing hardware support to develop a new server side hardware infrastructure for Wikipedia in Europe. At present time, the Kennisnet cluster of Wikipedia machines comprises 41 servers providing support for Wikipedia web traffic coming from European countries. Finally, Yahoo! donated a cluster of 23 machines, which is located at Seoul (South Korea), an it is currently providing support for traffic coming from Asian and Oceania countries. Figure 1.5 summarizes the deployment schema for this

[11] http://meta.wikimedia.org/wiki/Wikimedia_servers/hardware_orders
[12] http://meta.wikimedia.org/wiki/Image:Floridaserversfront1.jpg

Figure 1.4: A photography showing the Wikipedia cluster of servers at Tampa, Florida (USA)

server side infrastructure as of May 2006 [13].

According to a technical report published by Domas Mituzas on 2007 [68], the software server infrastructure in Wikipedia is essentially based on a LAMP (Linux, Apache, MySQL, PHP) infrastructure, enhanced with many add-ons and performance wise tools:

- *Linux*: All Wikimedia servers executes Fedora and Ubuntu as their base operating system.
- *PowerDNS*: Distributes DNS requests geographically, among the 3 different server locations worldwide.
- *LVS*: Linux Virtual Server provides load-balancing of requests to application servers and cachés.
- *Squid*: Proxy and web traffic caching service to speed up load of contents on the client side.
- *lighttpd*: Static files provisioning.
- *Apache*: Web server application, renders the wiki pages to send them back to clients.
- *PHP5*: Dynamic construction of web pages.
- *MediaWiki*: Core software package implementing the logic behind the wiki site.
- *Lucenne, Mono*: Speed up contents searches.

[13] `http://meta.wikimedia.org/wiki/Image:Wikimedia-servers-2006-05-09.svg`

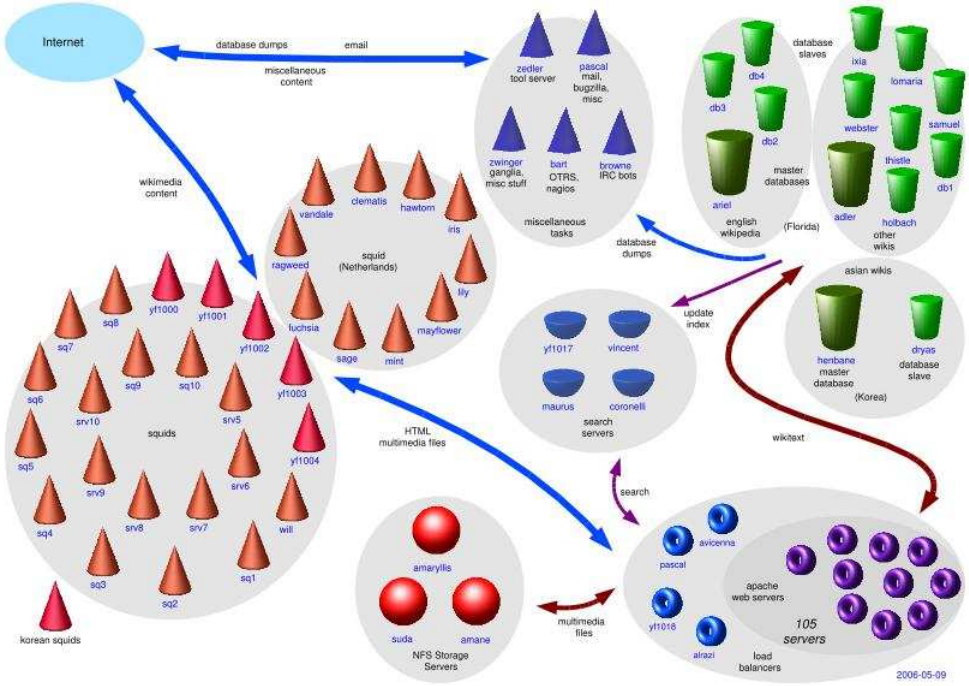

Figure 1.5: Schema summarizing the server side infrastructure in Wikipedia

- *Memcached*: Responsible for caching specific objects to further increase performance of the whole infrastructure.

According to this, the whole software infrastructure is based on free open source software, since the cost saving of this approach is fundamental to continue following the self-financed policy of the Wikimedia Foundation (otherwise, it would be completely unfeasible to sustain such deployment facing the licenses costs of non-free software solutions).

The set of Squid servers scattered over the 3 locations speeds up content provisioning to final users, by means of a thorough, aggressive web caching strategy and a fine-tuned Content Delivery Network infrastructure, which of course, is geographically smart (based on PowerDNS as it was stated above). Database servers at Tampa also execute fine-tuned MySQL servers, to enhance the database response performance. MySQL runs on top of a federated cluster infrastructure of database servers (comprised by the 12 aforementioned dedicated nodes). There also exist 4 additional tool servers, dedicated to pre-production tests of software tools.

So far, Wikimedia Foundation has succeeded in maintaining a low-cost server side infrastructure, capable of attending high level traffic demands from worldwide users. Mediawiki developers have been capable of providing a fully-featured, simple and usable wiki interface to interact with the whole infrastructure. However, the open nature of the contents development process in Wikipedia also plays a central role in this project. In the next section, the organization and policies of this content creation process is presented, stressing its main differences in comparison with other similar initiatives.

1.4.4 The content creation process in Wikipedia

Any person coming to Wikipedia for the first time may wonder how a completely open contents creation site can find its way to produce coherent encyclopaedic articles. The obvious answer to this

question is that although Wikipedia exhibits a complete open policy for accepting contributions, its community of users also enforces a strict policy to manage those contents. On the main page of every language edition, we find a link (*About Wikipedia*[14]) that takes us to a wiki page presenting the project itself, the Wikimedia Foundation, as well as many guidelines oriented to prospective contributors that want to initiate their way as Wikipedia editors.

On one hand, regarding the articles editing policy, the Wikipedia community enforces some concrete points regarding the acceptance of new material:

- *Respecting the Neutral Point of View (NPoV)*[15]: The NPoV guarantees that every Wikipedia article should provide unbiased, accurate information, if possible taken from reliable sources. This is a key policy to resolve possible disputes raised within the contents creation process. The main objective pursued by this policy is to ensure that Wikipedia contents are always refined through *consensus*, avoiding particular opinions of individual users. As an encyclopaedia, Wikipedia contents must not be offensive, they must respect all significant views and provide unbiased information, not influenced by political ideologies, religion beliefs, etc.

- *No original research is allowed*: The main goal of Wikipedia is creating encyclopaedic articles. By definition, an encyclopaedia presents contrasted facts, that should be accurate and verifiable. As a consequence, there is no room at all in Wikipedia for original research. Research papers, thesis works etc. present innovative analyses, new perspectives, methodologies, approaches and hypothesis, possibly sustained by quantitative results. But that information must be first validated by the scientific community, by means of additional research on those topics. Once the research results has been extensively validated and documented, they turn into factual knowledge that may be accepted in encyclopaedic works.

- *Contents must be verifiable*: As it was previously stated, the third pillar of the encyclopaedic work is to provide verifiable contents. In this sense, the Wikipedia community of users always tries to provide complete references to document the articles contents. High quality articles, recognized as such by the community, always provide extensive references to additional information sources, to ensure that Wikipedia users can contrast articles information.

- *Following the manual of style*: There are additional style recommendations that new editors are encouraged to follow, in order to contribute to Wikipedia articles in an efficient way. Every language edition provides their own guidelines for new editors[16].

On the other hand, the community of users itself does not have a completely plane organization. There exists a hierarchy of special users, who hold special privileges within the project, such as article deletion and protection, banning controversial or vandal users, or promoting other users to a special privileged status. Tables 1.1 and 1.2 summarizes the distinct user levels that we can find in Wikipedia, along with their respective attributions[17].

[14] In the English language edition: `http://en.wikipedia.org/wiki/Wikipedia:About`

[15] `http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view`

[16] In the English language edition, those guidelines include a *Manual of Style* `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style` and a guide on *How to edit a page* `http://en.wikipedia.org/wiki/Wikipedia:How_to_edit_a_page`

[17] `http://en.wikipedia.org/wiki/Wikipedia:User_access_levels`

Table 1.1: Table summarizing regular users privileges and rights associated to each level in Wikipedia

User level	Description	User rights
Anonymous users	Those who have not created a user account in the system	Read pages; modify all pages except protected or semi-protected ones; create talk pages in any namespace; they must answer a *captcha* to perform edits containing at least one external link; they must click a confirmation button to purge pages
New users	Those who have just created a new user account in the system, but they still have not confirmed it using an email account	Create new pages; email other users if they configured an email account in their user profiles; mark edits as minor; purge pages without confirmation; they still must answer a *captcha* to introduce edits containing at least one external link; they can customize their Wikipedia interface and account options
Autoconfirmed users	Those who created a user account in the system and, additionally, confirmed it using an email account; the status is automatically granted by the software and cannot be removed subsequently; requirements are to have an account older than 4 days and making at least 10 edits for users with regular access; for users accessing through a Tor network, it is required an account older than 90 days and at least 100 edits	Move pages; edit semi-protected pages; upload files or a new version of an existing file

Table 1.2: Table summarizing regular users privileges and rights associated to each level in Wikipedia

User level	**Description**	**User rights**
Administrators (*sysops*)	Considered as system administrators in practice, this privilege is granted by a community upon individual request; the process includes careful examination of their editing history and participation level in the community	Page deletion; page protection; blocking and unblocking contributions from certain user accounts or IP addresses; modify protected pages; modify the MediaWiki interface; grant and remove rollback and ipblock-exempt rights to other users
Bureaucrats	This privilege level is also granted upon consensus among the community to exceptionally trusted users	Add other users to the `sysop` or `bureaucrat` groups (but with no deleting privilege); add or delete users from the `bot` user group; rename users (including themselves)
Stewards	Elected role, whose privileges span across all Wikimedia Foundation wikis	Grant and/or revoke any permission to any other user; this includes granting *administrator* or *bureaucrat* privileges to users in wikis with no local users with the required rights to do so; specifically, they may grant or revoke *oversight* or *checkuser* privileges, since no other role is capable of performing those changes
Rollback	Special user	They can revert editions from other users with the rollback feature
Ipblock-exempt	Special user	Not affected by autoblocks or blocking IP addresses
AccountCreator	Special user	Not affected by the 6 accounts per day per IP limit; they can also bypass the security checks on account creation steps

Uploader	Special user	Marked with a special flag, granted by a *steward* after discussion on the Village, which allows new users to upload files without waiting for the mandatory requirements to perform such action
Oversigth	Special user; they must be older than 18 and they must also have identified themselves to the Wikimedia Foundation	They have the right to hide certain revisions on any page from public users, as well as look up a log file of those actions and the hidden content
CheckUser	Special user; they must be older than 18 and they must also have identified themselves to the Wikimedia Foundation	They can access a list of all IP addresses utilized by any user account to edit the English Wikipedia; they have access to the list of edits performed from a certain IP address and a list of all user accounts that accessed from a certain IP; they may access log files of those petitions, as well
Bots	Automate or semi-automated programs	Perfrom massive changes in wiki pages, with a very well-defined objective and clearly identified actions allowed; they can be automatically blocked if they are detected to depart from their original goals or attributions
Developers	Software developers	Certain members of the development team may receive this special privilege to access some restricted development areas/functionalities
Founder	Jimmy Wales (only for English Wikipedia)	Special role created by Tim Starling for Jimmy Wales in the English Wikipedia; it is largely recognized as a mere status symbol; Wales actions are visible through the English Wikipedia rights log[18]

[18] `http://en.wikipedia.org/wiki/Special:Log/rights`

1.5 Organization of this thesis work

In this section, the content of subsequent chapters in this thesis work are described. The overall structure follows a traditional IMRAD (Introduction, Methodology, Results and Discussion) organization, with the exception of Results and Discussion, which has been merged in the same chapter to improve readability and comprehensiveness.

Therefore, the organization of chapters follows the schema presented in Table 1.3.

Chapter num.	Title	Description
1	Motivation	This chapter presents some background to contextualize the Wikipedia project, it explain why this initiative is extremely interesting from a research point of view and finally, it enumerates the research questions that I am going to tackle in this thesis work.
2	Related Research	It presents a detailed description of previous research works related to wiki projects in general, and Wikipedia in particular. It focuses mainly on quantitative analyses, community focused and content production studies related to Wikipedia as a whole, or any of its language editions in particular. It also presents a brief description of related research on adjacent topics such as web semantics and social networking, that may serve to further contextualize the current research work in the whole picture of research on Wikipedia.
3	Methodology	In this chapter, the methodology followed to implement the quantitative analysis of the Wikipedia community of authors is described. I also define key concepts, parameters and constraints applicable to all subsequent results and discussions presented in this thesis work.
4	Empirical analyses and results	This chapter presents and in-depth discussion of all results and metrics obtained from the empirical analyses performed in this thesis work. These include a general overview of the status and evolution of the top ten language editions, analyses on the concentration level of contributions within a certain community, demographic analyses, implications for content quality, as well as a thorough discussion on possible sustainability requirements from the project in due course .

Chapter num.	Title	Description
5	Conclusions and Further Research	Finally, I conclude this thesis with a brief summary of the most important conclusions that can inferred from our results, and explore possible research lines that we may follow in the future related to this research area.

Table 1.3: Organization of chapters in this thesis work

Chapter 2

Related research

> "There is no branch of detective science which is so important and so much neglected as the art of tracing footsteps". *A Study in Scarlet*, Arthur Conan Doyle, (1888).

In this chapter the most relevant research efforts conducted on Wikipedia and wiki projects in general are described in detail. As well, a thorough description of the state-of-the-art in this field is offered, along with pointers to additional information sources to find bibliography, tools and results related to Wikipedia and wiki research. Finally a description of seminal works and outstanding research publications focused on Wikipedia is provided. This presentation is aimed to contextualize the contribution of this thesis work, as well as to draw a general picture of current and future trends in Wikipedia and wiki research.

2.1 Collaborative open communities and wikis

Wiki communities are just one part of a broader, emerging new wave of collaborative networking paradigms, attracting the interest of both practitioners and researchers in this field. Though the first example of a wiki website can be found back in 1995, when Ward Cunningham installed the "WikiWikiWeb", on the *c2.com* Internet domain [1], wiki tools and wiki related projects in general did not attract the attention of the research community until some years later. However, the general concept of **collaborative authoring** has been a matter of research even before the creation of the first wiki, and of course, well before the arrival of the *Web 2.0* concept.

In 1990, Neuwirth *et al.* published one of the first exploratory studies on collaborative authoring on the Web [72]. This document presents the *PREP* editor, a platform which aims to support co-authoring and commenting of on line documents. The paper describes three different challenges that should be tackled to successfully design a networking computational system offering these capabilities:

1. Providing appropriate methods to support social interactions among participants in the process.
2. Integrating adequate support of cognitive aspects of co-authoring and external commenting.
3. Implementing practical tools provisioning both types of interaction.

[1] `http://c2.com/cgi/wiki?WelcomeVisitors`

The authors conclude in this document that effective technologies supporting collaborative authoring and commenting should address other important problems other than merely offering a sharable database and hypermedia capabilities. Eventually, one of the key factors influencing the rapid success of wiki-based collaborative authoring has been to provide easy-to-use, yet rich interfaces addressing the problems presented in this paper. In the same way, supporting social and cognitive interactions among collaborative authors (for instance, by means of *talk pages* or discussion mechanisms) has been another important factor contributing to the success of wiki platforms.

Later on, in 1992 Dourish and Bellotti [37] explored additional issues regarding networked collaborative authoring, focusing in practical aspects of *awareness information*, that is, details about the comments and activities that other authors are implementing on the shared document. The authors present the term *shared feedback*, applicable to those collaborative authoring systems that makes information about the activities performed by a certain author visible to other participants as well, by means of feedback on this activities displayed on the shared workspace. Again, this concept has been implemented in wiki platforms, presenting a history page that list all changes made on a certain wiki page chronologically, and also including comments associated to the content found on the current (and past versions) of a wiki page. This feature facilitates the creation of such a *shared feedback* environment, providing up-to-date information to other authors about the current status of the collaborative authoring process.

Franco *et al.* introduced in 1995 another crucial factor of collaborative authoring [42]. This research work presents the anatomy of a *flame*, a hostile and insulting message deliberately sent to a virtual community without any intention other than offending other users. This unfriendly, aggressive behavior may alter productive relationships among virtual community members. These interferences may be viewed as dangerous, pernicious and susceptible of generating divisions and disputes within the community. On the contrary, these authors maintain the thesis that these actions may trigger counterfeit responses from the community itself, helping other members to realize which ones are the actual values that they should preserve in order to ensure the stability of the whole virtual organization. The increasing interest on tools and frameworks for collaborative authoring resulted in the development of many different alternatives to support these initiatives. Noël *et al.* presented [76] a comparative compendium of 19 different tools for collaborative authoring, showing the state-of-the-art of this technology at that time, and describing useful hints for systems designers who confront the challenge of creating new systems of this kind.

In the same way, many collaborative projects supported by wikis naturally lead to the establishment of **open communities** of individuals who contribute to it. Nonetheless, the creation of such kind of virtual communities is a central factor for the successful development of other initiatives different from content authoring. The organizational structure of these communities and motivation of volunteers who join them has been thoroughly examined in previous research works. Raymond analyzes in its seminal work [91] the structure of a special type of collaborative virtual groups: FLOSS development projects. While traditional organization of software development groups follows the *cathedral* approach, centralizing the organization of the software development process, FLOSS development groups tend to organize following the *bazaar* style, in which independent producers (in this case of software code, documentation, translations, etc) organize their activities in a decentralized manner. Decentralization gives each contributor complete flexibility to decide which responsibilities she is willing to assume within the project, to dynamically adapt (if possible) their working effort according to her personal circumstances and to choose in which areas she wants to contribute. As in the bazaar, the aggregation of autonomous merchants results in the creation of a collective entity, empowered by sinergies established between individual participants. However, one of the questions still to be answered is whether this flexible working environment scales effectively, to assume the

development of large size FLOSS development projects. Butler *et al.* study in [19] different factors influencing the contributions of participants to virtual communities. They conclude that, as it might be expected, project leaders undertake much of the work towards the construction and maintenance of the virtual community, but also, that other members actively contribute in these tasks. On top of that, community members tend to contribute to the whole development process in different ways, according to their personal preferences about which are the most important values and benefits that should be pursued within the virtual community.

Another important factor influencing the current popularity of wiki-based projects is the **open content** concept, presented by Cedergren in [22]. In this paper, the author defines *open content* as that allowing users to redistribute and improve it, and/or which is produced without considering any possibilities of financial reward in the short term, oftenly created in a collective manner within a virtual community. This concept, similar to the *open source* concept in software development, makes it possible that content collaboratively created within virtual communities may circulate to other virtual communities, without any drawbacks imposed by IPR (Intellectual Property Rights) clauses. Favoring the open exchange, aggregation and improvement of content in distinct virtual communities rises the working factor spent in the collaborative creation process, promoting sinergies between projects resulting in more effective approaches to benefit from the authors creative effort.

As a consequence of the new avenues opened by wikis for collaborative authoring, their influence has reshaped the traditional way of creating content on networked environments. Gillmor [45] analyzes a paradigmatic example of the possible consequences of this new approach applied to collaborative working groups. "We the Media" questions the classical perception of centralized journalism (in the sense that the journalist is the main source driving the creation of a piece of news) as the unique approach to reflect the highly dynamic reality of our world today. In a global village connected by the Internet, readers become writers of interesting content, enriching the data, details, opinions and multimedia contents associated to any news event located in any part of our planet nearly in real-time. Some of these authors acquired enough influence and readers as to be considered as trustable information sources for professional journalists, completely changing the rules of this business towards a highly dynamic merge of different information pieces fusioned in a final result, collaboratively developed by different autonomous writers. The adoption of Wikipedia as a method for participatory journalism can be found in [61]. Désilets *et al.* present in [33] a methodology leading to the application of wiki technologies to translation.

In the same context, Surowiecki presents [114] the *wisdom of the crowds*, a term that has strongly influenced subsequent research on collaborative authoring in the following years. The author analyze the implications of collective intelligence and collaborative work for the organization of business, economies, societies and nations nowadays, and how this new wave of understanding and implementing social interactions is shaping our lifestyle resulting in globally affected environments, subject to more complex relationships between a growing number of geographically distributed participants. A quite similar perspective is covered in the book by Benkler "The Wealth of Networks : How Social Production Transforms Markets and Freedom" [15]. Krowne and Bazaz analyze [58] authority models for CSCW, focusing on answering the question of which authority models are the best ones for collaborative, commons-based content authoring. They conclude that, despite the author-centered model is the preferred one in such environments, the free-form model, in which several distinct authors equally account for the content creation effort, is the most productive one.

Given this general framework, with the widespread need of platforms supporting the creation and development of open communities for content fabrication, wikis quickly found their role as a central tool aimed to boost the effectiveness of this cooperative process. In 2003, an study published by Sébastien Paquet at the University of Montreal [85] was the first one confirming the emerging adoption

of wiki technologies to implement collaborative knowledge sharing platforms. Likewise, Wagner explores [125] how to use wikis as a tool to create virtual environments for knowledge management and collective collaboration. The same author further explores the adoption of *conversational technologies* (such as discussion forums, weblogs and wikis) to support knowledge management and knowledge acquisition systems [126], [127]. Buffa *et al.* present in 2004 a paper explaining three case studies about the application of wiki technologies for collaborative development of software, involving groups of users geographically distributed [17]. They conclude that students were able to acquire the necessary skills to effectively use the environment in a short period of time, a period that can be further minimized with focused training sessions. Ebersbach and Glaser also points out the advent of a new period in which wikis will drive the collaborative content creation process on the Internet [38], an argument further extended in their classic book "Wiki: Web Collaboration" [39]. Xiao *et al.* also study in [135] the adoption of wiki technologies for collaborative software development. Of course, Wikipedia is another specific example of the broad range of projects that can be built around wikis and open communities. Kolbitsch and Hermann explore in [56] the challenge of building virtual communities of users around encyclopaedic contents, focusing on the Wikipedia as the most prominent example (at the time) of how a really community-driven encyclopaedia should look like. Eventually, they demand some enhancements, such as contents transclusion and homogeneous links, that has been eventually implemented in Mediawiki in the following years.

There also exist examples of research on the application of the so called *Web 2.0* technologies to virtual communities built upon wikis. DiBona in [34] already studied possible applications of the open source working philosophy to other creative and social environments. Later on, O'Reilly defined in [79] the Web 2.0 concept, and Schoop published in 2006 "The Pragmatic web: a manifesto" [102], supporting the adoption of this new artifacts to create a more dynamic Web environment, focused on the importance of contributions from final users. Wu in [134] and Cosley in [27] analyze two distinct ways of adapting these solutions to enhance knowledge management in virtual communities.

This previous background confirms the important role played by wikis as a tool to support collaborative authoring projects, as well as to set up open virtual communities around them. It is time now to focus on more specific topics directly related to this thesis. In the following section, the most relevant active research lines about Wikipedia will be presented, along with remarkable publications revolving around this topic.

2.2 Wikipedia research: state-of-the-art

The following sections of this chapter present seminal research works that have influenced subsequent publications in this field, as well as outstanding research contributions providing relevant results, methodologies and conclusions that are key factors to understand the state-of-the-art in this field.

2.3 Classification of research works on Wikipedia and wikis

Generally speaking, research on Wikipedia and wikis is multidisciplinary by nature. Wikipedia and wiki projects have been able to attract a critical mass of researchers from quite distant areas such as Sociology, Computer Science, Linguistics and Education. All of them are finding new interesting and challenging aspects waiting to be analyzed, hypothesized and revealed to the research community worldwide.

One of the first areas of interest analyzed in wiki literature was the application of wiki technologies

to build knowledge sharing platforms (specifically, in industrial and production environments). The book by Leuf and Cunningham [60] was the first one to provide good examples on how wiki technologies could be applied by system administrators and managers to implement knowledge sharing systems with little deployment effort. Subsequently, the widespread adoption of wiki technologies in many projects, each one pursuing different goals, and involving quite distinct communities of users, began to create the appropriate atmosphere for interdisciplinary research on the effect of the adoption of wiki technologies in collaborative working environments.

Nevertheless, despite the obvious benefits of the adoption of wikis to support cooperation and distributed authoring on virtual groups, one of the most important concerns analyzed by researchers over the previous years was the quality level of the content created in such environments. In particular, totally open communities like the one found in Wikipedia raise some doubts about the potential quality of its articles. Lipczynska and Sonya explore in [62] the implications of giving the entire control over the content creation process to final users, analyzing again the example of the Wikipedia. They conclude that the project provides a source of unquestionable value as a reference, though their analysis is restricted to a reduced number of articles. On the other side, Denning *et al.* rise in [31] some critical questions about to what extent Wikipedia contents can be trusted, mainly due to the open nature of its collaborative writing process. Some factors like content accuracy, volatility, heterogeneous coverage, uncertainty of the contributors level of expertise and the use of unverified citation sources should be also taken into account when assessing the overall quality of the project. Nevertheless, focusing on a set of articles that has been further scrutinized by the community, it can be concluded that there is no reason for Wikipedia to be jealous of other traditional, renowned encyclopaedias like the Britannica. A very controversial article published on Nature [44] lead to the aforementioned conclusion after a blind review of a selected subset of articles taken from Britannica and Wikipedia, undertaken by a trustable committee of renowned experts in several distinct scientific fields. This article has been one of the most cited ones regarding the level of trustworthiness of Wikipedia content, and it even unleashed a brief public debate between Britannica and the Nature editorial board. In any case, Waters finally brought some good sense to this issue with another famous paper entitled "Why you can't cite Wikipedia in my class" [130]. Even though the achievements of the Wikipedia project must be recognized, any encyclopaedia (disregarding its virtual or traditional origin) should be treated as a *tertiary*, and thus, it should not be accepted as a trustable source of citation, even if it leads us to a trustable, citable source. Waters exemplifies his arguments with a real case study, in which he recommended the history department at Middlebury College not to accept any Wikipedia citation on academic works, an attitude renewed by the Wikimedia Foundation itself: Wikipedia, as any other encyclopaedia, cannot be used as a direct source of citation. All the same, Wikipedia still maintains an increasing level of popularity, and its value, at least for providing a quick introduction to an increasing number of topics is unquestionable. Nielsen even finds in [75] that the accuracy level of Wikipedia citations to outside sources of information is getting better, following structured citation markup policies and a good level of agreement with the citation pattern seen in the scientific literature, though it can be found a light trend towards citation of high-impact magazines like Nature and Science.

Another important aspect analyzed by previous research works on Wikipedia and wikis is the organizational structure of the community of individuals participating in this kind of initiatives. Two different approximations to this issue can be found. The first one explores the inner structure of the community and the implications of such structure in the behavioral and activity patterns found in the project over time. Emigh and Herring presents in [40] a genre analysis on Wikipedia and Everything2, two examples of collaborative working communities built around the creation of encyclopaedic knowledge. In this research work, these authors find that there exist strong correlations

between the level of post-production editorial control imposed by the project and the formality and standardization of the collaborative documents stemming from such virtual environment. The social, cultural and even economical implications of the Wikipedia project have been explored in [63]. Spek *et al.* analyze in [105] the inner organizational patterns of the Wikipedia community, that following a typical bottom-up approach also found in other open communities. The applicability of wikis to other specific environments involving collaborative authoring, such as academic courses and working groups have been also examined in this line of research. Forte and Bruckman present in [41] a study of the possible applications of Wikipedia on e-learning and education. It seems that collaborative authoring environments such Wikipedia can be easily adapted for integration in the learning process, adopting the adequate implementation policies. In particular, these authors conclude that perceived audience plays an important role in helping students monitor the quality of writing, though the acquisition of such perception of audience is not seamlessly identified by students.

The second approach followed in the study of the cooperative community of authors built around Wikipedia is to explore the motivational aspects influencing the participation of individuals in this project, from many different perspectives. Miller explores in [66] the notion of the *disappearing author*, regarding the collective nature of the content creation process in Wikipedia, which favors merging individual efforts of users into a common visible result, dissolving the immediate identification of each individual contribution. This line of research was also previously explored by Ciffolilli in its seminal work about *phantom authority*, the self-selection process for attracting new users to virtual communities, and existing methods for retaining collaborators in these virtual communities, focusing on Wikipedia as a case study [24]. All these previous works show that one of the key aspects that should be considered in the study of the Wikipedia community of contributors is that the concept of individual authoring is somewhat different to the traditional notion perceived in other areas, like traditional literature. Instead of clearly differentiating which contents where produced by each author, the goal of Wikipedia is to integrate all individual contributions seamlessly, creating a coherent product while offering (at the same time) the opportunity to track down each individual contribution if needed. Reagle analyzed in [93] and [94] different cases of social interactions that may be found in the Wikipedia virtual community, either as a mind-expanding tool or as a source of leadership roles within the community. Rafaeli also explore in a series of research papers motivations behind the knowledge creation process of Wikipedia [89], [90], [88]. Bryant *et al.* also explore in [16] the motivations behind collaboration on Wikipedia. [77] and [59] also add further considerations in this research line. Viégas *et al.* present in [121] a detailed analysis of the inner mechanisms to reach consensus about articles content in Wikipedia, identifying common situations of disputes, debates and vandalism that must be addressed by the Wikipedia community of users to achieve their goals. Some additional thoughts about these line of research can be found in [54]. Some members of the Wikimedia Foundation have also presented their own point of view regarding the motivations behind the great success of this project. Some examples are [10], [96], [13] and [129].

It is possible to find several different research tracks related to Wikipedia and wiki projects in this area. All the same, as far as this thesis is concerned, it will focus on those research works on Wikipedia circumscribed in the Computer Science area. Thus, the following taxonomy to organize research publications about Wikipedia in this context is proposed:

- *Quantitative analyses*: One of the most obvious research lines is to undertake quantitative analyses suitable for modelling the Wikipedia system behavior, activity patterns of authors, evolution of contents over the time, etc. Most of these research works comprise the application of statistical methods, data mining techniques and analysis of traffic and system log files to create these quantitative models.

- *Quality of contents*: One of the emerging research lines that is rapidly attracting the attention of many Information Systems researchers is the problem of measuring quality of contents in on-line data repositories. This problem becomes even more interesting in a collaborative data repository like Wikipedia, accepting contributions from any interested user without any restrictions. The impressive activity level experimented in the most active language editions of Wikipedia creates a extremely dynamic evolution of its contents, which poses a harder to solve challenge for quality metrics systems. Researchers have also tried to push this problem beyond traditional, human-reviewed methods, to build automatic systems to assess the quality of contents in Wikipedia or any other wiki project.

- *Social networks, web graphs and links models*: Collaborative systems and projects represents a natural target for Social Networking researchers. Again, the large size of the Wikipedia community of users, specially in some very active language editions, creates virtual communities of unparalleled size, presenting their own behavioral patterns and functional philosophies. Besides that, the structure of the linked data repository itself, represents a great opportunity for web researchers interested in analysis the structure of contents links, popularity of contents, and possible relationships between contents popularity and number of contributions received.

- *Semantic web and wikis*: The last research track regarding Wikipedia and wiki projects is trying to reorganize the system contents in a more efficient, accessible way, allowing users to search through these contents employing semantic technologies that facilitates finding more accurate results. This research line also includes semantic analysis of Wikipedia contents to build tag clouds, organization of contents based on user's ratings, exploring better ways to link related topics, automatic and user-driven generation of contents taxonomies and categories, as well as their validation.

As far as this thesis is concerned, most of the methods, models and results that will be presented fall in the quantitative analyses research track. Nevertheless, some of these results, specially those focusing on the definition of models to explain the content creation process in Wikipedia, and describing behavioral patterns of Wikipedia authors, may be applied to other research areas such as social networks or quality of contents evaluation and assessing.

2.3.1 Quantitative analyses of Wikipedia

Research works falling in this category presents analyses of Wikipedia based on quantitative results, statistical models and empirical tests. At the same time, these contributions can be classified in three distinct subcategories:

- *General descriptive studies*: Quantitative analyses focused on general descriptions of Wikipedia's features, such as total number of articles, total number of authors, size of contents, number of contributions received and so on, providing a general description of the whole infrastructure, some specific language edition or a comparison among overall quantitative parameters in different language editions.

- *Studies on specific aspects*: Sharing the same empirical approach, these research works address more concrete questions and topics, such as how many different types of contributions can be found, which is the level of inequality in author's contributions, describing behavioral

contribution patterns of Wikipedia authors or different types of vandalism that may suffer Wikipedia articles.

- *Forecasting studies*: This subcategory includes research works that go one step ahead of an overall quantitative description, presenting models to forecast the most probable evolution scenarios for Wikipedia in the next future. Most of these contributions apply time series analysis and similar forecasting methods that try to predict future trends of several parameters from the analysis of their previous history.

General descriptive studies

One of the first examples of quantitative analyses of Wikipedia is the research work published by Jakob Voss in 2005 [124]. In this paper, Voss presents some statistics and graphics modelling the evolution over time, and the activity patterns of the German language edition of Wikipedia. Among other interesting conclusions, he finds that the evolution over time of distinct quantitative indicators, such as the size of the database, the number of articles, the number of active Wikipedians (users who contributed more than 5 times to the project on a given month), the number of very active Wikipedians (users who contributed more than 100 times to the project on a certain month) and the total number of words and internal links, follow an exponential growing rate. Moreover, he also finds that the number of distinct authors per article and the number of distinct articles per author follow Lotka's Law. As a consequence, the activity patterns of the German language edition can be modelled using this approach, that can be frequently found in scientific and scholarly publishing too, and even in other collaborative open environments like FLOSS development projects, as stated in [73].

A seminal research work by Viégas *et al.* [122], presented a revolutionary method to visualize the evolution over time of the contributions made to a certain Wikipedia article.Many subsequent publications have followed the findings and conclusions presented in this paper to study the working patterns adopted by the Wikipedia community of authors. These authors developed a new software tool, named *History Flow*, to undertake this analysis . Loading the metadata corresponding to the editing activity of a certain article, this program is able to present a colored map showing the amount of content added, deleted or modified by every author who collaborated the article creation process. This is particularly useful to visualize the impressive growing rate experimented by many articles in recent years. Some visual examples demonstrating this fact can be found on a subsequent work by some of the same authors [121]. Viégas *et al.* [131] also studied the activity patterns followed by Wikipedia users, by means of a new visualizing technique called *chromograms*. Chromograms are simple graphs depicting the activity pattern of a certain author over time, coloring her contributions to the project according to the activity type performed in each one: typos correction, addition of contents, reverting vandalism, etc. In this way, it can be visually identified, with little effort, the activity patterns usually follow by a certain author within the project. Wilkinson and Huberman find in [132] evidences of a direct correlation between the visibility level of a certain article (measured in terms of its Google pagerank popularity level) and the number of edits received by that article. On top of that, they also find indicators supporting the thesis that there is a direct relationship between the number of edits received by a Wikipedia article and the quality of its contents. These results, however, are restricted to a set of articles in the English language edition of Wikipedia, and they have not been validated for other language editions yet. Ortega *et al.* present in [84] some quantitative results corresponding to a preliminary study of the top ten language editions of Wikipedia up to that date, according to the official count of the number of articles in each one. The findings presented in this research work are extended in this thesis conforming a more complete picture of the general quantitative parameters

describing the top ten language editions of Wikipedia. Among other important results, these authors find that the article population in Wikipedia can be divided into two different subpopulations:

- Tiny articles: Those with less than 200 bytes length, corresponding to *stubs* and *redirects* in most of the cases.
- Standard articles: Those with a length greater than 200 bytes, corresponding to standard encyclopaedic articles that received enough contributions as to depart themselves from the previous subpopulation of tiny articles.

These findings are consistent through all the top ten language editions analyzed in this paper. In the same way, another important result presented in this research work is that there exist a strong correlation between the number of edits per author and the probability mass ratio of the standard articles population, providing evidence of a direct relationship between the contribution level of authors in the top ten languages editions (measured by their number of edits) and the resulting length of articles in that language edition.

Still another contribution related to general descriptive studies on Wikipedia is the work by Almeida *et al.* [8]. They find that the evolution over time of Wikipedia and its number of updates follows a self-similar process, deviating from the general trend of the Web that usually follows a Poisson process. Furthermore, the show evidences that the exponential growing rate of Wikipedia is directly sustained by its rapidly growing number of users, and hence, that the success of the project is directly correlated with its totally open philosophy to accept contributions.

Studies on specific aspects

The concept of the *wisdom of crowds* [114] was already presented, previously in this chapter. An interesting research Work by Kittur *et al.* tried to find quantitative evidences sustaining this thesis in the English Wikipedia [53]. These authors sustain the thesis that a growing number of users with a very low number of contributions to the project were progressively taking over the main part of the content creation effort in Wikipedia. To illustrate their findings, they provide a comparative study of different types of Wikipedia users, clustered according to their total number of contributions to the project, and then comparing their activity patterns with the ones exhibited by the English Wikipedia *sysops*. Nevertheless, Ortega *et al.* [82] complemented these findings, pointing out that *sysops* cannot be considered as the most significant case of highly active contributors in many language editions of Wikipedia. Furthermore, though it is true that there exist a growing population of less active users contributing to the project, more than 90% of the total number of contributions each month has been carried out by a *core* of very active users, who have maintained such pro-active activity pattern in previous months. In addition to this, another research paper by the same authors [83] explores the inequality level of the monthly contributions of Wikipedia authors to the project, for the top ten language editions, finding quantitative evidence supporting the thesis that the monthly number of contributions to the project follows a self-regulated pattern, in which approximately 15 percent of the total number of authors performs between 80-85% of the total number of contributions in a certain month, when the language edition has surpassed its initial period of existence. Again, these previous findings will be extended in this thesis, with an in-depth examination of the inequality level of contributions within the Wikipedia community of authors, as well as analyzing the inequalities in the distribution of contributions among different articles in each language version.

There are also some quantitative research papers dealing with the Wikipedia system infrastructure, and possible ways to overcome specific problems such as improving scalability, availability and

resilience over failures. Following this approach, Urdaneta *et al.* describe in [119] some quantitative results characterizing the traffic workload attended by the Wikipedia system infrastructure, and present a preliminary taxonomy to classify possible user actions within the daily activity of the Wikipedia. These insights were also applied by the same authors in a previous research work, exploring viable approaches to decentralize the Wikipedia system infrastructure, in order to avoid some of the problems mentioned above [118]. The adoption of P2P strategies to improve the technical capabilities of wikis has been also explored by other research works like [70].

2.3.2 Quality of content in Wikipedia

The fact that the quality of Wikipedia contents has attracted the attention of a high proportion of the wiki research community in recent years is unquestionable. Stvilia *et al.* were the first to present a detail theoretical study of the distinct aspects that should be considered to successfully address the study of the quality level of Wikipedia contents, and the trustability level of the authors of the contributions. A complete study of these perspective can be found in [110], [109] and [113]. The same authors have further extended this preliminary research studies, founding a complete framework to analyze the quality of information in originated in complex data repositories and collaborative projects like Wikipedia [112], [3], [111]. This framework includes a detailed study of the different factors affecting information quality in Wikipedia articles, comparing featured articles (FAs), former featured articles and a random sample of standard articles. The analysis presents a descriptive statistical study, along with time series data describing some characteristic trends found in the content reviewing process of FAs. Finally, content analysis aspects influencing quality evaluation discussions are also examined. In the same way, Anthony *et al.* present in [9] a thorough revision of the quality level of contributions to Wikipedia, following an almost manual, exhaustive analysis of contributions from registered and anonymous authors to the English language editions. Their most valuable conclusion is the identification of two different sets of authors providing high quality contributions to the project. On one hand, there exist the so-called *zealots*, registered users with a high level of commitment to the project activities, and providing frequent contributions. Anthony *et al.* presents the thesis that those authors are seeking for reputation within the community, and that the higher number of contributions they provide, and the higher the quality of their contributed contents is, the better the perception of the rest of the community about this authors' reputation is too. On the other hand, we can also find what Anthony *et al.* call *good Samaritans*, that is, casual users with sporadic contributions to Wikipedia articles, but also providing high quality contents. It is more difficult to present sustainable arguments supporting this behavior, others than users with a strong background in certain topics providing valuable contents regarding their matter of expertise. It is also curious that the length of the contributions from these authors tends to be larger than those from registered users, a somewhat unexpected result if it is taken into account that we are considering passing-by users that did not subsequently came back again to contribute.

Given that the Wikipedia community is completely open to the contributions coming from any Internet user (with no restrictions whatsoever), one of the most attractive research topics studied in the literature is the analysis of the credibility of Wikipedia content. Frequently, one of the most disconcerting features of Wikipedia for the general public is the capacity of a totally open project to produce quality articles, even though it is not possible to enforce that contributors should demonstrate their competency or skills level about each contributed topic. In 2006, the online magazine First Monday published an interesting work about these issues, written by Thomas Chesney [23]. In this paper, Chesney examines Wikipedias credibility conducting a research survey among 258 researchers (with a 21% response rate), comprising 3 objectives: 1) assess the article credibility, 2) assess the

author of the article credibility and 3) assess the credibility of Wikipedia as a whole. Reviewers could be experts or not in the topic covered by the article they have to assess. The author found that, though there exist not very much difference regarding points 2) and 3) between expert and non-expert reviewers, there does exist such difference regarding their perception of 1). In essence, expert reviewers tend to perceive higher credibility of Wikipedia articles than non-experts. While the size of the sample may rise some objections from statistical savvy readers, this can be considered as a preliminary result proving the prospective capabilities of the Wikipedia project on the provision of high quality contents.

As a result of the ability of Wikipedia to produce quality content, (at least on a subset of the whole collection of articles in each language edition), the natural question that followed this finding was whether it could be feasible or not to define some metrics suitable for identifying high quality articles in an automated or semi-automated manner. This goal would affect not only the capacity of users to discriminate among search results, looking for trustworthy articles, but also the efficiency of the community regarding the development and control of the content production process itself. Hu *et al.* presents in [51] three article quality measurement models based on interactions between Wikipedia articles data and contributors, extracted from the articles revision history. The same authors present in [52] a proposal to implement two of these quality measurement models to improve Wikipedia search engines, refining results ranking according to the articles quality level. Other independent search engines, like Wikiseek [2] have already applied trust measurements to rank search results on Wikipedia articles more efficiently. Dondio *et al.* performed [35] an analysis to map generally accepted content quality parameters in collaborative environments to Wikipedia articles. The main objective was to compute trust values, suitable for identifying high and low quality articles in Wikipedia in an automated, transparent way. The analysis was conducted over 8,000 articles in the English language edition of Wikipedia, accumulating 65% of the total editing activity in that language edition up to that point in time. In the same line, McGuiness *et al.* presents in [65] an analysis of trust levels in collaborative information repositories, focusing in the case of Wikipedia.

Regarding the particular case of FAs in Wikipedia, Viégas *et al.* presents in [123] a detailed explanation of the review process of FAs. After considering the list of particular requirements that each language edition imposes to an article for being considered as a FAC, a nominator proposes an article to be included in the formal peer-review process. This process usually includes a voting process after which the article is promoted or demoted by the FA Director. Moreover, the FA status is not permanent, as the article contents may continue to vary over the time and the FA requirements may also be modified subsequently. Again, the demotion process of a FA is launched by a nominator, and the FA should follow a peer-review and voting process before finally maintaining or losing its status. Stein and Hess presents in [108] a good quantitative analysis of FAs in the German language edition of Wikipedia, introducing quality measurements for *featured articles* and *worth reading* articles in the German Wikipedia, according to the author's reputation. They propose to use the number of contributions in FA as a measure of the author's reputation, following a similar inductive approach as the one adopted in this thesis. This previous proposal will be used here as a measure for quality of contents in FAs, but this time, applying it to the top ten language editions of Wikipedia.

In WikiSym 2007, Wilkinson and Huberman [133] presented quantitative results characterizing the cooperation in the contents creation process in Wikipedia, and how this affects the quality of these contents. Their analysis included a comparison of Featured and non-Featured Articles ordered by their ranking value, a measurement of a web page popularity computed using a proprietary algorithm by Google named pagerank. It has been proved that the pagerank has a strong correlation with the

[2] *http://www.wikiseek.com*

number of times a Wikipedia page is viewed [107]. Again, this analysis was restricted to the English language edition of Wikipedia. Following the same research line, Zeng *et al.* presents [137] a model to compute article trust in Wikipedia, based on each article revision history and using dynamic Bayesian networks (DBN).

The open nature of the content creation process in Wikipedia also rises some concerns about the capacity of the community to preserve the quality of content already produced, specially with regards to the correction of possible acts of vandalism that may be performed by malicious attackers, willing to damage encyclopedic content on purpose. Priedhorsky *et al.* [86], and Viégas *et al.* [121] already studied the resiliency level of Wikipedia against acts of vandalism, and how the community of users develops rigorous strategy of contents surveillance to effectively revert these attacks. Furthermore, in [123], Viégas *et al.* presents an in-depth examination of the formal process in Wikipedia for high quality articles to became *Featured Articles*.

Following a completely different approach, other previous research works look at the evaluation of the quality of Wikipedia articles from a different point of view. Rather than focusing on the automatic evaluation of the quality of content itself, they try to assess the reputation level of Wikipiedia authors, based on the quality of the content produced by each of them. The rationale behind this approach is that, given a certain article, if some additional information is available indicating the trust level of the author who revised each section (given his/her overall level of trustworthiness on all previous contributions), we will have a good proxy to evaluate the quality level of the content included in that individual article. This alternative approach is not totally extent of some problems, for instance the direct correlation between the reputation of an individual author and the quality of the content provided by that author in a certain article, which still have to be demonstrated. Nevertheless, despite the possible limitations of this evaluation strategy the measurement of the reputation level of Wikipedia authors has produced important contributions improving the state-of-the-art on Wikipedia research. There exist other distinct sets of measures proposed for the evaluation of author's reputation. Korfiatis *et al.* in [57] measured reputation of Wikipedia authors according to the number of edits made by subsequent users that alter the contents introduced by previous authors. Subsequent modifications of a certain author contribution are taken as a disagreement, while maintaining contents edited by previous authors is considered as an approval.

Other important examples of previous research works focused on the automatic evaluation of author's reputation in Wikipedia can be found, as well. Adler *et al.* proposed in [6] another automated algorithm to assess the quality of Wikipedia articles contents, in accordance to the reputation level of authors who have contributed to them. The author's reputation is computed based on the number of subsequent revisions surpassed by his/her contribution without being altered by later editors reviewing the same article, following the same approach already presented in [57]. This is the first method to quantify the quality of Wikipedia contents (and the authors reliability, as a convenient side-effect) that has been successfully applied to any language edition without inconsistencies. In fact, a MediaWiki plug-in implementing this algorithm has been already developed and it is publicly available on the author's web page [3]. A demo site is also offered, with all English Wikipedia pages colored according to the level of trust of its contents. This demo corresponds to the complete snapshot of the English Wikipedia, as of February 6, 2007 [4]. The authors has further elaborated this model in [7] and [5].

[3] `http://trust.cse.ucsc.edu/`

[4] `http://wiki-trust.cse.ucsc.edu/index.php/Main_Page`

2.3.3 Social networks and web graphs in Wikipedia

Another interesting perspective within the Wikipedia research field is the application of web graphs and Social Network Analysis (SNA) to discover interaction patterns among authors and contents. Following this research line, Zlatić *et al.* [139] studied several language editions of Wikipedia from the web graph perspective, demonstrating the presence of many common network characteristics shared among distinct language versions of Wikipedia, such as their degree distributions, growth, topology, reciprocity, clustering, assortativity, path lengths, and triad significance profiles. Bellomi and Bonato also applied web graph analysis in [14] to analyze the internal citation structure of Wikipedia based on HITS (Hyperlink-Induced Topic Search). HITS is a network analysis algorithm that has been successfully used for ranking web pages related to a common topic according to their potential relevance. The authors developed a crawler that retrieved every internal link discover in each Wikipedia article to other Wikipedia articles, then applied the HITS algorithm to find lexical authorities within this context. Their conclusions brought valuable results, in the form of confirming that highest rank authorities in Wikipedia tend to represent particular and concrete contents, rather than universal and abstract entities. Korfiatis *et al.* applied in [57] SNA techniques to unveil authoritative sources of knowledge in Wikipedia.

Capocci *et al.* represented in [20] a content graph of the English Wikipedia, a visualization in which every node is an article and edges represent hyperlinks between articles, representing the whole encyclopaedia as a direct graph. The authors find that the growing process of Wikipedia strongly depends on preferential attachments between authors (despite these authors may be able to contribute to any given part of the project) and that the Wikipedia rate of growth can be reproduced by simple statistical models, as the topological properties of its graph closely follows those found for the World Wide Web. Buriol *et al.* subsequently presented an in-depth analysis of quantitative parameters driving the evolution over time of this graph [18]. Zesch and Gurevych presents in [138] and analysis of the Wikipedia category graph and article graph, assessing the usefulness of the first ones as a Natural Language Processing (NLP) resource. Besides that, Muller *et al.* undertake in [71] a SNA on a corporate wiki built on top of MediaWiki, utilizing their own Social Network Analysis (SNA) software tool named SONIVIS [5]. They focus on the study of possible applications of this methodology to asses the knowledge management process in corporate environments, through the analysis of collaboration and social interactions among the wiki authors.

Completing the different perspectives provided by previous research works in this field, Spinellis and Louridas [106] presented interesting findings about the structure of the web graph conformed by articles in the English Wikipedia. Considering each individual article as a node, and the internal links pointing to other articles in the same language version, these authors found that the resulting network is scale-free. This important result will be revisited again in the following chapter, but for now, it is enough to indicate that the main consequence of this finding is that there exists an underlying process, namely preferential attachment, driving the cooperative content creation in the Wikipedia community. Following this model, the principle governing this creational process is that the greater the number of links pointing to a certain article is, the higher the probability that a larger number of authors will contribute to that article. On average, this will generate a subset of very active articles, corresponding to the most linked Wikipedia entries in other encyclopedic pages, concentrating a significant proportion of the total number of revisions received in that language version. In this thesis work, additional results complementing these findings will be provided, confirming the same conclusions presented in this previous research work.

[5] http://sonivis.org

2.4 Information sources in wiki research

This section presents a brief summary of different information sources where interested researchers can find additional details about bibliography, conferences and tools in the wiki research area.

As a direct consequence of his inherent features, allowing its users to constantly update its contents, Wikipedia (and any other wiki project) can be well considered as a perfect example of a self-documented platform. Therefore, it is natural to find a comprehensive and complete compendium about Wikipedia research in Wikipedia itself [6].

This page serves as a central starting point from which to find references about research publications, papers, researchers, conferences, tools and any research activity regarding Wikipedia. It is a very complete repository about Wikipedia research, but unfortunately, it is still far from being exhaustive too. The main reason behind this lack of accuracy is the unstructured, dispersed, and unsupervised strategy to introduce and organize information in meta.wikimedia, which sometimes leads to replicated entries and comments, and most of times delegates to interested researchers themselves the task of updating the contents to reflect their most up-to-date research works (a task that, paradoxically, many of them do not undertake). On the other side, it provides useful pointers to other exhaustive information sources, which enhances its utility. This page provides the following sections and links:

- Research projects [7]: It tries to summarize relevant research projects about Wikipedia, though currently it is one of the most uncompleted pages of the whole set.
- Researchers: List and links to some researchers involved in the analysis of Wikipedia and wiki projects, from different points of view. It currently contains 43 indexed researchers, and again, it is far from being exhaustive.
- Research Resources: Undoubtedly, it is the most useful epigraph in this page. It contains several links to other information sources about Wikipedia and wiki research:
    1. *Wiki Research Bibliography* [8]: Comprehensive list of publications about Wikipedia and wiki research. Most of its entries provide links to the electronic version of each paper or publication.
    2. *Wikipedia in Academic Studies*: Alternative presentation of academic bibliography related to Wikipedia, organized in a table with columns *authors*, *title*, *conference*, *publication year*, *link to on-line version (if exists)*, *notes*, *abstract* and *keywords*. At the bottom of this page, it can be found links to thesis works, valuable articles in non peer-reviewed publications, reviews, books and books chapters, lectures, unpublished papers (available on-line) and additional external links about Wikipedia research.
    3. *Wikibibliographie ENCYCLEN*: Exhaustive database about Wikipedia and wiki research bibliography references (further description below).
    4. *Communities and groups*: Links to some research communities interested in this field, specially Wikimedia Research Network, the official group and mailing list for researchers interested in Wikipedia and Wikimedia Foundation projects.

[6] `http://meta.wikimedia.org/wiki/Research`
[7] `http://meta.wikimedia.org/wiki/Research/Research_Projects`
[8] `http://meta.wikimedia.org/wiki/Wiki_Research_Bibliography`

5. *Conferences and events*, mainly related to web technologies, collaborative projects, web semantics, wikis, computer human interaction and information systems, which have already published research works about Wikipedia.
6. Research tools, providing statistics, visualization methods and results to analyze Wikipedia (non exhaustive list).

However, Wikipedia is not the only one resource, neither the most complete one, to find information about Wikipedia and wiki research, specially regarding bibliography references. A link under *Research Resources* in `meta.wikimedia.org` Research page, lead us to **Wikibibliographie ENCYCLEN** [2]. This is a thorough, exhaustive and well-organized database, maintained by the *Veille Scientifique et Technologique - Institut National de Recherche Pédagogique*, at Lyon (France). As of April 2008, it includes *588 bibliographic references* to research papers and on-line publications and relevant essays about Wikipedia and wiki projects. Users can make complex queries, filtering results by author, relevance, timestamp, keywords and many more parameters. Filters can be also combined to obtain ever more refined results. Indexed bibliographic entries also include links to the on-line version of the publication (if exists and it is freely accessible), as well as any other relevant mention or comments about it.

In the same way, **Wiki Research Bibliography** [1] offers a very similar database containing bibliographic references about Wikipedia and wiki related research works. It presents an identical user interface to the one found in Wikibibliographie ENCYCLEN, though its contents are slightly updated in comparison with the French database site. Actually, both websites are based on WIKINDX [9], a free, collaborative tool to gather and manage bibliographic entries, quotations/notes and authoring references, designed either for single use (on a variety of operating systems) or multi-user collaborative use across the Internet.

As a result of this, collaborative wiki technologies, if applied with a minimum organized criterion to gather, organize and present its results, can be a very powerful tool to manage bibliographic databases. Just to get an approximate idea of the size of these databases, the 588 articles and references in Wikibibliographie ENCYCLEN can be compared with the 127 articles classified by the tag "Wikipedia" in **CiteULike** [10], which could be considered as another type of collaboratively built repository of bibliographic information, this time driven by bookmarks added by CiteULike users to their favorite papers. On the other side, **Google Scholar** [11], presents approximately 1,100 results having the word "Wikipedia" in the article title, many of which are just simply cite entries and not links to real scholarly documents.

To conclude, in 2006 the International Symposium on Wikis (WikiSym 2006) hosted a workshop on Wikipedia Research [12], moderated by Jakob Voss and Angela Beesley, showing the most relevant topics related to this area.

2.5 Conclusions: future trends in research on Wikipedia

Throughout this chapter, we have presented some of the most relevant research lines on Wikipedia and collaborative communities working with wikis. Large scale global projects like Wikipedia present to researchers a unique opportunity to study a novel phenomenon: the organization and activity patterns

[9] `http://wikindx.sourceforge.net/`
[10] `http://www.citeulike.org/tag/wikipedia`, consulted on April 13th, 2008
[11] `http://scholar.google.es`
[12] `http://www.wikisym.org/ws2006/program.html\#Workshops`

of a huge community, conformed by millions of users from many different cultures, collaborating towards the creation of the definitive encyclopaedic repository indexing the human knowledge. Although other collaborative communities have been studied previously, specially as far as FLOSS development initiatives is concerned, the size and impressive growing rate of the Wikipedia cannot be compared with any other project analyzed before.

We have shown that there are some preliminary results on the quantitative characterization of Wikipedia, in terms of the effort spent by authors in the project, the evolution of the size of the project (number of articles, length of articles), and useful techniques aimed to visualize this evolution over time. Broadly speaking, most of these studies undertake an exploratory analisis, looking for interesting activity patterns that can be identified within a certain language version. Almost all of them select the English Wikipedia as a case study, since it is the larger version both in the number of articles and the size of its community. As an overall conclusion, these previous publications show that the size of the English Wikipedia (and some other versions with a high number of articles) is growing exponentially (in number of articles, number of changes submitted to the project, size of the database, and so forth). Likewise, other research studies examined the inner structure of the Wikipedia community of authors, from a quantitative point of view. The main interest here is to demonstrate whether we can find in Wikipedia the perfect example of the so called "wisdom of crowds" effect, dynamizing the collaborative authoring process in each language version by means of the contributions received from a plethora of very infrequent contributors. However, some studies have shown that, even though the influence of these passing by authors is by no means negligible, a significant proportion of the authoring process in Wikipedia is performed by a small core of very active contributors.

Other studies have focused on the analysis of the quality of the content created in Wikipedia, specially focusing on the automation of the assessment process to identify articles with a high quality level from the rest of encyclopedic entries in a certain language version. Following a similar approach, some researchers concentrate on the automated evaluation of the reputation level of Wikipedia authors, based on the quality of the content provided by them. The general assumption here is that the higher the number of revisions that a certain contribution from one author remains unchanged, the better the quality of that contribution should be and thus the reputation of the author responsible for that contribution is improved. However, none of these preliminary studies has been able to reconcile the results obtained for the reputation of authors with the outcomes obtained from the assessment process performed by the community of authors to evaluate the quality level of each individual article. As a result, further research should be conducted in this field to explore more effective ways to identify potentially quality articles, and better explore the relationship between authors reputation and the quality of contents produced by them.

A third group of previous research works deals with the interpretation of the web graph conformed by Wikipedia articles, following the structure created by the internal links connecting pages within a certain language version. Alternatively, they also examine the social network conformed by the whole community of authors in a certain language version, or the subgroup created by a cluster of authors contributing around a set of more specific topics or categories. These previous studies show that the Wikipedia web graph is scale-free, leading to the identification of a preferential attachment process that drives the content authoring work flow, at least for the English language version.

Research on Wikipedia will continue for sure in years to come, since the need to better understand one of the biggest collaborative initiatives in the history of human race, and to discover new brilliant applications to exploit its vast content coverage will endure as well. In this context, the aim of this thesis work is to take a step forward on this thrilling research area, modelling the activity patterns of Wikipedia contributors found in quantitative data extracted from public Wikipedia database dumps.

In this sense, we can see that previous research works has been focused mainly of the English language version (because of its outstanding characteristics) or else, in a small subset of language versions. Therefore, there was a lack of comprehensive studies specifically aimed to search for similar quantitative patterns found in several language versions, to extract valid conclusions about general properties shared by all of them. We have focused on the top ten language editions of Wikipedia to provide this side-by-side comparison of our quantitative results.

In the same way, the absence of a rigorous statistical method to analyze the internal social structure of different communities of authors in Wikipedia at the same time prevented us to find common behavioral patterns that could be applied to describe large collaborative communities like these ones. Finally, most of these previous studies performed exploratory studies that simply looked for patterns, structures and traits that can be found in a certain language version, overlooking the implications of these findings for critical aspects like the sustainability of the Wikipedia production model in the following years. More concretely, there was no detailed information about the typical lifecycle of authors in Wikipedia, for how long they are expected to contribute to the project and the roles adopted by these contributors as they get more and more mature in the relationship with Wikipedia. In this thesis, we have tackled all these questions in order to find a more complete framework suitable for understanding the internal organization of the Wikipedia project from different points of view. Our main goal is to find out whether it is feasible to consider that Wikipedia will be able to maintain its outstanding growing rate in due course, or the project will face some limitations preventing it to keep on this expansion in terms of effort, number of contributors and number of encyclopedic entries created in each language version.

Chapter 3

Methodology

> "You know a conjurer gets no credit when once he has explained his trick; and if I show you too much of my method of working, you will come to the conclusion that I am a very ordinary individual after all". *A Study in Scarlet*, Arthur Conan Doyle, (1888).

This chapter introduces the methodology followed to develop this empirical analysis of the top ten language editions of Wikipedia. As it was shown in Chapter 2, there are many previous analyses focused on the Wikipedia community and its contents are based on empirical studies. However, most of these were focused on a certain language version, or a specific subset of pages or authors. The same practical approach is followed here to model the contributions performed by the members of the Wikipedia community, applying common Exploratory Data Analysis techniques to illustrate the general features of Wikipedia and its community of authors. However, previous research works are extended here, presenting novel statistical techniques that can be successfully employed to gain knowledge about Wikipedia dynamics. Moreover, all these techniques are applied to the complete database dumps of the top ten language versions of Wikipedia, in the first comparative analysis of such kind ever implemented, so far.

The next section is an introduction to the overall framework of this thesis work, providing a high level overview of our workflow, as well as some basic implementation aspects. Then, we introduce some general features of Wikipedia dynamics identified during the Exploratory Data Analysis stage in our workflow. After that, WikiXRay, the tool we have created to automate the quantitative analysis of Wikipedia database dumps as well as to produce all the results included in this thesis, is presented in detail. Later, we focus our attention on effective statistical models and tools suitable for being applied to characterize the internal organization of Wikipedia and its community of authors. To conclude, we analyze the evolution of the Wikipedia community of authors over time, which will led us to conclude several conditions that must be satisfied to ensure the sustainability of the project in the following years. Throughout this presentation, we will identify the specific research questions to be tackled in each section of our study. Likewise, we provide in-depth definitions of the corresponding metrics and terms applied in the study, as well as the graphics and results to answer every question along this path.

3.1 General overview of the methodology

In this section, a general overview of the methodology applied in this analysis of the top ten language versions of Wikipedia is offered, including:

1. A description of the data sources from which is obtained empirical data about the community.
2. A precise definition of recurrent terms that we will refer to in subsequent sections of this thesis.
3. Some remarks about implementation aspects affecting our analysis, that should be taken into account by other researchers trying to perform similar studies on their own.

At the end of this section, the reader should have a clear picture of the general roadmap followed in this empirical analysis, as well as a solid understanding of the basic aspects that define the baseline theoretical framework for our research work.

3.1.1 Data sources

The basis of the analyses performed in this work is the extraction of the complete history of changes performed on Wikipedia pages. That information is ready to be retrieved from the *Wikimedia Downloads* center [1]. On that page, the *Database XML and SQL dumps* section provides access to the web page containing the current status of the database dump process for each language edition.

This service provides public access to different versions of the content of tables in the MediaWiki database, for each language version. Hence, using these interface we can download the list of users with special privileges in the system, the current list of articles in a language edition, the dump containing only the last revision for every article, or even the log of special actions performed in the system (like uploading files or images).

For our purposes, we need to retrieve only two files from the set of available dump files in each language version:

- `pages-meta-history.xml.7z`: An XML file containing the full history of all changes performed on every wiki page in that language edition.
- `user_groups.sql.gz`: An SQL file, containing the list of all users with special privileges in the system, along with their corresponding roles.

The first one is provided in XML format, since this is the standard dump style for any file containing information about contributions to wiki pages (*revisions* in Wikipedia jargon). This XML format follows the structure presented in Table 3.1.

[1] `download.wikimedia.org`

```
<mediawiki xmlns="http://www.mediawiki.org/xml/export-0.3/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.mediawiki.org/xml/export-0.3/
http://www.mediawiki.org/xml/export-0.3.xsd" version="0.3" xml:lang="fur">
  <siteinfo>
    <sitename>Vichipedie</sitename>
    <base>http://fur.wikipedia.org/wiki/Pagjine_princip%C3%A2l</base>
    <generator>MediaWiki 1.14alpha</generator>
    <case>first-letter</case>
      <namespaces>
      <namespace key="-2">Media</namespace>
      <namespace key="-1">Speciâl</namespace>
      <namespace key="0" />
      <namespace key="1">Discussion</namespace>
      <namespace key="2">Utent</namespace>
      <namespace key="3">Discussion utent</namespace>
      <namespace key="4">Vichipedie</namespace>
      <namespace key="5">Discussion Vichipedie</namespace>
      [...]
    </namespaces>
  </siteinfo>
 <page>
    <title>Pagjine principâl</title>
    <id>1</id>
    <restrictions>edit=autoconfirmed:move=autoconfirmed</restrictions>
    <revision>
      <id>1</id>
      <timestamp>2005-01-25T06:55:26Z</timestamp>
      <contributor>
        <ip>24.251.243.233</ip>
      </contributor>
      <text xml:space="preserve">'''Benvignût al Vichipedìe furlan!'''
      [...We find here the text of the revision]
      </text>
    </revision>

    </revision>
    [...Rest of revisions in this page...]
</page>
<page>
[...Rest of pages in this dump...]
</page>
</mediawiki>
```

Table 3.1: Excerpt of XML format in Wikipedia dump files (sample taken from `furwiki-pages-meta-history.xml`)

The dump begins with some information about the XML schema employed, the name of the site, and the corresponding identifiers for each namespace in that language version (which may be a very useful information to perform filtering actions later on). In the following lines, the actual dump proceeds hierarchically, presenting the information for every page, then for every revision in a certain page. Attributes and subsequent sub-elements in each node are indented to the next level, as usual. A clear inconvenient of this format is that it does not enforce unique XML tags. Thus, we may come across the same `<id>...</id>` pair of tags in different levels, identifying either the user ID, the page ID or the revision ID, for instance. To solve this problem, our software tool has been programmed to infer the precise meaning of identical tags, based on the contextual information provided by the current reading position within the dump file.

To give a rough idea of the sheer size of the database dumps processed in this thesis, Table 3.2 summarizes some descriptive parameters of the dumps of every language included in this study. We refer to subsection 3.1.2 for a precise definition of the metrics presented in this table. All measurements have been computed from the dump archives retrieved, except for the official count of the total number of articles, retrieved from the Wikipedia stats page maintained by Erik Zatche [2]. The English version is the biggest one, with more than 2,1 million articles. The 2 million mark was surpassed on September 9, 2007 [3], with an article entitled "El Hormiguero" about the popular Spanish TV show of the same name.

Table 3.2: Summary data for the top ten language versions of Wikipedia, as of end of December, 2007. The top ten list has not suffered notorious modifications until present time. Perhaps the most relevant is the introduction of the Russian version, in the 10th position as the time of this writting, replacing the Swedish version.

Language	Num. wiki pages	Num. articles (off. count)	Num. logged users	Num. revisions
EN	11,4M	2,18M	1,82M	167,4M
DE	1,92M	695K	226K	37,3M
FR	2,37M	597K	127K	25,8M
PL	811K	456K	51K	10,4M
JA	1,16M	453K	90K	17,5M
NL	936K	391K	60K	10,6M
IT	1,17M	392K	62K	12,8M
PT	1,37M	346K	64K	8,9M
ES	969K	311K	132K	14,2M
SV	613K	266K	27K	5,5M

A word of caution is in order here. Because of the large duration of the dump process for the biggest editions of Wikipedia, available information for each language extends until different final

[2] (`http://stats.wikimedia.org`). Since there was no available data on that web page for the English Wikipedia, we estimated the total number of articles as of December 2007 from our own sample data, directly extracted from the database dump file

[3] `http://wikimediafoundation.org/wiki/Wikipedia_Reaches_2_Million_Articles`

dates. Moreover, the dump process of the largest language versions has experimented some problems over the past 2 years, precisely due to the huge size of the database for each language version, along with some technical problems derived from the software architecture developed in the system. As a consequence, the last available dump for the English version of Wikipedia was retrieved on the first week of March, 2008. The duration of the dump process for this version exceeded one month, and the timestamp for the last logged operation recorded is `2008-01-06 02:18:52`. As the time of this writting, a new dump version for the English Wikipedia is not available yet. Therefore, although available data in other versions surpasses that limit date, we would not be able to compare the top ten languages in a uniform way. We decided to restrict our analysis up to the last complete month available in all these versions, which resulted to be December 2007.

The process of retrieving the appropriate information, load a database and undertake the necessary pre-processing tasks to facilitate subsequent analyses soon became a time-consuming task. To deal with this problems, we have created an efficient software package to automate the whole analysis process. Its name is *WikiXRay*, and those readers interested in the implementation details of this tool are referred to section 3.2 to find additional information. On subsection 3.1.3, we comment only on those implementation aspects that are mandatory to understand our working methodology.

3.1.2 Definitions

Throughout this thesis work, we will refer to several terms and concepts that are fundamental to understand the workflow of our empirical analyses. To improve readability and facilitate comprehension of subsequent descriptions and discussions, we provide here a definition for those key terms, so that the reader should refer to this section at any time to clarify their precise meaning. Whenever a concept has been already defined within the Wikipedia community (like data elements describe in the MediaWiki database design) we try to stick to that existing definition as much as possible [4].

Namespace : Each of the logical areas in which the content of any wiki based on MediaWiki is classified. The database stores, for each wiki page, a numerical identifier indicating which namespace it belongs to. For the purpose of this analysis, we will focus primarily on pages stored in the `main` namespace, which corresponds to Wikipedia articles. Whenever a different namespace is consider in our analysis, it will be explicitly indicated. Thus, unless it is stated otherwise, we assume by default that all results apply to articles stored in the `main` namespace. Please, refer to Table 3.3 later on in this chapter for a complete list of the namespaces that will appear in this study.

Page : Any wiki page, disregarding the namespace in which it is stored in the system database, whose information can be edited by a Wikipedia contributor. This includes encyclopedic articles, user pages, discussion pages associated with each article (`talk_pages`), etc. Any page can be ***uniquely*** identified by their corresponding page ID in the database.

Article : Wiki page containing information of encyclopedic articles in a certain language edition. Articles are stored under the `main` namespace in the MediaWiki database of the corresponding language version.

[4] In Appendix A the reader will find a more complete glossary of terms used throughout this thesis

Featured Article (FA) : An article that has deserved to be nominated as one of the top quality articles produced in a certain language edition of Wikipedia. The nomination takes place after an exhaustive reviewing process performed by all interested members of the community, upon an open call issued by the corresponding responsible in that language version. Candidate articles are proposed by community members to enter this reviewing process. The result of a voting process, reflecting the opinions of all reviewers involved, determines whether the article is promoted to this new status or not. The promotion is not permanent, that is, as soon as community members detect that the quality of the article has lowered, they can suggest the article as a candidate for a new reviewing process. In case that the FA does not pass this examination, it can be demoted again to its original non-FA status.

Redirect : A special type of article, with no content at all, simply providing alternative encyclopedic entries for a certain term.

Stub : An article considered so short as to be considered as a useful encyclopedic article. Stubs are usually new articles recently opened, providing a seed upon which to create a longer, more complete encyclopedic entry. They are usually marked as such using special *templates*, customized in each language edition (sometimes, even for distinct categories in each language version). There is no official policy regarding the minimum length an article must attain to avoid being considered a stub.

Talk page : Wiki pages containing discussion about the contents of an encyclopedic article in a certain language version. Each talk page is presented next to its corresponding article in the MediaWiki interface. However, newly created articles does not automatically come with a talk page, so it may or may not exist (until a user decides to create it). They are stored under the `talk_page` namespace, with the exception of talk pages associated to *user pages* (see below), which are stored under the `user_talk_page` namespace.

User page : Wiki pages presenting information of a *logged author* in a certain language edition (see below). They are stored under the special `user_page` namespace in MediaWiki.

Revision : Any individual modification on a wiki page in a certain language edition of Wikipedia, that is registered in the database as such and identified by a ***unique*** numeric ID.

Author : An individual who belongs to the community of a certain language edition in Wikipedia, and who performed at least one revision in that language edition[5]. An author is identified by a numeric ID in the database, associated to every revision attributed to her.

Logged author : Any author registered in a certain language edition of Wikipedia, by creating a user account. Logged authors can be ***uniquely*** identified by either their user identifier (`rev_user` field) or their login (`rev_user_text` field) in the `revision` table of the corresponding database. Therefore, authors must log in the system before performing revisions, to let the database register their identity.

[5]Otherwise, it will not appear in the corresponding table of the database, registering each revision performed in the system. For privacy reasons, access to the database table containing the full user list in each language version, along with sensible information like email accounts and so forth, is not allowed

Anonymous author : Any author who does not create a user account in a certain language edition of Wikipedia, and performs revisions under anonymous identity. Anonymous authors are identified in the database by a common user identifier `rev_user` $= 0$ and the IP address from which the author contacted the system, which is stored in the `rev_user_text` field, both corresponding to the `revision` table in the database.

As a consequence of the previous definition, *anonymous authors* can not be individually traced from the database dumps. That is, whenever an anonymous user performs a revision, that revision is associated with the common identifier `rev_user` $= 0$. Despite that, we may think that the IP address stored in the `rev_user_text` field could allow us to uniquely trace individual users. This is not feasible, though. The application of certain networking technologies (like *proxies* or *Network Address Translation*) can masquerade an undefined number of individual users behind the same IP address when connecting to the system. As a result, anonymous users *will be consistently filtered out* throughout this thesis work, unless it is stated otherwise.

In the same way, the unique login and numeric identifier of a Wikipedia author is linked to the author's account in a certain language edition, and it is not shared among multiple accounts corresponding to the same individual in different versions. Due to this limitation, we cannot trace revisions of the same author in different version of Wikipedia. Likewise, we can not rely on finding the same login identifier in another version as an evidence that the same author is contributing in that language edition, either. It is perfectly possible that another author may have chosen the same login identifier [6].

Privileged author : Any logged author who received certain special privileges within the system, which are stored in the `user_groups` table of the database. Tables 1.1 and 1.2 already presented the different user levels that we can find in Wikipedia. Unless stated otherwise, whenever we refer to *privileged authors* in this analysis, we focus exclusively on the group of *administrators* of a certain language edition of Wikipedia.

Bot : Small software programmes that performs revisions on a certain language edition of Wikipedia in an automated way. Many bots can be uniquely identified due to their special privileged status `'bot'` associated with their correspondent `rev_user` unique identifier. This relationship is reflected in the `user_groups` table in the database. Since bots are not real human users, *we systematically filter them and their revisions* in this analysis, unless we indicate otherwise. Again, we need to be cautious here. It is known that not all existing bots in Wikipedia have been identified in this table, as it should have been. As a consequence, some bots may have been introduced in the subset of logged authors, which is the main target of our empirical analysis. We have tried our best to filter out some clear cases of supposedly human authors that, in fact, resulted to be bots. Nevertheless, the exceptional productivity of some very active authors may lead us to confuse them with bots, and whenever in doubt, we decided it was better not to compromise our sample with incorrect deletions. All the same, these are so infrequent cases as to have any significant effect on the validity of our results.

[6] This has been specially problematic for very active and well reputated authors in some language versions, who have witnessed how their famous logins in the community has been adopted by individuals in other Wikipedias, sometimes with pernicious purposes. The MediaWiki development team is now actively working on a new unified login system, that will solve many of these problems, providing common credentials for the same author in all language versions of Wikipedia. This might lead to a completely new research line, allowing us to compare the contributions of individual authors on different language versions of Wikipedia, an unfeasible task at the time of this writing

Birth : This is a term introduced in the survival analysis of the Wikipedia community of logged authors. We consider that a new birth has occurred in the community of logged authors in a certain language version, when a new author contributes with a revision for the first time in the history of that language version.

Death : Again, this term is introduced in the survival analysis of Wikipedia logged authors, for each language version. We consider the death of a logged author as the time at which this author performed her last revision in the history of that language version, and thus, she never come back again to contribute (as far as the activity registered in the database dump file can show).

3.1.3 Additional implementation details

To undertake this quantitative analysis of the evolution of Wikipedia and its community of users, we have retrieved the corresponding database dump files of the top ten language editions of Wikipedia, according to their official number of articles. The top ten list is always displayed at Wikipedia's main page `http://wikipedia.org`. The official count of the number of articles in every language edition introduces slight additional requirements to include an article in the grand total [7]. It must fall in the `main` namespace, redirects are not considered and the article must also include one internal link. All the same, we realize that this criterion (choosing the top ten Wikipedias) may not be the fairest one, since many of the largest editions may have many stubs, for instance. However, it is difficult to establish an equitable criteria to select the most "popular" editions, since other metrics have their own downsides. For instance, ordering the language editions by their total number of logged authors would not probably reflect the popularity of each version, either, since many authors, even logged ones, does not perform a high number of revisions. Furthermore, the order established with the official article count is the same one obtained if we list the Wikipedias ordered by their number of internal links (excluding redirects), which is traditionally considered as a fairer popularity metric.

The automation of the data mining process of a huge data repository like the one considered in this thesis is crucial to successfully achieve our goals. As we previously mentioned, in the following section the reader may find additional information about inner implementation details of our tool, WikiXRay, and the different sub-processes involved in this study. In this subsection, we discuss only those implementation details that are required to understand the subsequent analyses and results presented in this thesis.

Table 3.3 summarizes the principal namespaces adopted in the Wikipedia database to organize wiki pages. We highlight in this table those namespaces that will be of interest in our analyses. Since our study is mainly focused on the Wikipedia community of authors, we will concentrate on those namespaces reflecting critical activities within the community: encyclopedic content development (`main`), discussions about articles (`talk`), and pages containing users profiles (`user`), as well as their corresponding discussion pages (`user_talk`).

[7] `http://en.wikipedia.org/wiki/Wikipedia:What_is_an_article\%3F`

Table 3.3: List of most relevant namespaces in the Wikipedia database (for each language edition, though we present the English version names)

Namespace ID	Namespace
-2	Media
-1	Special
0	**Main (blank name)**
1	**Talk**
2	**User**
3	**User_talk**
4	Wikipedia
5	Wikipedia_talk
6	Image
7	Image_talk
8	MediaWiki
9	MediaWiki_talk
10	Template
11	Template_talk
12	Help
13	Help_talk
14	Category
15	Category_talk

WikiXRay organizes information retrieved from the database dumps of each language edition in tables, stored in a local MySQL database. In the first step of the analysis, the database dumps are parsed and the information loaded into 4 different tables. Table 3.4 presents a brief description of these baseline tables, from which we will extract information to perform our statistical analyses. Actually, these baseline tables are exactly the same database tables defined in MediaWiki, and so are most of their respective fields and data types (though we have defined some new additional fields in WikiXRay, to support some of our own analyses). Table names are displayed in bold, while the name of fields in each table are displayed in italics. Since each language edition in Wikipedia uses its own database (with a separate list of users, pages and so forth) we have to recreate those tables for each language edition analyzed.

Table/Field name	Description
Page	**It stores information for every wiki page created in a certain language edition**
page_id	Unique numeric identifier for every page
page_namespace	Numeric identifier of the namespace in which the page is stored
page_title	The title of the page (string)
page_latest	The numeric identifier of the last revision registered for that page

Table/Field name	Description
page_len	Length in bytes of the page (filtering out wiki markup and additional HTML tags)
page_is_redirect	A flag value indicating whether the page is a redirect
page_is_stub	A flag value indication whether a page is a stub
page_random	Random number assigned to each page to facilitate random sampling
page_is_new	A flag value indicating whether the page is new
page_restrictions	Stores any editing restrictions imposed to the page (e.g. protected or semi-protected pages)
Revision	**It stores information for every revision performed in every wiki page in a certain language edition**
revision_id	Unique numeric identifier for each revision
revision_page	Numeric identifier of the page on which this revision was performed
revision_user	Numeric identifier of the author who performed the revision (0 for anonymous users, their unique ID for logged users)
revision_user_text	In case the revision was undertaken by an anonymous author, it stores the IP address from which the user connected to the system; if the revision was performed by a logged author, then it stores the unique login of the author (both values stored as strings)
revision_timestamp	Timestamp value of indicating when it was registered this revision in the database (in format `YYYY-MM-DD HH:MM:SS`)
revision_len	Length (in bytes) of every revision (filtering out all wiki markup and additional HTML tags)
[*revision_num_letters*]	Number of letters in the article, considering only readable text (thus, excluding wiki tags and other HTML code)
[*revision_num_words*]	Number of words in the article, considering only readable text (thus, excluding wiki tags and other HTML code)
[*revision_num_inlinks*]	Number of internal links (those pointing to other wiki pages in the same language version) in the article
[*revision_num_outlinks*]	Number of external links (those pointing to other wiki pages outside this language edition, or to other web pages in the Internet) in this article

Table/Field name	Description
revision_parent_id	Numeric ID of the previous revision performed on this article (allowing us to construct the complete *revision tree*, with a chronologically ordered list of all revisions undertaken on a certain article).
revision_is_redirect	A flag value indicating whether, at this revision, the article is a redirect
revision_is_stub	A flag value indicating whether, at this revision, the article is a stub
revision_minor_edit	A flag value indicating whether the author marked the *minor edit* checkbox before confirming the revision
revision_comment	A text value storing the comments of the author who performed this revision (if any).
User_groups	**Provides information about special privileged authors in a certain language edition (bots, administrators, etc.)**
ug_user	Unique numeric identifier of the author
ug_group	Special privilege granted to that user. It must correspond to one of the special privileges listed in Tables 1.1 and 1.2; administrators are given a `sysop` value in this field

Table 3.4: List of baseline tables in WikiXRay (same as in MediaWiki), along with their corresponding fields and accompanying descriptions.

We will frequently refer to these tables and fields to describe the data sources employed in each individual analysis undertaken in this thesis work. Nevertheless, we will not calculate values for all fields presented in Table 3.4. Currently, WikiXRay is capable of calculating all these values, but some of them (marked in squared brackets in the Table 3.4) entail complex and time-consuming operations. This prevent us to efficiently execute those tasks on large language versions, like the ones included in this thesis work.

After obtaining the relevant data sets from the database dumps, and organizing these data sets in our local database, we need a statistical software package implementing (if possible) all the analyses and techniques required for this thesis work. WikiXRay utilizes *GNU R* [87], a powerful and easy-to-use statistical package which is also libre software (released under GNU GPL license). Among the most important advantages of R, from a practical perspective, we can cite the following:

- Massive coverage of almost any conceivable statistical technique for data analysis, thanks to the massive package library, (the Comprehensive R Archive Network, abbreviated CRAN [8]), populated with pre-built packages ready to be plugged in the baseline installation to expand its capabilities. It currently provides more than 1,000 different statistical packages, also released under GPL license.

- Simple syntax, concise format of functions implementing even the most complicated analyses.

[8] `lib.stat.cmu.edu/R/CRAN/`

- Support for crafting high quality graphics, even advanced types of pictures like 3D-mesh, 3D-surface, and trellis-like graphs.

- Seamless integration with LaTeXenvironment, through a technique called *literate programming*, allowing us to create automatic, professional reports summarizing our statistical analyses and findings [9].

Regarding the statistical techniques applied to the examination of the community of authors in each of the top ten Wikipedias, 3 different aspects will be explored in this thesis:

- *Social structure*: The first step in our analysis of the Wikipedia community of authors is to find indications of social structures and organizational patterns. These structures may have emerged spontaneously, or as a consequence of explicit organizational guidelines imposed by the community itself.

- *Demographic analysis*: The Wikipedia community of authors can be viewed as a virtual social group of human contributors who first join the project, then perform their content creation activity over a certain period of time, and eventually leave the project, due to quite disparate reasons. This process can be modelled in demographic terms, looking at the births and deaths rates, and estimating the *expected survival time* of users in a certain language edition. In this way, we can delimit the typical lifecycle of Wikipedia users, a useful information that will complement other aspects of this study.

- *Author reputation*: The third approach of our analysis is the study of author reputation in Wikipedia. In previous chapters, we have seen that the production of quality contents in Wikipedia is becoming a matter of concern of both researchers and practitioners. We apply a simple, feasible set of metrics proposed by Stein and Hess in [108] to analyze the reputation of authors in Wikipedia. We also explore the adequacy of this metrics to forecast which articles in each language edition have the highest probability of being promoted to Featured Articles in the near future. We also examine differential metrics of FAs with respect to non-FAs, that may have implications about the future capability of Wikipedia to create more quality articles.

To conclude this section, we recapitulate the standard guidelines that we will apply to our data sets prior to perform our analyses. With the exception of the general analysis of global features of Wikipedia (which is outlined in Section 3.3), in all subsequent analyses we apply the following conditions unless we state otherwise:

1. All statistical procedures and metrics will be applied exclusively to the set of revisions performed on Wikipedia articles, in each of the language editions under study.

2. We will consistently eliminate redirects and their associated revisions from the data sets.

3. Revisions from bots will be systematically filtered out from our analyses.

4. Revisions from anonymous users will not be considered in our analyses.

Nonetheless, in case we would need to include any of the aforementioned subsets in a certain study, we will clearly indicate so in the corresponding introduction.

[9] `http://www.stat.umn.edu/~charlie/Sweave/`

3.2 WikiXRay

WikiXRay [10] is a software tool that automates the quantitative analysis of the public database dumps available for each language version of Wikipedia [11]. The automation of this process, specially the initial steps of retrieving and parsing the huge database dump files to extract relevant information that could be stored in a local database for easy management, was a long-awaited request for many researchers on Wikipedia over the past years. Therefore, once I identified this need it was clear that one of the main contributions of this thesis should be a software tool aimed to:

- Resolve the difficulties and technical challenges imposed by the efficient processing of Wikipedia database dump files.
- Eventually construct a public repository with pre-compiled quantitative information for each language version of Wikipedia, ready to be used by any researchers interested in avoiding the complexities of the initial steps required to extract the relevant information.
- Facilitate the reproducibility of numerical results, graphs and statistical procedures performed in this thesis, as well as to offer an easily extensible tool that could integrate new future modules implementing novel statistical techniques or different analyses on quantitative data from Wikipedia.

Presently, the tool has succeeded to accomplish the first and third goals in the above list, while still further development work has to be performed to automate the creation and frequent update of a public repository with pre-compiled information for each language version. Nevertheless, the autonomous functional behavior exhibited by the current version of the tool paves the way to accomplish this remaining goal without significant development effort.

Figure 3.1 presents the outlines of the functional architecture followed in the development of WikiXRay. According to this schema, we can see that the process is quite straightforward, though the most interesting features are precisely those related to the actual implementation procedures to circumvent the obstacles found inside each step. The whole process begins specifying the language version (or set of language versions) the user wants to analyze. In the next step, the tool start to retrieve all needed files from the Wikimedia download website, storing them in the local disk. Once the required dump files have been store locally, the tool launches the first critical step found in this workflow, namely the uncompression and parsing of the dump files. In the current version, we need to retrieve only two files for each language version processed:

1. `pages-meta-history.xml.7z`: This is an XML dump file containing the complete information for all activity registered in the database about every single revision performed on every wiki page in that language version. The file is compressed using 7zip for better efficiency of the compression algorithm, as well as to reduce the final size of the archive (between 10-100 times lower than the original one, thanks to the really high efficiency of this compression algorithm when applied to plain text files).
2. `user-groups.sql.gz`: SQL file containing the individual dump for the *user groups* table in the database of that language version. The file is compressed in gzip format before publication, since its size does not usually represents a problem in terms of required transfer time over the Internet.

[10] `http://wikixray.berlios.de/`
[11] `http://download.wikimedia.org`

The first of these dump files is by far the most problematic one, because of its huge size. Just to give an illustrative example, the uncompressed dump of the English Wikipedia exceeds the 2 TB limit, something that prevents even the local storage of this file without a high capacity server. Moreover, the efficient parsing of this file to extract relevant quantitative information for each page and revision is by no means a trivial task. In fact, to overcome these problems a completely new parser was developed from scratch, in order to provide a performance-wise alternative to other available options at that time.

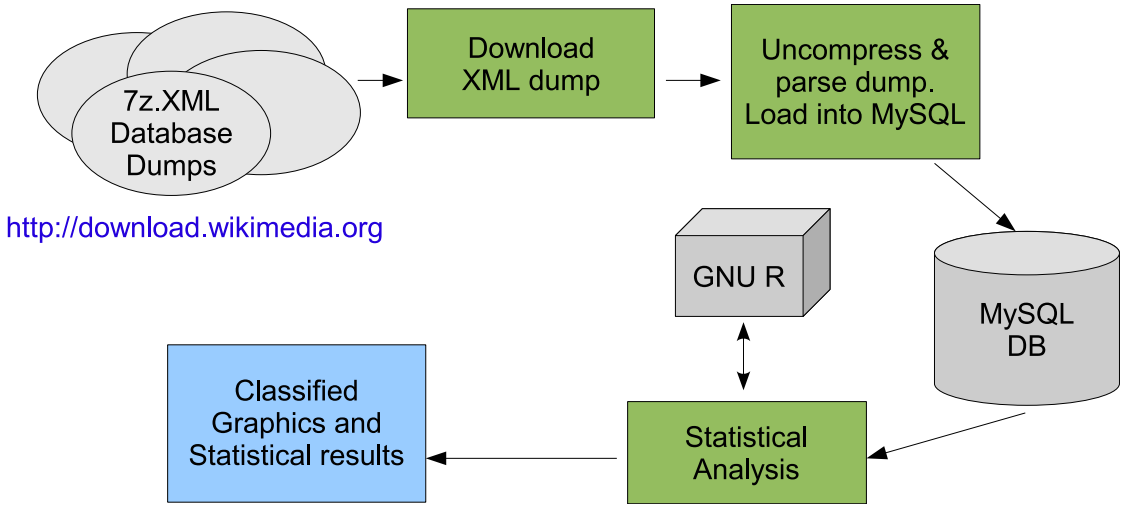

Figure 3.1: Schematic respresentation of the functional architecture of WikiXRay

The parser [12] has been built to process the information flow directly as it comes from the uncompression utility, compute and extract all relevant quantitative information on the fly, store the information in a local MySQL database, and then delete all processed data chunks to avoid running into problems regarding the storage capacity of the underlying hardware system. This approach has several drawbacks in terms of performance, but nonetheless, it is still more efficient than other available options, also computing information that we can not find in other solutions (like the length in bytes of every revision after filtering all templates and superfluous metadata). The parser has been built so that it can be used as an independent tool, either to reproduce the standard usage to retrieve only relevant summary information from each revision/page, or else, importing the complete text for every revision to create an exact replica of the database for that language version.

Once the parser module has finished to extract all relevant quantitative information from the dump file, a new module is executed to build new tables in the database to make it easier to perform several specific types of statistical analyses (including evolution over time of relevant descriptive metrics, and the survival analysis presented in this thesis). When this module concludes its execution, we obtain a local MySQL database ready to develop data mining analyses on this language version of Wikipedia.

Finally, a set of different quantitative analyses and statistical techniques are applied on this local database. Each individual module acts as a plug in, providing a new set of analyses to be undertaken. A central module is in charge of monitoring the execution of each individual plug-in, reporting any anomalies that it may encounter during the execution. All analysis modules have been prepared to organize the numerical and graphical outcomes of each analysis in a convenient way, classifying resulting files in an organized hierarchy of directories in the local filesystem. In general, all current modules make heavy use of the statistical software package *GNU R* [87] to implement all statistical

[12] `http://meta.wikimedia.org/wiki/WikiXRay_Python_parser`

analyses and techniques efficiently. Currently, numerical results are provided in summary text files, with appropriate names to identify the statistical analysis and the language version they come from. Graphics are created in EPS format, which is suitable for either direct display or direct use in scientific publications that must meet high quality standards. The GNU R scripts can directly access the local database to retrieve the data for these analyses, through the database connection API provided by the *RMySQL* library.

The current version of WikiXRay provides the following implemented analyses:

- General statistics: Detail analysis of the evolution per month of key parameters describing the activity of authors, the activity registered by wiki pages, their length and distribution on different namespaces, as well as a detailed analysis of descriptive parameters per namespace, and statistics broken down by different types of contributors (logged authors, anonymous authors, sysops, bots, etc.).

- Social structure: It includes all the statistical analyses presented in the corresponding section of the same name in chapter 4. The analyses comprise the application of a number of techniques to obtain the best fitted distributions for key descriptive parameters informing about the distribution of contributions among authors and articles. The statistical libraries applied in this module include some scripts provided by Clauset *et al.* to accompany their article on this topic [25], as well as some libraries retrieved from the Comprehensive R Archive Network (CRAN) repositories (*Hmisc* and *MASS*). The last one is provided by Venables and Ripley to illustrate their book on statistical methods available in GNU R [120].

- Inequality: This module implements the analysis of the inequality level of contributions from authors and revisions received per article, allowing us to understand the inner distribution of effort among the population of authors and articles. This module makes use of the *ineq* library for GNU R, retrieved from the CRAN repository.

- Demography: This module comprises all survival analyses presented in section 4.4 of this thesis. We need the *survival* and *muhaz* libraries for GNU R to implement these analyses in an efficient way.

- Reputation and quality: A module implementing some basic metrics focused on discovering common statistical patterns of FAs, as well as trying to measure the reputation of authors using the method previously proposed by Stein and Hess [108].

- Evolution: This module undertakes a more in-depth analysis of the evolution in time of the statistical distributions fitted to our empirical data in the social structure module. It also performs the 3D analysis of the evolution of contributions from the core group of very active users in each month over the remaining history of each language version. Finally, the module also produce a graph for the evolution in time of the monthly Gini coefficient for contributions from logged authors in that language version.

WikiXRay has been written in the Python programming language, and it follows a completely modular approach, with each module implementing a clear subset of tasks as independent as possible from the remaining modules of the architecture. Moreover, the current architecture also facilitates the easy addition of new modules implementing different statistical methods or new analyses, since the tool searches for available plugins and there is a special module which centralizes the operation of launching new add-ons once detected. As well, the tool offers enough flexibility to specify all

necessary options by means of a central configuration file, that is read in just after we launch the program from the command line. Nevertheless, in order to avoid adding too much complexity to the configuration process, most of the internal parameters for each module already provide sensible default values, which are valid for the standard procedures usually performed in a wide variety of situations. Interested readers can find additional information about performance exhibited by the current version of WikiXRay in the corresponding wiki page on [13].

3.2.1 Design aspects

In this section, we provide some background information to understand the design and strategies adopted to implement WikiXRay. This information could be useful for researchers and developers confronting the same problems in similar applications, as well as for developers interested in expanding their knowledge about the inner organization of WikiXRay. However, those readers interested solely in their role as standard users to work with WikiXRay can safely skip this section, since they will find detailed information on the installation and use of WikiXRay on the project page in BerliOS [14].

The baseline database schema utilized in WikiXRay replicates the database schema defined in MediaWiki, with some new columns to store additional quantitative information for each page and revision. Table 3.4 already presented a summary of the different fields included in each table. In addition to this, we have selected the MyISAM storage engine to build all our tables in MySQL, since it is expected that we will run read-only queries once we have populated the table with quantitative data from the dump files.

Regarding the parsing process and the insertion of data from the dump files, several considerations were taken into account to design the current version of WikiXRay. In the first place, we have used a SAX compliant Python library to implement the XML parser, since DOM libraries impose the requirement of loading all information about XML nodes found in the document in memory, and thus it is not suitable to implement the parser due to the huge size of our uncompressed XML files. The parser filters out all undesired information pieces from the text included in each revision, applying regular expressions. It also has to deal with some problems regarding the non-unique tag identifiers utilized inside the XML dump files, constructing a dynamic cacheé to identify the precise meaning of each tag without consuming too many resources from the system. Finally, the parser employs extended inserts to insert larger data packets into the database tables, allowing us to boost the performance of the tool by means of adjusting the size of these chunks.

Finally, we cannot stress enough the importance of an adequate fine tuning of the configuration variables for our MySQL server. Most of the baseline and intermediate tables created by WikiXRay make heavy use of indexes built on select field to improve the execution performance of our statistical analyses significantly. However, the best combination of values for these configuration parameters always depend on the actual hardware resources exhibited by our underlying server infrastructure. As a general recommendation, we should allocate as much memory space as possible for MyISAM key cachesés, and also increment the allowed maximum size of the sort cache and the query cache. In case we select a large value for the size of the chunks employed in our extended inserts (to populate the database tables initially) we may also have to adjust the size of the max_allowed_packet variable in order to avoid transfer errors due to an excessively large size of the information packet sent to our server.

[13] http://meta.wikimedia.org/wiki/WikiXRay_Performance
[14] http://wikixray.berlios.de/

3.2.2 Roadmap for future improvements

Even though the current version of WikiXRay offers a complete set of features that will address a broad range of typical analyses that can be performed on Wikipedia dump files, there exist a number of potential improvements that should be undertaken in future development of this tool. These improvements include supporting new statistical analyses, extracting new information from the dump files and adapting or redesigning current working algorithms to leverage the performance of some critical modules (like dump uncompression and parsing). The following list includes some of the most important development tasks that should be undertaken in the following months for future releases of WikiXRay:

1. Include support for alternative DB engines: Currently, WikiXRay only supports the creation and usage of MySQL, and specifically MyISAM tables. It would be desirable to include support for alternative engines (InnoDB) or even alternative DB servers (like PostgreSQL), to broaden the potential audience of users. In addition to that, this would permit us to compare the performance exhibited by the tool with different configurations and DB engines, facilitating the selection of the most appropriate combinations to boost the performance of the tool.

2. Standarization of output dump files: One of the goals of WikiXRay is to create a local MySQL database with quantitative information extracted from dump files. However, we can also dump this local database and export this information, in order to allow other researches to avoid the initial, time-consuming parsing process. Thus, it would be preferable to standardize a common data model for similar tools aimed to analyze Wikipedia dump files, in order to favor the interoperability between the outcomes produced by each tool.

3. Parallelization of uncompression and parsing: One of the top priority requests for future releases of WikiXRay is parallelizing the uncompression of dump files, and then process each independent data flow in different instances of the parser. This would help to reduce the time needed for this stage, one of the most problematic ones regarding this. Nevertheless, it would be difficult to parallelize the uncompression process until we can extract the original XML file for the first time, since uncompression with 7zip or any other data compression program is relatively slow. One possible option would be to slice the original big XML file in chunks, then compress these chunks again (for storage efficiency) and finally upload the set of chunks in a public repository, so that anyone else can speed up the process again in the future.

4. Extracting per-revision link and content information: In the current version of WikiXRay, a very simple process of the actual content in each revision is undertaken, in order to filter out all non-text information to calculate the actual size of the revision. Future versions should try to identify the links stored in each revision (both internal and externals), together with additional info about relevant metadata (like templates) included in each revision. Information about links could open the doors of advanced research lines to explore the evolution in time of the web graph of internal and external links in each language version.

5. Include support for SNA: This another top request received for future versions of WikiXRay. In its most basic form, a new module implementing the calculation of some basic metrics in SNA, either for the complete language version or for specific subsets of authors (core authors, sysops, etc) should be provided. In addition to that, it would be worthy to explore efficient ways to create simple graphs depicting the network of authors, articles etc., specially focusing on the identification of clusters of related nodes within the global network.

6. Include support for multilevel hierarchical models: This branch of linear statistical models deals with the estimation of significant statistics describing a certain population, but stratifying the results among meaningful subgroups. It would be interesting to explore whether we can infer additional information from stratified analyses considering some of the parameters that we already identified in this thesis.

7. Linear models, GAM and non-parametric models: A whole set of parametric and non-parametric statistical models is still to be applied, in order to discover significant relationships between distinct descriptive parameters in each language version (for both authors and articles). Once support for internal links is added, it should be feasible to explore with this methods whether there exist a direct correlation between the visibility of a certain article (in terms of the number of links pointing to it) and the number of revisions received by that article. Some results recently published by Spinellis and Louridas [106], suggest that there exist a preferential attachment process driving the creation of content in Wikipedia.

8. Integration with other tools for quality and reputation metrics: The added value provided by other relevant tools in this area, specially WikiTrust from UCSC Wiki Lab [15] could leverage the current capabilities of WikiXRay, creating a comprehensive framework for an in-depth study of the quality of content in Wikipedia, the reputation level gained by Wikipedia authors and, specially, the key quantitative parameters affecting quality and reputation in each language version.

9. GUI for interaction and monitoring: One of the current strengths of WikiXRay is that it is a command line tool, easy to use in production servers for periodic execution an integration in more general work flows. However, the absence of a GUI to interact with the program, as well as the lack of proper graphical interfaces to monitor the entire analysis process may have been reducing the audience interested in it.

10. Automatic generation of summary reports: WikiXRay makes heavy use of GNU R and numerous statistical packages to undertake statistical quantitative analyses on each language version. However, some nice packages has still to be integrated in further releases. One of these is Sweave [16] which offers a convenient way to embed GNU R commands and LaTeX commands in a single file, using a strategy known as *literate programming*. Once the file is ready to be used, R commands are interpreted and executed, and the output (either numerical or graphical) is integrated within the LaTeX file, which is then compilated in the final step. Therefore, we have a simple way to create professional summary reports for our tool.

11. Automatic creation of web-based reports: Another combination that may deserve our attention for future releases of WikiXRay is the combo GNU R plus GGobi [17] [26] to make attractive graphs to report our summaries on web pages, comprising all relevant statistics for our wiki website.

12. Integration with advanced visualization frameworks: Finally, it would also worth the time to explore efficient ways of interacting with advanced visualization frameworks like Processing [18]. In particular, it would be useful to profit from the ability to build dynamic visualizations to see

[15] `http://trust.cse.ucsc.edu/`
[16] `http://www.stat.umn.edu/~charlie/Sweave/`
[17] `http://www.ggobi.org/`
[18] `processing.org`

in action the evolution of web graphs, curves and other visual summaries for key descriptive parameters over time.

As we can see, there is still a long, exciting path ahead for future improvements on WikiXRay. My intention is that the tool should continue to evolve, integrating new functionalities and, hopefully, receiving the attention from other developers interested in collaborating with me in this challenging endeavor.

3.3 General features of Wikipedia dynamics

Our first analysis presents general aspects characterizing the global dynamics of the community of authors found in each language edition, and comparisons among behavioral patterns found in different versions. These general properties will help to contextualize the operational environment of Wikipedia authors, how they interact with the system as well as with other authors. We also obtain activity parameters and metrics to quantify the authors' effort sustaining the global content creation process. In addition, we examine related processes (like content discussion and creation of user pages) that permit us to identify idiosyncratic customs and behaviors in each language community.

At the same time, we are also interested in analyze the influence of different types of users in the project, particularly bots and anonymous users, who will be filtered out in subsequent analyses of this thesis work. Determining the importance of the role they play in the global dynamics of the community, we will be able to conclude the scope to which we will be able to extend our conclusions regarding logged users.

Finally, we also present an analysis of the Wikipedia project from the articles side. We look for singular patterns of content production, allowing us to gain additional knowledge about the social dynamics of the Wikipedia community of authors, through the inspection of their outcomes. To achieve this goal, we examine the distribution of Wikipedia pages in different namespaces for each language version. This will help us to better understand the most active collaboration areas in every community.It is known that one of the differential factors of Wikipedia are lively discussions about articles content on their associated talk pages. Previous research [121] has demonstrated that the Wikipedia community of authors places strong emphasis on group coordination, enforcing editing and behavioral policies and process control. The most evident product of these concerns is the election of privileged users (administrators, bureaucrats, and so on), who have deserved to be considered as trusted members enforcing Wikipedia good practices on this ecosystem.

3.3.1 Exploratory Data Analysis

Broadly speaking, for these analyses concerning general characteristics and dynamics of the Wikipedia activity we will directly employ most of the data included in our baseline tables in WikiXRay, already presented in Table 3.4. The majority of our calculations include obtaining proportions and percentages, counting data (number of articles, number of revisions) as well as finding monthly figures to depict the evolution over time of some interesting behavioral patterns. Of course, as we indicated in the previous section, this general analysis of Wikipedia features will also include a study of the influence of anonymous users, bots and redirects in the overall activity patterns, since from this point on, we will filter these data sets in most of our analyses (including them again only for comparison purposes).

The statistical techniques that we apply in this analyses can be classified under the scope of Exploratory Data Analysis [64], [29], a collection of data display techniques that allow researchers to look for interesting patterns and properties in data, before or as part of the implementation of a formal analysis. More precisely, we will make use of the following statistical tools:

- **Summary**: In GNU R, the summary function (applied on a raw data set) computes the following descriptive values from a distribution: minimum value; first quartile; mean value; median; third quartile; maximum value.

- **Histogram**: Basic, well-known statistical graph, displaying the frequency (alternatively, the probability density) of a data set. It is usually constructed as a set of side-by-side rectangles, which height is proportional to the number of observations in the dataset whose values fall in the range comprised by the rectangle. Instead, if we depict the probability density values, the height of each rectangle is adjusted according to the width of its interval, in order to ensure that the sum of the areas of all rectangles in the graph equals 1.

- **Kernel Density Estimation**: Also known as smooth density estimation, is an alternative procedure to represent the distribution of a certain data set. We plot the probability density function (continuous random variables) or the probability mass function (discrete random variables), estimating the shape of the function (usually denoted by $f(x)$). Unlike the histogram, it requires the choice of adequate bandwith parameters that controls the smoothing level. Broadly speaking, we can see the KDE as a smooth outline of the shape of the histogram graph. This also let us to compare several curves in the same graphs for more comfortable inspection of differences between distributions (cf. [64], pages 44-46).

  Figure 3.2 shows an example, depicting the histogram for a sample of 3,000 elements taken from a Normal population with mean $\mu = 0$ and standard deviation $sd = 1$, along with the KDE curve for the same data set.

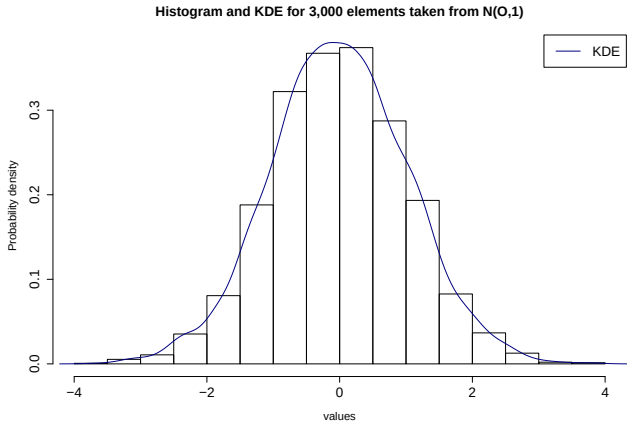

Figure 3.2: Example graph showing the histogram and KDE curve for a random sample of 3,000 elements from a N(0,1) distribution, plotted with GNU R

In order to make easier the comparison of graphical summaries and models obtained, we will use an advanced graphical package included in GNU R, called `lattice` [99]. It implements a new style

of statistical plots, called *trellis graphs* [100]. It allows us to depict a number of individual plots in the same graph, illustrating different patterns for each of the subgroups included in a data set.

In this section, we will obtain empirical results and graphs to find out the answer for the 3 first research questions addressed in this thesis:

- **Q1: How does the community of authors in the top ten Wikipedias evolve over time?**: To answer this question, we examine the evolution in time of the monthly number of revisions received by every language version. Then, we concentrate our study on relevant subsets, looking at the evolution of the monthly number of revisions received by articles, from both logged authors and anonymous authors (for comparative purposes). Once we examine the overall evolution of the effort spent by authors, we turn to analyze the population of authors and articles. Finally, we also calculate the monthly rate of contributions by bots, as well as the number of active bots per month, to delimitate the percentage of the total effort attributed to non-human authors.

- **Q2: What is the distribution of content and pages in the top ten Wikipedias?**: To address this question, we will make use of histogram plots to graphically show the shape of the distribution of the length of pages and articles, and its evolution over time. We will also probe the relationship between the length of every article and the number of different logged authors who revised it.

- **Q3:How does the coordination among authors in the top ten Wikipedias evolve over time?**: Our main target in this case will be the subset of talk pages, where authors concentrate their discussions about contents and their organization. We apply similar statistics and graphs on this subset, examining the evolution of the monthly number of revisions received by talk pages, the evolution of the number of active logged authors per month contributing to these pages. In the same way, we analyze the evolution of the monthly number of active talk pages, trying to infer whether discussion about articles content is broadening its scope. To complement our study, we inspect the percentage of the total number of pages by namespace in every language version, looking for unusually high percentages in those namespaces that play a critical role in the organization and coordination of community members (talk pages, user talk pages) as well as in the content classification (category and category talk namespaces). To conclude this section, the KDE of the length in bytes of talk pages will show the shape of this distribution for all language versions, and we study its evolution over time, looking for interesting trend that may indicate organizational changes in the next future.

This set of graphs and empirical results will help us to create and overall picture of the status of Wikipedia and its evolution until the end of 2007. The main conclusions extracted from this initial descriptive work will serve as the baseline to elaborate the remaining analyses and deductions extracted in subsequent sections of this document.

3.4 Social structure of Wikipedia

The completely open philosophy of Wikipedia is often a matter of very active discussion. Leaving aside possible implications about the quality of its contents (that we will also visit in section 3.6), the question of who is authoring the majority of Wikipedia content has been a source of debate in multiple

forums. Analyzing the social structure of Wikipedia will help us to understand the complex activity patterns showed by its community of users.

Actually, over the last years there has been a great controversy about this point. For example, on September 4th, 2006, Aaron Swartz included a quote from Wikipedia's founder, Jimmy Wales, in his blog [115]. Wales argued that the majority of the total number of contributions to the Wikipedia came from a small group of authors. Swartz used a different metric, counting the number of characters in each contribution (rather than the number of contributions), and searched for the text blocks retained until the final version of the article. He then applied this metric to several randomly picked articles, and showed that less frequent contributors were actually providing much of the articles content. Despite that, articles continue to evolve and change over time. if we focus only in the final revision of an article we will inevitably miss some contribution effort . Perhaps some content is removed later in the article's life, so that it is ruled out in its final revision. Nonetheless, we should take it into account if we want to have a complete picture of the content creation patterns found in the Wikipedia.

Our study of the social structure of Wikipedia will focus on finding answers to the 4th question posed in this thesis work:

- **Q4: Which are the key parameters defining the social structure and stratification of Wikipedia authors?**

To resolve this challenging task, we analyze the community of logged authors in the top ten Wikipedias from three different perspectives:

1. We explore the distribution patterns of the total number of different articles per author, as well as the total number of distinct authors per article. These metrics will complement the previous ones about inequality, producing a clearer explanation of the distribution of work among Wikipedia authors. We will use rigorous statistical procedures to obtain the best fitted curves to our empirical data, thus obtaining the first stratification patterns characterizing the Wikipedia community of authors, and the set of Wikipedia articles in each version.

2. We also complement the previous study fitting the distribution of the number of revisions per logged author and the number of revisions received by each article, to show the stratification of the community of editors from an alternative point of view, using effort metrics.

3. Finally, regarding the question about who is contributing the majority of content to Wikipedia, we will analyze the inequality patterns in author's number of revisions. Our previous research work in this area [83], [82] shows that we can identify a *core* group of very active authors who are responsible for the majority of revisions performed in the encyclopedia. We will show the highly unequal distribution of revisions in authors by means of the Lorenz curve and Gini coefficient for each language version. At the same time, we also present a novel application of the same statistical method to analyze the inequality level of contributions received by Wikipedia articles. In this case, the results of our analysis show that there also exist a subset of very active articles receiving more contributions. Later on, we will identify that this group of very active articles contains the majority of top quality articles in all language versions.

In summary, the whole set of graphs and numerical results in this section provide valuable insights and conclusion about the internal structure of the Wikipedia community of authors and the group of encyclopaedic articles found in each language version. These descriptive analysis may serve as a starting point for other researchers who need to know some details about the stratification of the Wikipedia community, according to the effort spent by volunteers in the project, so that they can

effectively identify the appropriate subset of individuals for their own studies. As we will see, most of the distributions follow a Pareto-like shape over a significant proportion of their whole range. This will derive important conclusions about the type of generational process responsible for the current shape of the Wikipedia work flow.

3.4.1 Statistical model

Apart from the previous analytical tools already presented in subsection 3.3.1, we will apply two additional techniques to explore the stratification of Wikipedia authors according to their activity patterns and their privilege status:

CDF and CCDF : The Cumulative Distribution Function (CDF) of a random variable (usually represented by $F(x)$), represents the probability of a random variable (either continuous or discrete) of taking values below a certain threshold: $F(x) = P(X < x)$. Therefore, $F'(x) = f(x)$, $F(x) \in [0, 1]$ and $F(x_1) \geqslant F(x_2) \forall x_1, x_2 / x_1 > x_2$. Sometimes, the CDF is useful to fit empirical distributions found in our samples, comparing the CDF of our sample with all possible candidate distributions. However, we can also use an alternative tool, the Complementary CDF (CCDF), which is simply $1 - F(x)$. The CCDF is preferable in some situations, for instance, fitting highly unequal distributions (Pareto, Log-normal, etc.), since they produce a well-known CCDF shape. We will use it to fit distributions of some of our results in this analysis.

Gini coefficient : This is a normalized measure of inequality, that we will apply to study the distribution (or dispersion) of contributions from Wikipedia authors. It was first proposed by Corrado Gini in [46], and it is a very well-known measure of the inequality of distribution of income and other quantitative factors among the members of a certain population.

To calculate the Gini coefficient, first the Lorenz curve for contributions has to be created. The Lorenz curve is a graphical representation of the cumulative sum of contributions, where we sort contributors on the horizontal axis by their amount of contribution. Then, we plot in the vertical axis the accumulated contribution, normalized to 1. A perfectly equal distribution would result in a straight diagonal line that divides the first quadrant of a Cartesian plot in two equal halves. This line of perfect equality is usually represented in addition to the Lorenz curve for comparative purposes. The Gini coefficient represents the area between the line of perfect equality and the empirical Lorenz curve, obtained from sampled data. Figure 3.3 presents a graphical example of this procedure, where the shadowed section represents the area given by the Gini coefficient.

We consider a population comprising n individuals. Let $p(i)$ be the cumulative percentage of the population represented by all contributors up to the i-th individual (sorted by their amount of contribution, in ascending order). Let $q(i)$ be the cumulative percentage of the parameter under study contributed by all previous individuals up to the i-th subject (included). Then, the value of the Gini coefficient is given by the following equation:

$$G = \frac{\sum_{i=1}^{n-1} [p(i) - q(i)]}{\sum_{i=1}^{n-1} p(i)} \tag{3.1}$$

Possible values for equation (3.1) are restricted to the closed interval $G \in [0, 1]$. A Gini coefficient $G = 0$ represents perfect equality, that is, a situation in which the contribution from every member

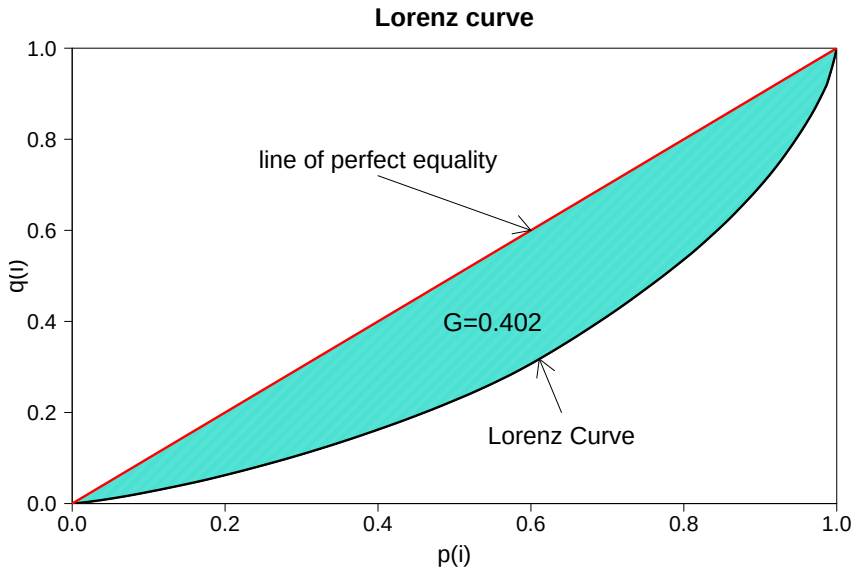

Figure 3.3: Illustration of the meaning of the Gini coefficient. The figure represents an hypothetical Lorenz curve along with a line of perfect equality. The Gini coefficient represents the colored area between the two curves

is exactly the same as the contribution from the previous one. A Gini coefficient $G = 1$ represents the extreme inequality (only one member concentrates all contributions, while the rest ones does not contribute at all). Therefore, Gini coefficients next to 0 shows a high equality level, while Gini coefficients tending to 1 presents a high level of inequality.

Consequently, values of the Gini coefficient close to 0 correspond to equal or almost equal distributions (lower departures from the line of perfect equality), while values close to 1 are good indicators of high inequalities.

The Gini coefficient has been successfully applied to measure inequality in different research areas such as Economics [36], [67], Education [117], and Health Sciences [128], [21]. It is known that the Gini coefficient, as well as other alternative statistical measures of inequality in populations (like the *skewness coefficient* or the *coefficient of variation*), presents some undesirable properties. In spite of this, these properties are usually irrelevant when we apply them on real data, following the usual two-parameter lognormal model for the distribution of the variable under study [92]. A more in-depth presentation of the different parameters available to measure inequality and concentration (with special attention to the distribution of income and wealth) can be found in [11]. In GNU R, we use the `ineq` package [136] to compute the Gini coefficients and the Lorenz curves for all the graphs included in this thesis.

3.5 Demographic analysis

Another interesting point of view for the analysis of the Wikipedia community of authors is the study of demographic parameters of this population. More precisely, we are interested in modelling the average lifetime of Wikipedia authors, measured as the total period of time in which they contribute to the project. In this way, we will explore whether the Wikipedia community of authors is a young group

of contributors, providing sustained support to the project over time, or it is difficult for Wikipedia to retain authors for long time periods. We are also interested in analyzing possible correlations between the authors lifetime and significant qualitative factors, like participating in talk pages discussions, reviewing featured articles, or reaching the *core* group of very active users.

This demographic analysis offers the elements to find out the answer for the 5th research question tackled in this thesis:

- **Q5: What is the average lifetime of Wikipedia volunteer authors in the project?**.

We will specifically obtain the following graphs and numerical results:

- Evolution in time of the monthly number of births and deaths: A careful examination of the monthly rates of births and deaths in each language version will give us a basic, yet clear picture of the current trend in Wikipedia with respect to the ability of the community to attract new contributors and maintain the old ones.

- Survival functions, that graphically describe the expected lifetime of Wikipedia authors in each version. We also apply this graphical method to analyze the time that it takes very active logged authors to join the core of each community, for how long they can maintain their membership and how much time elapses since they leave the core until they finally abandon Wikipedia to never come back again. These functions are complemented with the analysis of the instantaneous risk of death in logged authors and core authors, by means of the hazard function (more on this in the following subsection, explaining the precise meaning and applications of this statistical techniques.

- We also probe the influence level of relevant parameters on the survivability of Wikipedia authors. To achieve this, we apply the Cox proportional hazard model (which will be also presented in the following subsection) to measure the effect of editing talk pages and editing FAs in the lifetime of Wikipedia authors. The results unequivocally reveal a strong positive influence of both factors to extend the lifetime of Wikipedia authors.

- To conclude, we present several KDE plots summarizing the restricted mean and median survival time of logged authors in Wikipedia, then focusing on the same values for those authors who reached the core of very active contributors in each language version.

At this point, we should have a clear picture of what is the demographic situation of the community of authors in all language versions under study. With this data, we will infer important conclusions regarding the maintainability of the current structure of the Wikipedia community to produce high quality content for encyclopaedic articles.

3.5.1 Statistical model

To undertake the demographic analysis of the Wikipedia community of authors we apply a statistical technique called *survival analysis*. It is a potent, yet conceptually simple, statistical methodology that allows us to build empirical models for data analyses in which the variable of interest can be formulated in terms of *time until an event occurs*. In GNU R, we can use the extensive set of tools and procedures included in the `survival` package [116], written by Terry Therneau and ported to *R* by Thomas Lumley. We will also make use of the `muhaz` *R* package [80], written originally by Kenneth Hess and ported to *R* by R. Gentleman. [55], [30], [64] and [120] provide good introductions to the

theory of survival analysis and practical examples using *R*. We present here a very brief introduction to the basic theory behind survival analysis, just to provide a minimum background framework to understand the results and conclusions that we will draw in Chapter 4 of this thesis work. We will refer to the following definitions in subsequent sections of this document:

Event of interest :We identify the **event of interest** in our survival analysis as the **last registered revision** from a Wikipedia logged author to an article in its language version. Thus, we define the **lifetime** of a Wikipedia logged author as the time (measured in days) elapsed from her first registered contribution to the project to the last one logged in the dump files[19].

Let T denote a random variable, describing the *lifetime* (in days) of logged authors of Wikipedia articles. Its values are contained in the interval $T \in (0, \infty)$, and its continuous distribution is specified by its cumulative distribution function $F(t)$, (expressing the probability of any developer or contributor of having a lifetime value $T \leq t$), with probability density function $f(t)$.

Survival function :We define the **survival function** $S(t)$ as:

$$S(t) = 1 - F(t) = P(T > t) \tag{3.2}$$

Thus, it expresses the probability for a certain author to stay in the community longer than some specified time t.

Hazard function :We define the **hazard function or force of mortality** $h(t)$ as a function measuring the risk of dying within an infinitesimal time interval Δt, given that the subject is alive at time t. The mathematical expression for this function is:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < (t + \Delta t) \,|\, T \geq t)}{\Delta t} \tag{3.3}$$

In our study, it would give us the (infinitesimal) risk that a Wikipedia author leaves the project, ceasing to contribute to it.

One of the most important advantages of survival analysis is that we do not need to wait until all subjects included in the trial reach the event of interest to estimate $S(t)$ and $h(t)$. Instead, we simply define the limits of our observation period, and then assign a boolean value indicating, for each subject, whether she was "dead" or "alive" at the end of the study. In survival analysis, this is called **censoring** data (more precisely, *right censoring*). In other studies, we may have other types of censoring as well. However, Wikipedia dump files include a complete list of all contributions performed within the period of analysis, so we do not have to deal with other distinct types of censoring here.

As a result, when dealing with right censoring we have to slightly modify our definition of lifetime:

Author's lifetime :To deal with *right censoring* information, we define the **observed lifetime** for a certain author as the time period elapsed from her first logged contribution in the registry archive to either her last logged contribution (if $censoring == True$) or else, until the end date of the study (if the individual is still alive at the end date of the study, so that $censoring == False$).

Finally, once we have computed the observation time for each subject, and the censoring information, we calculate an estimation of $S(t)$, using the *survfit* function included in the *survival* library of *GNU R*. We define here the mathematical model that it applies to estimate the survival function:

[19]This definition of lifetime does not deal with intermediate idle intervals, in which an author may cease her contribution just to take it up again later on. Nevertheless, we do not consider that this limitation may affect the validity of our results

Survival probability : Let $r(t)$ be the number of cases at risk before time t, i.e., those subjects that are still alive in the trial before time t. If we define a set of intervals $I_i = [t_i, t_{i+1})$, covering $[0, \infty]$), then the probability p_i of surviving an interval I_i is:

$$p_i = \frac{[r(t_i) - d_i]}{r(t_i)} \tag{3.4}$$

Where d_i is the number of deaths counted in interval I_i. Hence, the probability of surviving until t_i is:

$$P(T > t_i) = S(t_i) \approx \prod_{0}^{i-1} p_j \approx \prod_{0}^{i-1} \frac{r(t_i) - d_i}{r(t_i)} \tag{3.5}$$

We note that the fraction in the last productory will be non-unity only for intervals in which deaths occur.

Kaplan-Meier estimator : We define the **Kaplan-Meier estimator** of a survival curve $S(t)$, denoted as $\hat{S}(t)$, as a maximum likelihood estimator obtained with:

$$\hat{S}(t) = \prod \frac{r(t_i) - d_i}{r(t_i)} \tag{3.6}$$

The estimation of survival functions using our data samples from Wikipedia dump files will conform the core of our demographic analysis of Wikipedia authors. To explore the influence of qualitative parameters on the lifetime of authors, we will apply the *Cox proportional hazards model*, which allow us to measure the significance level of these parameters on the hazard function. In Appendix C, we provide an extended introduction of these techniques for those readers interested in learning additional details about these methodologies.

As far as we know, this is the first attempt to apply survival analysis to study open collaborative projects like Wikipedia. Therefore, our aim is to contribute to previous research work, providing a novel approach suitable for obtaining complementing information that will give us additional insights about the social dynamics of open collaborative projects.

3.6 Author reputation & quality of content

The 6th research question proposed in our thesis is:

- **Q6: Can we identify basic quantitative metrics to describe the reputation of Wikipedia authors and the quality of Wikipedia articles?**

Regarding the analysis of author reputation and quality of content in Wikipedia articles, we perform a side-by-side comparison of FAs in the top ten language editions, from two different points of view. On one side, we analyze quantitative parameters focusing on articles. We calculate the number of revisions per article, the distribution of the length in bytes of revisions, as well as the number of different authors who edited every FA. In addition to that, we calculate the *age* of every FA, as the difference (in days) between the timestamps of the first and last revision of that FA. This parameter measure how long did it take to FA to reach its current special status [20]. Finally, we

[20] Although we cannot know on which date a certain article was promoted to the FA state, the continuous reviewing process ensures that the current level of quality of the FAs still remains acceptable to deserve this distinction.

compute the *recentness* of FAs, which measures the difference (in days) between the timestamp of the last edit received by a certain article, and the latest timestamp stored in the database snapshot. We will also compute all the above parameters also for non-FAs. In this way, we will be able to identify characteristic behavioral patterns suitable for differentiating FAs and non-FAs from a quantitative point of view.

On the other side, we are also interested in quantitative parameters focusing on authors, and more precisely, on logged authors. Although we cannot identify individual contributions from the group of anonymous users, we do compare aggregate contributions from logged-in and anonymous users, to answer the question: do logged-in users contribute more to FAs than anonymous ones? As well, we calculate the average length of contributions from logged and anonymous authors, in order to compare different behavioral patterns in both subgroups.

On top of that, we also calculate the *age* of logged authors in the system, computing the difference (in days) of the timestamps of the first and last contributions made by that author. We then compute the *recentness* of every author, which is the time interval between the last edit from that user and the last timestamp stored in the database snapshot. This quantity determines whether the author is still an active community member or she has ceased its contribution work at present time. In the following subsection we offer precise definitions for these parameters. We also compare the activity level of human and bots authors in FAs to answer the question: do human logged authors contribute more to FAs than bots? Finally, we consider that it is also of great interest to offer some comparison about the authors' reputation levels in different language editions. In the following subsection, we present our approach to measure this parameter in the top ten Wikipedias.

3.6.1 Metrics for reputation and quality

As we previously mentioned in Section 2.3.2, Stein and Hess proposed in [108] a metric for authors' reputation in Wikipedia, based on their contributions. We have followed the same methodology to compute, for the top ten language editions of Wikipedia two different measures:

- *Page-based author contribution* (pb): Measures the contribution of an author a based on the number of distinct pages that she edited (regardless of the number of edits). For instance, 10 edits on the same page only counts as 1 in this measure.

- *Edit-based author contribution* (eb): Measures the contribution of an author a based on the number of edits that she made, whereby each revision counts. For example, 10 edits on the same page counts as 10 in this case.

So that we can define the following measures for this quantitative analysis:

Age of authors : Let a be a logged author in a certain language version of Wikipedia. Let $rev\_ts(a)$ be the timestamp value storing the date and time of a revision made by author a in that language version. We can define the *age* of this author in the system, $age(a)$ as:

$$age(a) = max(rev\_ts(a)) - min(rev\_ts(a)) \quad [days]$$

Age of articles : Let p be an article, identified by a unique numeric ID and title string in a certain language edition of Wikipedia. Let $rev\_ts(p)$ be the timestamp value storing the date and time of a

revision received by article p in that language edition. We can define the *age* of this article in the system, $age(p)$ as:

$$age(p) = max(rev\_ts(p)) - min(rev\_ts(p)) \quad [days]$$

Recentness of authors : Let a be a logged-in author, identified by a unique numeric ID and log-in name in a certain Wikipedia language edition. Let $rev\_ts(a)$ be the timestamp value storing the date and time of an edit made by author a in that language edition, and let $max\_ts(lang)$ be the maximum timestamp value store in the database dump of that language edition (the timestamp of the last edit store in the snapshot). We can define the *recentness* of this author, $recent(a)$ as:

$$recent(a) = max\_ts(lang) - max(rev\_ts(a)) \quad [days]$$

Recentness of articles : Let p be a page, identified by a unique numeric ID and title string in a certain language edition of Wikipedia. Let $rev\_ts(p)$ be the timestamp value storing the date and time of an edit received by page p in that language edition, and let $max\_ts(lang)$ be the maximum timestamp value store in the database dump of that language edition (the timestamp of the last edit store in the snapshot) We can define the *recentness* of this page in the system, $recent(p)$ as:

$$recent(p) = max\_ts(lang) - max(rev\_ts(p)) \quad [days]$$

Page-based author reputation : If a is a logged author in a certain Wikipedia language version, we can define the *page-based reputation* of this author, $rep_{pb}(a)$ as:

$$rep_{pb}(a) = \frac{\text{\# of edits of } a \text{ on } FAs}{\text{total \# of edits by } a} \tag{3.7}$$

Edit-based author reputation : If a is a logged author in a certain Wikipedia language version, we can define the *edit-based reputation* of this author, $rep_{eb}(a)$ as:

$$rep_{eb}(a) = \frac{\text{\# of distinct } FAs \text{ } a \text{ edited}}{\text{total \# of articles } a \text{ edited}} \tag{3.8}$$

As Stein and Hess mentioned in their previous work, this measure may be vulnerable to deliberate attacks from users that can artificially inflate their personal reputation, by making a lot of small edits in FAs. However, as we are quantifying past interactions within a closed snapshot of each database, users would not have any direct motivation to try and alter their reputation by this measure. The metric can also be applied to compute the quality rating of articles in a straightforward manner. The quality rating of a page p is computed based on the reputation of its authors. This reputation level can be computed either on a *per-author-basis* (ab) or on a per-edit-basis (eb):

Author-based article reputation : If p is a Wikipedia article, in a certain Wikipedia language version, then we can define the *author-based* rating of this article, $rat_{ab}(p)$ as:

$$rat_{ab}(p) = \frac{\sum_{a \in authors(p)} rep_{pb}(a)}{|authors(p)|} \tag{3.9}$$

Edit-based article reputation : If p is a Wikipedia article, in a certain Wikipedia language version, then we can define the *edit-based* rating of this article, $rat_{eb}(p)$ as:

$$rat_{eb}(p) = \frac{\sum_{e \in edits(p)} rep_{eb}(author(e))}{|edits(p)|} \tag{3.10}$$

Before we continue, we need to point out some additional remarks regarding the computation of all these parameters. WikiXRay creates GNU R scripts to plot all these graphs and perform statistic analyses on the results. For *age* and *recentness*, we take the log10 of the corresponding values to represent the statistical graphs summarizing the results (histograms, correlations, etc.). As $log10(0) = -\inf$, we previously filter out all authors and pages with $age = 0$ (those which performed/received only one edit, or those which performed/received several edits in the same day, and did not come back again later), as well as authors and pages with $recentness = 0$ (those who performed/received an edit on the last day stored in the database snapshot). In the first case, the majority of filtered users were those who only made one edit in the system, clearly, a noise component for our study. In the second case, we are missing users and pages which performed/received edits on the last day included in the snapshot, but this is a very low percentage of the total number of pages and authors considered in each top ten language edition, so it does not affect the validity of our results.

To compute the correlation level between two quantitative parameters, we employ the Pearson's correlation coefficient, which provides a convenient numerical result in the interval $[-1, 1]$, summarizing the relationship between two data vectors, containing samples of two given quantitative variables. The Pearson correlation coefficient is represented by r and we can compute it using the following expression:

$$r = cor(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Finally, we would also like to remark that this is not the only metric propose to quantify the reputation of Wikipedia authors. One of the most notorious examples is the alternative metric proposed by de Alfaro and Adler in [6], [7] and [5]. In our opinion, these metrics provide additional background information about Wikipedia authors reputation, and therefore, further research should be conducted in due course to explore the most efficient way to combine these metrics and create a unified common approach to study this parameter.

3.7 Evolution of the Wikipedia community

This analysis of contribution patters of the Wikipedia community of authors would not have been complete if we did not include a study of the evolution in time of some of the relevant parameters identified in previous sections. In this way, we have to find an answer for the final research question in our list of interests:

- **Q7: Is it possible to infer, based on previous history data, any sustainability conditions affecting the top ten Wikipedias in due course?**

Careful examination of possible changes in these factors over time will reveal meaningful trends in the past history of Wikipedia. These previous behavioral patterns in the community of authors may

also be useful to infer most probable trends in the next future, with special interest in sustainability conditions that should be achieved to ensure the continuity of the Wikipedia project in due course.

3.7.1 Analyses and metrics

In the first place, we need to choose some parameters identified in previous sections that will reveal additional information about the community behavioral patterns. We are interested in tracking the following parameters over time:

- Monthly Gini coefficient over time: The Gini coefficient of the total number of contributions per author is a good estimator of the inequality level of contributions in the Wikipedia community. However, following the changes of the Gini coefficient on a monthly basis, we can discover interesting trend patterns to characterize the evolution of the inequality of contributions over time.

- Mean and median lifetime: In a similar manner, analyzing the changes of the mean and median lifetime of Wikipedia authors on a quarterly basis, we may learn additional insights about the capacity of the project to retain contributors and maintain its impressive growing rate in the next future.

- Evolution of the core group: Finally, we study the evolution of the group of core contributors over time. We are interested in checking whether this group is very stable over time, or the individuals included in this group changes over time. We are also interested in exploring the contribution effort of the core group over time, to check whether they tend to increase their number of revisions or it remains stable over time. Finally, since we have showed in previous research work that this group of very active authors is supporting the major part of the activity in Wikipedia, we will calculate the proportion of new users that must join the project to ensure the sustainability of the core group with new users that take the relay of previous core members.

The statistical techniques involved in these analyses include the organization and plotting of longitudinal data, tracking its evolution over time. Though there exist several formal techniques to analyze and model longitudinal data, we will restrict our analysis in this thesis to smoothing techniques revealing meaningful trends in our samples. The application of time series analysis and other tools to model the evolution of relevant parameters in the Wikipedia community of authors will be left for further research work in this area.

Finally, to study the evolution of the core of very active authors in each language version, we will make use of a 3D plot displaying the changes of the number of revisions from the most active group of authors in each month over the remaining months in the history of that language version. This 3D plots has already been successfully applied in previous research papers covering this topic [82]. The same approach has been successfully applied in the study of the evolution of cores of very active contributors in other types of open communities, as well, like FLOSS development projects [98].

Chapter 4

Empirical Analyses and Results

> "You see, but you do not observe. The distinction is clear". "It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts". *A Scandal in Bohemia*, in *The Adventures of Sherlock Holmes*, Arthur Conan Doyle (1892).

4.1 Introduction

Throughout this chapter, we will present the results of our empirical study of the Wikipedia community of authors, following the different methodologies presented in Chapter 3. Results are presented in the form of tables summarizing numerical proportions, coefficients, slopes of our fits etc., as well as colored figures that will help us to characterize the Wikipedia community of authors from different perspectives. All these empirical results are accompanied by appropriate discussions of their implications for understanding the content creation process in Wikipedia.

Along this presentation, we will try to focus on suitable metrics to characterize the activity of authors, and the distribution of effort among community members. We will follow with special interest the evolution of activity metrics over time, looking for patterns and trends in the activity of authors, suggesting distinct stages along the whole history of our sample data. We will also inspect the activity patterns and evolution over time of talk pages, since they are the central element to implement the authors coordination mechanisms. Our findings will be complemented by an in-depth survival analysis of Wikipedia authors, to study the speed of change on relevant parameters that may affect the transition between successive stages of the project and the community. In addition to this, we will examine common descriptive parameters shared among those authors providing the majority of quality content to the project. Since the main objective of Wikipedia has focused now on the production of good quality articles, our findings may be useful to find high reputation authors, and candidate articles that may be nominated to undertake the reviewing process towards the FA qualification in the future. In this regard, we will explore the applicability of the reputation metrics proposed by Stein and Hess [108], comparing the results obtained for the set of language versions covered in this thesis.

The chapter will conclude analyzing the evolution of the most critical parameters and distributions identified in previous sections. We also examine possible future scenarios for the evolution of the project over the following years. We focus specially on sustainability conditions that must be satisfied for the project to maintain the same production level in due course. Our findings support the necessity for adopting measures to attract new users and maintain current authors in Wikipedia, in case the

project is willing to recover the steady growing rate and productivity effort exhibited in earlier periods of its activity history.

Before going into the nitty-gritty details of our empirical analysis of the top ten Wikipedias, we present some overall statistics that illustrate the huge information repository that we had to analyze in this thesis. As we stated previously, the size of our databases, even after filtering out every revision content to obtain its most relevant descriptive metrics, forced us to look for performance-wise methodologies, algorithms and database queries, in order to obtain the empirical results presented in this document within a reasonable period of time. In the first place, Table 4.2 summarizes some size metrics for the size of the language versions under study. As we can see, all language versions store activity information for a time period exceeding 2,000 days (calculated as the difference in days between the maximum and minimum timestamps found in the *revision* table of every language version).

Some of these results deserve a more detailed inspection. For instance, it is remarkable that the English Wikipedia accumulates a total number of more than 11 million pages in its database. This is almost 6 times more wiki pages than in the German language version, and almost 5 times more pages than the total number of pages produced in the French Wikipedia. The reason of this outstanding number of pages is connected with the role assumed by the English Wikipedia as the "central" language version in the project. If we examine the distribution of wiki pages according to their namespace in the English version, we can see some clear indicators of this role. For instance, besides the high number of pages in the `main` namespace (more than 4.5 million), we count more than 1,7 million talk pages; more than 2.7 million of `user_talk` pages (usually employed for leaving messages to individual users); more than 300K pages in the `wikipedia` namespace (describing topics related to the Wikipedia project itself); more than 780K `image` pages; more than 134K `template` pages and finally, in excess of 312K `category` pages. Without any doubt, the perception of many users about the English Wikipedia as the "main" language version contributes to boost the number of pages created in some namespaces (like images, wikipedia and template). On the other hand, it is paradoxical that the number of discussion pages associated to individual users exceeds in almost 1.5 times the total number of registered users along the complete history of this language version. This gives us a clear picture of the infrequently high activity level of English Wikipedians regarding user discussion pages. However, as we will see later on, the total number of logged authors in this language version overwhelms that of any of the rest of versions in this study, which contributes to reach this unusually high number of user discussion pages.

If we focus on the number of pages in the `main`, we can see that they conform the largest proportion of the total number of pages, in all versions. The 3rd and 4th columns in Table 4.1 show the distribution of these pages between ordinary articles and redirect pages. In general, the number of redirect pages in each language version accounts for a significant proportion of the whole number of pages in the `main` namespace. We have 3 important exceptions to this general trend, though. The first one is the Polish Wikipedia, with a comparatively low number of redirect pages. We will explore in subsequent sections of this thesis the causes behind this deviation of the Polish version from the general trend. On the other side, we have more redirect pages than articles in two version, namely the French and Portuguese Wikipedias. Again, this is an unusual result, indicating a very special interest in these language versions on providing alternative encyclopaedic entries. Finally, if we turn now to talk pages, we find that the top 3 Wikipedias are quite far away from the remaining versions, and actually than the French version stores more than 150K more pages than the German one. In this case, French Wikipedians seem to support more active discussion about article content than German ones.

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 4.1: 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.

We can examine the overall metrics about the number of authors and their overall activity in the top ten Wikipedias, as well. Table 4.2 provides information about the total number of logged authors and the total number of revisions performed in the top ten Wikipedias. We will see later that logged authors are responsible for the largest proportion of contributions to each version, so the total number of revisions is a valid metric for the purposes of this comparison. Finally, we also include the total number of bots officially registered as such in each versions, to complete our comparison (since bots are accountable for a significant proportion of contributions in some languages, as we will also see in subsequent sections). The disproportionate difference between the English version and the rest of the top ten Wikipedias is quite evident: 8 times more users than the second Wikipedia in the list (German version); 4,5 times more revisions (up to an impressive total number of more than 167 million revisions that we had to process for the English version); and 2.3 times more registered bots. But we can learn some additional remarkable insights form this Table. According to this information, if we list the Wikipedias in descending order according to their total number of revisions, the top 3 positions remain the same. However, the 4th position is occupied by the Japanese Wikipedia, with more than 17 million revisions, followed by the Spanish version (in excess of 14 million) and the Italian version (exceeding 12 million revisions).

Also remarkable is the total number of logged authors in the Spanish Wikipedia, doubling the total number of authors registered in its sourrounding language versions in the top ten list, and even surpassing the French Wikipedia to reach the top 3 position. If we stop to think about this result for a moment, we realize that the Spanish Wikipedia is very far from the head of the list, following the official count of the total number of articles produced. Therefore, we have a language version with an infrequently high number of authors (along its entire history) that has not produced an equivalently large number of articles. The main conclusion is that Spanish authors tend to build on existing content, rather than creating new encyclopaedic entries. This peculiar behavioral patter of the Spanish community concords with some previous findings [84], showing that the second largest community of readers in Wikipedia is the Spanish one. This suggests radical difference between the popularity of Wikipedia among Spanish readers, and the number of those individuals who eventually contribute to the project.

Table 4.2: Additional overall descriptive parameters for the top ten Wikipedias: total number of logged authors, number of revisions and number of bots in the top ten Wikipedias

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

Finally, the number of bots in all language versions other than the English Wikipedia is fairly similar, with the exception of the Japanese and Portuguese versions. The Portuguese Wikipedia is another unusual anomaly in the general trend, since we will show later that the proportion of revisions performed by bots (over the total number of revisions received) in this language version is quite significant.

4.2 Analysis of General Features and Dynamics in Wikipedia

We begin our empirical analysis of the Wikipedia community of authors presenting an overall characterization of the evolution of some general metrics, describing the effort spent by the Wikipedia community of authors. We also inspect the organization of Wikipedia content over the different namespaces defined in the system. This will give us some insights about the most active collaborative areas in each of the language versions under study. Finally, we study the participation of authors in discussion pages and other coordination mechanisms, which may indicate a more intense implication of these authors in the content creation process. In the same way, the analysis of coordination pages may also show whether Wikipedia authors are spending more effort in collaboration forums rather than in content pages themselves. As we will see, this is important since, as article content becomes more and more rich, we would expect an increment of the coordination activity to deal with the growing complexity of the reviewing process (larger text sections, more multimedia content, higher number of different users participating in the article, and so on).

To start with this section, we look at the evolution in time of the total number of revisions from logged authors to the top ten Wikipedias, disregarding the namespace of the revised pages. The graph is displayed in Figure 4.1. Unless stated otherwise, all graphs like this one in this section use a logarithmic scale for the vertical axis in order to facilitate the comparison of the different versions, despite their different range of number of revisions received. A first evaluation promptly focus our attention on the steady growing trend maintained by this statistic in all languages, just until summer 2006. In 2007, the total number of contributions becomes stable in all versions, remaining at a approximately constant level that follows the maximum number of contributions reached at the end

of 2006. This is an interesting finding, one that was already mentioned by Voss in its research article about Wikipedia measurements, back in 2005 [124], when he hypothesized about possible evolution scenarios for Wikipedia. These kind of systems may not grow indefinitely as time goes by, and eventually find and upper limit for the total number of contributions received, due to multiple causes. Apparently, at that time Wikipedia still managed to find alternative ways to maintain its growing level of productivity. However, our results shows that the global growing trend has broken for the top ten language editions. Our purpose in this thesis is to find likely causes for this behavior, thus offering a model that could explain this change of tendency in the contribution pattern found for the Wikipedia community of authors.

Other interesting episodes can be also identified in this graph. For instance, we have the remarkable growth experimented by the Japanese Wikipedia in the first quarter of 2003, coinciding with the first appearance of Wikipedia in the Japanese electronic online magazine Wired News. This has been reported to be the first coverage of the Japanese Wikipedia in mass media [1], showing that there is a positive, and strong correlation between the number of contributions received by the project and its popularity level.

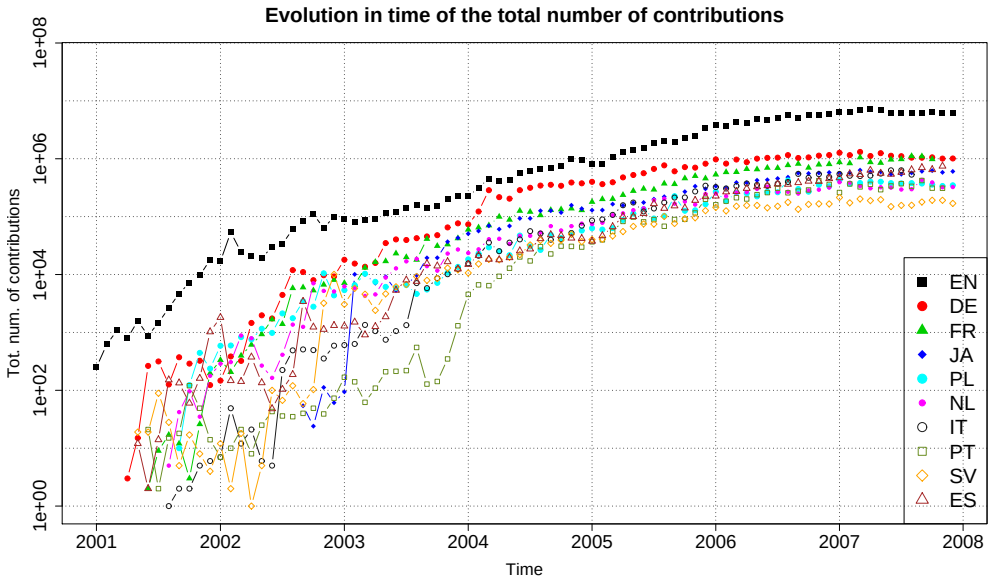

Figure 4.1: Evolution of the total number of revisions performed in all pages of the top ten Wikipedias by logged authors. The vertical axis follows a logarithmic scale. The graph clearly shows that the number of contributions received from logged authors has stabilized over the last year, breaking the constant growing rate exhibited by all language editions since their creation.

After looking at the overall trend of revisions of logged authors for all namespaces, the natural question is whether the same trend applies for the contributions performed in Wikipedia articles solely (that is, for wiki pages falling in the `main` namespace). Figure 4.2 presents the results for the monthly number of revisions performed by logged authors in articles. The interesting finding here is that the shape of the graph is quite similar to the previous one, thus showing that the baseline contribution to the total number of revisions in each month is performed on Wikipedia articles, as we might

[1] `http://en.wikipedia.org/wiki/Japanese_wikipedia`

have expected *a priori*. In other words, the majority of revisions from Wikipedia logged authors is performed on articles, demonstrating that this is the main task in the community workflow. Since summer 2006, the same change to a steady-state of the monthly number of contributions is found in this graph as well. Since this is the activity consuming most of the effort spent by Wikipedia authors, the causes producing this change of trend must have to do with any of the parameters affecting the work of logged authors in the `main` namespace.

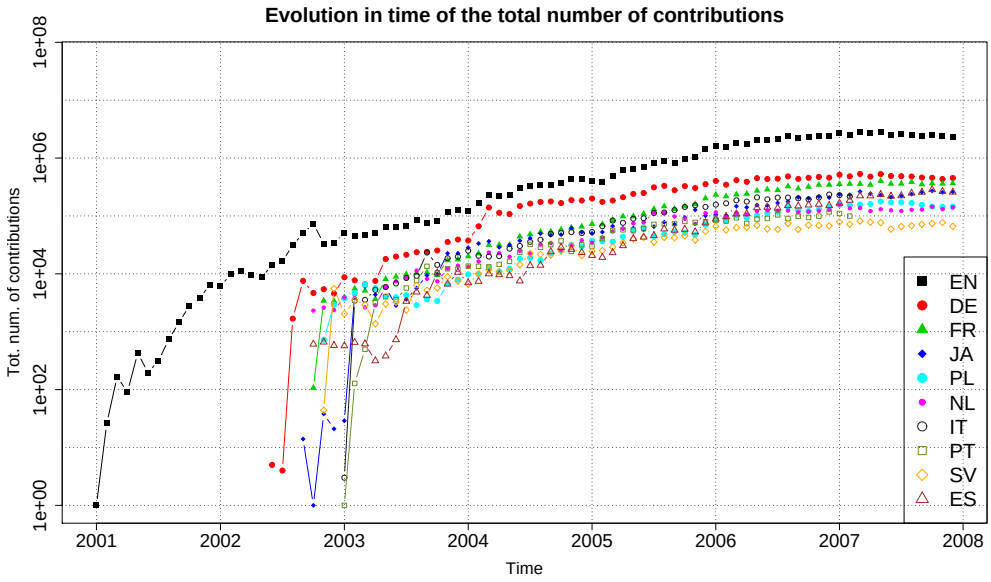

Figure 4.2: Evolution of the total number of revisions performed in articles of the top ten Wikipedias by logged authors. The vertical axis follows a logarithmic scale. Redirect articles have been filtered out. The graph shows again that the number of revisions in articles has break on the last year the growing tendency followed by all language versions in their previous history.

One question that may arise at this point is: what about the contributions from anonymous authors? Do we find the same leverage effect in the late months of history of the top ten Wikipedias? Though we are not including anonymous authors by default in this analysis, the answer to this question may provide valuable information for extracting relevant conclusions about the possible causes behind this deceleration in the monthly number of revisions. Figure 4.3 presents the answer. We find the same stabilization effect, breaking the continuous growing rate shown by all languages in their early history. This is an additional hint for our investigation. The possible causes behind the new stabilized trend in the number of revisions per month must also affect anonymous authors.

The first and most obvious cause that me be producing this approximately constant rate of the monthly number of contributions is that the number of logged authors has simply ceased to grow as well. Looking at Figure 4.4 we can confirm our suspects. The figure presents the monthly number of active logged authors for each language version. Since summer 2006, and specially through 2007, we can see an abrupt change of the steady growing trend in past years for all versions. Clearly, this sudden alteration in the pattern followed by the project in its early history is directly responsible for the stabilization in the monthly rate of revisions. A stable population of logged authors can not maintain the previous growing rate of revisions, and thus the monthly effort becomes constant over time. Moreover, this change of trend can have unexpected consequences for the future sustainability

of the project in due course, since the project will not be able to maintain its log-linear growing rate with a unchanging number of authors. In particular, there also exist the risk that, should the number of users abandoning the project exceeded the number of new users joining Wikipedia, we would eventually see decreasing trends in the number of monthly revisions received in subsequent years. If we focus more closely on the behavior exhibited by all curves on Figure 4.4 over 2007, we can see that this is precisely what happened in brief periods within this year, though the drop has not been very significant yet.

Again, if we measure the number of different logged authors who contributed to Wikipedia articles, exclusively, we can also see the same break of the growing trend followed by all language editions in previous years. Unfortunately, we can not measure the number of monthly anonymous authors accurately, but the hypothesis that the same effect is also happening in that group of authors seems quite reasonable, given the results of monthly contributions previously presented in this section.

Our analysis would have been incomplete without looking to the number of monthly active bots, performing revisions and maintenance tasks in each version. After all, though we are systematically filtering bots revisions in this thesis, their presence may have also been affecting the system in a global sense. Figure 4.6 present these results, and shows different patterns. Whilst the 3 largest Wikipedias (English, German and French) also present levelled off results in their last year of history, other versions like the Japanese still maintain a growing rate well into 2007. It is interesting to notice, though, that the Japanese Wikipedia has systematically presented the lowest number of active bots per month, and possibly, they have started to consider the inclusion of new bots to automate some routine maintenance tasks, like in other language editions.

Likewise, Figure 4.7 shows the monthly share of revisions from logged authors due to bot activity. It is important to remark that in some language versions, like the Polish, Portuguese and Dutch Wikipedias, the share of revisions attributed to bots is by no means negligible. Thus, we need to filter out these contributions to eliminate noise from our analyses. The curious case of the `tsca.bot` in the Polish Wikipedia stands out clearly on July 2005, accumulating up to 60% of the total number of revisions by logged authors in that month. This bot received a new task on that month, consisting on automatically downloading and incorporating to the Polish Wikipedia articles statistics from official government pages about French, Polish and Italian municipalities. This new task was responsible for more than 40,000 revisions in subsequent months, and clearly determined the number of revisions received in July 2005, its first month of activity.

4.2.1 Distribution of Wikipedia articles and pages

Complementing our previous analysis of the contributions from logged authors in Wikipedia, we undertake in this section a study of the organization of content in the top ten Wikipedias, with special interest to the activity level registered in the `main` namespace, as well as its evolution over time. We start with Figure 4.8, which shows the evolution of the monthly number of active articles, that is, those who received at least one revision from logged authors in the corresponding month. The plot shows the same stabilization trend in 2007 already identified in the previous section, for the monthly number of contributions from logged authors, due to the leverage in the monthly number of active logged authors. According to this results, we can see that there also exist a direct influence of the lack of new logged authors in the last year and the number of articles receiving new contributions. This is a logical consequence, since we would expect that a stabilized community of users were not be capable of revising different articles beyond a certain limit, given by the size of this community in each language version. Figure 4.9 shows that redirect pages are not an exception to this general trend, and the have also experimented a deceleration in their number of active pages per month.

At the same time, it is also interesting to analyze the product of the creation effort undertaken by Wikipedia authors. Figure 4.10 shows the KDE of the log10 of length of the English Wikipedia pages, in bytes, according to their namespace. Likewise, Figure 4.11 presents the same KDE plot for the length in bytes of articles in the top ten Wikipedias. The first graph shows us another interesting pattern. Pages in talk namespaces usually present unimodal distributions for their length values, while pages in, say it, standard namespaces (at least, those not directly focused on discussions about page content) usually present bimodal distributions. As we can see in Figure 4.11, the bimodal distribution for the length of pages falling in the `main` namespace is shared by all language versions under study. As we already presented in a previous research work [84], the left side subpopulation in each KDE plot correspond to the group of stubs and redirect pages in each language version (therefore, redirect pages has not been ruled out in this graph, to evidence this interesting pattern). The right side of each KDE curve correspond to more standard articles, that have reach longer versions.

It is quite remarkable that, in all language versions, the most common length of standard articles is situated around 1,5KB, indicating that this might be considered as the expected length of a Wikipedia article. We can also identify two different shapes in the KDE curve for standard articles: a sharper version, found in the Polish and Portuguese versions (and, to a lesser extent, in the Italian and Dutch Wikipedias); and a smoother version, found in the rest of languages. If we recall the previous results found in Figure 4.7, the sharper histograms correspond to language versions with a higher share of monthly revisions coming from bots, most notably the Polish Wikipedia. Therefore, we can infer that language versions with a higher proportion of human activity produce smoother versions of the KDE curve for the length of articles, while those versions with a higher proportion of bot activity create sharper KDE plots, as a result of this strong influence.

Another interesting facet that we should analyze is the evolution of the KDE curve for articles' length over time, in order to check whether or not we can identify common patterns in the progression of all language versions. As we can see in Figure 4.12, this is the case. In all language versions we can see a clear evolution pattern, in the form of a bias towards longer articles as time goes by. Nevertheless, the increment experimented in this length is not quite significant, and the median of standard articles has remained quite close to the 1,5KB value already identified in the previous graphs. Nonetheless, if we take a closer look to this curves, we can see that the increment in the median values of all language versions has become lower with the course of time, possibly showing the influence of the leverage in the monthly number of active logged authors.

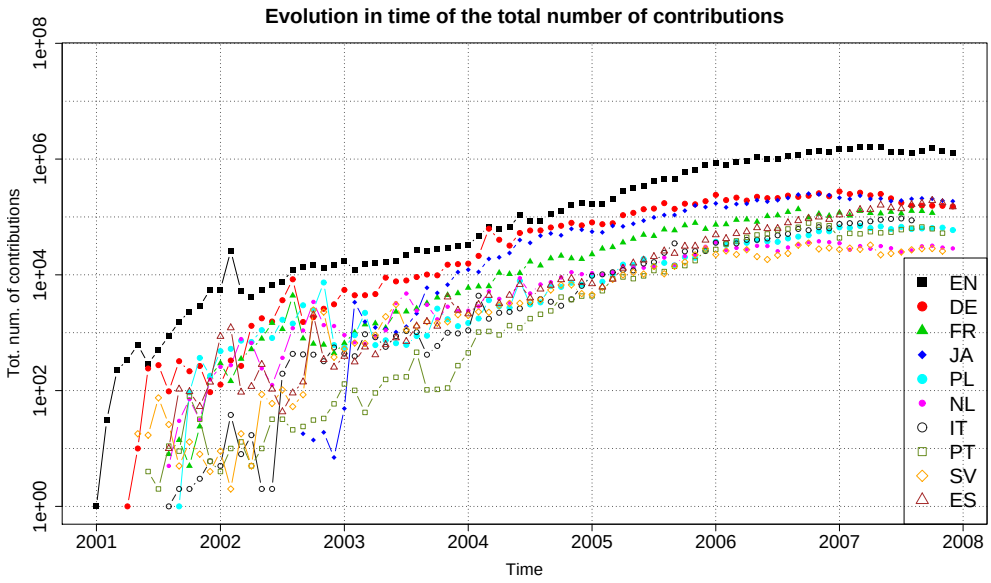

Figure 4.3: Evolution of the total number of revisions performed in articles of the top ten Wikipedias by anonymous authors. The vertical axis follows a logarithmic scale.Redirect articles have been filtered out. Once again, in the last year the number of contributions seems to have reach a stabilized level

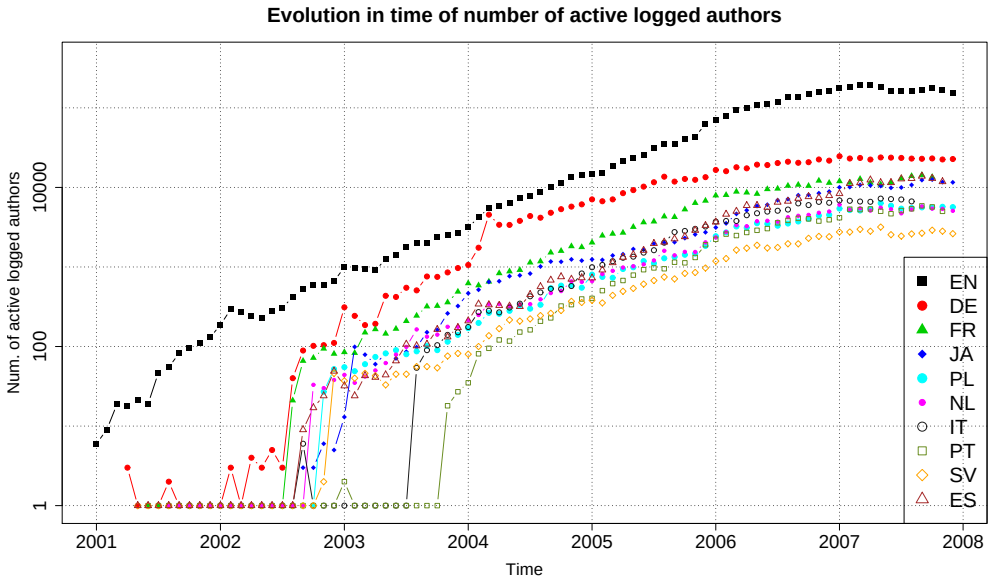

Figure 4.4: Evolution of the total number of number of active logged authors per month in the top ten Wikipedias. The graphic exhibits the same leveraging effect already identified for the number of contributions over the last year, offering a possible cause for this effect

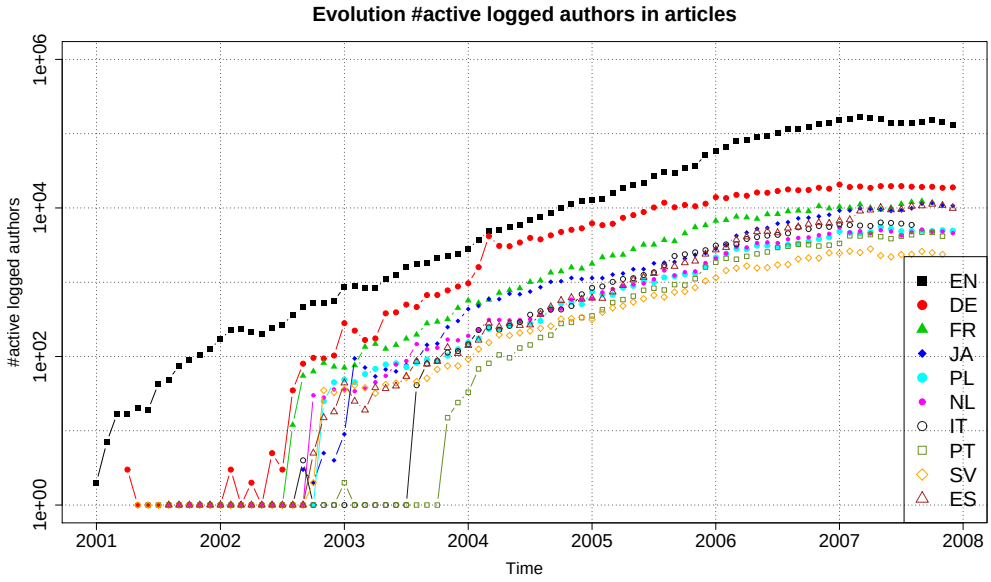

Figure 4.5: Evolution of the total number of number of active logged authors per month reviewing articles in the top ten Wikipedias. Revisions of redirect articles have been filtered out. Again, the break of the growing tendency followed on the early history of all language editions can explain the leverage of the number of revisions performed in articles over the last year

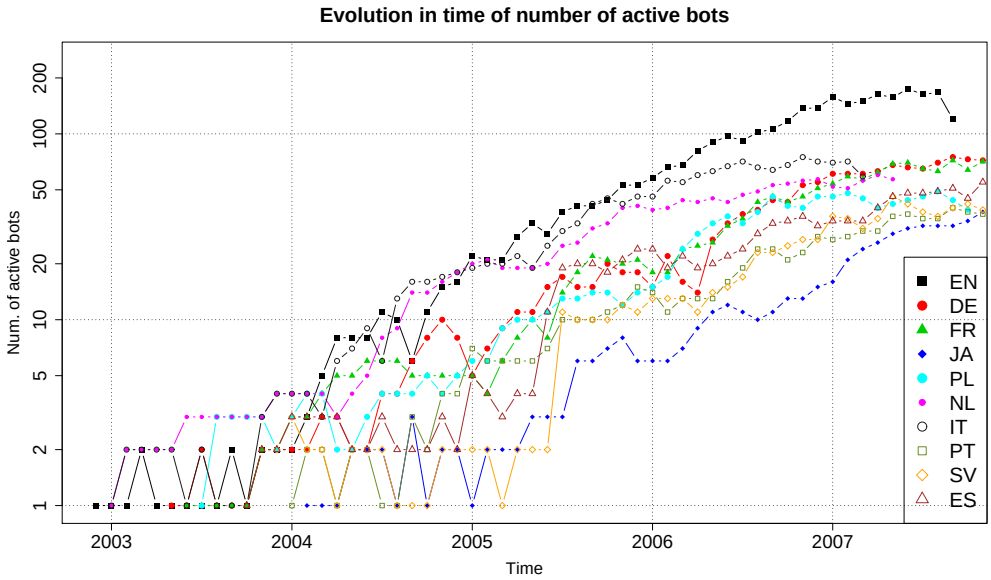

Figure 4.6: Evolution of the total number of number of active bots per month modifying pages in the top ten Wikipedias. All namespaces have been considered. The number of active Wikipedia bots has also reach a somewhat constant rate over the last year, thus contributing to reveal the leverage of the total number of contributions in all language versions.

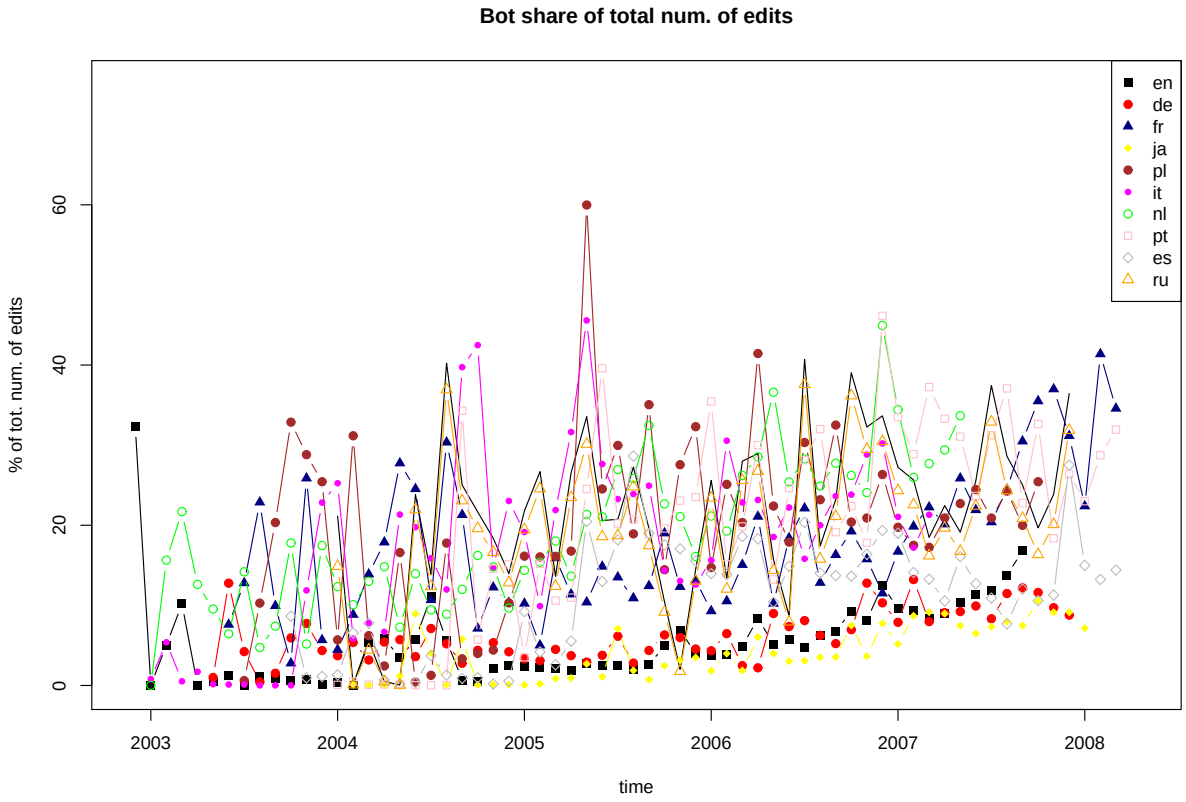

Figure 4.7: Evolution of percentage of the total number of revisions per month performed by bots in the top ten Wikipedias. The proportion of revisions attributed to bots in each month is quite significant in some language versions, like the Dutch, Portuguese, Italian, and specially the Polish Wikipedia. On the contrary, large versions like the English, German and Japanese Wikipedias systematically present lower sharing of the number of revisions from bots, showing that the majority of their content production activity is undertaken by human authors

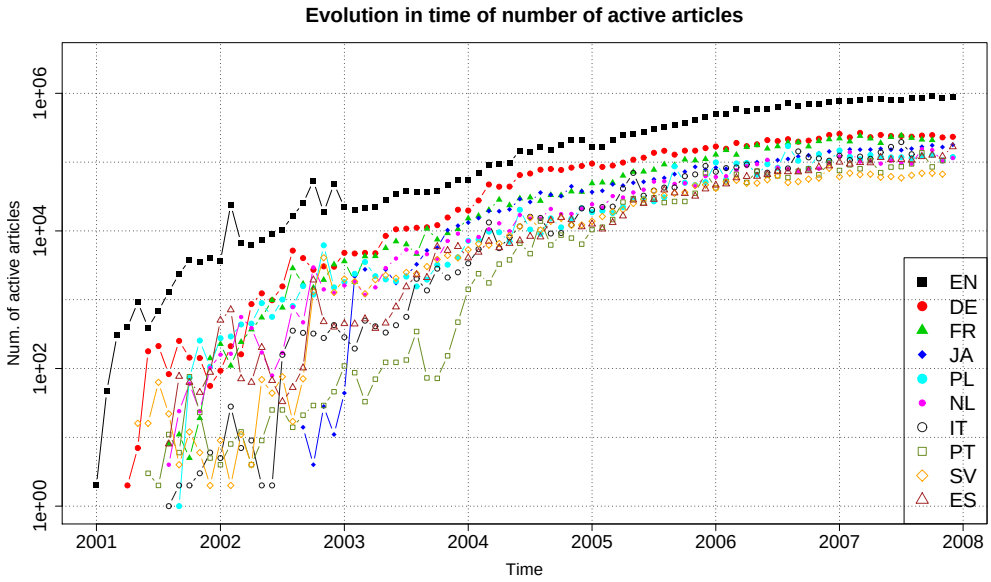

Figure 4.8: Evolution of active articles per month in the top ten language versions of Wikipedia. We can appreciate the same leverage effect in 2007 already identified for the contributions from logged authors, thus demonstrating the influence of the leverage in the number of monthly active logged authors in this statistic, as well

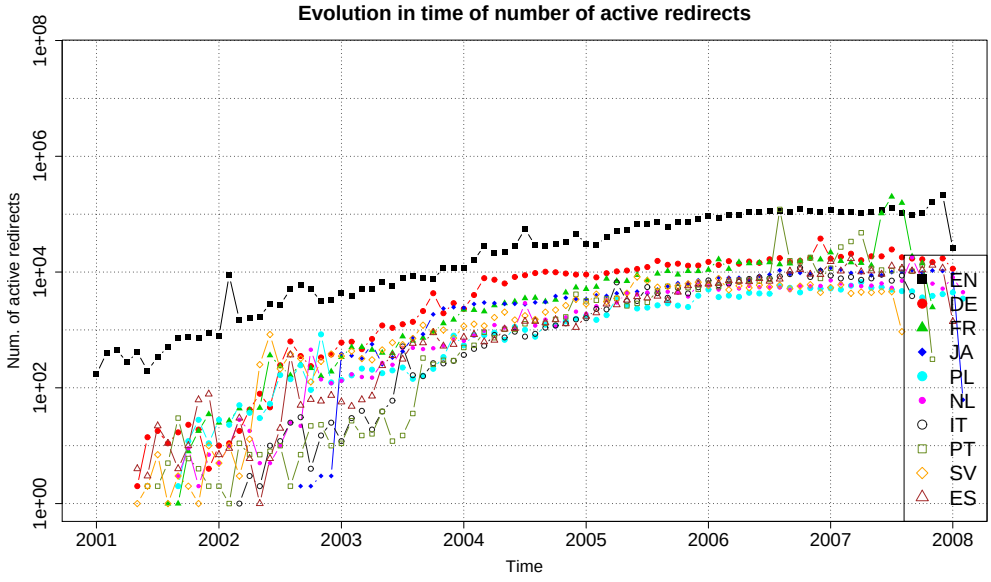

Figure 4.9: Evolution of active redirects per month in the top ten language versions of Wikipedia

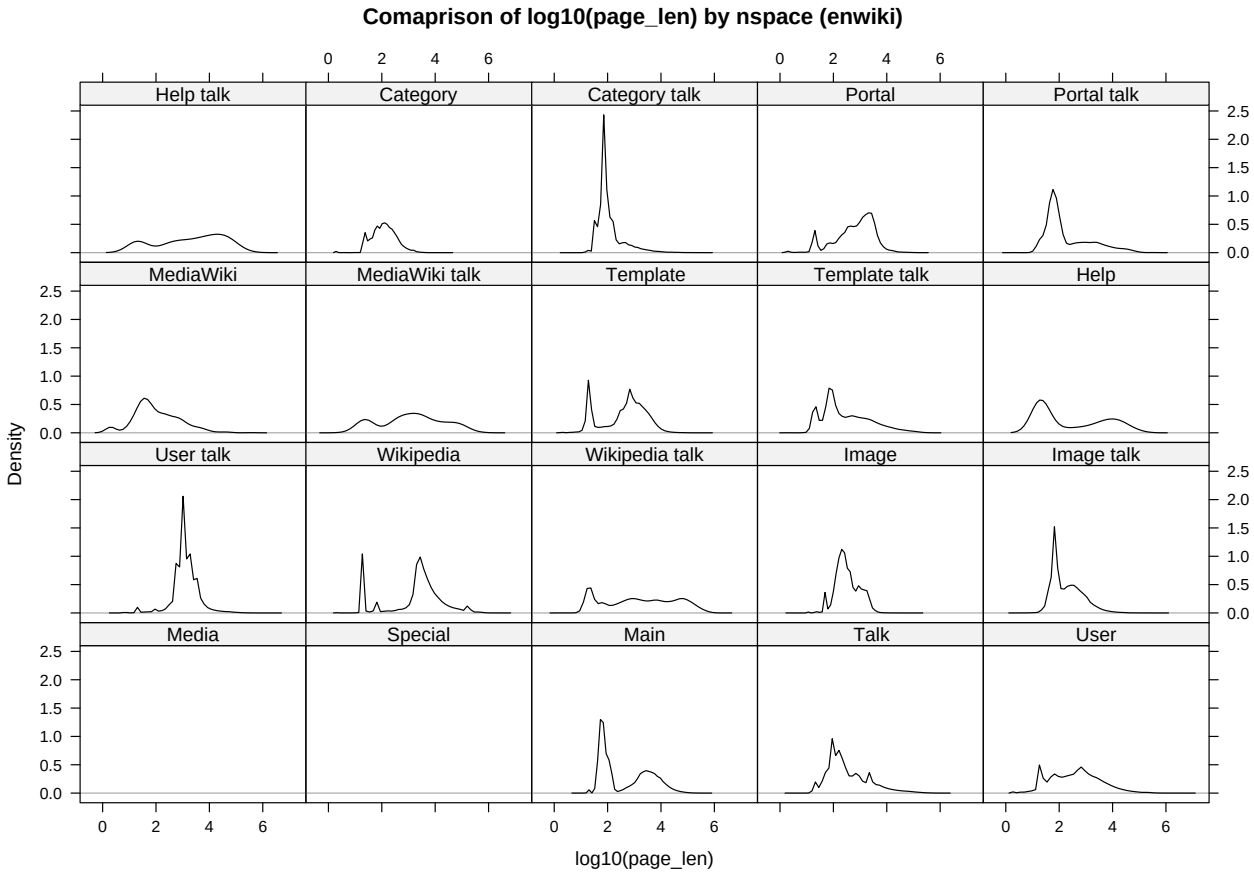

Figure 4.10: KDE of log10 of length of pages in bytes in the English Wikipedia, according to their namespace. In general, content pages present bimodal length distributions, while talk and discussion pages tend to present unimodal patterns

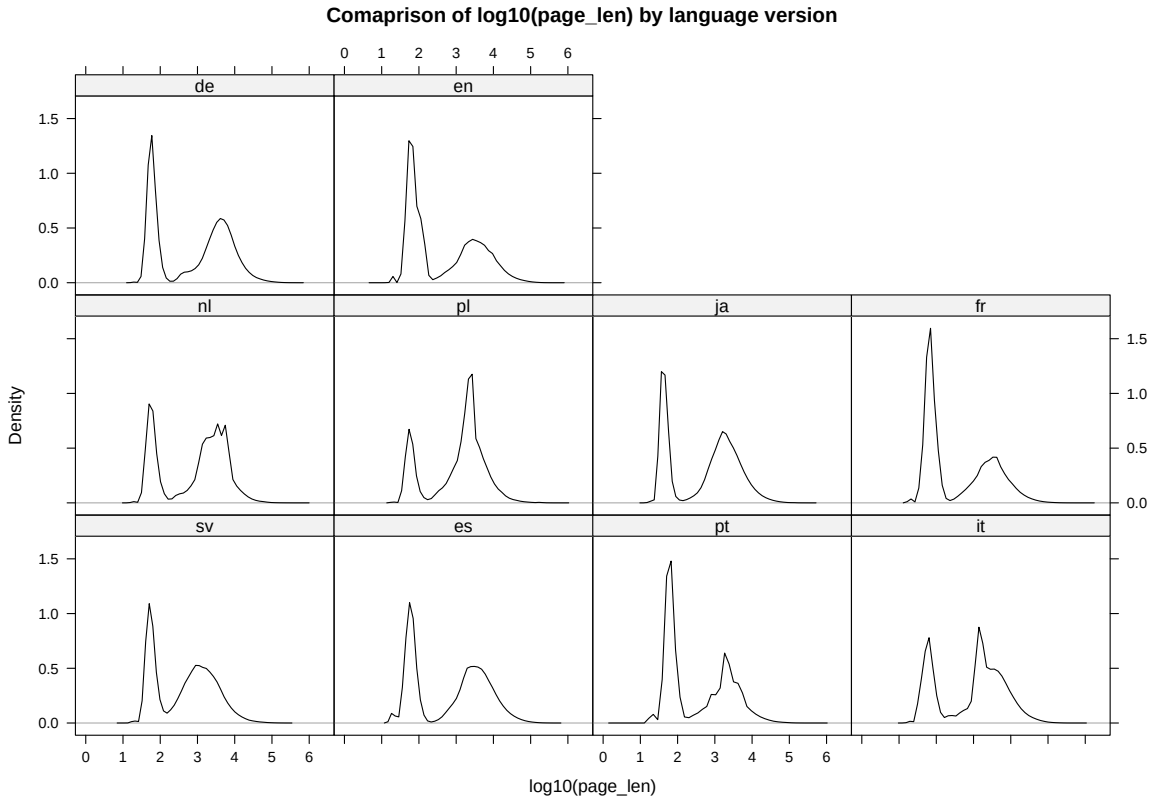

Figure 4.11: KDE of the log10 of length in bytes of articles in the top ten Wikipedias. The same bimodal pattern is found in all language versions. The left side population is composed by redirect and stub articles, while the right side population in each histogram comprises standard articles, which median length tend to be situated around 1,5 KB

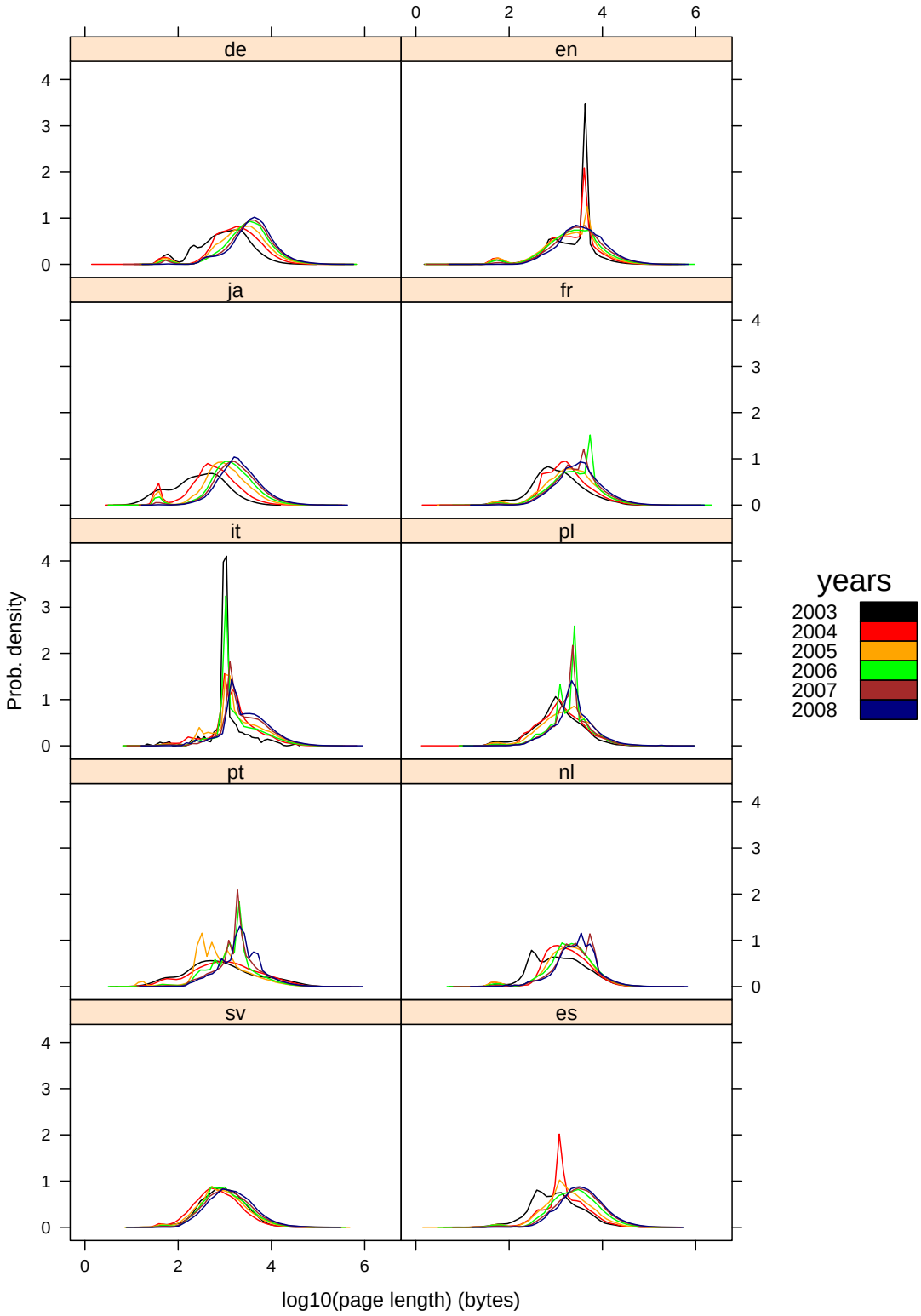

Figure 4.12: Evolution of KDE of the log10 of length in bytes of articles in the top ten Wikipedias. Globally, we can see a clear pattern of the median of the length increasing as time goes by in most language versions, with the notorious exception of the English Wikipedia. Therefore, articles tend to become longer over time, as they receive a higher number of revisions

However, if we look at the influence of the different number of authors in the length of each individual article, we can obtain a plot similar to that presented in Figure 4.13 for the French Wikipedia, for all language versions under analysis. The graph shows that, indeed, the length of articles is directly correlated with the number of different authors who participated in their creation, but this relationship is not so clear, since the highest length values are situated in the top left area, corresponding to articles produced by a small number of different authors. This is why we can still appreciate an increment in the length of articles even in the last year of history of our data samples, since the influence of the number of active users is not so strong as in previous cases.

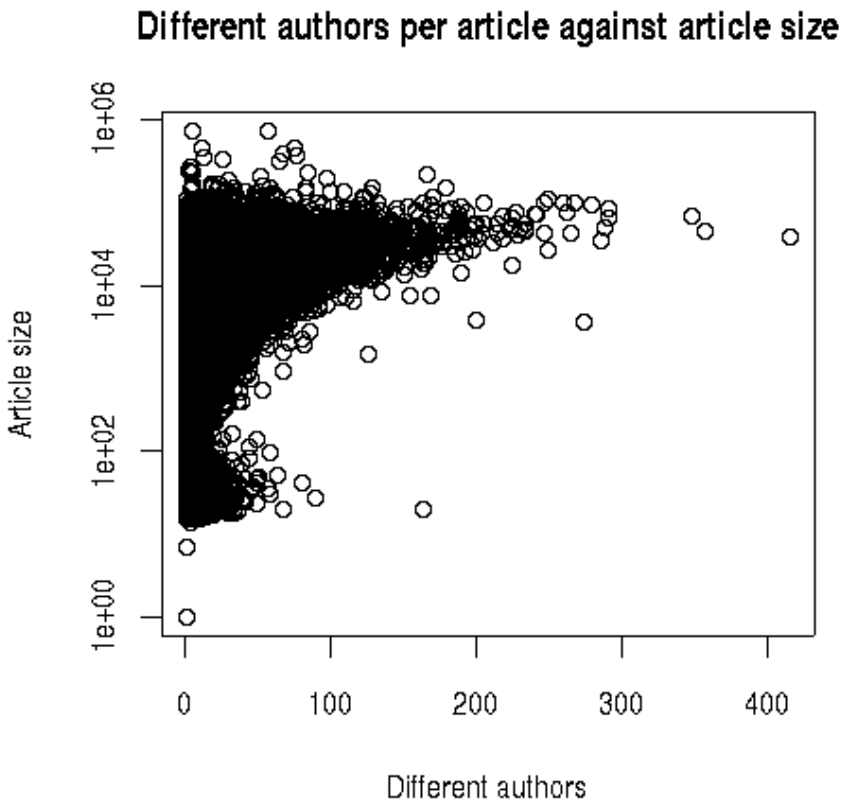

Figure 4.13: Scatterplot of length of articles against number of different authors per article in the French Wikipedia. The same pattern can be found in all remaining language versions under study. Indeed, the length of articles tend to increase as more different articles revise them, but there exists no direct correlation between these two variables, contrary to what we might expect intuitively.

4.2.2 Coordination and implication of authors

To conclude this global characterization of the top ten Wikipedias, we turn to analyze the contribution of logged authors to talk pages, the heart of the organizational process and discussions about articles content. Figures 4.14 and 4.15, show respectively the evolution of monthly number of revisions

received by talk pages, and the number of active logged authors in talk pages per month. The same deceleration in the steady growing rate of early years, already found in previous graphs, happens here too. Nevertheless, if we inspect Figure 4.16, depicting the number of active talk pages, we find a different situation. The number of active talk pages in all language versions has continued its steady growing trend, even in 2007. This a completely new, and unexpected pattern. According to these results, though the activity in the content creation process has suffered a leverage effect in the last year of history, and revisions to talk pages have followed the same pattern, logged authors seem to have concentrated their organizational effort on the creation of new discussion pages. This is specially significant for the French Wikipedia, which registered an even stepper increment in the beginning of 2007, that serve it to reach the same level found in the German language version.

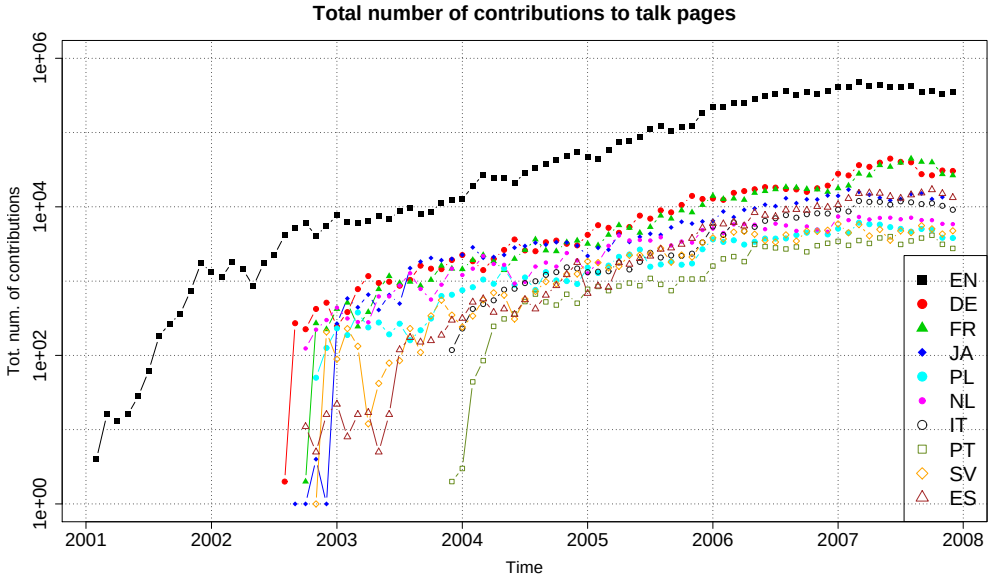

Figure 4.14: Evolution of the total number of revisions performed in talk pages of the top ten Wikipedias by logged authors. The number of contributions to discussion pages has also suffered the same leveraging effect, as the number of contributions in articles becomes approximately constant over the last year in all language versions.

The interesting aspect of this finding is that it complements the previous result already identified for logged authors. The content creation process of articles have lost vigority, but it seems that the interest of authors has turned to strengthen the organizational infrastructure of the project, opening new discussion forums for those articles that did not have one, or participating in existing talk pages in a more active way. The main consequence of this fact is that, though according to Figure 4.14 the community can not increase its number of contributions to talk pages (since the size of the group is not growing), it is extending their organizational interest to a growing number of articles. As we stated in Chapter 3, previous research works have pointed out the relevancy of talk pages as the core of the enhancement process of articles content, to obtain results of higher quality [121], [113], [54]. As more and more articles are included in the discussion process, this trend paves the way for improving the quality of a wider group of articles in Wikipedia. This is a chief goal for the project right now, and it seems that the repeated calls of Wikipedia's founder Jimmy Wales to concentrate authors effort on improving the quality of content are finding an answer from the community.

The consequence of this trend towards concentrating the community efforts on other namespaces different than the `main` one becomes even more clear in Figure 4.17. It shows the proportion of pages falling in each namespace, for the top ten Wikipedias. The impressive growing rate in the number of active talk pages in the French Wikipedia has produced a bunch of new discussion pages, helping this version reach the level found in the largest edition, the English Wikipedia. After them, the German and Japanese language versions exhibit the largest proportion of talk pages. But this is not the unique interesting finding that we can discover in this graph. User talk pages also deserve our attention, specially in the Portuguese Wikipedia, presenting and impressive proportion of this kind of pages, in comparison even with the largest Wikipedias. In the case of category pages, the smallest Wikipedias in this list, the Spanish and Swedish versions, surprisingly present the highest proportions of this kind of pages. This may indicate a more prolific effort in smaller language editions towards categorization of content and the creation of new taxonomies. The Swedish Wikipedia shows an interesting combination of a big proportion of article pages, the highest proportion of category pages, and a significant proportion of user talk pages, indicating a split interest within the community on both the creation of encyclopedic articles, as well as coordination aspects like content categorization and direct communication among users through discussion pages associated to their user page.

In fact, the analysis of the ratio of user pages per total number of logged authors can also render interesting conclusions. Table 4.3 shows these results for the top ten Wikipedias. The figures show that the typical ratio found in these language versions varies between 40 and 50%, that is, approximately one personal page every two authors. The most relevant exceptions for this general rule are the French and German Wikipedias, exhibiting a substantially higher proportion, and the Spanish version, which seems to be lagging behind the overall trend.

Language	Num. logged authors	Num. user pages	Ratio
EN	1,824,439	543,431	29.79%
DE	226,912	129,650	57.14%
FR	127,767	78,280	61.27%
PL	51,796	25,062	48.39%
JA	90,828	37,195	40.95%
NL	60,749	26,029	42.85%
IT	62,690	27,346	43.62%
PT	64,994	28,769	44.26%
ES	132,239	46,554	35.20%
SV	26,972	11,034	40.90%

Table 4.3: Total number of logged authors, user pages and ratio (number of user pages per user) in the top ten language versions of Wikipedia

Likewise, we can analyze the ratio of talk pages per total number of articles found in all language editions under study. We can also learn interesting aspects from these results, presented in Figure 4.4. Obviously, the first result that focus our attention is the outstanding proportion of talk pages per article found in the English Wikipedia, well beyond any limits reached by the other language versions, establishing an astonishing highest ratio of 80.8%. Behind it, the French and German Wikipedias outstand from the rest of language versions. It seems that the faster rate of growth of monthly active talk pages in the French versions has been translated in a strong increment in the total number of new talk pages, almost getting up to a 60% of talk pages per total number of articles. On the other side, the extremely low ratio found in the Polish language version sets out interesting theories about the

source of efforts in this language version. The combination of a very active cohort of bots, together with the very low ratio of talk pages, indicates that the Polish language version is not following the same organizational pattern found in other language editions. Such a low ratio of talk pages points out the little effort undertaken on coordination actions and discussion about article contents in the Polish version.

Language	Num. articles	Num. talk pages	Ratio
EN	2,183496	1,764252	80.80%
DE	700,032	219,520	31.36%
FR	629,927	366,512	58.18%
PL	475,428	40,061	0.084%
JA	476,457	92,712	19.46%
NL	412,994	48,898	11.84%
IT	416,694	83,707	20.09%
PT	363,552	84,174	23.15%
ES	338,792	73,562	21.71%
SV	273,968	41,701	15.22%

Table 4.4: Total number of articles, talk pages and ratio (number of talk pages per article) in the top ten language versions of Wikipedia

Finally, we can also study the patterns found in the KDE curves for the log10 of length of talk pages in the top ten Wikipedias, along with their evolution over time. These results are depicted in Figures 4.18 and 4.19. In the first of these graphs, the pattern presented by the French Wikipedia focus our attention immediately, due to the sharp and high peak in its KDE curve. Recalling that this is the language version which has experimented the strongest growth rate in its number of monthly active talk pages, and that it presents the second highest proportion of talk pages per article, the KDE curve suggests that a significant proportion of talk pages in this version have been created recently. This hypothesis is corroborated by the evolution of the KDE curve depicted in Figure 4.19. In the early years of its history, the KDE curve is smoother, until we reach January 1, 2008. The blue colored curve confirms that the sharp peak was not found in previous years, focusing our attention on the sudden creation of a high number of talk pages in this last year of history. Another interesting general pattern in the evolution of the length of talk pages is the opposite trend found for the median of length, with respect to that found for the evolution of the length of articles. The exception to this global trend is the German Wikipedia, which seems to maintain a quite stable KDE curve for this statistic. The interpretation of this results is twofold. In the first place, we know that talk pages periodically suffer from a recycling process, archiving old discussions and preventing talk pages to extend their length indefinitely. In the second place, the more new talk pages are opened in each language version, the greater the density of shorter talk pages in the KDE curve (since we do not expect new talk pages to suddenly reach the median length just before their inception). Together, both points of view indicate that the global trend followed in the majority of language versions is towards the creation of new talk pages for articles, expanding the coverage of the discussion process about articles content.

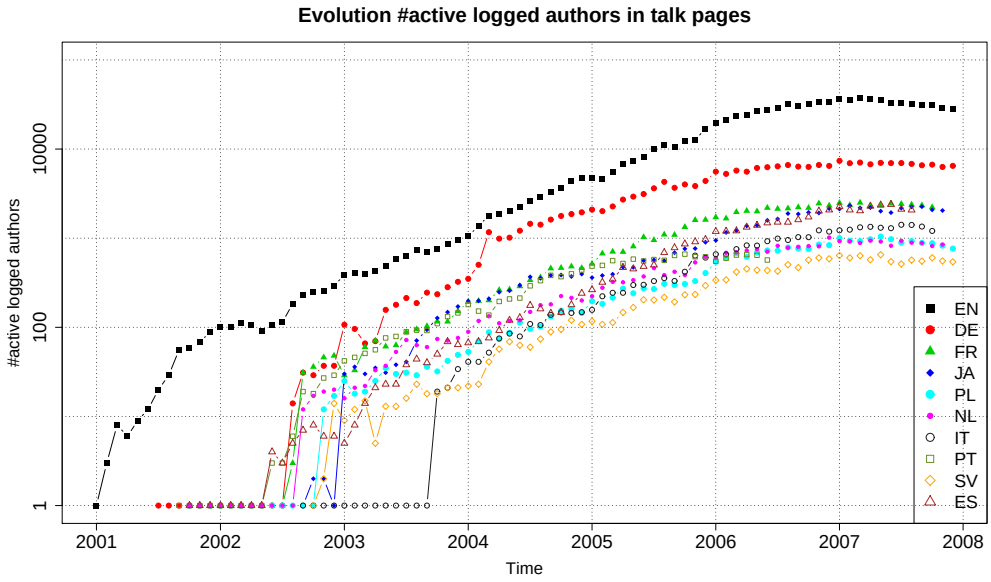

Figure 4.15: Evolution of the total number of number of active logged authors per month participating in talk pages in the top ten Wikipedias. Once again, the leverage of the number of contributions to talk pages over the last year of all language versions can be explained by the stabilization of the number of active authors contributing to talk pages

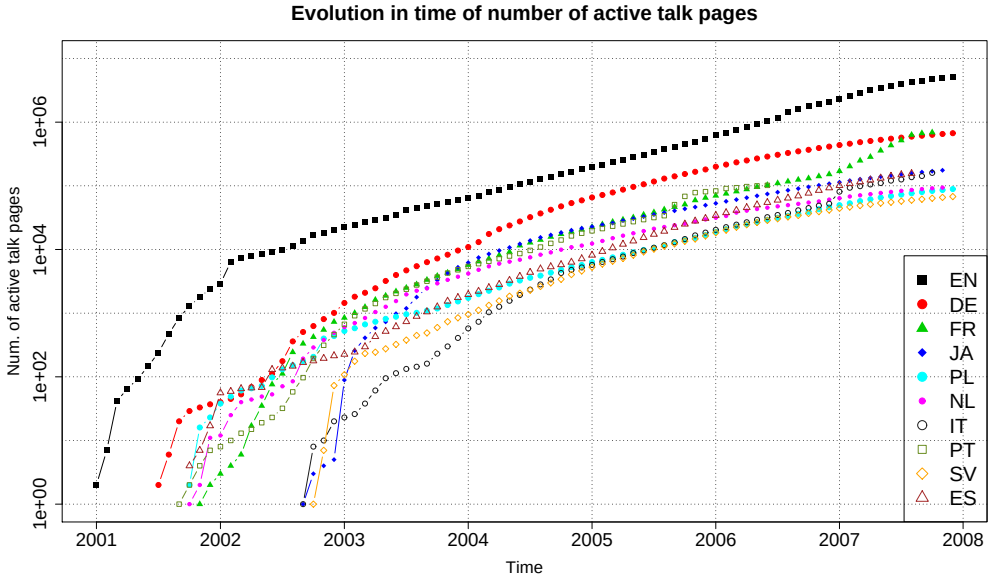

Figure 4.16: Evolution of active talk pages per month in the top ten language versions of Wikipedia

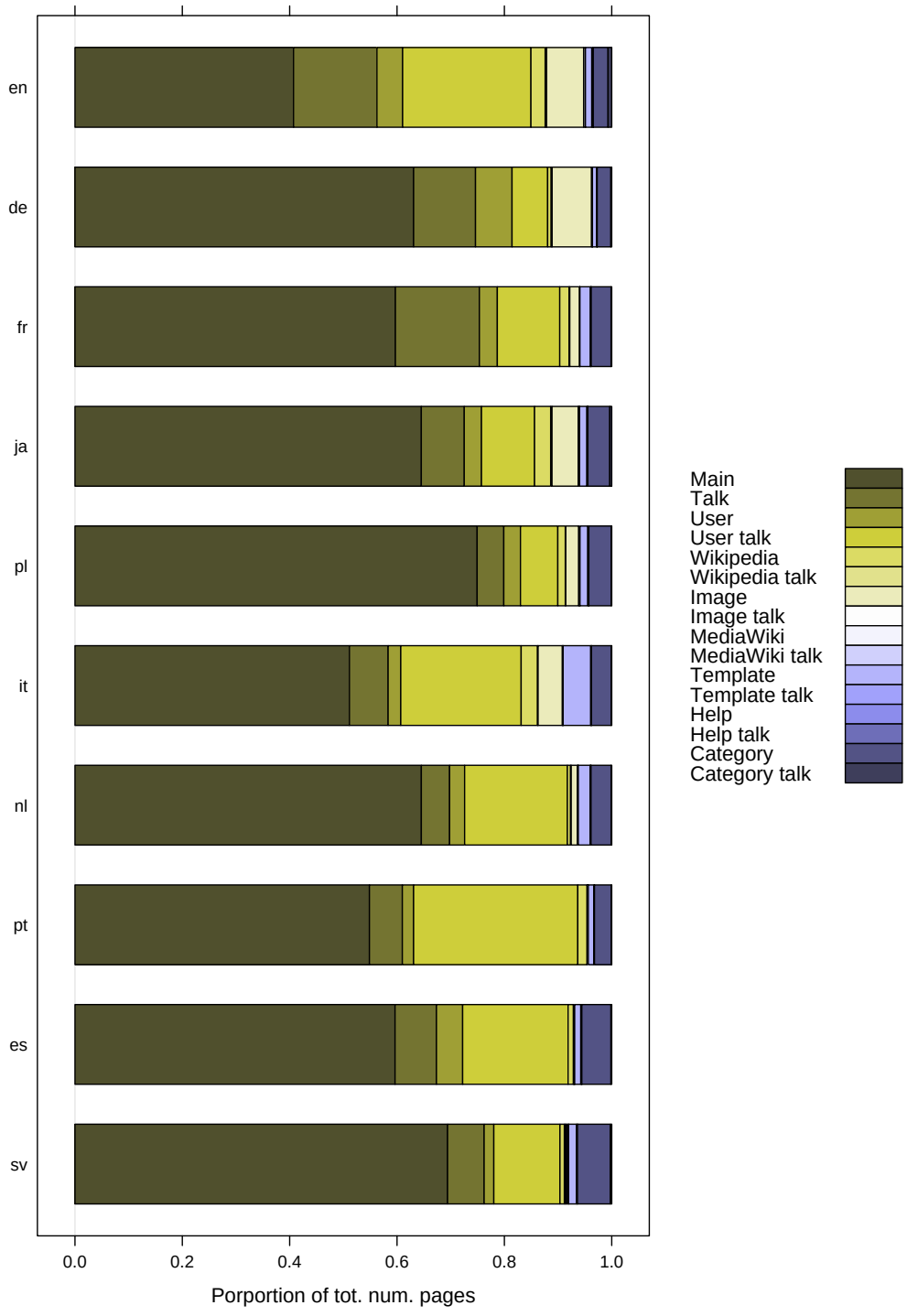

Figure 4.17: Proportion of total number of pages falling in every namespace for the top ten Wikipedias. In some language editions, the proportion is strongly biased towards the `main` namespace (which stores articles), while other versions present a strong bias towards discussion pages, like talk pages (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)

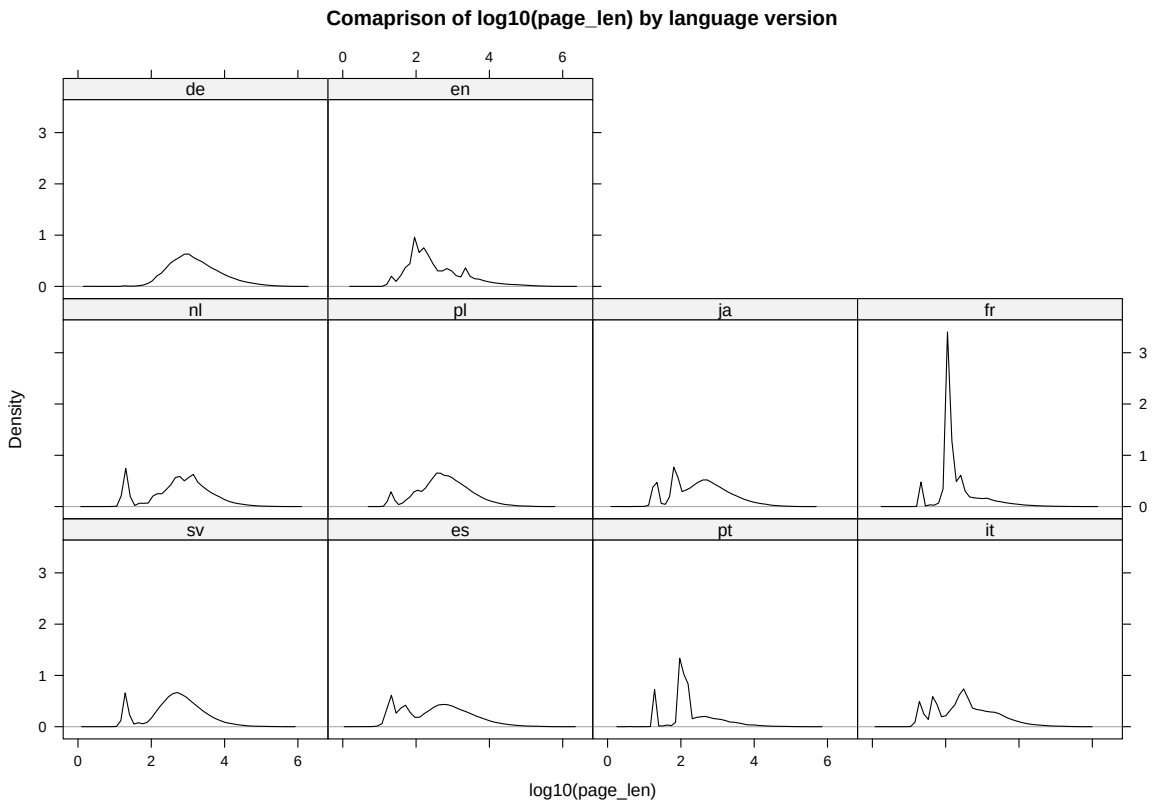

Figure 4.18: KDE of the log10 of length in bytes of talk pages in the top ten Wikipedias. The extremely high peak in the French language version focus our attention on the sudden creation of a large number of talk pages in this Wikipedia

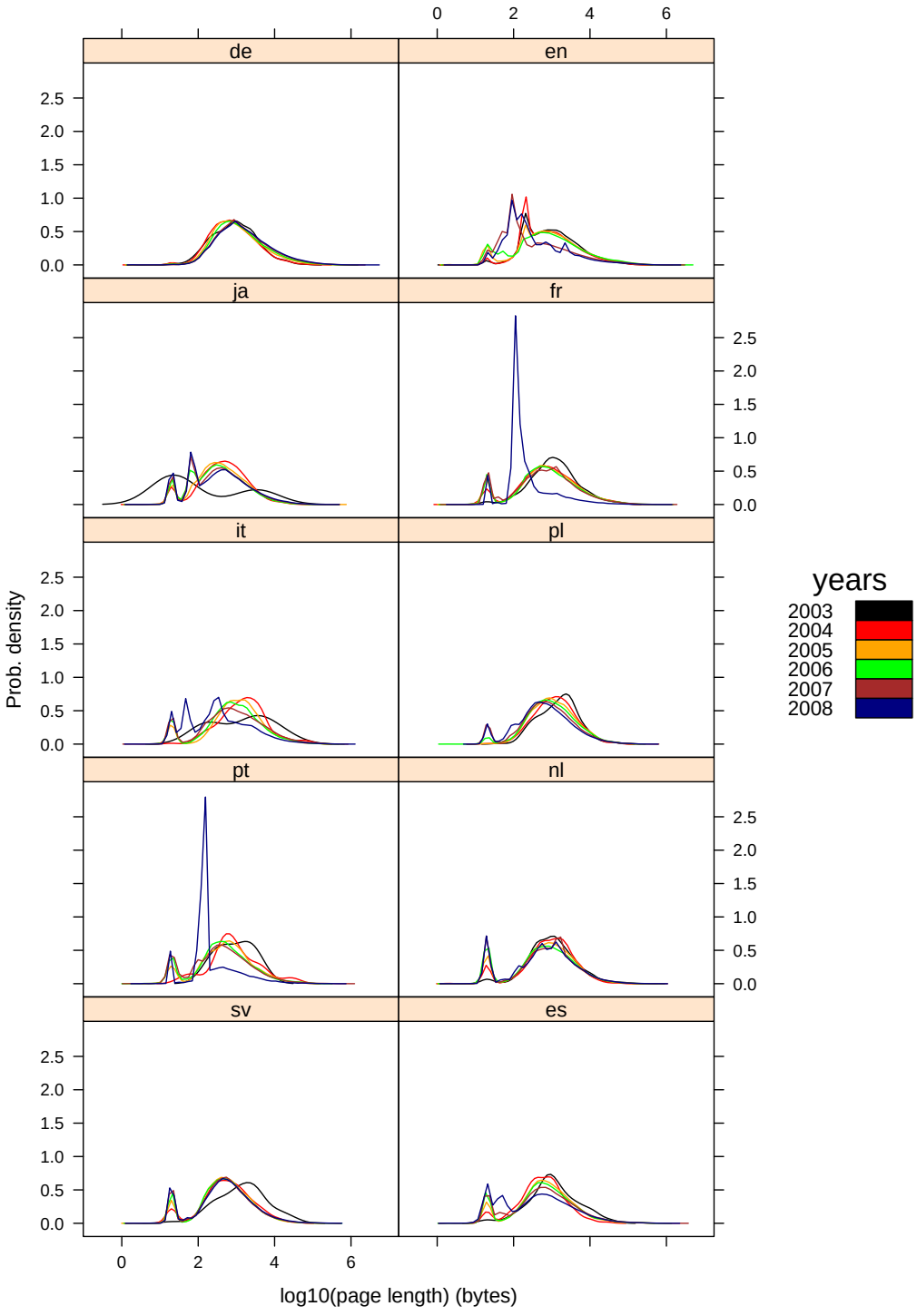

Figure 4.19: Evolution of KDE of the log10 of length in bytes of talk pages in the top ten Wikipedias. Contrary to what we saw in the case of articles, the median of the length of talk pages tend to become lower as the language versions evolve over time. The most relevant exceptions are the German and Spanish Wikipedias, which have maintained a similar density graph over their whole history

4.3 The Social Structure of Wikipedia

The second section of our empirical study of the top ten Wikipedias is devoted to the analysis of the social structure found in their communities of logged authors. In general, we will be interested in measuring the inequality level of the distribution of activity among logged authors, as well as the different distributions exhibited by relevant statistics that will help us to characterize the work of logged authors in each version. To implement this analysis we follow the common approach of using continuous approximations of known probability distributions to fit discrete variables in the corresponding cases, in particular, when dealing with statistics that follows Pareto like models, that present a characteristic log-linear pattern in their CCDF curve.

The process of adjusting a theoretical probability distribution to empirical data may become quite inaccurate if we do not follow the appropriate preventive measures and strong methods to fit our curves to the data found in our samples. In this respect, we try to comply as much as possible with the invaluable approaches and recommendations found in [25], one of the best papers written about this topic. The analysis of power laws and Pareto fits has also deserved an ample coverage on previous research works. Most notable examples are [74] and [48] and [47], which show that the most secure approach to fit parameters of these distributions is through M.L.E (*Maximum Likelihood Estimators*). This approach is also shown to produce unbiased estimators of these coefficients. As well, they present how to calculate their standard deviation. The authors of [25] have also written several useful library files for GNU R that implement these methods. We have applied these algorithms to obtain the fits presented in these thesis work [2].

Throughout this analysis, we come across 3 different types of probability distributions. All of them will be examined and fitted using the CCDF plot, thus using the better known method for this purpose, so far. The 3 distinct types of theoretical distributions found in our data are:

- Pareto distribution: This distribution follows a straight line all along the entire range of its CCDF plot, when we use a logarithmic scale for both axes. The mathematical properties and other interesting characteristics of this distribution can be found in [74]. The slope of the line α is the characteristic parameter of this distribution. Sometimes,the empirical data only follows the straight line shape from a minimal value, which is usually known as `xmin`. The methodology presented by [25] and followed in this thesis work produces the M.L.E. of both parameters, along with the maximum distance from the fitted line to the empirical data.

- Upper truncated Pareto distribution: Similar to the previous one, it follows a straight line along its lower values, but it suddenly drops off from a certain upper limit value. The algorithm presented in [4] is applied in the `VGAM` library of GNU R to find the M.L.E. of the lower, upper limits and the slope of the distribution, using generalized linear models.

- Lognormal distribution: The lognormal distribution presents a characteristic curved shaped all along its CCDF curve, without any straight line throughout its range. The `fitdistr` function, included in the `MASS` library provides a good tool to fit this family of theoretical distributions to empirical data. The logarithmic mean and standard deviation are the characteristic parameters defining this distribution.

A word of caution is in order here. Researchers trying to fit Pareto like distributions would

[2] At the time of this writing, the source code retrieve to perform this analyses was retrieved from Aaron Clauset's web page about power laws, on `http://www.santafe.edu/~aaronc/powerlaws/`

benefit a lot from reading the interesting explanation provided by Lada A. Adamic [3] about the correct correspondence between the slopes of the straight line found in the CCDF function of Pareto distributions, and the slope of the power law line followed by the distribution of the statistic. In the case of the method followed by Clauset *et al.*, the library adjusts the slope of the power law function, and therefore, the slope that we must use to represent the fit in the CCDF graph is $\alpha + 1$. On the contrary, the `VGAM` library will produce the estimator of the slope in the CCDF curve, and we must perform the opposite operation to obtain the slope of the power law distribution. In the tables included in this section, we will present the results directly obtained from running the corresponding algorithms for each case. We provide these advise to avoid any misunderstandings for those readers interesting in subsequently validating our findings.

A second useful remark must be added at this point. Since $\alpha < 2$ for all our fits, it means that the process have neither a finite mean, nor a finite variance or standard deviation. Therefore, these values can not be calculated in our case, and this is the reason for not including them in subsequent tables. In section D of [74] we can also find an interesting explanation of the relationship of the Pareto distribution and the Lorenz curve, a tool that we will use to measure the inequality of contributions from authors in the following subsection. Finally, Section E on the same paper talks about a remarkable property of Pareto like distributions, which is that they are scale-free. Since the shape of the distribution curve is a straight line, no matter which point we select the slope is constant, and so is the percentage variation of the statistic under study.

Those readers interested in learning additional details about the different probability distribution functions utilized throughout this thesis are referred to Appendix B. In it, we introduce the formulae and most relevant properties concerning power laws, Pareto distributions and and Zipf's law. We also introduce the basic properties and formulae corresponding to the lognormal probability distribution.

Once we have a clear roadmap to perform our analysis, we start studying several statistics that characterize the contribution of Wikipedia logged authors. Figure 4.20 shows the CCDF of the different number of articles edited per author in the top ten Wikipedias. As we can see, all language editions follow an upper truncated Pareto distribution. The curve drops off at the natural upper limit reached in each language edition for the maximum number of different articles that a single author can manage in each language edition. This is an interesting result, specially when combined with the stabilization in the number of active logged authors per month during 2007. Human authors have a limited capacity to attend a certain number of different articles. Though some of these authors can develop an impressive coverage capability (namely, one human author in the Spanish Wikipedia was capable of revising more than 80,000 different articles over 3 years), most logged authors can not do so. Therefore, these graphs evidence that Wikipedia needs to attract new authors, of any condition (both super active authors and standard ones), in case the project wants to avoid the uncomfortable situation of watching how the number of contributions received per month gets bogged down. Since each individual author has a limited coverage, we need to constantly renew the population of authors in each language version to maintain the creative capacity of the community.

Figure 4.21 shows a frequency scatterplot of the same results, along with the optimal fit of a power law line, obtained with the aforementined algorithms. We can check in this graph that the fit is quite accurate over most of the range of the upper truncated Pareto distribution. However the rightmost side of all graphs always poses some problems for the algorithms to adjust the curve, specially when the number of different articles revised by some super active authors is disproportionately high. Though we have done our best to try and filter out as much bots as possible, there may still exist some bots not properly identified as such in the `user_groups` table of its language version. Despite

[3] `www.hpl.hp.com/research/idl/papers/ranking/ranking.html`

these difficulties, we believe that the results obtained are reasonably accurate, looking forward to have a more precise list of bots for each language version. Table 4.5 summarizes the characteristic coefficients obtained with our fitting procedures.

Language	**Slope**	**Lower**	**Upper**	**Loglik.**
EN	0.5836	1	103,266	-2241236
DE	0.4859	1	97,301	-370835.7
FR	0.4978	1	59,496	-200126.6
PL	0.4820	1	28,203	-85809.3
JA	0.4397	1	36,932	-179696.9
NL	0.5298	1	61,710	-83501.6
IT	0.5077	1	30,182	-95904.4
PT	0.5817	1	27,884	-72800.57
ES	0.5611	1	29,131	-150446.4
SV	0.4835	1	31,671	-46663.28

Table 4.5: Characteristic coefficients of the fitted upper truncated Pareto distributions for the total number of different articles revised per logged author in the top ten Wikipedias: the slope of the CCDF, the lower and upper limits and the Loglik. value obtained in the process

In the same way, we also fit the number of different logged authors who have revised every article in the top ten Wikipedias. Contrary to the results obtained for the previous statistic, the number of different authors per article does not follow any Pareto-like distribution, not even an upper truncated one. The CCDF plot shows a clear curved shape all along the graphic for all language versions, and actually, we confirm this hypothesis fitting a lognormal distribution to our data, and proving that the fit is accurate for the major part of each curve. Previous research works stated that this statistic followed a power law [18]. However, that claim does not hold for the *current* version of the data samples, though we will check later that the claim was valid for the early years in the history of all versions. Figure 4.23 shows the results of our fit in the CCDF domain, and Figure 4.23 depicts the frequency scatterplot of the same statistic. Looking at Figure 4.23, we might be tempted to fit a power law straight line to the upper values of the scatterplot. Nevertheless, the CCDF clearly shows that this is not at all a good idea, and thus demonstrates the utility of this graph to avoid possible mistakes during the fitting process. All fitted coefficients can be found in Table 4.6, including the standard errors of both parameters and the log-likelihood value for each fitting procedure.

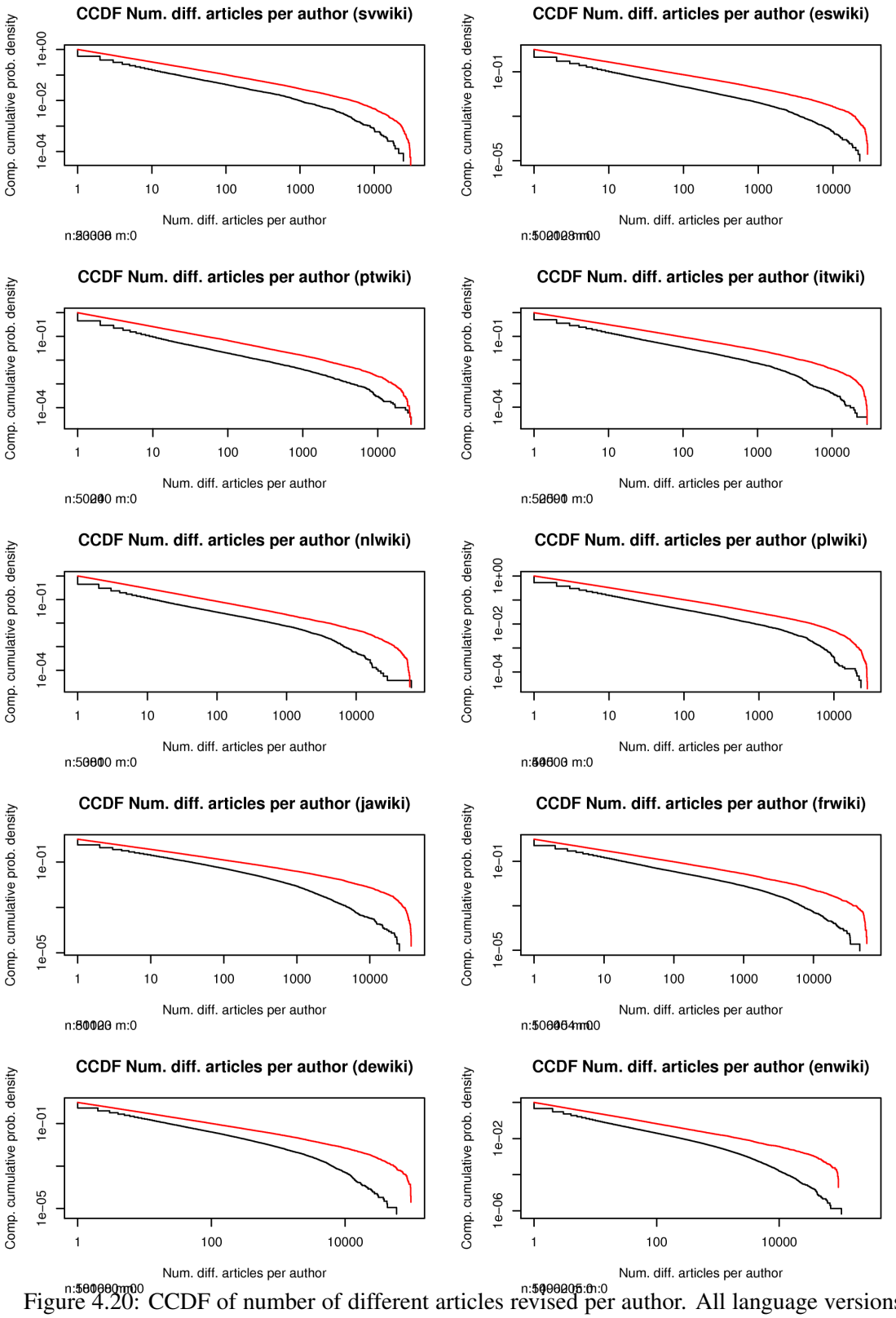

Figure 4.20: CCDF of number of different articles revised per author. All language versions seem to follow an upper truncated Pareto distribution. There exists a natural higher limit established by the maximum number of different articles that can be revised by a human author

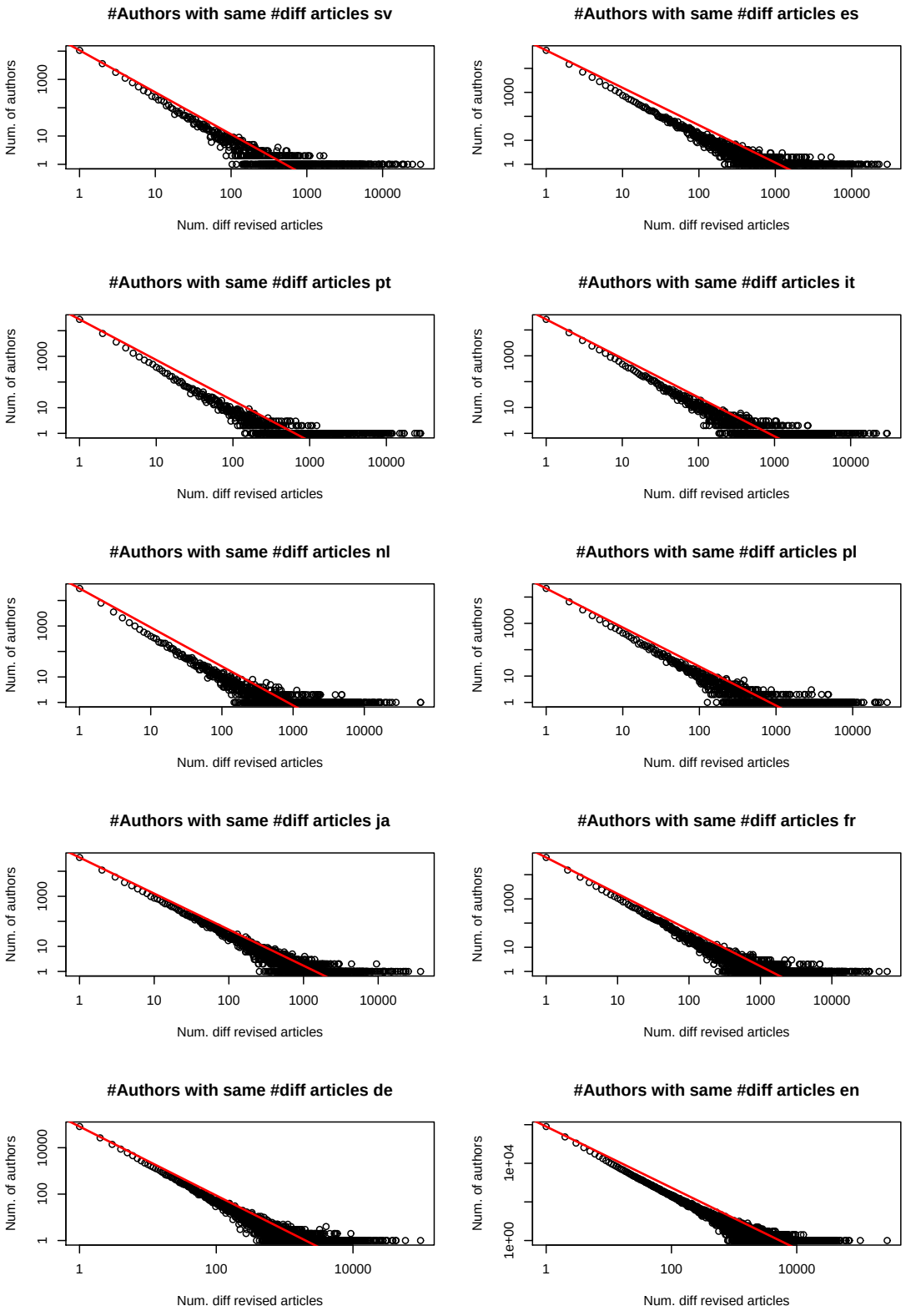

Figure 4.21: Scatterplot showing the number of authors sharing the same number of different articles revised per author, in the top ten Wikipedias. We also draw the best fit line, which follows a power-law, in all language versions for comparative purposes

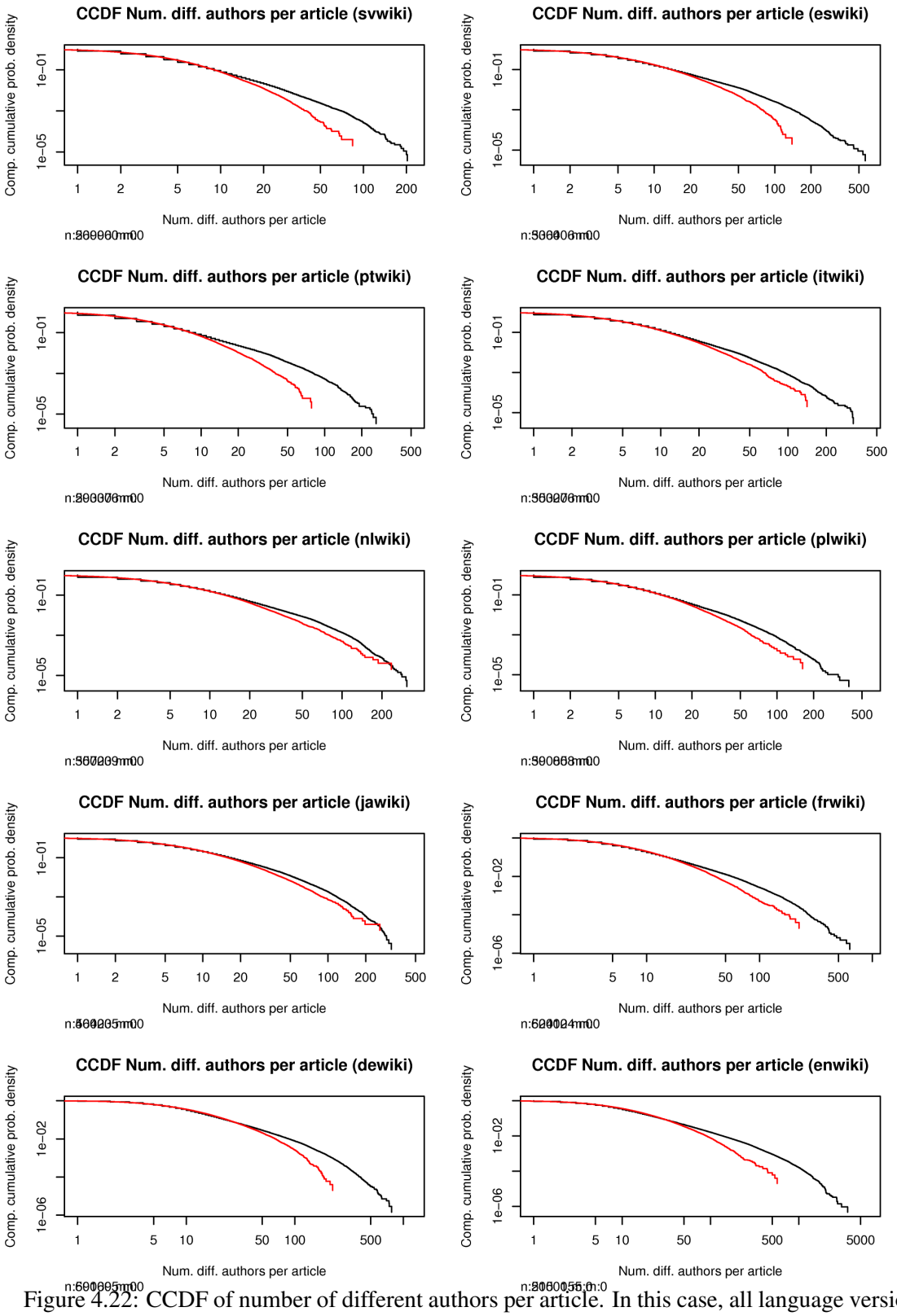

Figure 4.22: CCDF of number of different authors per article. In this case, all language versions seem to follow a standard lognormal distribution.

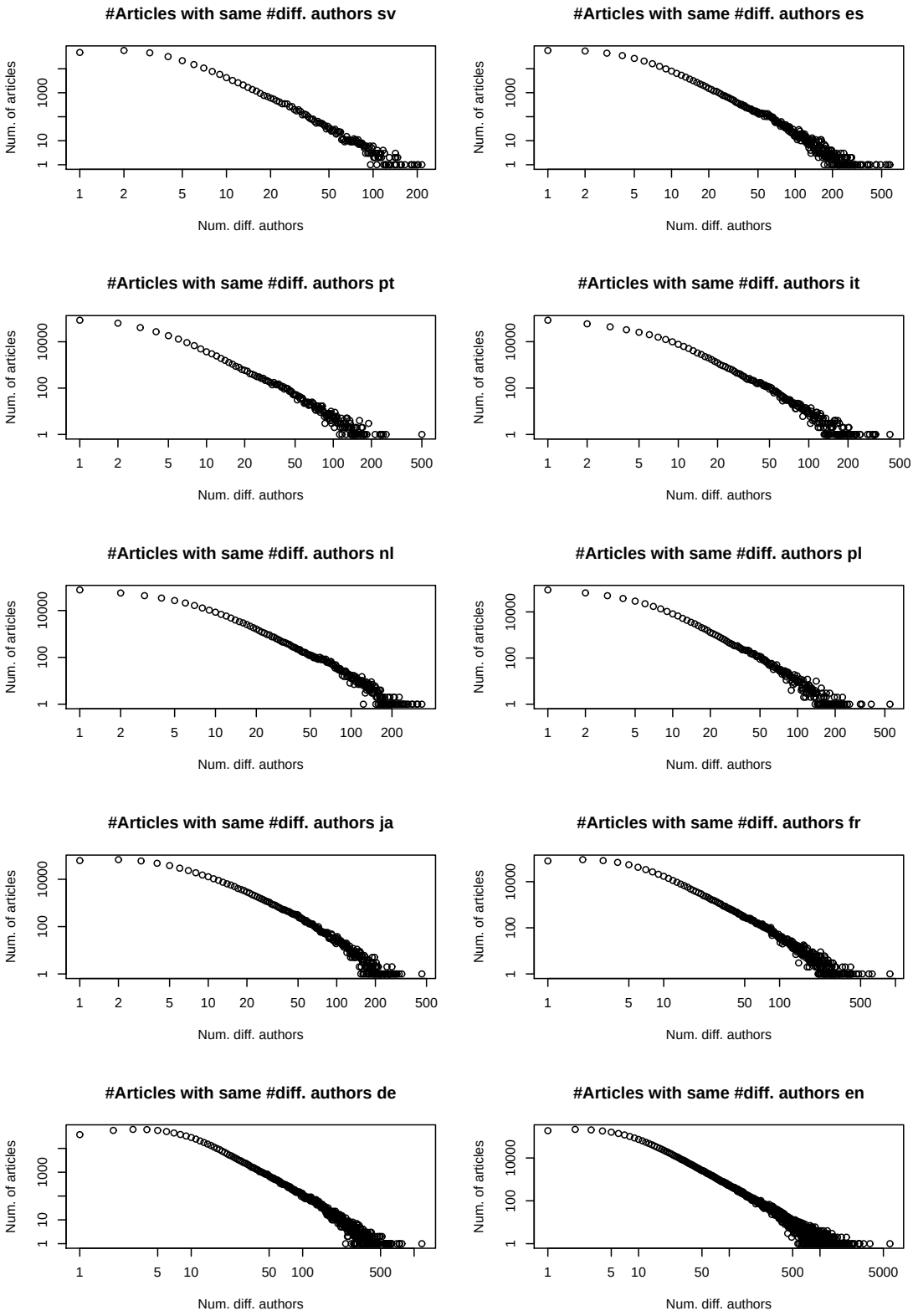

Figure 4.23: Scatterplot representing the number of different authors per article for the top ten Wikipedias. In this case, the graphs do not include a power law fit, since the CCDF shows that the statistic follows a lognormal distribution for all language versions under study

Language	Meanlog	sd(meanlog)	Sdlog	sd(sdlog)	Loglik.
EN	1.9007	0.00076	1.1078	0.00053	-7,357,916
DE	1.9456	0.00116	0.9622	0.00082	-2,300,611
FR	1.5064	0.001198	0.9460	0.000847	-1,791,203
PL	1.2157	0.00149	0.9308	0.00105	-1,001,737
JA	1.5220	0.00143	0.9710	0.00101	-1,351,652
NL	1.3104	0.00164	0.9775	0.00116	-966,905.2
IT	1.2122	0.00159	0.9451	0.00112	-909,562.8
PT	0.9907	0.0016	0.8669	0.00113	-665,016.5
ES	1.3844	0.00167	0.9688	0.00118	-932,400.7
SV	1.1691	0.00156	0.8089	0.0011	-641,398.4

Table 4.6: Descriptive coefficients of the fitted Lognormal distributions for the total number of different logged authors per article in the top ten Wikipedias: the mean and sd deviation of the curve in log scale, the standard errors of both parameters and the loglikelihood value obtained in the fit procedure

However, if we make a slight change in the selection of our statistics, we can obtain even more interesting results. Instead of plotting the CCDF of the number of different articles per author, we can depict the CCDF of the number of authors sharing the same number of different articles revised. In the same way, instead of analyzing the CCDF of the different number of authors per article, we can plot the CCDF of the number of articles sharing the same number of different authors. Figures 4.24 and 4.25 plots these graphs for the top ten Wikipedias. The remarkable property of these statistics is that both follow a Pareto law, a standard Pareto in the first case and an upper truncated pattern in the second place. What is the meaning of these results? In the case of authors, this statistic shows that the distribution of different types of authors according to their willingness to revise articles is *scale-free*. That is, no matter the scale we look at on this distribution, the shape of the curve is unchanged, with the exception of an overall multiplicative constant [74]. This scale-free property is only applicable to a power law probability distribution.

In the case of articles, the upper truncated Pareto fit suggests a similar interpretation, except for the fact that there seems to be a small group of quite popular articles accumulating an extremely large number of different authors, thus deviating the curve from a perfect scale-free behavior. As we will see later, the are evidences that within these group of quite popular articles, we can find the FAs produced in each language version.

In fact, Pareto-like (or power law) behavioral patterns have been frequently found in diverse scientific areas [74], [25]. Usually, in nature we find power law distributions having $2 \leqslant \alpha \leqslant 3$, but this is not the case for the number of logged authors per article in the top ten Wikipedias. The slopes are milder, ranging from 1.6 to 1.7 depending on the language version selected. It is notorious the short range covered by the slope values, showing quite similar behavioral patterns for all language versions, disregarding factors like the total number of articles produced or the size of the community.

It is remarkable that an self-organized, totally open community project like Wikipedia also follows a Pareto-like pattern, proving that human driven projects naturally tend to produce unequal effort distributions unless they are explicitly forced otherwise. In the case of Wikipedia, the patterns found for these statistics confirms our previous research findings about the presence of a *core group* of very active authors in each language version, responsible for a substantial proportion of the total creation effort in articles content. Regarding the Pareto shape area of the CCDF graphs, they clearly

demonstrate that the stratification of authors in each contribution level is scale-free, facilitating the classification of articles showing different contribution levels, since the slope of the CCDF curve in that region is constant.

Language	**x_min**	α	**D**
EN	4	1.61	0.0116
DE	4	1.65	0.00933
FR	7	1.63	0.025
PL	5	1.66	0.0191
JA	2	1.7	0.0184
NL	5	1.65	0.0248
IT	5	1.63	0.0299
PT	2	1.66	0.037
ES	2	1.67	0.0198
SV	3	1.67	0.02935

Table 4.7: Characterization coefficients of Pareto fits for the number of authors sharing the same number of different articles revised. The table includes the `x_min` value for the Pareto fit, the slope of power law line corresponding to this statistic, α and the maximum distance `D` from the empirical CCDF curve to the fitted Pareto distribution

Language	**Slope**	**Lower**	**Upper**	**Loglik.**
EN	0.2812	1	215,353	-4745.785
DE	0.1972	1	62,781	-2071.932
FR	0.2056	1	90,155	-1483.798
PL	0.1794	1	89,472	-929.5833
JA	0.1546	1	67,528	-1197.091
NL	0.1746	1	75,803	-1041.346
IT	0.1874	1	84,541	-920.9753
PT	0.21395	1	85,407	-793.5393
ES	0.2199	1	57,358	-1214.447
SV	0.1891	1	57,929	-701.4283

Table 4.8: Descriptive coefficients for the upper truncated Pareto fit for the number of articles sharing the same number of different authors. The slope of the Pareto fit, lower and upper limits of the adjusted curve and the loglikelihood ratio are provided

If we turn now to the analysis of the number of revisions per author and per article, we check that the distributions followed by these statistics are exactly the same as in the previous case. This demonstrates the close correlation between both types of metrics in the case of the top ten Wikipedias. The CCDF of the number of revisions per author, depicted in Figure 4.26 follows an upper truncated Pareto fit, and the number of revisions per article, displayed on Figure 4.28 follows a lognormal probability distribution. The corresponding scatterplots of the frequency distribution of these statistics are plotted in Figures 4.27 (along with the optimal power law fit) and 4.29. The accompanying tables provide all descriptive coefficients for the fitted curves in each language.

According to [69], there exist several possible generative models that explain Pareto behavioral patterns in natural processes. In the case of Wikipedia, recent research work published by Spinellis

and Louridas [106] reveals that the network of links in Wikipedia also conforms a scale-free network, in the English Wikipedia, following a Pareto distribution. These findings offer a viable hypothesis for the cause of this phenomenon, namely the influence of a **preferential attachment** creation pattern. Given that the behavioral pattern found in authors and articles exhibits the same properties, we can conclude that there also exists a preferential attachment process in the way that revisions are distributed among authors. Therefore, the number of revisions performed by articles is influenced by the underlaying preferential attachment process of the Wikipedia network of links in articles, and viceversa, demonstrating that both processes are tightly coupled in this language version. As far as we know, this is the first indication found about a link between the visibility of Wikipedia articles and the subsequent distribution of effort spent by Wikipedia authors. These findings also support the development of further research work to find stronger evidences of the correlation between more visible articles (in terms of the number of links appearing in other Wikipedia articles) and the decision of authors selecting which article to revise.

Looking beyond these findings for Wikipedia, future applications of these metrics to other open collaborative projects is straightforward. It will be interesting to check wether these creation models can also explain the behavioral patterns found in other different collaborative communities, such as FLOSS development projects and open multimedia repositories. Further exploration should also be conducted to analyze the relationships between behavioral patterns of contributors in these projects and the structure of the network of content contributed by these contributors.

Language	Slope	Lower	Upper	Loglik.
EN	0.5009	1	147,696	-3738725
DE	0.4205	1	138,458	-557195
FR	0.4208	1	79,107	-315560.8
PL	0.4207	1	52,625	-131048.9
JA	0.3866	1	73,683	-255536.6
NL	0.4734	1	180,590	-136725
IT	0.4125	1	56,942	-159001.9
PT	0.4762	1	63,493	-132464.2
ES	0.4634	1	64,965	-268717.6
SV	0.4298	1	65,407	-67477.85

Table 4.9: Characteristic coefficients of the fitted upper truncated Pareto distributions for the total number of revisions per logged author in the top ten Wikipedias: the slope of the CCDF, the lower and upper limits and the loglikelihood value obtained in the process

Table 4.10: Descriptive coefficients of the fitted Lognormal distributions for the total number of different logged authors revising each article in the top ten Wikipedias: the mean and sd deviation of the curve in log scale, the standard errors of both parameters and the loglikelihood value obtained in the fit procedure

Language	Meanlog	sd(meanlog)	Sdlog	sd(sdlog)	Loglik.
EN	2.5156	0.00085	1.2406	0.00059	-8,923,395
DE	2.4721	0.00131	1.0877	0.00092	-2,749,514

FR	2.1177	0.00140	1.1039	0.00098	-2,268,985
PL	1.7223	0.00177	1.1069	0.00125	-1,267,461
JA	1.9777	0.00165	1.1248	0.00117	-1,631,481
NL	1.8247	0.00194	1.1628	0.00137	-1,212,643
IT	1.8533	0.002	1.1894	0.0014	-1,217,280
PT	1.4544	0.00202	1.0930	0.00142	-869,079.1
ES	2.0332	0.00199	1.1534	0.0014	-1,209,342
SV	1.6369	0.00187	0.9706	0.00132	-816,883.5

4.3.1 Measuring inequality of contributions with Gini coefficients

Another interesting analysis, to quantify the inequality level of contributions performed by logged authors on Wikipedia articles, can be conducted using the Lorenz curves and Gini coefficients. Figure 4.30 presents the empirical Lorenz curves found for the distribution of revisions among logged authors in the top ten Wikipedias. None of these language versions has experimented significant changes from the situation presented in our previous research work about this topic [82]. The inequality levels found for all versions indicate a strong bias towards the contributions performed by the *core group* of very active authors. Table 4.11 summarizes the value of Gini coefficients, as well as other inequality metrics for the top ten Wikipedias. The remaining statistics are provided for comparative purposes, and interested readers can find additional information about the respective coefficients in Chapter 2 of [11]. All coefficients have been calculated using the standard arguments provided for the respective functions found in the *ineq* library of *GNU R*. RS is the Ricci-Schutz coefficient (also known as Pietra's measure); Atkinson and Kolm are the respective measures of inequality in distributions, while Theil denotes Theil's entropy measure.

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

As we can see from the results in this table, the top ten language editions maintain a very skewed distribution, with less than 10% of the total number of authors performing more than 90% of the total number of contributions received by each version. Therefore, there is a heavy dependency on

the work of this core of very active authors to maintain the revision activity in Wikipedia. In case that the number of authors in the core diminished, or the number of contributions from core authors began to decrease, Wikipedia could not keep on the same producitivy level, since new editors recently incorporated did not have enough time to reach the same performance as former core authors. As we will see in short in the following sectionthe demographic analysis of the community of logged authors will raise some implications in this direction.

At the same time, we can follow a similar procedure to analyze the inequality of revisions received by Wikipedia articles. As we can see in the Lorenz curves and the Gini coefficients, the distribution of revisions among articles is much more equal. However, there is still some bias towards a group of very active articles, suggesting that the preferential attachment process has also some influence in the way Wikipedia articles attract the attention of logged authors. Further analysis should be conducted to clear up the concrete parameters producing this pattern, for instance, the appearance of articles on the *Did you know...?* column on the main page of each language version. As far as this thesis is concerned, our preliminar study of the quality of Wikipedia articles reveals that all FAs belong to the group of articles receiving a notoriously higher number of revisions than the average encyclopaedic entries.

Language	Gini	RS	Atkinson	Theil	Kolm
EN	0.6963	0.5332	0.4186	1.2284	28.8950
DE	0.6092	0.4533	0.3155	0.8377	19.8341
FR	0.6211	0.4643	0.3274	0.8778	13.8759
PL	0.6073	0.4493	0.3121	0.8168	8.6391
JA	0.6163	0.4614	0.3175	0.8097	11.8197
NL	0.6149	0.4560	0.3179	0.8043	10.2734
IT	0.6333	0.47170	0.3387	0.8789	11.3338
PT	0.6375	0.47997	0.3451	0.9519	6.8242
ES	0.6456	0.4862	0.3542	0.9636	13.9214
SV	0.5648	0.4170	0.2703	0.7057	6.3291

Table 4.12: Gini coefficient and alternative inequality metrics found in the distribution of total number of revisions per article in the top ten Wikipedias

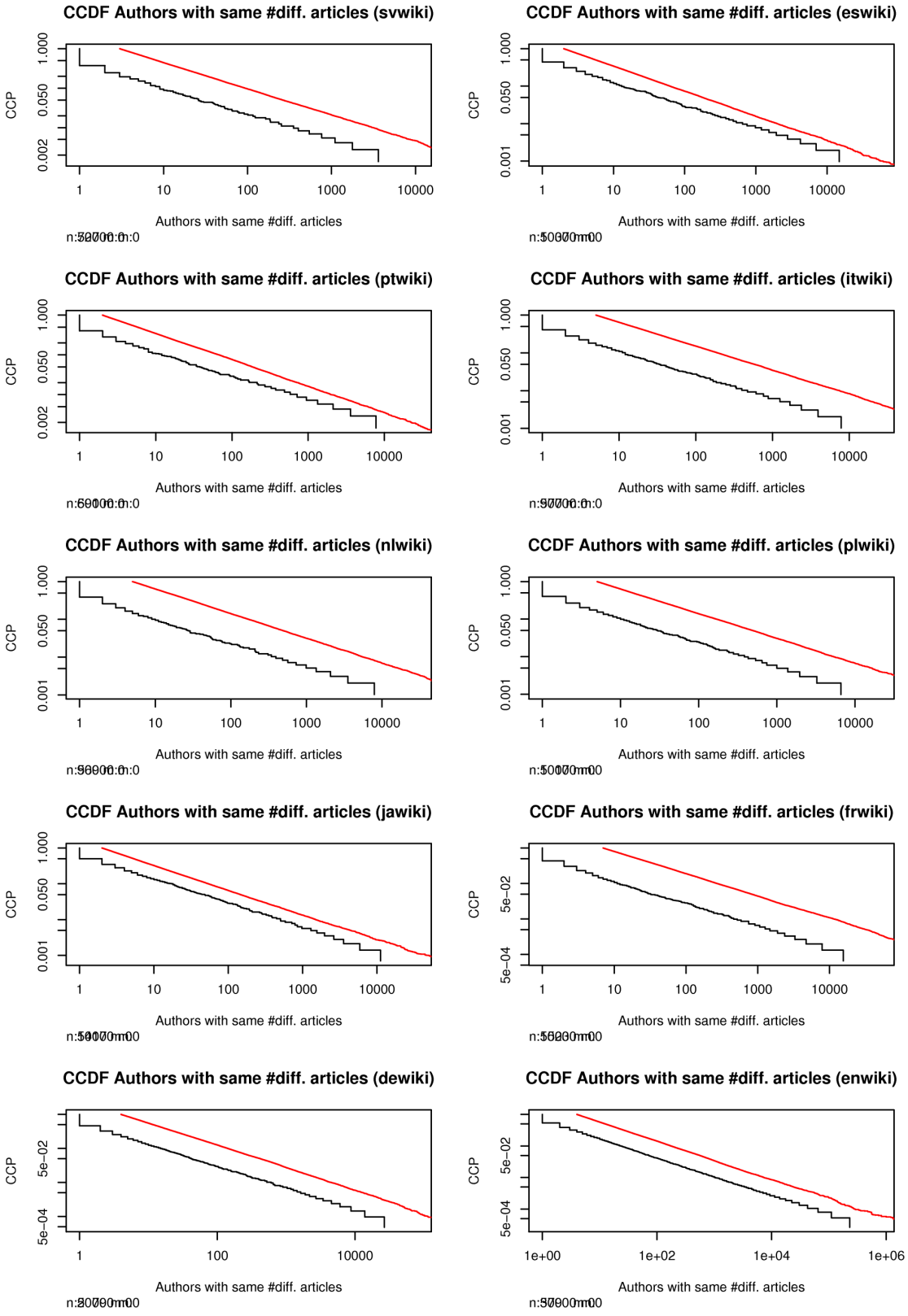

Figure 4.24: CCDF of number of authors sharing the same number of different articles revised in each language version. As we can see, in all versions the distribution perfectly follows a Pareto law. The best fitted Pareto line is also drawn in each graph for comparative purposes

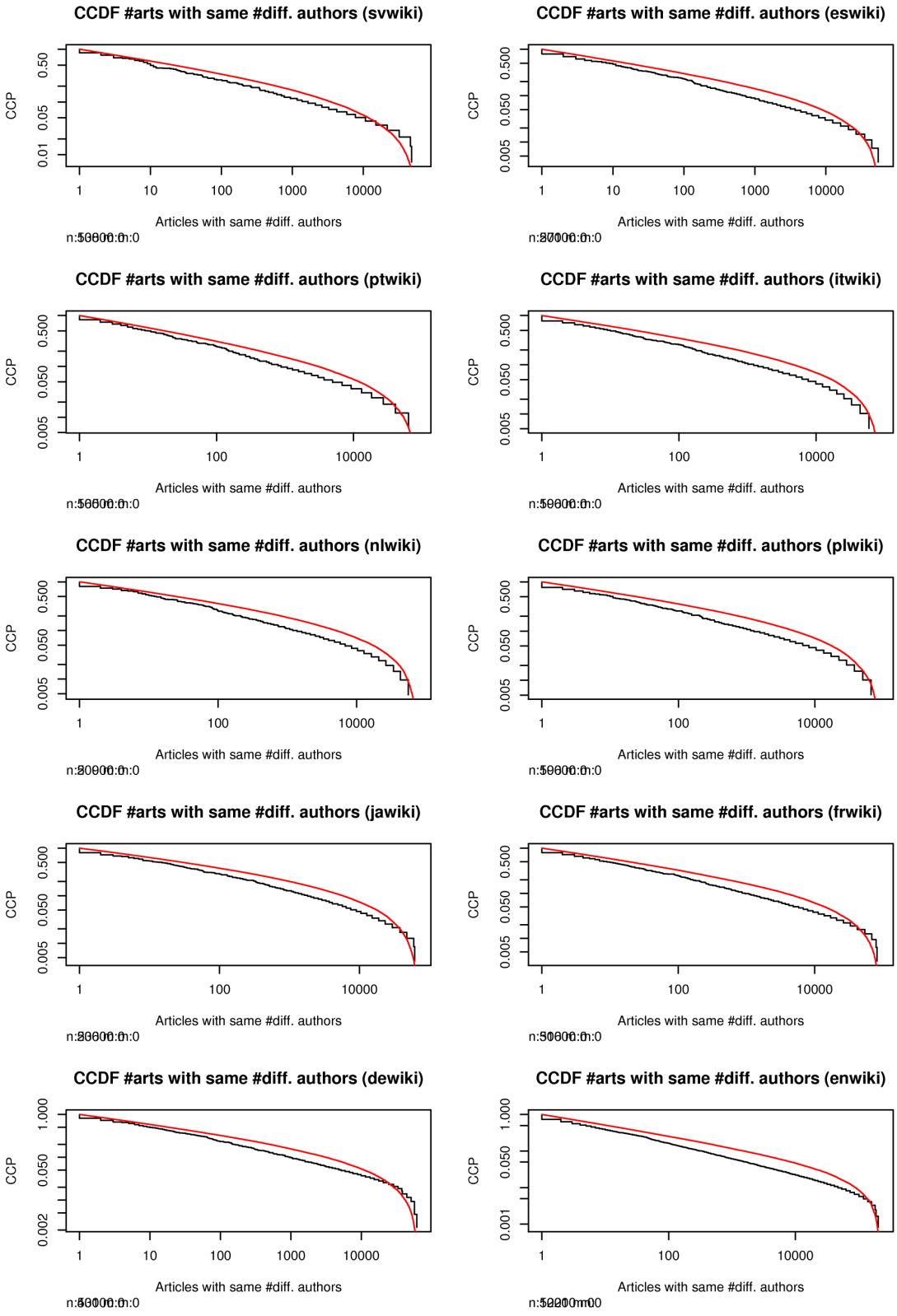

Figure 4.25: CCDF of number of articles sharing the same number of different logged authors. The distribution of this statistic follows an upper truncated Pareto in all language versions

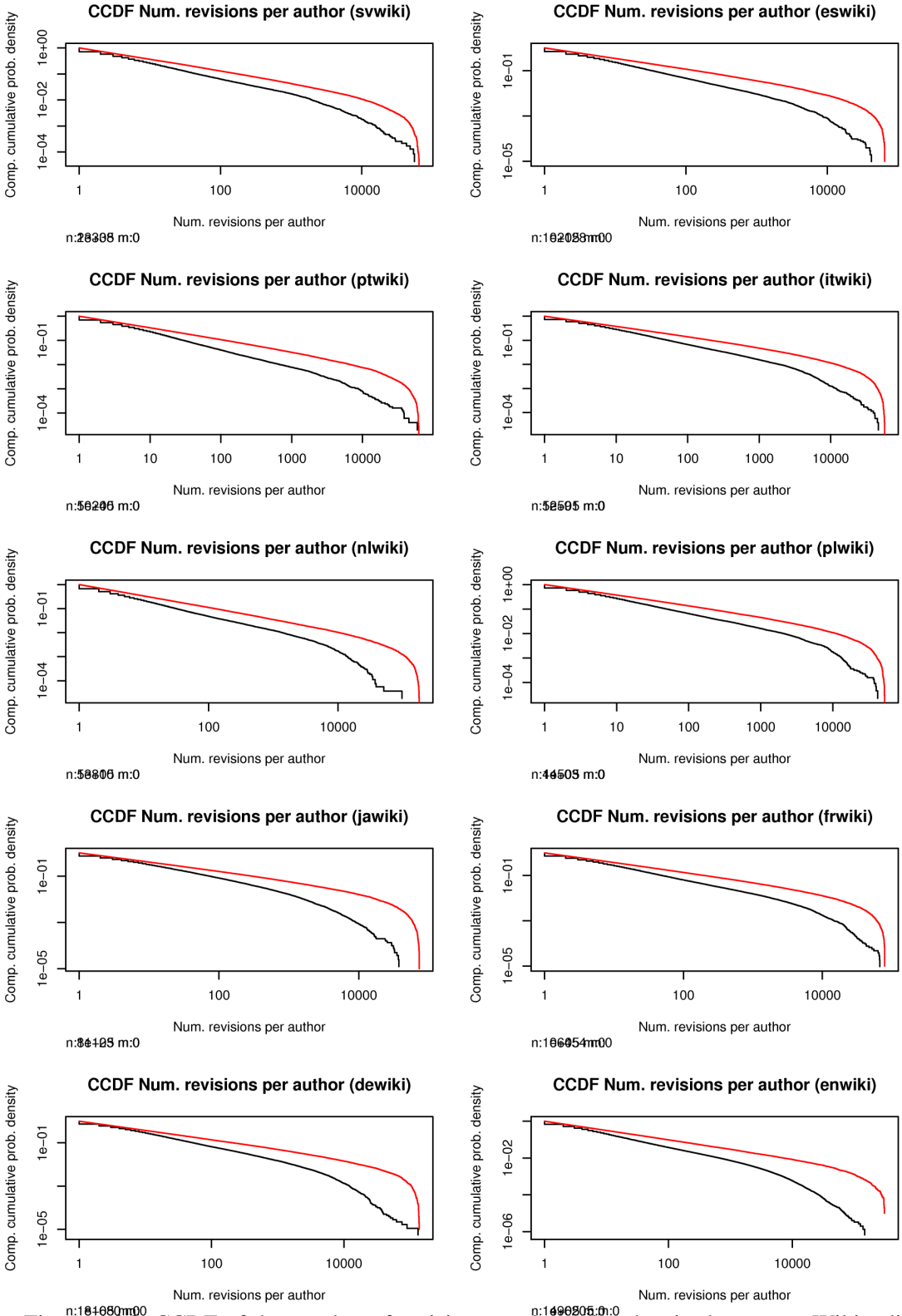

Figure 4.26: CCDF of the number of revisions per logged author in the top ten Wikipedias. As we can see, all languages follow an upper truncated Pareto distribution. The best fit is also displayed in each individual plot for comparative purposes

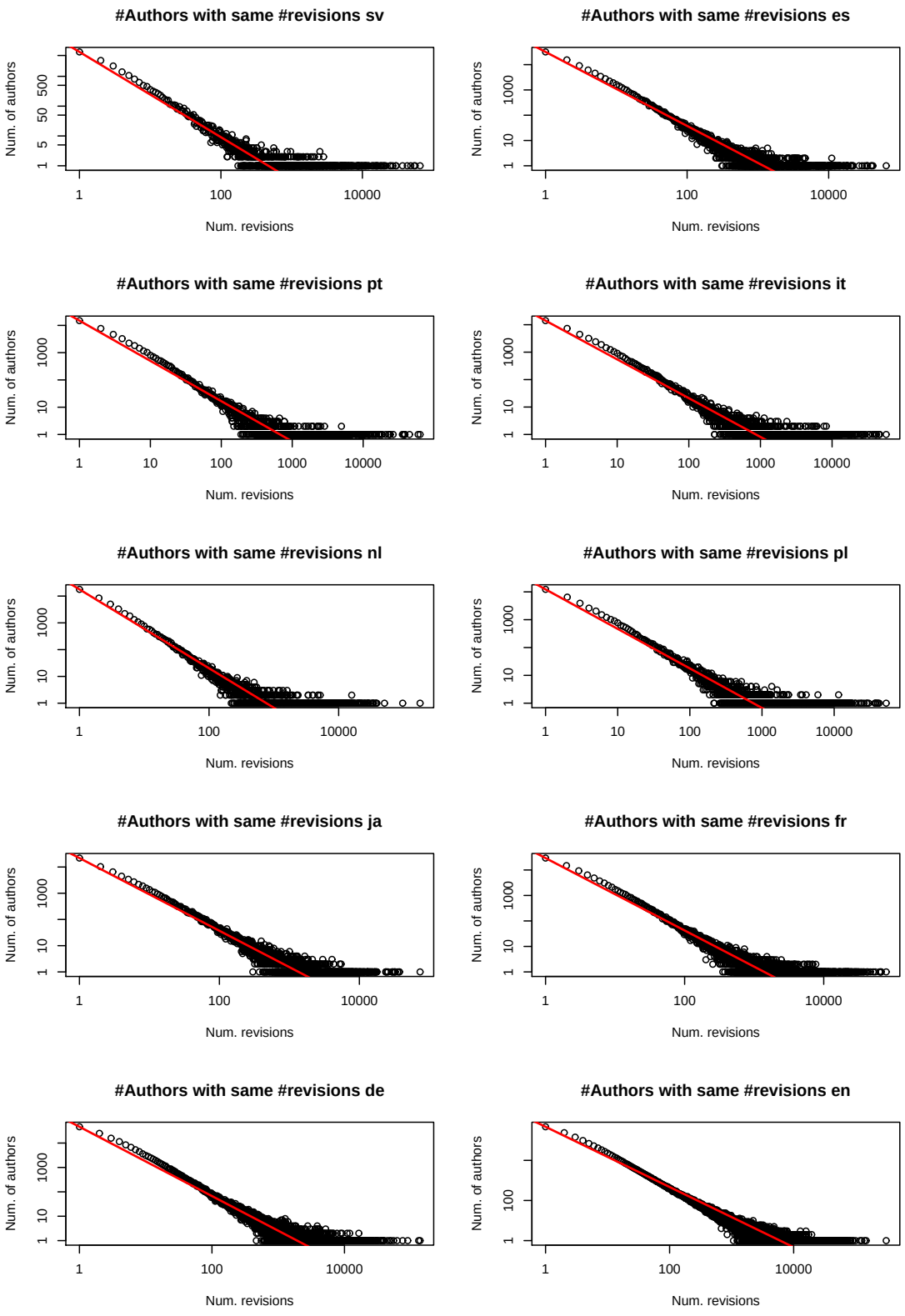

Figure 4.27: Scatterplot of number of revisions per author in the top ten Wikipedias. Since the distribution of this statistic follows and upper truncated Pareto distribution (as we have shown in the previous graph), we also provide the best fitted power law line in each individual plot for comparative purposes

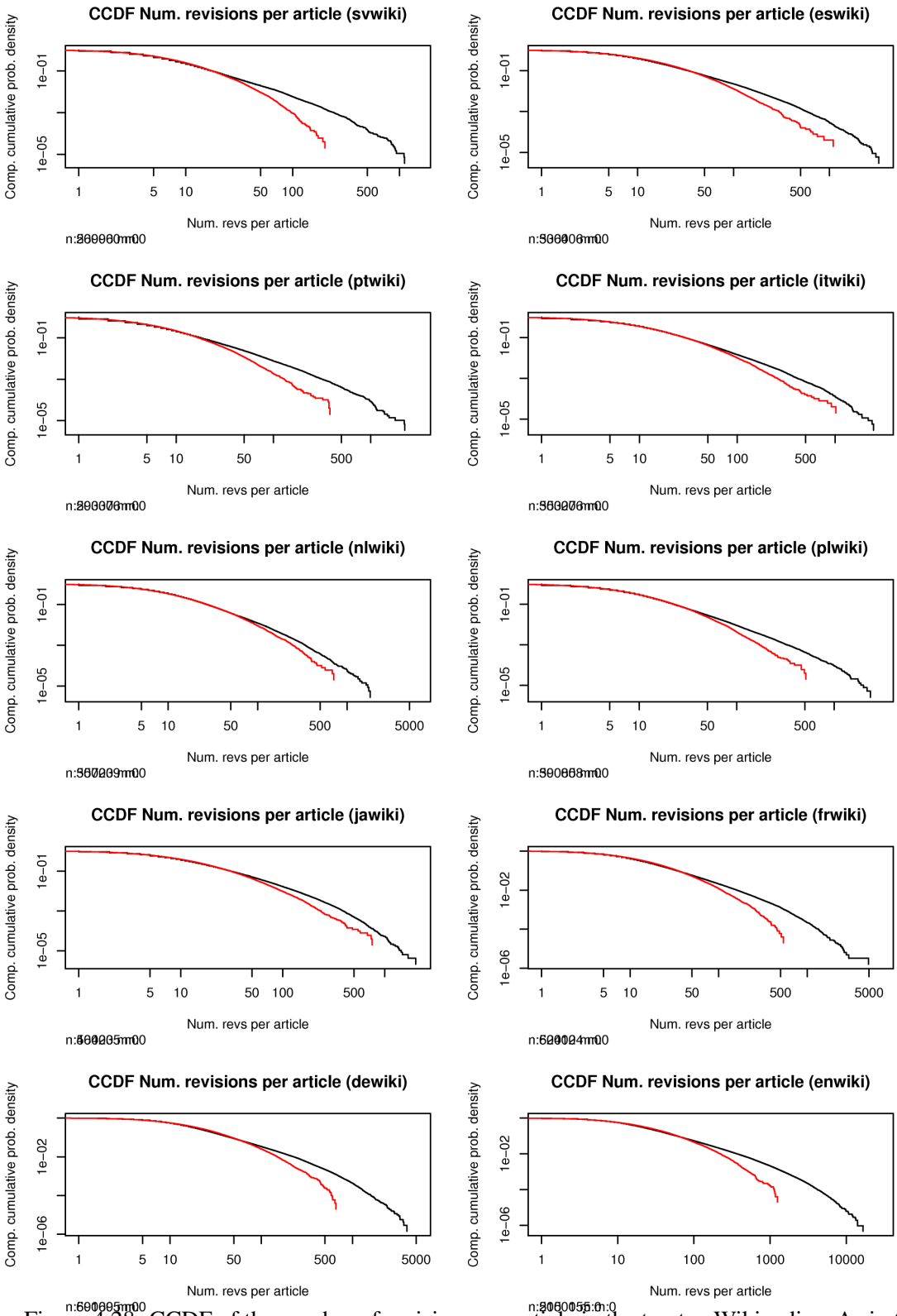

Figure 4.28: CCDF of the number of revisions per article in the top ten Wikipedias. As in the case of the different number of logged authors per article, the distribution of this statistic follows a lognormal pattern as well, which is also displayed in each individual graph for comparative purposes

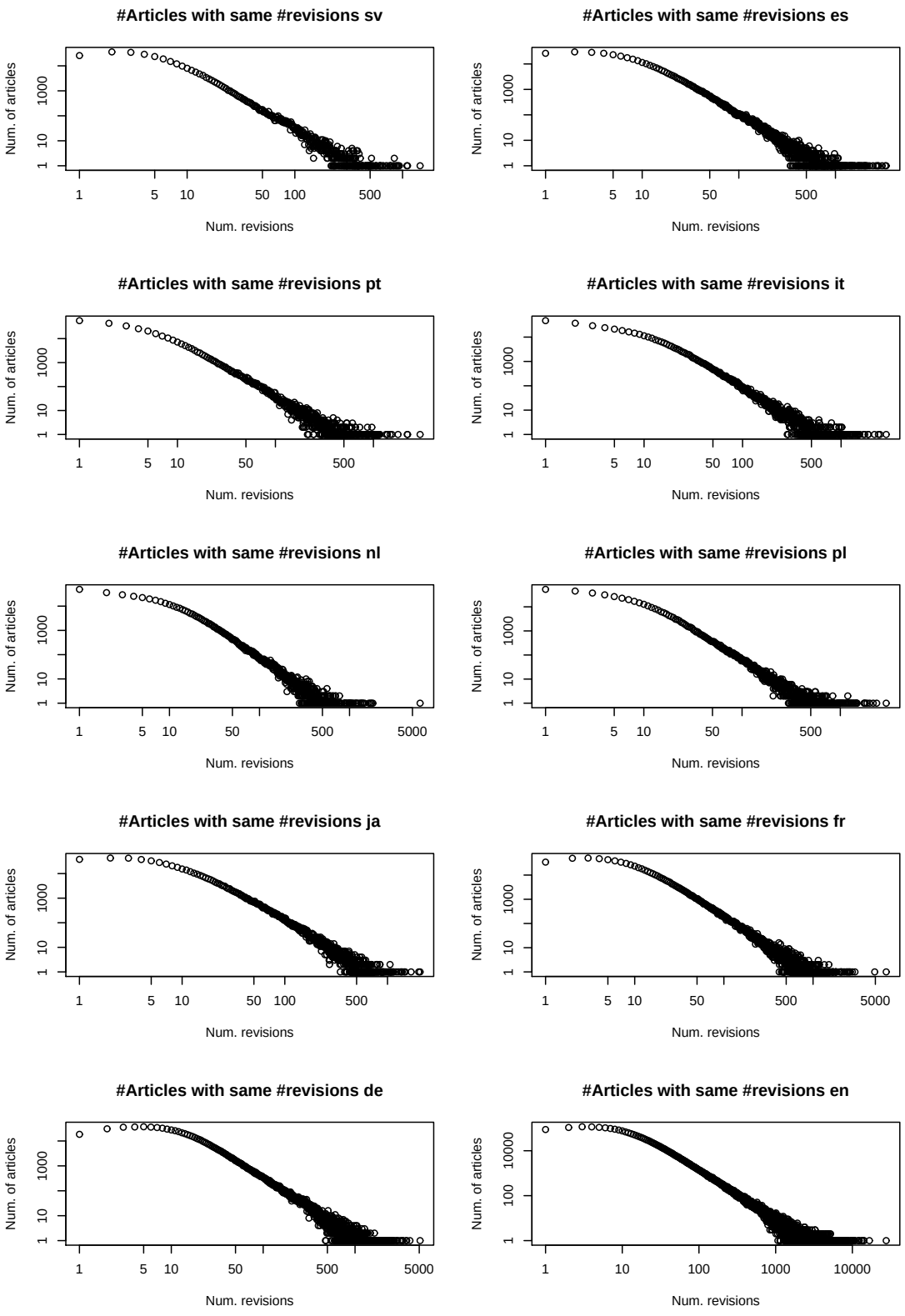

Figure 4.29: Scatterplot of the number of revisions from logged users received by each article in the top ten Wikipedias. Again, the statistic follows a lognormal distribution, so it does not make sense to plot a best fitted power law line, even for the cloud of points on the right side of each graph.

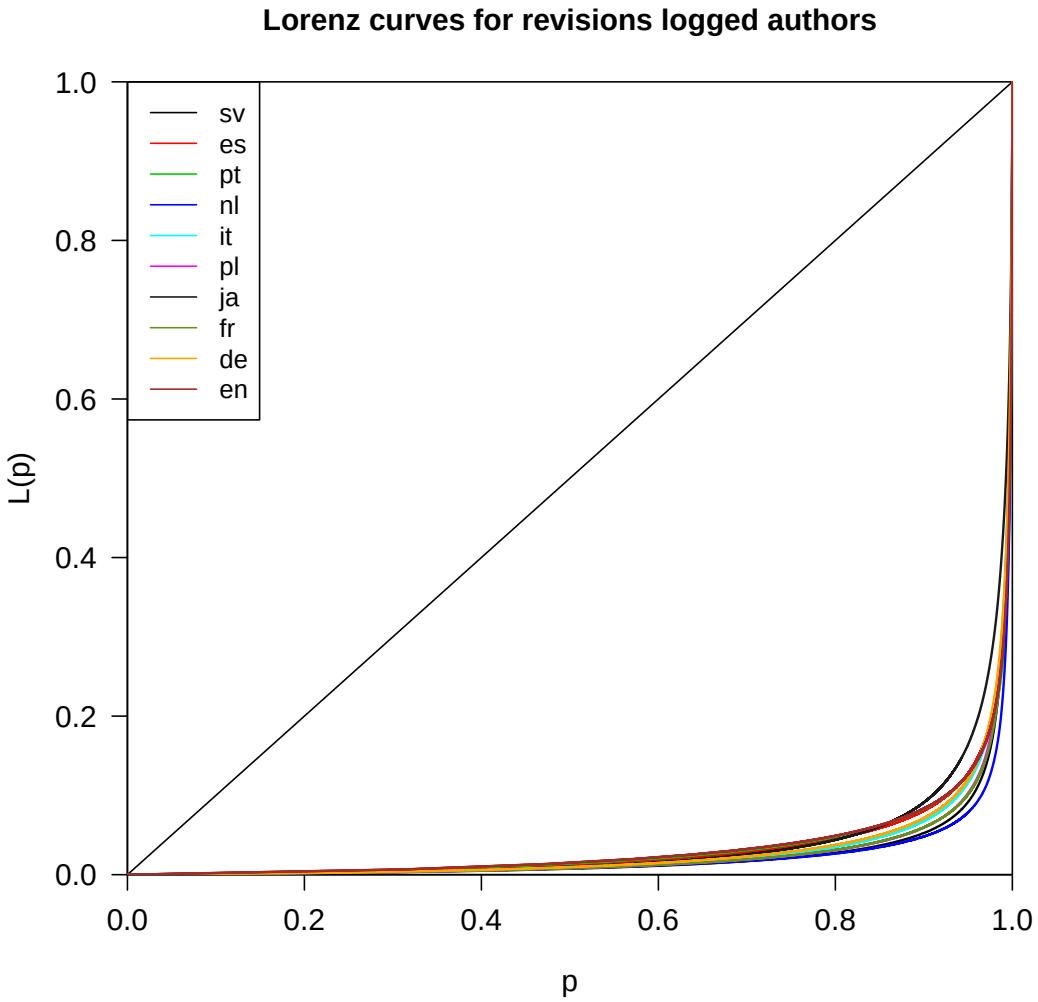

Figure 4.30: Lorenz curves showing the distribution of the total number of revisions performed by each logged author among the community in the top ten Wikipedias. The graph shows that there exist very little differences in the inequality level exhibited by all communities under study, showing highly biased distributions towards a small core of very active logged authors in each language version

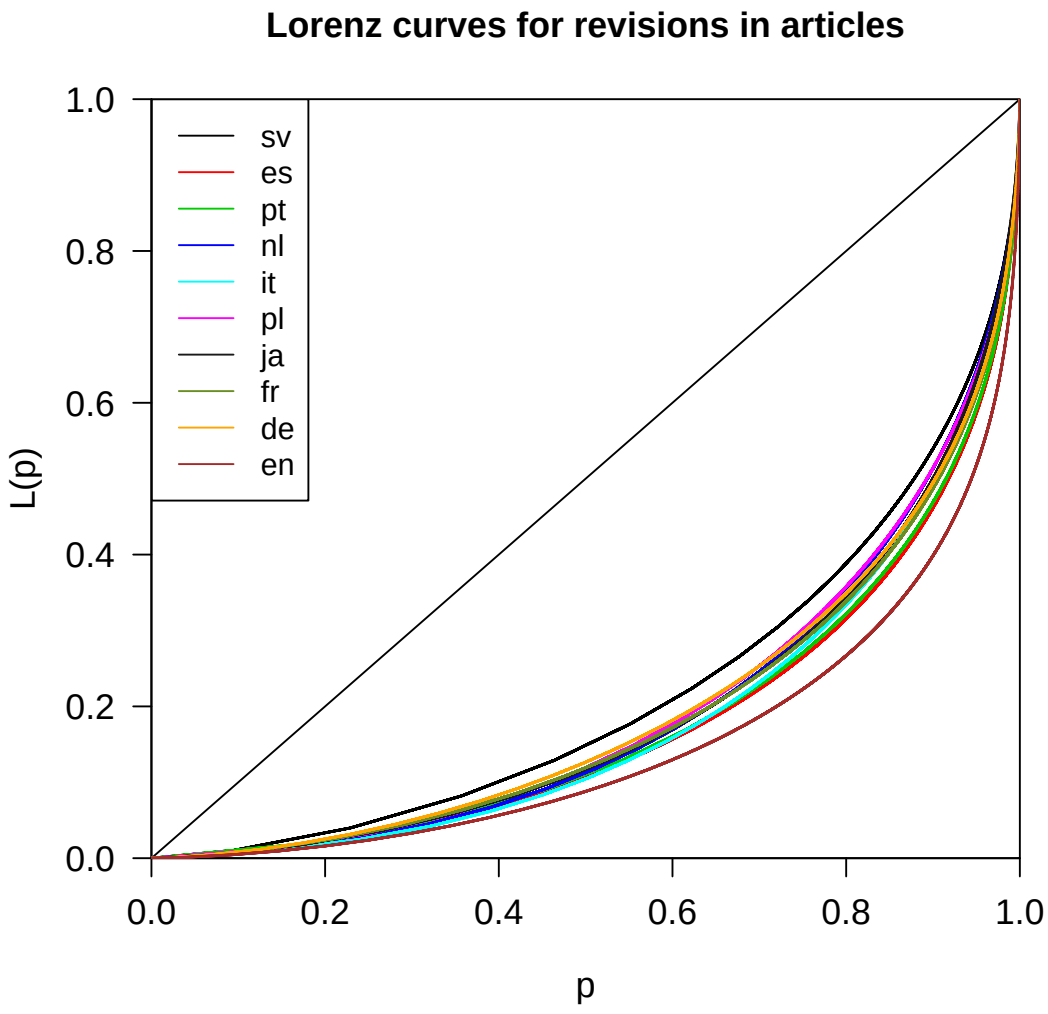

Figure 4.31: Lorenz curves showing the distribution of the total number of revisions from logged authors accumulated by each article in the top ten Wikipedias. The situation is different that the one previously presented for logged authors. The distribution of revisions among articles is more balanced, though there is a slight bias towards a group of more popular articles, represented by the right side area of the Lorenz curve for each language version

4.4 Demographic Analysis of the Wikipedia Community

So far, our empirical analysis of the top ten Wikipedias has revealed that the stabilization of the number of contributions from logged authors in Wikipedia during 2007 has influenced the evolution of the project, breaking down the steady growing rate of previous years. This effect deserves an more in-depth study, which is the purpose pursued in this section. A demographic analysis of the community of logged authors will reveal the origin of that leverage in the number of users, and may provide useful insights to explore possible evolution scenarios for the project in the following years.

Figure 4.32 presents the monthly numbers of births and deaths happened in the community of logged authors. Births are counted as new logged authors that did not contributed to the project before a certain month. Deaths are counted as logged authors that ceased to contribute in the corresponding month. Indeed, the plot of this statistics for all language versions clearly reveals the nature of the change in the trend of contributions from logged authors. The reason behind this phenomenon is the *increment in the monthly rate of deaths*, which has overcome the number of births per month, from 2007 on. In previous years, the rate of deaths closely followed the rate of births in all language versions, but the latter was always above in the graphs.

Unfortunately, this results raise several important concerns for the Wikipedia project. Though we do not have empirical data from 2008, the change in the trend of births and deaths will clearly decrease the number of available logged authors in all language versions, thus cutting out the capacity of the project to effectively undertake revisions and improve contents. Even more serious is the slightly decreasing trend that is starting to appear in the monthly number of births of most versions. The rate of deaths, on the contrary, does not seem to leave its ascending tendency. Evaluating the results for 2008 will be a key aspect to validate the hypothesis that this trend has changed indeed, and that the Wikipedia project needs to put in practice more aggressive measures to attract new users, if they do not want to see the monthly effort decrease in due course, as a result of the lack of human authors.

Following a more statistically formal approach, we can apply survival analysis techniques to explore interesting characterization parameters in the Wikipedia community of logged authors. Thanks to the flexibility of this methodology, as we already explained in section 3.5, we can not only measure deaths of users, as we have defined above, but also time elapsed to other relevant events. In particular, it would be interesting to analyze the time that took core authors to become members of this important group. We can also measure the time spent by these authors within the core group, and finally, we can also measure the time elapsed from their last revision as core members until their definitive death in the system. In this study, we consider that a logged author has reached the core group of a certain language edition when she was included in the top 10% of most active contributors in a certain month. The author is kept in this group until she definitely leaves it, not being included for the rest of her lifetime in the core group for any subsequent month. It is obvious that this might introduce some inconsistencies for authors that leave the core and join it again at some point in the future, but since we are interested in computing very active authors, that fine grain details does not affect the validity of our results. According to Priedhorsky et al.[86], the top 10% of authors in the English language version is responsible for 86% of the most read content in this language version. Therefore, using this criterion to select the core authors in each month for the top ten language editions seems to be a reasonable choice.

We start with Figure 4.33, presenting the survival curve $S(t)$ of logged authors in the top ten Wikipedias, measuring the survival time from their birth date to their death date in terms of revisions performed in articles. The first pattern that attracts our attention is the fairly high death rate of young authors, since less than 40% of them in all language editions achieve to reach an age of more than 500 days contributing to the system. Nonetheless, some interesting details attract our attention in

this graph. In the first place, we note the remarkable difference between the English and the German language versions. The first one presents one of the worst survival curves in this series, along with the Portuguese Wikipedia, whereas the German version shows the best results until approximately 800 days. From that point on, the Japanese language version is the best one. In fact, the German, French, Japanese and Polish Wikipedias exhibits some of the best survival curves in the set, and only the English version clearly deviates from this general trend. The most probable explanation for this difference, taking into account that we are considering only logged authors in this analysis, is that the English Wikipedia receives too contributions from too many casual users, who never come back again after performing just a few revisions. In that case, the curve would be reflecting the shorter lifetime of that significant group of ocasional authors. Given the current results, that suggest a strong, positive correlation between the size of the community and the lifetime of authors, this is one of the few possible explanations to this pattern.

Table 4.13 summarizes the results directly obtained using the *survival* package in GNU R. The first column indicates the number of events analyzed, that is, the size of the population under study until the end of 2007. The reason why these numbers are much lower than the total number of logged authors presented, in the first section of this chapter, is that we have to filter out all authors who only contributed once to the corresponding language version. The pre-conditions impose us an open range for lifetime values, $T \in (0, \infty)$. The second column presents the number of events ocurred in the whole period of time under study. The meaning of "event" varies depending on the analyzed incident. For instance, in this case counted events refer to the total number of registered deaths of logged authors during the whole time period under scrutiny. On the right side of prior columns, we have the *restricted mean* of the lifetime of logged authors, and its associated standard error. In case the last observation in the set of individuals at risk is not a death, then we could not calculate the mean survival time. To avoid this problem, the mean is calculated restricted to the last censoring registered in the population. According to the help files in the survival package, "any randomness in the last censoring time is not taken into account in computing the standard error of the restricted mean". The *median* column shows the median value of the authors lifetime. This metric is complemented by the lower and upper limits of the 95% confident interval of the median value. This results can be obtained graphically, plotting an horizontal line at the 0.5 value of the survival curve, and looking for the intersection points with the survival curve and its confidence intervals. In our case, the lower and upper curves of the confidence interval for each language version are ommitted, since this is the default behavior in the *survival* package when we plot several survival curves at the same time. In any case, the values of the confidence intervals are so close to the survival curve that they could not be differentiated with naked eyes.

In the first place, we would like to highlight the great accuracy of our results. Looking at the s.e. values for the restricted mean lifetime, as well as at the 95% confident intervals for the median values, our estimation is very precise, with errors typically showed in the order of days. In the second place, we have another proof of the skewed distribution of descriptive parameters in the community of logged authors. The restricted mean and median lifetime values are quite separated, showing the significant difference between the lifetime of many older authors in contrast with the short period of contribution of the 50% of the total population of logged authors. We discussed earlier about the effect shown by the restricted mean, with the English language edition almost showing the worst results (only exceeded by the Portuguese language version, but for a very narrow margin). However, the results for the median values are even more interesting. Looking at this column, we find that the Spanish Wikipedia presents the same estimator found in the English version (we look at the values delimitting the 95% c.i. to compare both estimations appropriately). This is astonishing, given that the community of authors in the English Wikipedia is 13.8 times larger than the Spanish one. The

difference in the restricted mean values are even greater. The values for the Swedish and Italian editions defy any intuitive guess we may have done *a priori*. From these results, we can conclude that, in some language versions, the size of the community is not the only parameter influencing the lifetime of logged authors, and other factors (like the stronger commitment of authors in certain communities to contribute to the project) should be taken into account.

```
> print(wkpfit, show.rmean=T)
Call: survfit(formula = wkp_surv ~ lang)

                n events rmean se(rmean) median 0.95LCL 0.95UCL
lang=en    715267 615397   274     0.577    100      99     101
lang=de    111466  90008   403     1.688    204     201     208
lang=fr     54089  39599   404     2.778    188     182     193
lang=ja     40705  28831   390     3.457    147     142     152
lang=pl     22850  16948   385     4.101    167     159     174
lang=nl     23290  18052   355     3.866    135     129     142
lang=it     28278  21157   336     3.635    140     134     147
lang=pt     22382  17800   267     3.723     82      77      87
lang=es     45485  34301   310     2.774    102      98     105
lang=sv     11295   8833   345     5.130    151     143     159
```

Table 4.13: Output of *survfit* call in GNU R on Wikipedia data for all logged authors contributing to the top ten language versions

In contrast with these results, Figure 4.34 demonstrates that very active users proceed rapidly to reach the core, since more than 60% of them achieve their membership after less than 120 days. This is good news for the Wikipedia community, since it confirms that a significant proportion of top contributors need little more than half a year to jump into the core which sustains the main percentage of the content creation effort. Indeed, an agile process to incorporate new members to the core of very active authors is a key factor for the sustainability of the community of contributors, since we need a new wave of young top contributors to replace former core members.

A more careful examination of the numerical results presented in Table 4.14, reveal that Japanese and Spanish Wikipedians are the fastest ones to reach the core of very active authors. 50% of new core members joined the group of top contributors 84 and 97 days before entering the system, respectively. Again, the accuracy of our estimators is high, though they are not as precise as in the previous case. The total number of censoring events registered, along with the size of the population under study play a decisive role with respect to this isse.

Reaching the core group of most active authors does not warrant superior longevity to Wikipedia authors, though. On the very contrary, only 30% of them (approximately) remain alive as core members after the 500 days threshold. It seems that the so-called "burning effect" on very active contributors (they eventually become exhausted and end up leaving the core) has a significant influence in the demography of the core in all language versions, as shown in Figure 4.35. The

```
> print(wkpfit, show.rmean=T)
Call: survfit(formula = wkp_surv ~ lang)

                n events rmean se(rmean) median 0.95LCL 0.95UCL
lang=en     49601  48067   208      1.11    119     117     121
lang=de      6944   6716   217      3.25    112     105     119
lang=fr      3455   3156   229      4.95    118     108     128
lang=ja      3246   2962   196      5.11     84      75      90
lang=pl      1559   1453   224      7.08    117     107     137
lang=nl      1430   1349   228      7.70    120     107     141
lang=it      1691   1553   195      5.76    114     101     126
lang=pt      1354   1236   188      6.19    104      91     119
lang=es      2959   2660   191      4.64     97      88     108
lang=sv       741    697   206      9.49    110      88     135
```

Table 4.14: Output of *survfit* call in GNU R on Wikipedia data for all logged authors who eventually joined the core in the top ten language versions

Spanish, Portuguese and Italian Wikipedias present the lowest survival curves all along the considered range, while the French, Polish and Sweden language versions hold the better survival curves for authors in the core. The 3 largest versions also hold the record of top long-lived authors (something natural in the English edition, since it has runned for a longer period of time). Crossing points in the Kaplan-Meier curve of the English Wikipedia (marking deaths of users in that point of time) shows that some former core members lasted in the core for more than 2,000 days (or 5.4 years). The median of the lifetime values, shown in Table 4.15 reflect the high mortality rate of core authors. A surprising result is that 50% of core authors in the English Wikipedia only maintain their top active position for less than a month, in contrast with German, Polish and Swedish authors, 50% of whom maintaining their core status for almost 3 months. The remarkable difference between the values of the restricted mean and the median in this case are noticeable. This clearly shows that there exist a subpopulation of long-lived core authors maintaining their high contribution rate to the project for more than a year, in many cases.

However, Figure 4.36 shows that English core members are the fastest abandoning the project after leaving the core. Again, the Portuguese and Spanish language versions also come up in this lower survivability area. On the opposite side, former core members of the German still remain very active once the left the very active group of logged authors, with the French, Polish and Italian Wikipedias presenting good survival curves as well (given the general trend identified in other versions). Numerical summaries in Table 4.16 show that, contrary to what we may think intuitively, former core authors does not quickly abadon the project. On the other side, in many language versions 50% of them carry on revising articles for more than a year since they left the core. The restricted mean values in this case present shorter departure from the median values, indicating a much less skewed distribution, and thus demonstrating that the behavioral pattern of former core authors is more

```
> print(wkpfit, show.rmean=T)
Call: survfit(formula = wkp_surv ~ lang)

                n events rmean se(rmean) median 0.95LCL 0.95UCL
lang=en     86219  75709   218      1.74     28      28      29
lang=de     10553   8677   341      5.45     81      73      87
lang=fr      5323   3960   384     10.17     75      60      86
lang=ja      4884   3662   336      9.35     77      60      85
lang=pl      2217   1645   379     14.92     86      61     104
lang=nl      2190   1682   369     14.31     75      57      88
lang=it      2567   1920   317     10.35     78      59      88
lang=pt      2597   2149   210      8.42     26      25      27
lang=es      5101   4007   266      8.69     37      30      43
lang=sv      1112    855   364     18.93     86      60     118
```

Table 4.15: Output of *survfit* call in GNU R on Wikipedia data for all logged authors within the core of any of the top ten language versions

uniform than in previous analyses. Again, we must stress the difference between the median values for the English and German Wikpiedias, more than two times greater in the second case. This make even clearer the different behaviour of German authors, that still demonstrates strong commitment to the project, even once they left the core. We should also take into account that some of this former core members may have been surpassed by new, more active contributors. Otherwise, they may have remained in the core of top contributors for a loger period. As we will see in a 3D graph showing the evolution of each group of monthly core members over the rest of months, there is a clear trend towards much higher number of revisions in subsequent generations core authors. This also reveals a tendency of incrementing the inequality level of contributions on the core of very active authors with respect to average users.

A different evaluation of the same situation can be obtained plotting the hazard function $h(t)$, which gives the instantaneous risk of death at any point in time given de survivability data collected in each language edition. We have computed the values of this function for the whole lifetime of logged authors, as well as for the members of the core group of very active authors in each language. We use a logarithmic scale in the horizontal axis to facilitate the visualization of hazard function patterns. Figures 4.37 and 4.38 shows the results for the whole lifetime of authors, and core authors, respectively. The results confirm a disproportionately high risk of death in the first days of Wikipedia authors life in the system, following a log-linear descending trend until the first 12 years of age. The hazard then becomes much lower and remains somewhat constant over the rest of the range. It is remarkable that the hazard rate for the Swedish language version deviates from this general trend, since it maintains a slightly increasing trend to change again after the first 12 days of authors contributing to the system.

Regarding the results for core authors, the hazard rate is, in general, much lower, showing that

```
> print(wkpfit, show.rmean=T)
Call: survfit(formula = wkp_surv ~ lang)

                n events rmean se(rmean) median 0.95LCL 0.95UCL
lang=en     58041  35701   315      2.32    217     214     220
lang=de      7534   3834   556      7.62    453     435     474
lang=fr      3414   1528   557     15.58    433     404     457
lang=ja      3045   1380   516     17.68    350     330     371
lang=pl      1424    641   519     20.22    394     364     450
lang=nl      1407    635   459     22.46    344     320     379
lang=it      1668    677   525     21.68    375     343     416
lang=pt      1659    892   357     12.79    271     249     297
lang=es      3172   1542   411     13.74    290     270     307
lang=sv       744    350   511     29.49    362     316     435
```

Table 4.16: Output of *survfit* call in GNU R on Wikipedia data for all former logged authors in the core who eventually left any of the top ten language versions

core authors usually maintain their implication in the project for longer time periods, as we might expect intuitively. The Spanish, English and Portuguese Wikipedias present the higher hazard rates before the first 50 days of stay in the core. Another valuable lesson learned from these graph is that, once the core author has surpassed the 100 days threshold, the probability that she remains in the core for a longer period of time raises dramatically, in all language versions.

In the same way, we can use the Cox proportional hazard model to study the influence of important parameters. In this case, we probe the influence of revisions performed on FAs and talk pages on the longevity of logged authors. The hypothesis is that authors demonstrating higher implication level in coordination activities and discussions about contents, and authors providing quality content on FAs, will have better survivability than those who did not contribute to this special type of pages. Figure 4.39 confirms this hypothesis, but only partially. As we can see, the individual effect of contributing to FAs or talk pages do have some influence in enhancing the longevity of logged authors in the system. But the definitive improvement in authors lifetime is only registered for authors who both edited in FAs and participated in talk pages, at the same time. Therefore, we conclude that participation of logged authors in coordination activities and FAs implies that those authors have better chances to maintain their relationship with the project for a longer time period. The fact that the combination of both parameters represents the main raise in the survival curves for Wikipedia logged authors indicates that isolated participation on coordination tasks within a certain language edition does not warrant an enhancement in the survivability of Wikipedia authors by itself, though we can appreciate that there exists a leverage in the lifetime of such users in the project.

Interestingly, in the case of the English Wikipedia we can observe that there is practically no difference at all between the effect of contributing to FAs and the survivability of users participating in talk pages. Predicted survivability curves are superimposed almost exactly, thus revealing that for this

language version, each of these control statistics provides the same effect. Furthemore, the combined effect of both parameters seems to produce an additive effect situation, enhancing the survivability curve of authors as if it were produced by the sum of the leverage effect provided by each individual curve by itself.

Finally, Figures 4.40, 4.41 and 4.42 summarize the restricted mean and median values of the lifetime of users in the top ten Wikipedias, and the summary values for the lifetime of core authors in all language versions. In the case of the global community of logged authors, the mean and median values are quite disparate, showing the strong bias of the distribution of lifetimes due to the high mortality rates in young authors. The modal value of the median lifetime is around 150 days whereas the distribution for the restricted mean lifetime of authors is centered around 375 days. Thus, as a result of the extreme bias found in the lifetime values of Wikipedia authors in all language versions under analysis, the distribution of the restricted mean and median values of authors' lifetime does not almost present any overlap at all.

An even more interesting pattern shows up when we see the graph of the restricted mean lifetime values for authors until they reach the core, and then for authors maintaining their membership to this group in all language versions. Amazingly, all language versions seem to present values quite concentrated around 200 days, thus revealing an interesting common pattern regarding the number of days that a certain author need to reach the core of top contributors in the top ten Wikipedias. On the other side, the curve showing the distribution of the restricted mean lifetime of authors in core reveals a broader distribution, encompassing the range between 200 and 400 days of membership. Again, the median values of these two statistics show that 50% of users leave the core much earlier, typically less than 100 days after obtaining membership, while the median value for the time to reach the core is strongly concentrated around a modal value of between 100 and 120 days.

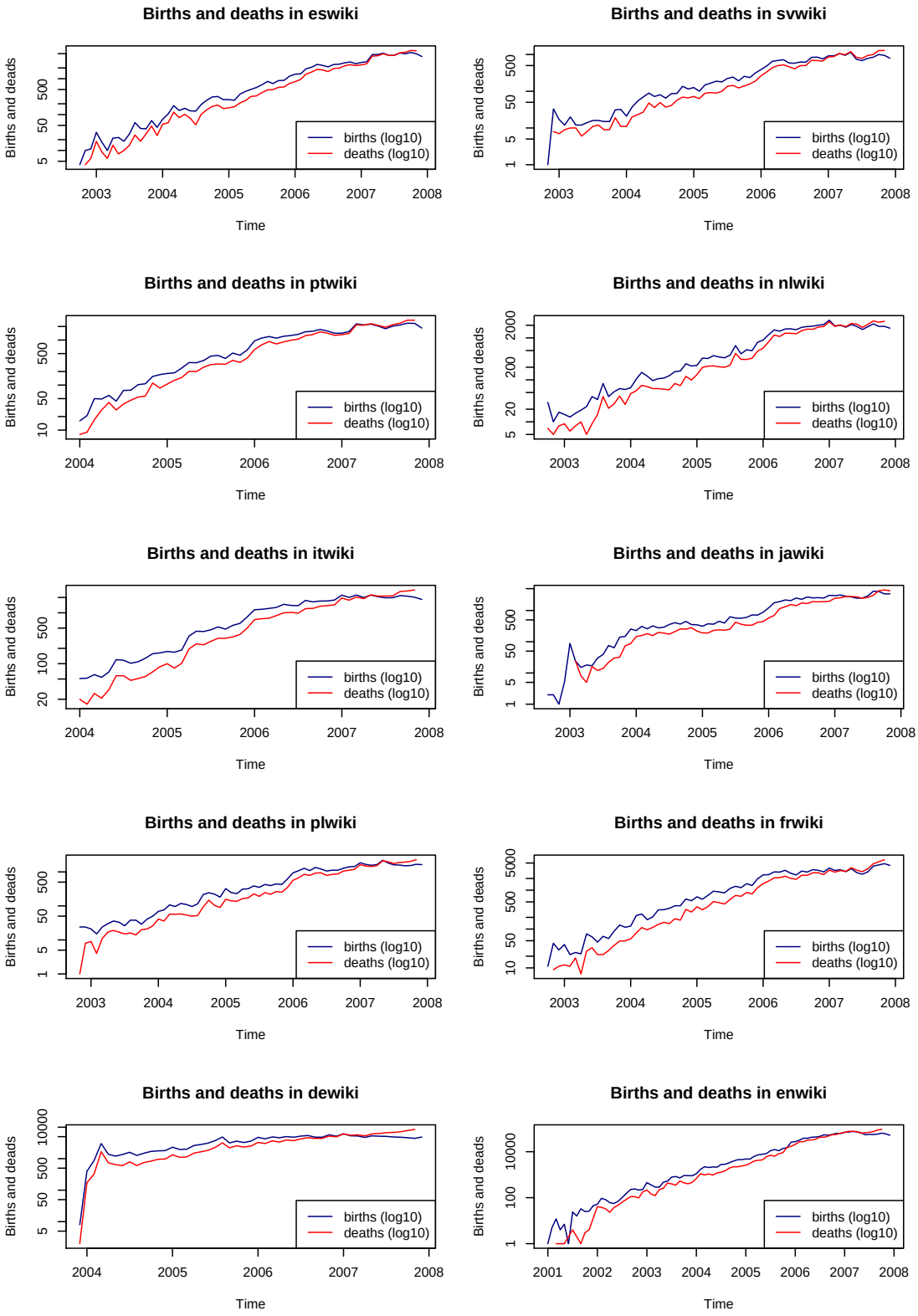

Figure 4.32: Monthly number of births an deaths of logged authors in the top ten Wikipedias. Both axes have a logarithmic scale. The graph shows that the number of deaths per month closely follows the number of births, suggesting a high mortality rate that prevents the population from growing at an exponential rate. We can also appreciate that in summer-Fall 2006, there was a dramatical 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

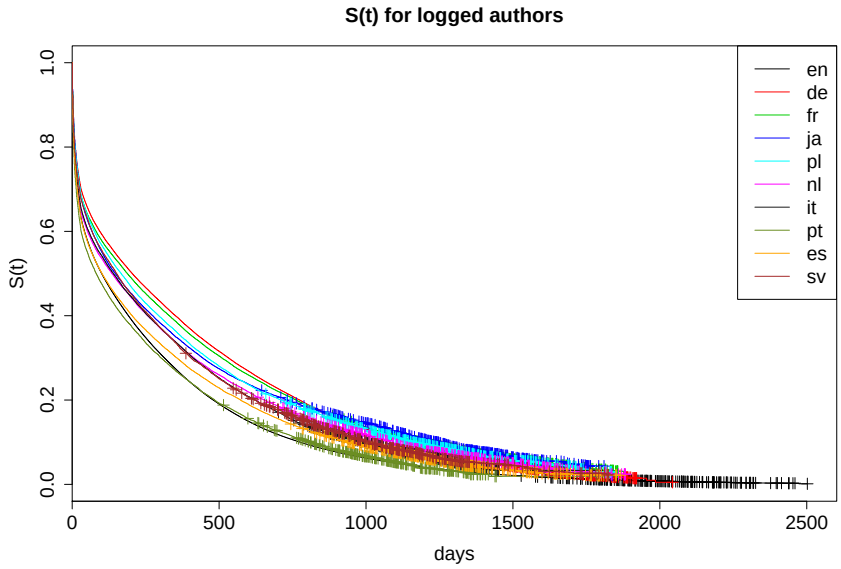

Figure 4.33: Survival functions of logged authors contributing in the top ten Wikipedias. The graph shows that the mortality level among young contributors (less than one year of participation in the project) is substantially high. It is also remarkable that less than 40% of authors in all language versions continue to participate in the project once they reached an age of more than 500 days

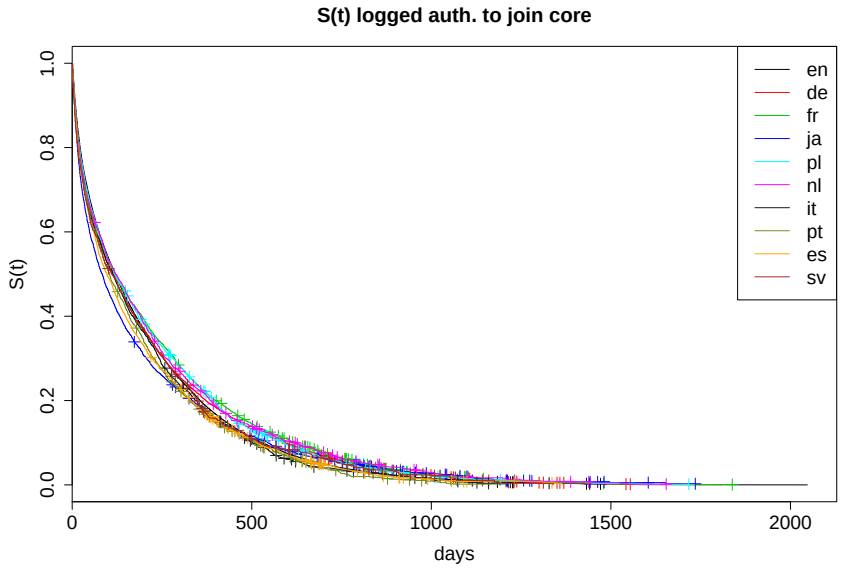

Figure 4.34: Survival functions measuring the final event of logged authors eventually joining the core of very authors (top 10% of most active authors in any month). The graph shows that those authors who finally reached the core did it fastly. More than 60% of them joined the core after less than 200 days of participation in the project

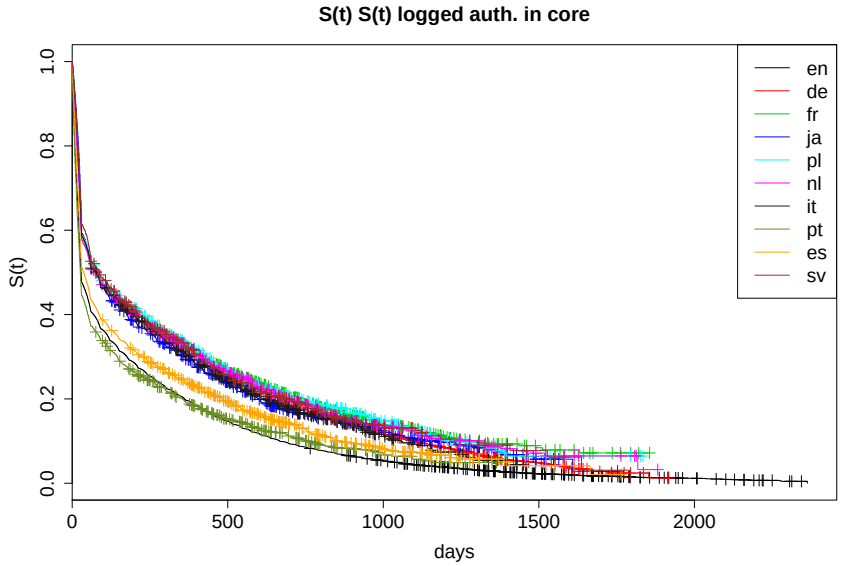

Figure 4.35: Survival functions measuring the survivability of logged authors in the core group of the top ten Wikipedias. The mortality of core users in the Portuguese, Spanish and English Wikipedias is higher than in the rest of versions under study. All the same, the mortality rate is very high, since only 30% of authors (approx.) remain alive after 500 days of core group membership.

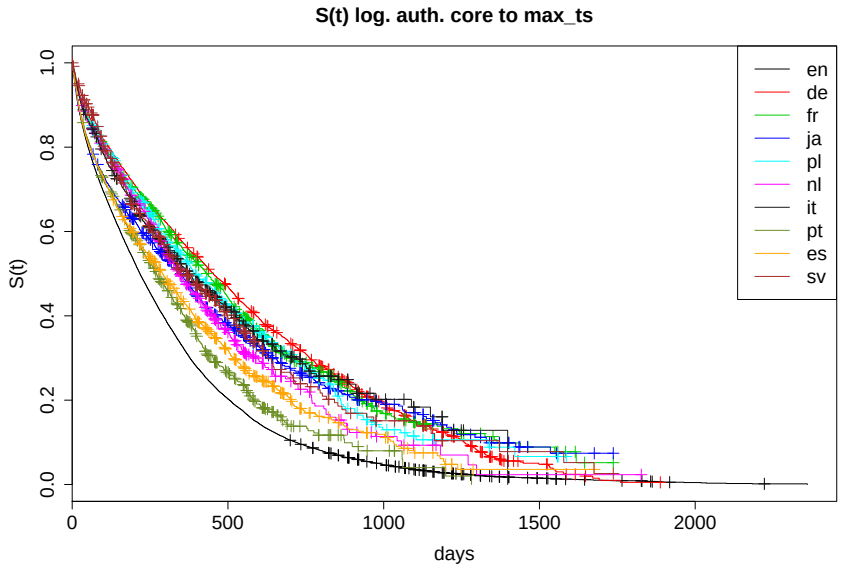

Figure 4.36: Survival functions measuring the time elapsed from departure of very active logged authors form the core group to their definitive death in the project. Contrary to the previous graphics presented before, we can see that a significant percentage of authors leaving the core still maintain their activity for a substantially longer time interval (more than 500 days for more than 40% of these authors) with the exception of the English Wikipedia, which shows higher mortality rates in this statistic, as well.

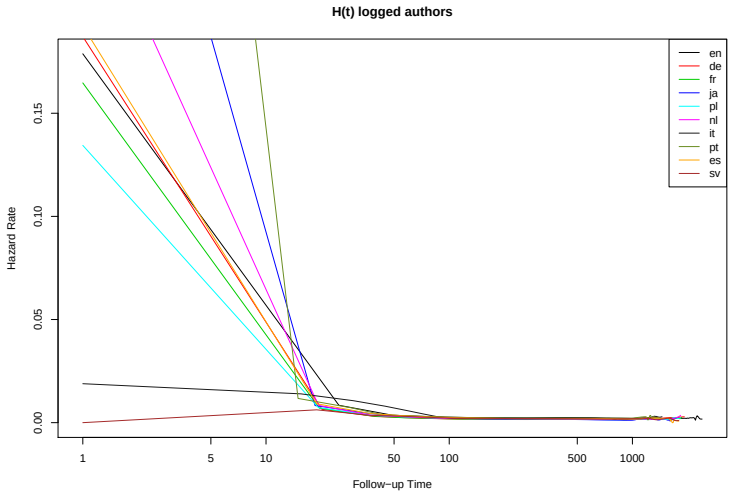

Figure 4.37: Hazard function of logged authors contributing to the top ten Wikipedias. The horizontal axis has a logarithmic scale, since otherwise the disproportionally high risk of death for young logged authors would have prevented us to appreciate the values for older authors. It is interesting to see that the risk for young authors decrease following a log-linear pattern, for an age of less than 15 days, except for the English and German Wikipedias. The remarkably lower risk for young authors in these language editions might be a logical cause behind the more active production pattern in these language versions, leading the top ten list.

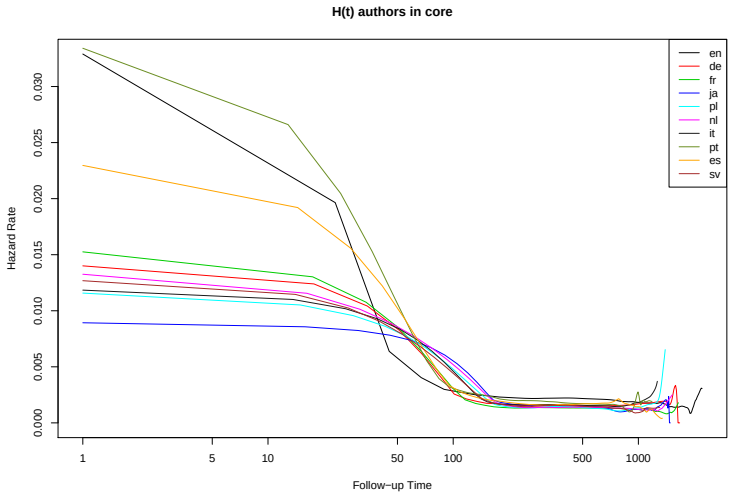

Figure 4.38: Hazard function of logged authors that reach the core group at any moment in the history of the top ten Wikipedias. Again, the risk of death for authors with less than 100 days in the core of top contributors is higher, and decrease following a log-linear pattern, but only for the English Spanish and Portuguese versions. The rest of versions under analysis show a higher, but somewhat constant, hazard rate for younger core authors, supporting the hypothesis that, in these language versions, core authors tend to maintain a longer relationship with the project

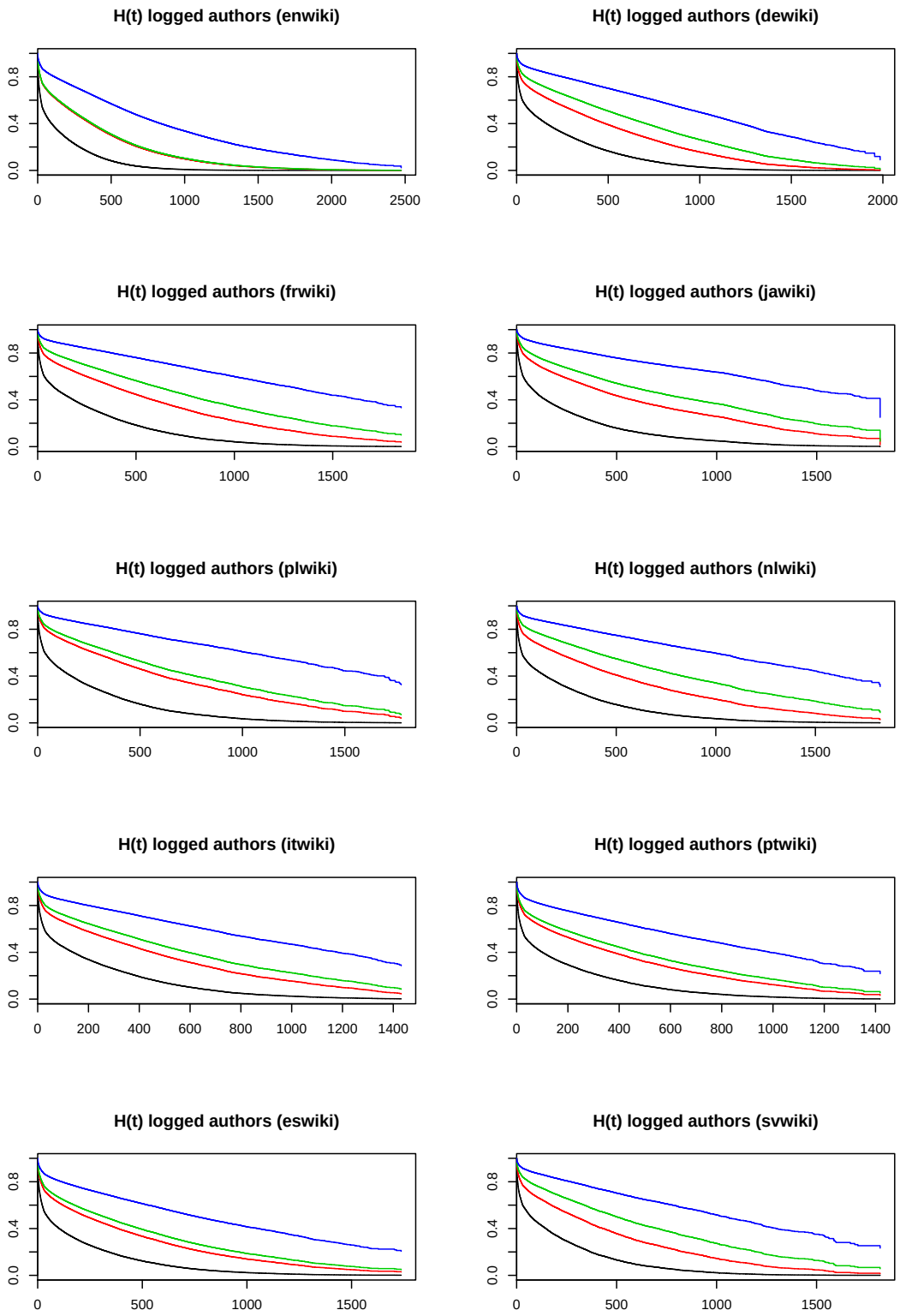

Figure 4.39: Survival functions plotted using the Cox proportional hazard model, for logged authors in the top ten Wikipedias. Control variables where: 1) authors who did not revised either FAs or talk pages (black); 2) authors who revised talk pages but did not contributed to FAs (red); 3) authors who performed revisions in FAs but not in talk pages (green); 4) authors who revised both FAs and talk pages (navy blue). The graph shows that the survivability of logged authors who revised FAs and participated in talk pages is substantially higher than that of authors who did not participated in FAs or discussion processes

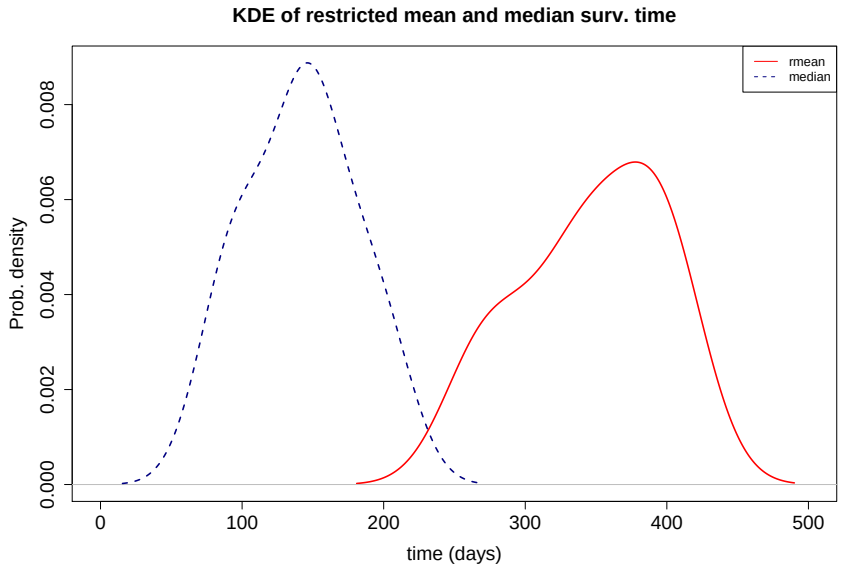

Figure 4.40: KDE of the restricted mean and median survival time of logged authors in the top ten Wikipedias. Given the high mortality rate of younger logged authors, it is not surprising that the median survival time is lower, for all language editions, that the restricted mean, which despite trying to avoid excessive influence of oldest individuals is still notably higher

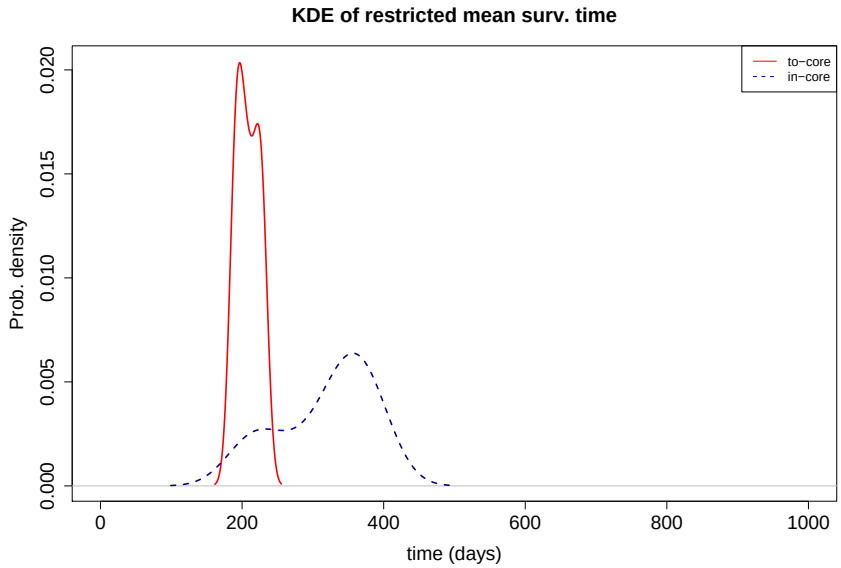

Figure 4.41: KDE of the restricted mean survival time of logged authors in the top ten Wikipedias who reached the core. The red solid line shows that, for all language editions, the restricted mean survival time to reach the core of very active contributors is concentrated around 200 days. Once the reached the core, the restricted mean of the time of membership (blue discontinuous line) varies between 200 and 400 days

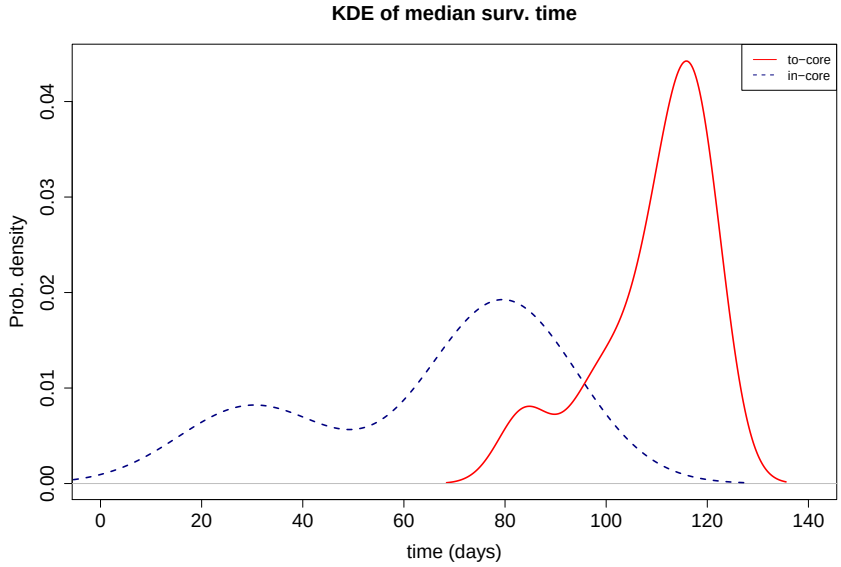

Figure 4.42: KDE of the meadian survival time of logged authors in the top ten Wikipedias who reached the core. As we can see, 50% of authors who eventually joined the core in the top ten Wikipedias needed between 80 and 130 days to reach that status. On the other side, 50% of core members left the core after a period oscillating between 20 and 110 days

4.5 Author Reputation and Article Quality

After the analysis of collaborative patters found in the Wikipedia community of authors, we now turn our attention to the characterization of authors participation in Wikipedia quality content. Exploring possible distinctive patterns that may be found in authors collaborating in Wikipedia FAs would provide us with valuable insights about how high quality content eventually stands out over regular quality articles in the language versions under analysis. To achieve this goal, we will analyze the *age* and *recentness* of both authors and articles, comparing the results of populations in FAs and non-FAs. Finally, we also explore the applicability of the quality metrics presented in previous research work by Stein and Hess in [108], trying to validate they usefulness to characterize quality articles and content in Wikipedia. As a future research work, we are interested in studying the application of these metrics to predict which current non-FAs may have better chances to be promoted to FAs in due course, as a result of their quality level, measured in this way.

In the first place, we present some quantitative measurements focused on articles. Table 4.17 summarizes the percentage of FAs in each of the top ten Wikipedias. A relevant result from this graph is that the total number of articles is not correlated with the percentage of FAs in a certain language edition. The Spanish language edition, in the 9th position according to its total number of articles, is the 1st Wikipedia by its percentage of FAs. It is also remarkable that the German language edition presents a higher percentage of FAs than the English Wikipedia (almost the double), even though the English language edition holds more than 3.5 times more articles than the German one. In general, it is remarkable the low percentage of FAs that we find in all language versions. It is also notirious that the Spanish language version almost doubles the ratio of FAs found in the Spanish Wikipedias,

whilst in the previus analyses we found very similar results for the lifetime of authors in both language versions. This clearly indicates that Spanish Wikipedians have found the correct combination to create a higher proportion of FAs from the set of encyclopaedic entries available in their language version.

Lang.	#Articles	# of redirects	#FAs	% FAs
EN	2,360,799	2,440,315	1,998	0.0846
DE	745,336	501,377	1,166	0.1564
FR	654,306	768,514	307	0.0469
PL	496,945	129,255	169	0.0340
JA	488,683	271,377	46	0.0094
IT	448,558	176,888	367	0.0818
NL	435,990	186,463	156	0.0357
PT	376,097	389,124	330	0.0877
ES	357,781	229,987	602	0.1683
SV	282,090	151,087	203	0.0720

Table 4.17: Num. of articles, num. of redirects, number of FAs and % of FAs in the top ten language editions of Wikipedia. It is remarkable the very low percentage of FAs reached in any of these language editions, despite their relatively long running time.

Table 4.18 presents the mean and median of the number of edits received by FAs and non-FAs in the top ten Wikipedias. To compute the number of edits, we also included redirects in the set of non-FAs, to compare our results with those obtained by Stein and Hess in 2007. We can see that the mean of the number of revisions in non-FAs for the German Wikipedia has not varied significantly. On the contrary, to compute the length of edits received by FAs and non-FAs, we filtered out redirects, as they have a much lower length that could distort our results. It is clear that FAs receive many more contributions from logged-in authors than non-FAs. The difference between both sets of articles is even more evident in some language editions like English (51 times more edits in FAs than in non-FAs) and Italian (33.8 times more edits in FAs). In this way, these results sustain the hypothesis that FAs in Wikipedia requires a much larger number of revisions by the community of authors to achieve their top quality level

Lang.	Type	Mean # revs.	Median #revs.	Mean len(revs.)	Median len(revs.)
EN	FAs	867.60	464	5,860	2,122
	non-FAs	16.77	3	1,994	596
DE	FAs	307.10	213	9,166	2,501
	non-FAs	14.70	5	2,651	880
FR	FAs	337.50	251	5,967	1,936
	non-FAs	11.69	4	1,967	606
PL	FAs	192.80	121	7,883	1,794
	non-FAs	9.02	4	2,223	1,256
JA	FAs	192.60	124.5	2,826	685
	non-FAs	10.54	4	1,148	522
IT	FAs	338.70	231	11,780	2,452
	non-FAs	10.00	3	1,910	654
NL	FAs	181.10	139	9,625	2,811

Lang.	Type	Mean # revs.	Median #revs.	Mean len(revs.)	Median len(revs.)
	non-FAs	9.24	3	2,109	880
PT	FAs	179.30	103.5	9,177	1,462
	non-FAs	6.21	2	1,976	838
ES	FAs	285.60	189.5	7,666	1,857
	non-FAs	10.85	3	2,756	1,242
SV	FAs	104.70	77	6,646	1,066
	non-FAs	6.42	3	1,117	376

Table 4.18: Mean and median of # of revisions and length of revisions (logged-in authors) for FA and non-FA in the top ten language editions.

Table 4.19 show the mean and median values of the number of different logged-in authors in FAs and non-FAs from the top ten Wikipedias. These figures demonstrate that FAs are written by a much higher number of distinct authors than non-FAs, showing that high quality contents in Wikipedia are a product of merged points of view from many different authors, compared to average quality articles. The joint conclusion that we can extract from the numeric summaries in these tables is that the production of quality content in Wikipedia presents a strong correlation between both a high number of authors and a large number of different revisions. In other words, Wikipedia needs to sustain, and increase as much as possible the number of different authors and the number of revisions received in case the project wants to ensure maintaining a process that is able to create top quality content.

Lang.	Type	Mean #logged auth.	Median #logged auth.
EN	FAs	216.4	113
	non-FAs	13.97	6
DE	FAs	80.17	57
	non-FAs	11.57	7
FR	FAs	58.27	41
	non-FAs	7.493	4
PL	FAs	38.62	27
	non-FAs	5.473	3
JA	FAs	60.65	49.5
	non-FAs	7.661	4
IT	FAs	50.35	37
	non-FAs	5.491	3
NL	FAs	46.03	37
	non-FAs	6.301	4
PT	FAs	37.75	28
	non-FAs	4.242	2
ES	FAs	49.99	32
	non-FAs	6.828	4
SV	FAs	30.4	23
	non-FAs	4.654	3

Table 4.19: Mean and median of number of distinct logged authors in FAs and non-FAs in the top ten language versions of Wikipedia

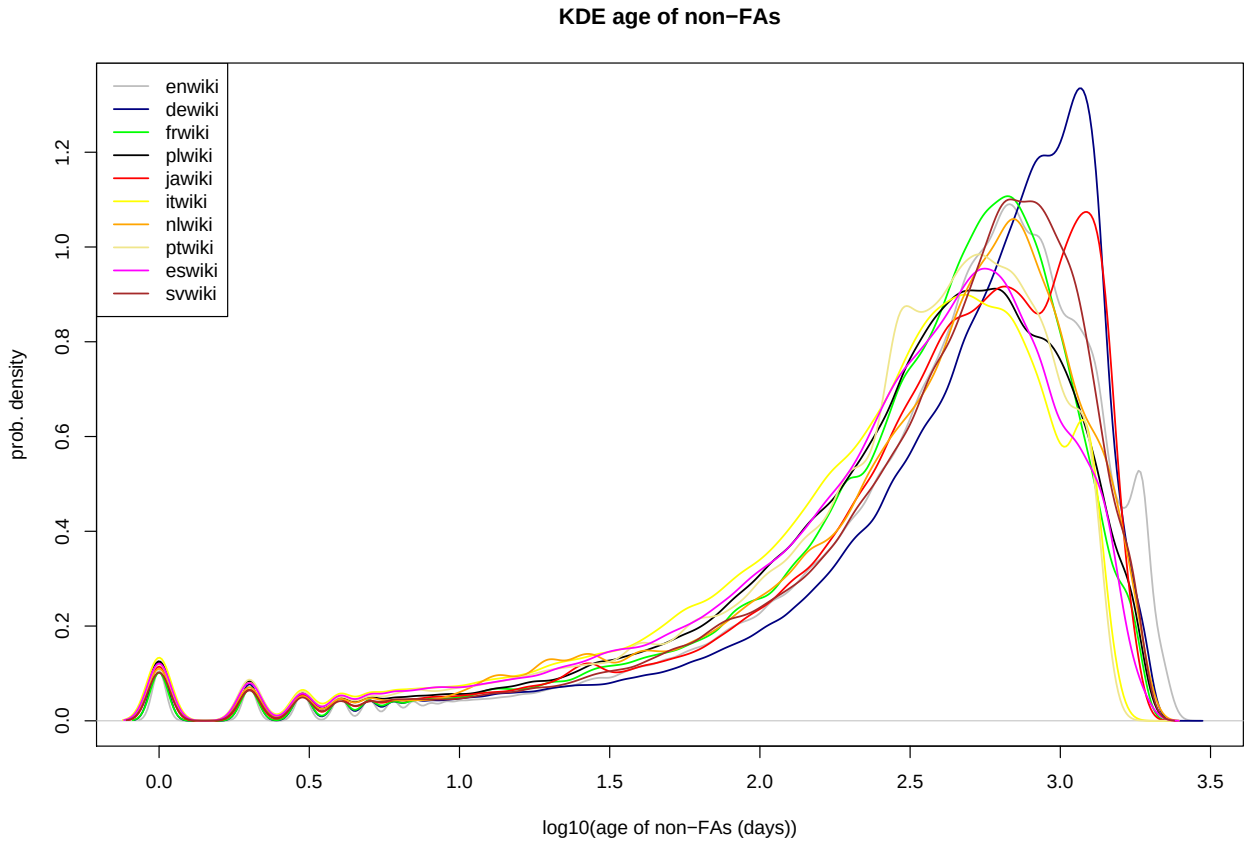

Figure 4.43: KDE of age of non-FAs. Typically, the age of standard articles in the top ten Wikipedias is situated around 750 days, except for the German and Japanese Wikipedias, with a clear majority of articles that are 1,000 days old

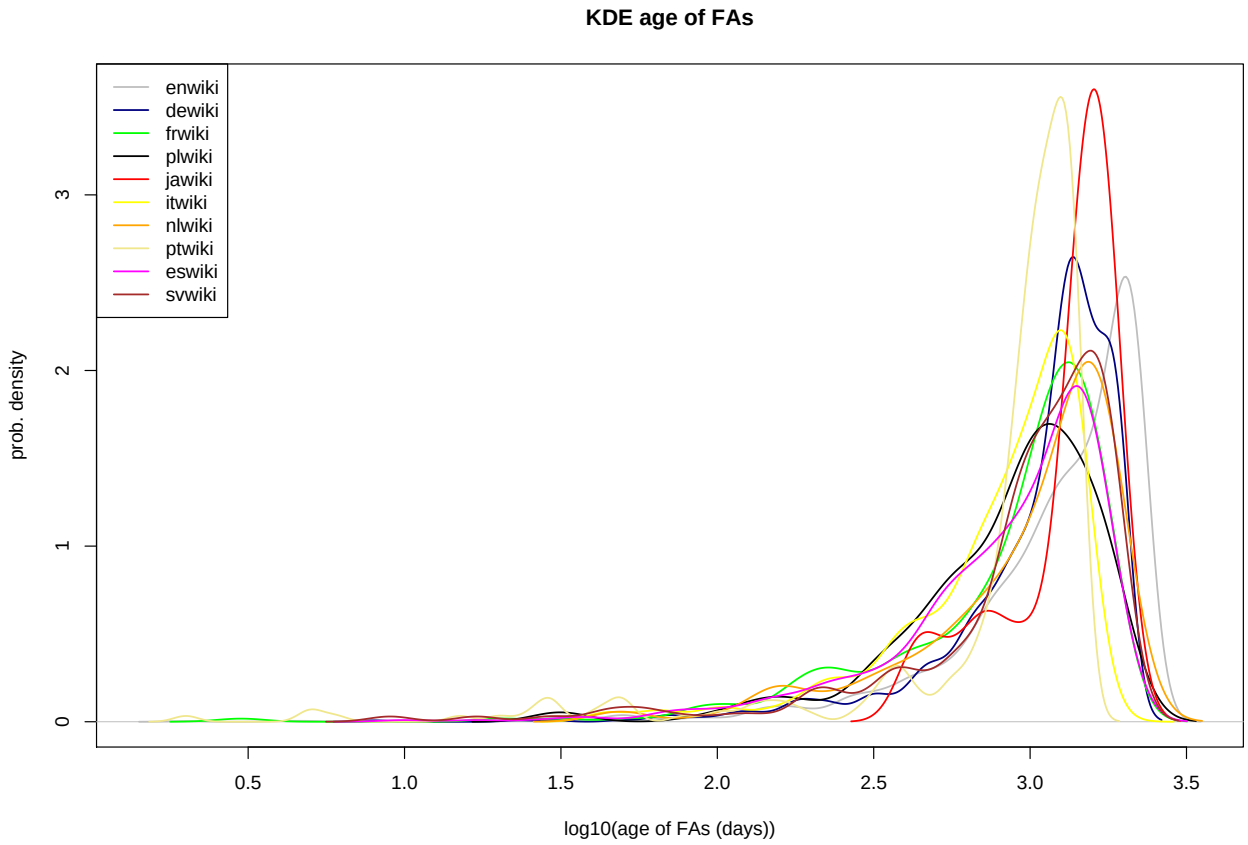

Figure 4.44: KDE of age of FAs. The proportion of FAs with an age lower than 500 days is very low, almost inexistent. The Polish Wikipedia has the largest proportion of younger FAs (most frequent value around 1,000 days). On the other side, the German and specially the English Wikipedia (the oldest one) have the largest proportion of older FAs. All the same, all language editions present quite similar patterns for this statistic in featured content, and these values are clearly higher than those found in non-featured content (as we might expect intuitively)

Regarding the age of non-FAs and FAs, we have to take into account that, unless they are deleted from the system, articles can not die in the same way that authors. In case a certain article is deleted, all its associated revisions are eliminated, as well, so we can not trace it using our database dump files. Therefore, these graphs will show the typical time an article needs to become a FAs, in comparison with the expected age of available articles in a certain language version. In this sense, Figure 4.43 shows that for most of the language versions, the modal value of age is around 750 days. Nevertheless, in the German and Japanese language versions we find a greater modal value for their age, around 1,000 days. Since neither of these versions is the oldest one (the English Wikipedia has more that one year of advantage in its running time), these results indicates that articles in the German and Japanese versions tend to last more in the system. If we focus on Figure 4.44, we can see that the distribution of values is quite biased toward the right area, with most FAs presenting an age of more than 1,000 days for all versions. As a consequence of this, we can conclude that FAs need longer time periods to be reviewed, processed and enhanced in order to achieve their privileged status, as we might have expected intuitively.

Table 4.20 summarizes the numerical results for the mean and median values. As we can see, all results are consistent with negatively skewed distributions (as shown in the probability density pictures) with median values above mean values in all language versions under study. In general, the oldest Wikipedias also present the higher median values of age of articles, both for featured and non-featured ones. The notorious exception to this general trend is the Japanese version, with even higher figures than the English Wikipedia, despite having a running time more than 500 days lower than the largest version. Thus, the general rule is that, with the exception of the Italian Wikipedias, 50% of the total number of FAs were topics with more than 1,000 days of presence in the system. If these terms were created so early, this is a clear indication about the interest of the community of authors about the topics covered by them. Moreover, since the number of FAs in each language version is so small, the figures presented for non-FAs are quite similar to the summary values corresponding to the whole articles population. In this respect, we can see for instance that 50% of the total number of non-FAs in the English Wikipedia where created less than 1,5 years from the end of 2007, approximately before the stabilization effect in the number of active authors and revisions per month. As a result, despite the general steady-state effect identified in all versions, all Wikipedias nearly doubled their number of articles in the supposedly not-so-active period. This strongly suggest that the stable number of contributions per month has been focused on opening new articles. The periodic competition to reach the next round-numbers threshold in the total number of articles for each language version may have contributed to maintain this trend, in spite of the stabilization in the authors activity level.

Lang.	Set	MEAN$\{age(p)\}$	MEDIAN$\{age(p)\}$
EN	FAs	1,169.50	1,435.49
	non-FAs	367.28	512.86
DE	FAs	1,064.14	1,306.17
	non-FAs	386.36	590.20
FR	FAs	812.83	1,086.43
	non-FAs	300.61	431.52
PL	FAs	820.35	981.74
	non-FAs	263.63	381.07
JA	FAs	1,261.82	1,496.24
	non-FAs	325.09	472.06
IT	FAs	762.08	952.80
	non-FAs	215.28	322.11
NL	FAs	931.11	1,253.14
	non-FAs	283.79	437.52
PT	FAs	762.08	1,071.52
	non-FAs	239.88	358.92
ES	FAs	809.10	1,013.91
	non-FAs	239.88	355.63
SV	FAs	895.36	1,140.24
	non-FAs	326.59	490.91

Table 4.20: Mean and median of age of FA and non-FA ($age(p)$) in the top ten language editions.

Figures 4.45 and 4.46 present respectively the *recentness* distribution for non-FAs and FAs. It is known that the nomination of a certain article as a FA candidate automatically produces a barrage of revisions on that article, due to the attention focused on it by a significant proportion of experienced logged authors. These figures show perfect agreement with this known trend, showing that the modal value of *recentness* for FAs is less than 50 days, while the modal *recentness* of non-FAs is around 100 days, thus doubling the values found for quality content. Again, the preferential attachment process takes place, influencing the number of revisions received by FAs in all language versions, since authors will tend to focus on working on top quality contents, specially the most experienced member of the community, as we will see shortly. Moreover, some of these FAs appear on the Featured Article section of the main page of every language version, thus boosting the visibility of this content and incrementing the possibilities of receiving even more contributions from a broader audience.

Regarding the *age* of logged authors in non-FAs and FAs in the top ten Wikipedias, Figures 4.47 and 4.48 present the results found for these versions. There exist a clear difference between the age distribution exhibited by non-FAs and FAs, namely the largest proportion of younger logged authors contributing to the former. The modal value for the age of authors is situated around 500 days. If we recall the results obtained in section 4.4, less than 40% of all logged authors had a lifetime longer than this value, while less than 30% of logged authors in the core of the language versions under study achieved to surpass that lifetime. As we can see, in Figure 4.48 the lower population of younger authors has almost dissappeared completely, revealing that older authors are the main creation force behind the content revision process in FAs for all language versions.

Table 4.21 summarizes the mean and median of the *age* of authors who revised FAs and non-FAs respectively. The extraordinary low values of the median age of authors in non-FAs in the Portuguese,

English and specially the Italian language versions shows a direct consequence of the lower lifetime of authors in these Wikipedias. However, even more interesting is the disproportionated difference between the mean or median age values of authors in FAs and non-FAs, of one order of magnitude in most cases. This is a particularly useful result in the context of this thesis work. It clearly demonstrates that FAs, that is top quality contents in Wikipedia, were mostly revised and contributed by long-lived authors within each community. Remembering the summary statistics for the typical age of authors in the system, we conclude that only a small proportion of the total number of logged authors will be able to reach such a longer lifetime, and now this numbers tell us that those scarce authors conform the elite of contributors of Wikipedia. Given the current tend of births and deaths, it will be more and more difficult to count on this kind of authors in the future, if the current trend is maintained over the following years. Unless newer authors were specifically skilled to perform more effective contributions to articles in less time (speaking in terms of content quality), the production of FAs in Wikipedia could become menaced by this negative demographic tendency.

Lang.	**Set**	**MEAN**$\{age(a)\}$	**MEDIAN**$\{age(a)\}$
EN	FAs	1,420	1,454
	non-FAs	254.3	39
DE	FAs	1,263	1,332
	non-FAs	429.9	268
FR	FAs	1,074	1,161
	non-FAs	323.8	149
PL	FAs	1,068	1,012
	non-FAs	405.4	293
JA	FAs	1,425	1,585
	non-FAs	379.3	193
IT	FAs	956.1	994
	non-FAs	237.8	25
NL	FAs	1,205	1311
	non-FAs	347.4	171
PT	FAs	1,206	1128
	non-FAs	235.4	68
ES	FAs	1,064	1,061
	non-FAs	311.2	157
SV	FAs	1,168	1,238
	non-FAs	386,5	203

Table 4.21: Mean and median of $age(a)$ in FA and non-FA in the top ten language editions.

Complementing these results, Figures 4.49 and 4.50 present the distribution of *recentness* for logged authors in non-FAs and FAs in Wikipedia. The influence of the change of trend in the ratio of births and deaths in all language versions, starting on summer 2006, can be found here, as well. The majority of authors contributing to non-FAs in Wikipedia left the project more than 500 days ago. Considering that the final date of our samples is January 1, 2008, that value concords with the period in which we found that the monthly number of deaths overtook the number of births. The influence of this factor is high enough as to lead to the absence of relevant differences among the curves for authors in non-FAs and FAs. Thus, it seems that the "big crack" in the monthly number of active authors since summer 2006 affected both populations in a similar manner.

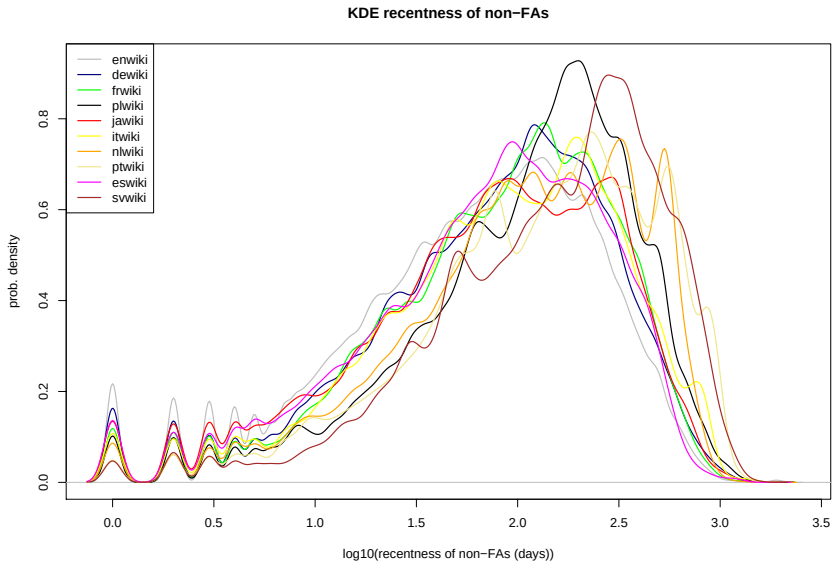

Figure 4.45: KDE of recentness of non-FAs. The graph shows that the largest proportion of non-featured content in the top ten language editions received their last contribution more than 100 days before the limit date in our data samples (January 1, 2008). The fact that non-FAs receive less frequent contributions from logged authors seems to be a clear explanation of their poor quality content

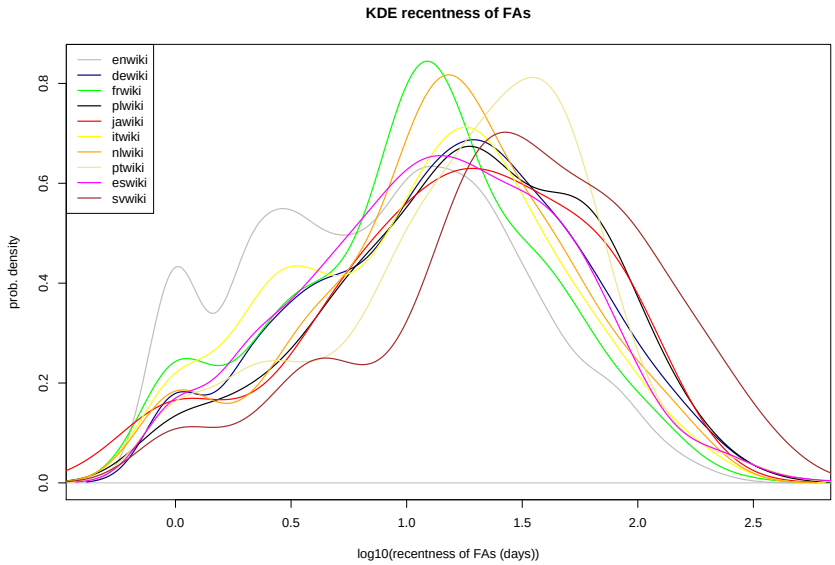

Figure 4.46: KDE of recentness of FAs. This graph is eloquent by itself, showing that featured content in the top ten Wikipedias tend to suffer more frequent modifications and revisions to further augment their high quality level. The majority of FAs in any of the top ten language versions received their last revision before the last 50 days of history stored in the analyzed archives

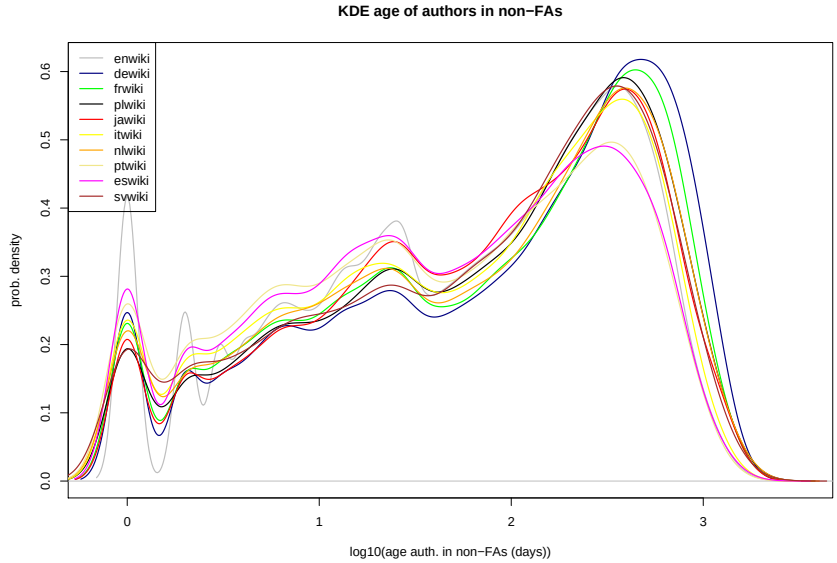

Figure 4.47: KDE of age of logged authors who revised non-FAs. We have a significant proportion of younger logged authors contributing to non-featured content, though the majority of revisors still have an age around 500 days for all language versions under analysis

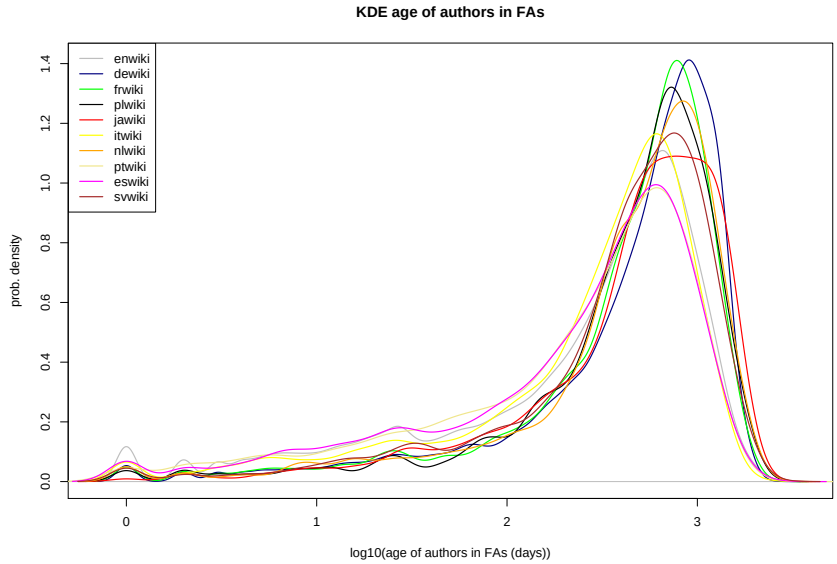

Figure 4.48: KDE of age of logged authors who revised FAs. It is interesting to notice that the group of younger authors has virtually dissapeared, for all language versions, showing that the creation of quality content is mainly devoted to fairly experienced authors in the top ten Wikipedias

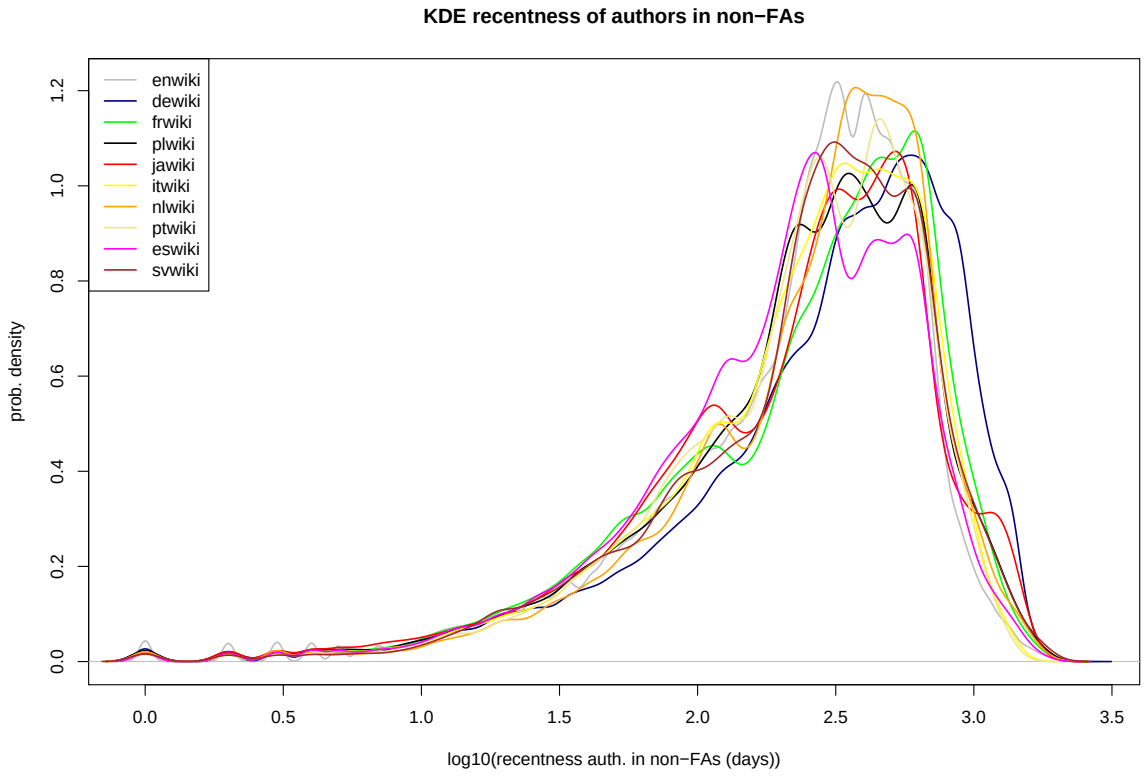

Figure 4.49: KDE of recentness of logged authors who revised non-FAs. The majority of logged authors contributing to non-featured content leaved the project more than 500 days ago (considering January 1, 2008 as the limit date of our data samples). Therefore, it is interesting to notice that Wikipedia lost the majority of their logged before getting to the steady-state region showed in the monthly number of contributions received during the last year of our samples

To conclude with this section, we present Tables 4.22 and 4.23, summarizing the reputation level of authors and ratings of articles according to the definitions proposed by Stein and Hess in [108] [4]. For comparison purposes with these previous results, we also considered redirect articles in these calculations. Our results confirm that: 1) the rating levels of FAs and non-FAs (both author-based and edit-based) are higher in all computed language editions than those corresponding to non-FAs; 2) there exist a strong correlation between the average reputation of authors and the number of FAs in every computed language edition, with the exception of the German Wikipedia (if we remove that sample, the correlation coefficient raises up to $r = 0.8848$). Probably, the higher number of articles in this language edition counterbalances the reputation of many of its authors, thus lowering the average value of authors' rating.

[4] At the time of this writing, the last values corresponding to the English language edition of Wikipedia could not have been computed before the submission deadline, though for the camera-ready version of this paper, these results will be available.

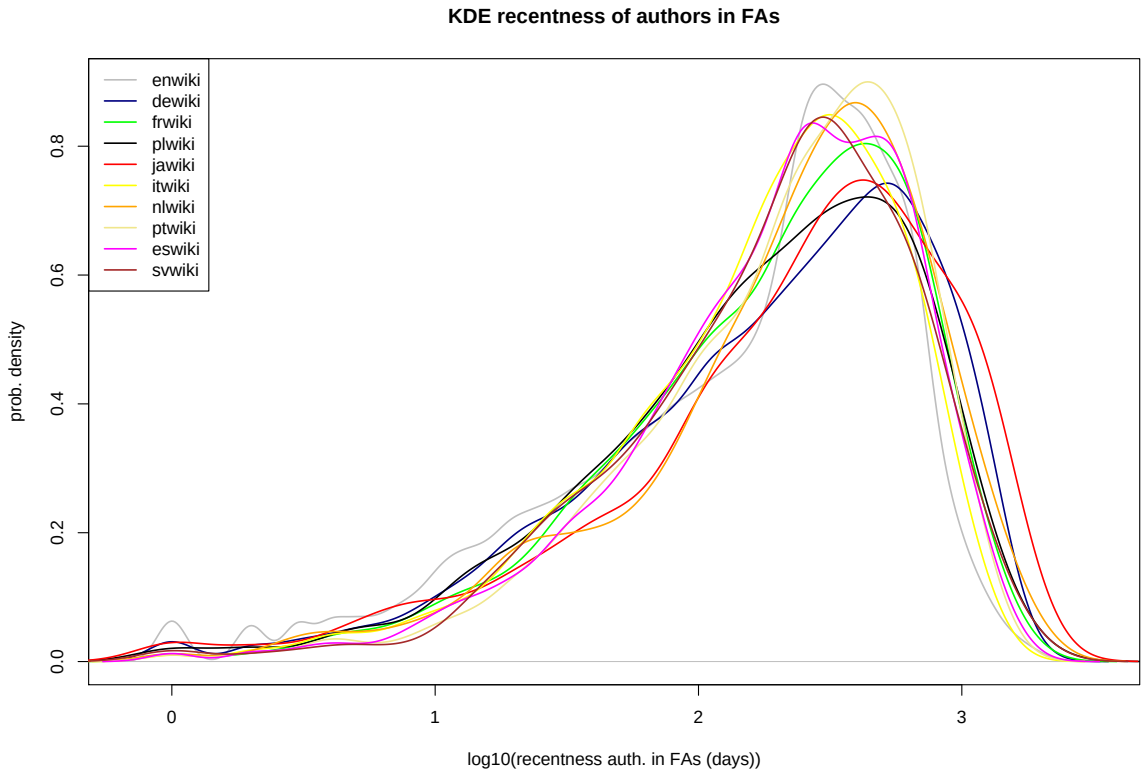

Figure 4.50: KDE of recentness of logged authors who revised FAs. In this case, we can not see any major difference between authors contributing to featured and non-featured content. It seems that the "big crack" in the logged authors population, in summer 2006, had also a great impact in revisors of FAs as well

Table 4.22: Mean values of authors' reputation in the top ten Wikipedias

Lang.	**MEAN**$\{rep_{pb}(a)\}$	**MEAN**$\{rep_{eb}(a)\}$
EN	0.02009	0.02004
DE	0.01150	0.01163
FR	0.00515	0.00533
PL	0.00408	0.00428
JA	0.00114	0.00116
IT	0.01477	0.01494
NL	0.00315	0.00336
PT	0.01339	0.01346
ES	0.01596	0.01590
SV	0.00685	0.00705

It is also noticeable the very low rating values corresponding to non-FAs in the Japanese language edition of Wikipedia. We should remember that this language edition has the lower number of FAs in the list of the top ten Wikipedias. Hence, it seems that the non-FAs in the Japanese Wikipedia have a longer way to reach the FA status than non-FAs in the rest of language editions (though Japanese FAs do present average rating values more in concordance with the ones found in the rest of top ten Wikipedias).

Lang.	**Set of articles**	**MEAN**$\{rat_{ab}(p)\}$	**MEAN**$\{rat_{eb}(p)\}$
EN	*FAs*	0.071583	0.159752
	non-FAs	0.007443	0.014103
	all	0.007470	0.014166
DE	*FAs*	**0.036102**	**0.107535**
	non-FAs	0.00842	0.0153545
	all	0.008452	0.0154469
FR	*FAs*	**0.023597**	**0.081127**
	non-FAs	0.0017562	0.004061
	all	0.001761	0.004078
PL	*FAs*	**0.018640**	**0.052306**
	non-FAs	0.001868	0.004396
	all	0.001872	0.004410
JA	*FAs*	**0.034484**	**0.050225**
	non-FAs	0.000588	0.001043
	all	0.000590	0.001046
IT	*FAs*	**0.03903**	**0.108078**
	non-FAs	0.005760	0.014344
	all	0.005782	0.014402
NL	*FAs*	**0.019544**	**0.05500**
	non-FAs	0.002062	0.004177
	all	0.002067	0.004190
PT	*FAs*	**0.051913**	**0.081945**
	non-FAs	0.003961	0.009451
	all	0.003982	0.009483
ES	*FAs*	**0.044994**	**0.119105**
	non-FAs	0.008629	0.021184
	all	0.008668	0.021290
SV	*FAs*	**0.028713**	**0.049068**
	non-FAs	0.003450	0.006975
	all	0.003463	0.006996

Table 4.23: Mean values of articles rating in the top ten Wikipedias

Stein and Hess also proved that a pruned version of this measure shows significant differences in favor of average ratings of FAs, as well. This way, they already showed that this measurement presents a strong resiliency against possible *flash crowd* effects, maybe produced by a greater number of authors whose contributions are attracted to FAs candidates, starting from the date they were nominated by the community.

Our quantitative analysis of FAs and author reputation in Wikipedia leave some interesting conclusions. First and foremost, there exist common quantitative patterns in FAs of the top ten language editions of Wikipedia, according to their total number of articles. FAs in these language editions are longer, present a higher number of different authors and longer revisions than non-FAs. They are also older articles (in average), according to our definition of *age*, and they have a much lower average *recentness* value. Finally, FAs presents higher average rating values, computed following Stein and Hess proposal.

So that, according to this results it would be possible to create an automated tool to assess the quality of articles for any language edition of Wikipedia, from a quantitative point of view. This could be some kind of mash-up gadget, that could retrieved pre-computed values of all these parameters for any given page, and then display them in a sidebar, at the time of rendering the article in the client host. This metadata could be very valuable for a certain author to rapidly get a first notion of the current quality level of that article. It could also be utilized to identify possible candidate articles to be promoted to the FA status.

In combination with other quality assessment plug-ins like, for instance, the content driven reputation system proposed by Adler et al. [6], we may end up with a simple, yet powerful set of tools to facilitate recognition of high quality contents in Wikipedia. Possible applications of this set of tools are endless, beginning with direct assessing of quality of contents for Wikipedia readers, along with serving as a useful, automated way to help in the identification of articles suitable for being included in "stable", revised versions of any language edition of Wikipedia.

On the other hand, we have also found shared behavioral patterns among authors contributing to FAs in Wikipedia. Firstly, FAs in Wikipedia received contributions from a higher number of distinct authors, which confirms the hypothesis that high quality contents (respecting the NPoV principle, as well as other similar top quality requirements) are produced mixing the point of view of many different editors, thus providing a more balanced coverage of a certain topic. Another relevant result here is that we demonstrate that logged-in authors contributing to FAs are older than authors contributing to non-FAs. Hence, as we might have thought, more experienced authors with a deeper knowledge of inner details of Wikipedia workflow can contribute more effectively to produce higher quality articles than novel users with less experience contributing to Wikipedia. In the case of authors, *recentness* values does not show significant differences between authors contributing to FAs and non-FAs, though FAs present a heavy tail in the low region of recentness values, indicating that, at least, a noticeable number of authors do maintain a more frequent rate of contributions to FAs. Further work in this research line should explore if this group of frequent contributors to FAs presents some connections with the core of very active contributors that has already being identified in previous research works like [82].

Finally, our quantitative analysis has served to validate and extend the proposal previously presented by Stein and Hess in [108], of measurements authors reputation and articles quality ratings, based on the number of FAs, and contributions to FAs made by Wikipedia authors. We prove here that those measurements are valid for any language edition in the list of top ten Wikipedias, if we want to distinguish FAs from non-FAs. An interesting sequel of this research work would be to verify is this model is adequate to identify non-FAs that could be candidate for promotion to FAs, as that would help to accelerate the identification process of FAs, one aspect that should be rapidly improved in Wikipedia in due course. As a first sample of potential applications of this measurements, we have also shown that we can use the average value of $\text{MEAN}\{rep_{pb}(a)\}$ to predict the number of FAs that we should find in 8 from the 10 language editions of Wikipedia under study. As a result, the average value of $rep_{pb}(a)$ may act as a proxy to estimate the capacity of a certain community of authors in Wikipedia to produce high quality contents.

4.6 Evolution of Wikipedia

The final section of this thesis work covers the analysis of the evolution over time of some critical parameters that we have identified so far. Understanding the evolution of this parameters is also critical to characterize the behavioral patterns found in the Wikipedia community of authors, since some of them may exhibit at present time different patterns than those ones found in the early history days of their log archives. We are specially interested in finding out possible changes in descriptive parameters measuring the inequality level of contributions of Wikipedia authors over time, paying special attention to the patterns found after the shift in the trend of monthly deaths and births in all language versions.

We start with the evolution over time of the different CCDF of several parameters analyzed on the previous Section about the social structure of the Wikipedia community of logged authors. Figure 4.51 shows the evolution in time of the number of different articles edited per author in the top ten Wikipedias. We can observe that the trend followed by these curves over time is to become stepper, though the differences among distinct years are not quite significant. Naturally, the upper limit of the upper truncated Pareto distribution has raised in all language versions, since authors progressively have more time to contribute to articles. It is remarkable, though, that this growth in the maximum number of different articles edited per author yields quite similar values in all language versions, despite their disparate number of authors and articles, and also disregarding the running time of each language version.

Figure 4.52 shows de total number of different logged authors who edited a certain article in Wikipedia, proving that the distribution of this statistic has never abandonned its lognormal shape as time goes by in any language version. Likewise, Figures 4.53 and 4.54 depict the evolution in time of the number of revisions per logged author and the number of revisions per article respectively. The evolution of the former statistic does not provide any further interesting information, but the evolution of the number of revisions accumulated per article does show a relevant shift in the pattern of this statistic. In earlier years, the statistic presents in all language versions a pattern closer to a Pareto distribution, specially for the upper values of the graph. Nevertheless, starting from 2007 the best fit is not anymore a Pareto-like distribution, but a lognormal one. Therefore, the main point to remark is that the distribution of the number of revisions per article is becoming less biased, evolving towards a lognormal shape as new revisions begin to fall in articles that did not have got so much attention previously.

As we have seen, although we found a change in the monthly number of births and deaths in the top ten Wikipedias from summer 2006 on, and despite the steady-state reached by the monthly number of contributions by all type of authors, the evolution over time of the parameters presented above indicates that the community of users is trying to broaden the coverage of the number of articles revised, and authors that remain alive in the system are intensifying their workload. This findings concords with the evolution of the Lorenz curves for authors and articles, presented in Figures 4.55 and 4.56, respectively. The trend exhibited by the Lorenz curves as time goes by is that the core group of most active authors in each language edition is gradually accumulating a higher proportion of the total number of revisions performed. At the same time, the most popular articles in each language version (a group that includes all FAs as we already demonstrated in the previous section) is also receiving a growing number of contributions, skewing the Lorenz curve towards the right side of the graph.

Nevertheless, in previous sections we have also confirmed that we should pay attention to the evolution of the monthly figures of certain statistics in order to undercover interesting patterns that may have remain shadowed by annual subtotals otherwise. Such is the case of the monthly Gini

coefficient for logged authors in the top ten Wikipedias, shown in Figure 4.57. Previous research work conducted by our group revealed this interesting pattern in the Wikipedia community of authors [83]. Given the results we have presented in this thesis, the explanation of this apparently self-regulation behavioral pattern is twofold. In the first place, we have the leverage in the monthly number of active logged authors. In the second place, the graphs showing the evolution in time of the community tell us that the group of core authors is increasing their workload as time goes by. Therefore, the approximately constant level of the monthly Gini coefficient for logged authors is being obtained at the cost of a substantial raising of the monthly number of contributions that the most active authors perform in the system. As a consequence, the distribution of these revisions is becoming more and more unequal. On top of that, we also have to recall that there exist an natural upper limit in the capacity of human authors to perform revisions during a fixed time period.

This creates a potentially risky situation for the sustainability of the whole project. If the current trend is maintained over the following months, we will eventually see the month in which the core of very active contributors is not capable of increasing their workload enough to maintain the current activity level, and the monthly Gini coefficient will begin to decrease, not because there are more less active authors also contributing to the project, but because the core of top active authors will progressively lose some of their members.

Complementary results already presented in other previous research works by our group gives additional insights about this phenomenon. Figure 4.58 presents the evolution over time of the number of contributions performed by the top 10% of most active users in each month. That is to say, for each month we calculate the top 10% of the list of most active authors. Then, we depict the evolution of the monthly number of revisions performed by these authors over the remaining months, thus creating a 3D grid. In this grid we can see the effect of developers turnover in the core of very active authors. The left side of the graph shows that earlier members of the core of very active authors have not increased their number of contributions in the last months. On the contrary, new very active members (responsible for the highest, warmer area on the top right of the graph) are relatively new users who did not contributed so much in previous months. The consequence of these results is that newer members of the core contribute more actively than previous ones, thus helping to skew the distribution of the number of contributions per author towards the group of core users over time.

Now, the question turns out to be whether or not we are beginning to register a drop in the number of core authors in each language version. Figure 4.59 gives us the answer to this question. Indeed, we can not appreciate significant falls in the number of authors in the core for any language version, though we can clearly see that this statistic as also reached a steady-state since summer 2006, coinciding with the shift in the trend of monthly births and deaths. As a result, the increasing number of deaths in the system is also affecting the capacity of the community to attract top contributors, as well. Since the Gini graphs demonstrate that this core of very active authors is responsible for the major part of the content creation process, its evolution in the following months will be critical to see whether Wikipedia gets into an hypothetical recession period.

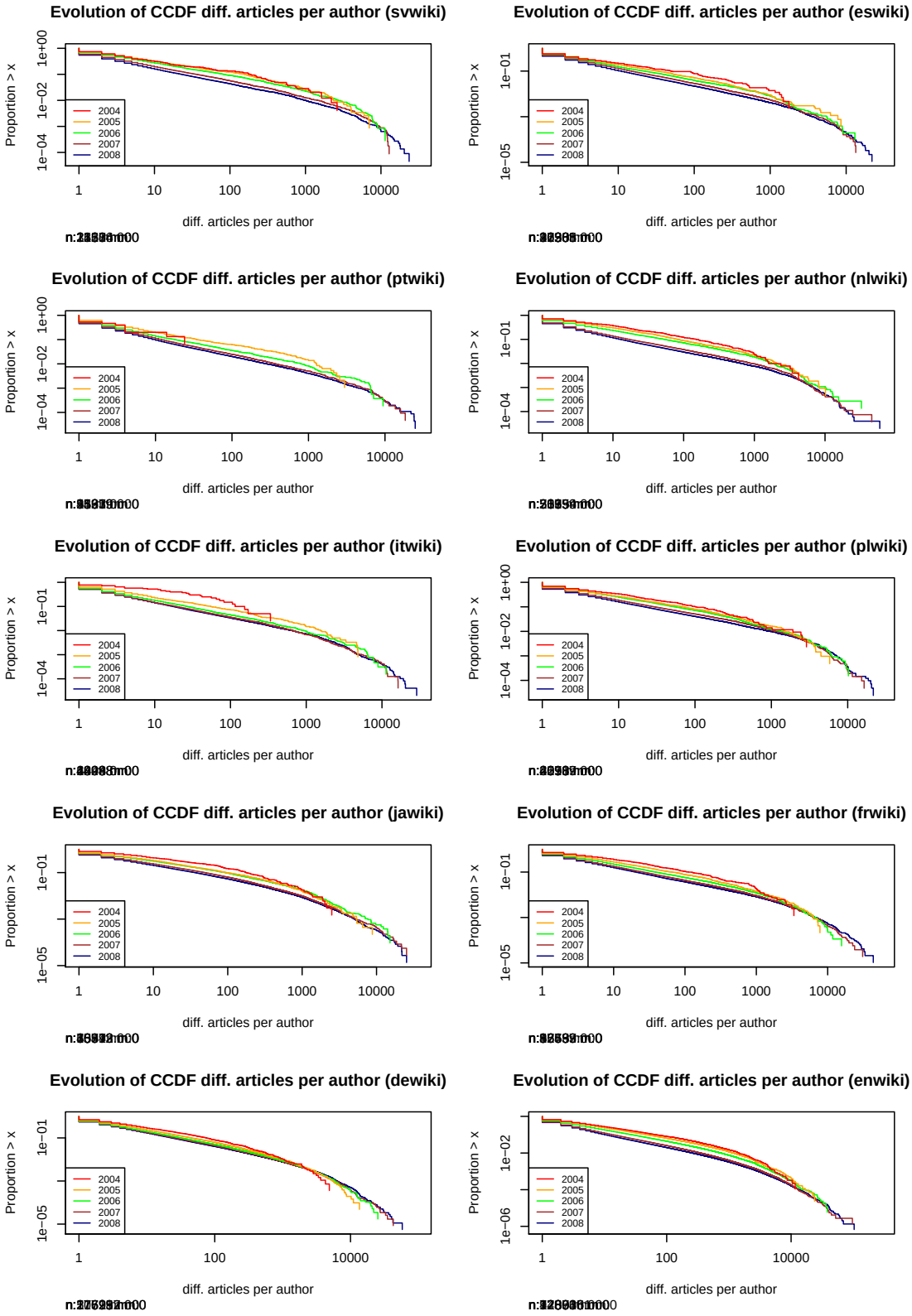

Figure 4.51: Evolution in time of the CCDF of the number of different articles revised per logged author in the top ten Wikipedias. We recall that this statistic follows an upper truncated Pareto distribution in all versions. The graphs show that the slope of the Pareto area tend to become stepper as time goes by, and the number of articles revised by the most active contributors helps to expand the upper limit of the distribution to the right of the plot. This also allows the Pareto-like area to expand its influence over a larger proportion of the logged authors community in each language

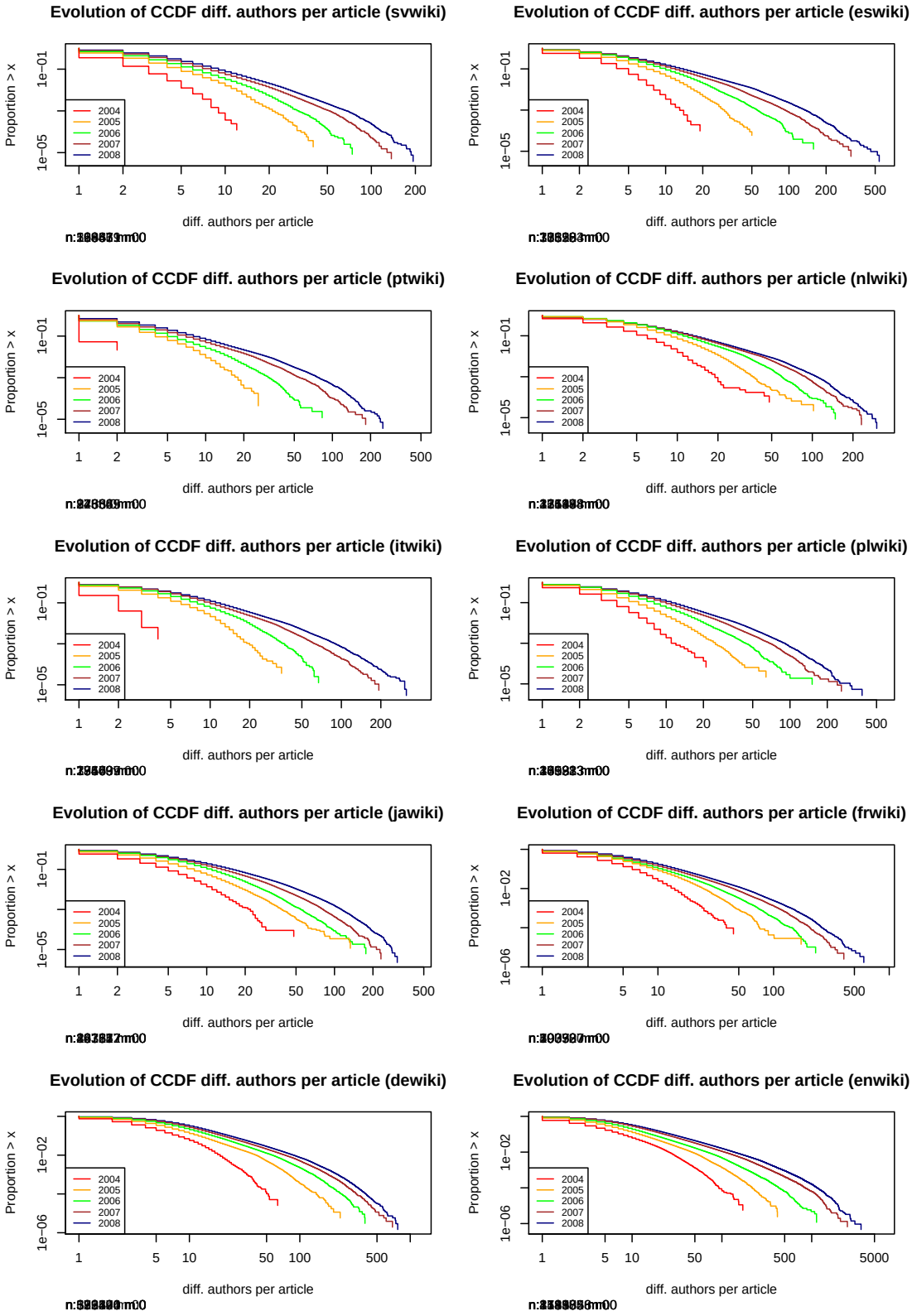

Figure 4.52: Evolution in time of the CCDF of the number of different logged authors who revised articles in the top ten Wikipedias. We can see that the lognormal patter have been present all along the history of this statistic for all language versions. The maximum number of different logged authors per article has grown until the range 200-500 in all language versions, even the largest ones, with the sole exception of the English Wikipedia, with an upper limit one order of magnitude higher than the remaining versions

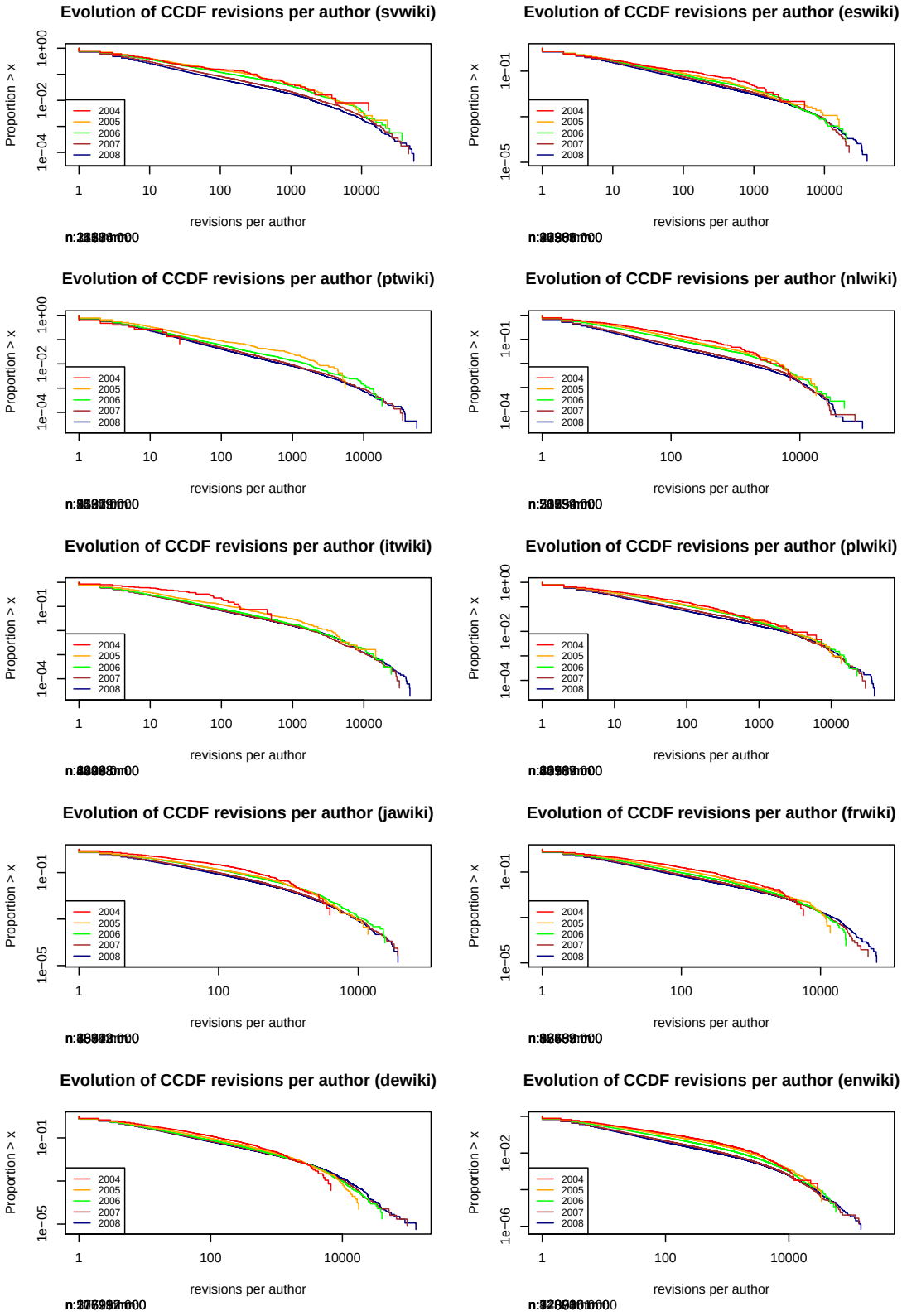

Figure 4.53: Evolution in time of the CCDF of number of revisions per author in the top ten Wikipedias. The graph exhibits a similar behavioral patter than the one found for the number of different articles revised per author, with an upper truncated Pareto distribution whose higher limit grows rapidly over time

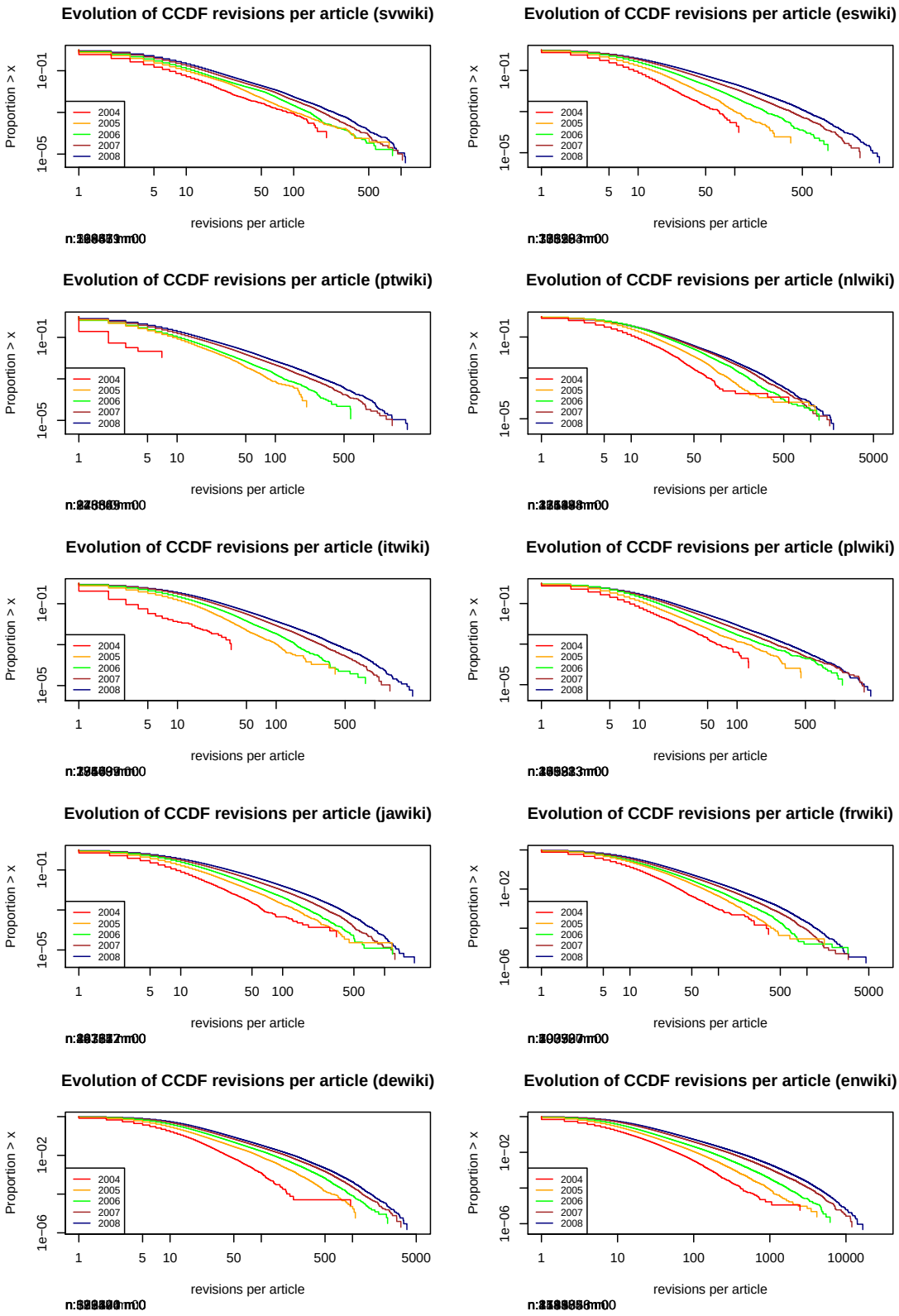

Figure 4.54: Evolution in time of the CCDF of number of revisions received per article in the top ten Wikipedias. The shape of the upper values of the graph in previous years (2005 and 2006, in particular) may suggest a Pareto fit for those region, as reported in previous research works. However, the graph also shows that in 2008 the shape of the graph can not be approximated by a straight line, and the curvature corresponding to a lognormal distribution provides a better fit for this statistic

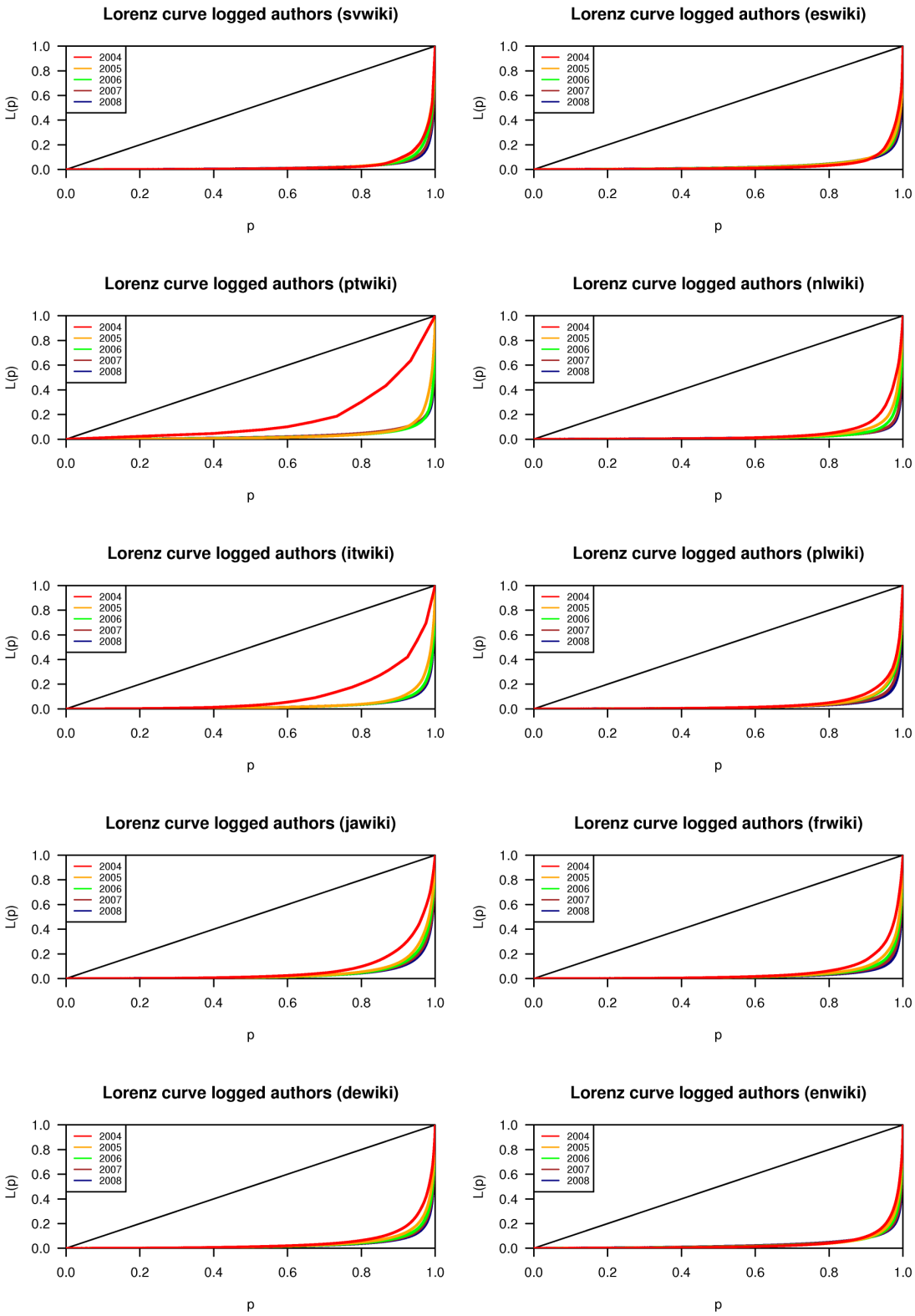

Figure 4.55: Evolution in time of the Lorenz curve showing the distribution of revisions among the logged authors community in the top ten Wikipedias. There exist a clear tendency towards higher inequality levels as the project evolves, showing that the core of very active authors in progressively taking over a larger proportion of the creation process, though the differences among distinct years is not very significant

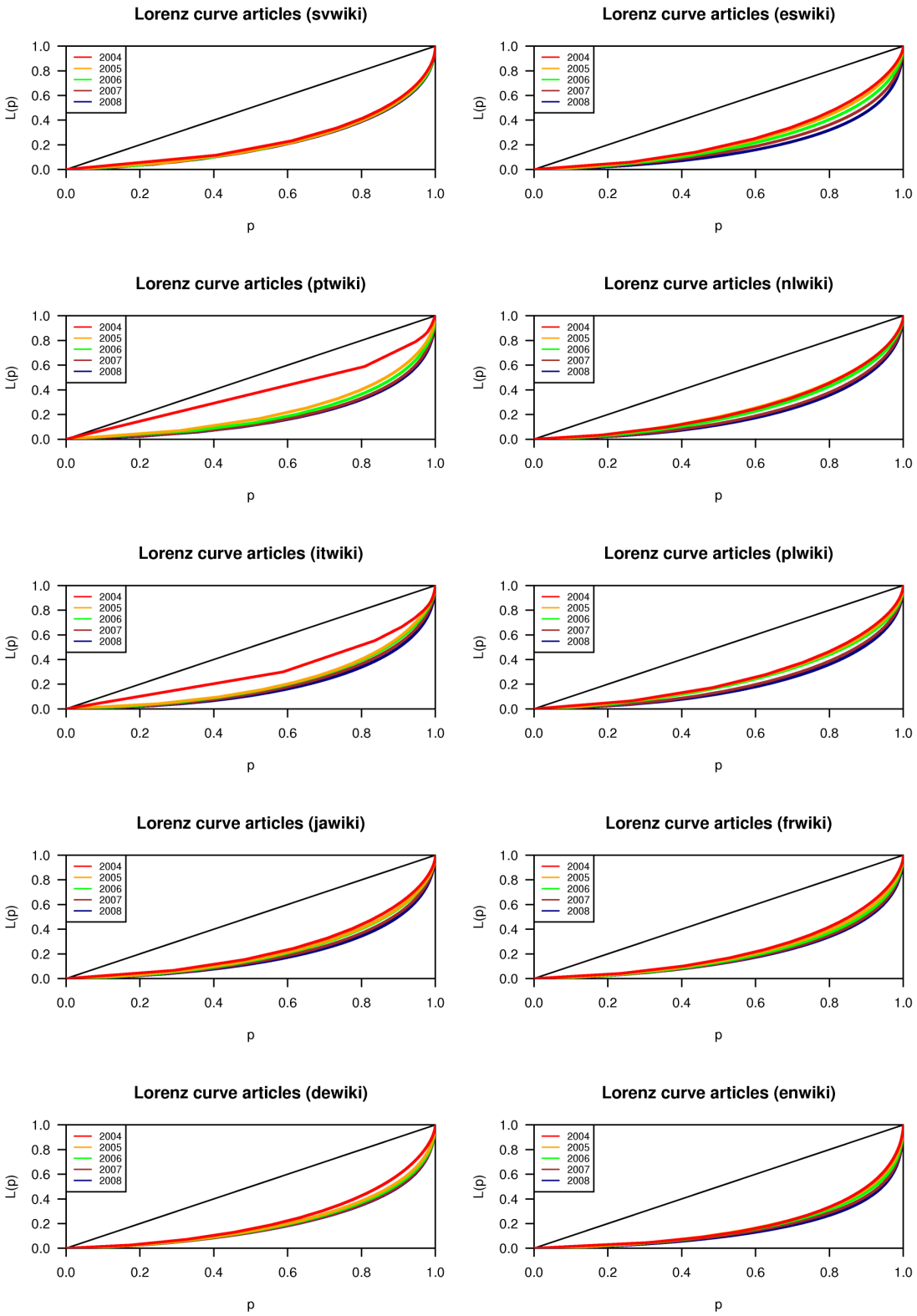

Figure 4.56: Evolution in time of the Lorenz curve showing the distribution of revisions among articles in the top ten Wikipedias. Again, there is a clear tendency toward higher inequality levels as time goes by, focusing on a group of articles that progressively become more and more popular in each language version. Recalling the results previously presented for the recentness of FAs, we can infer that they are situated in the most popular right side of these graphs

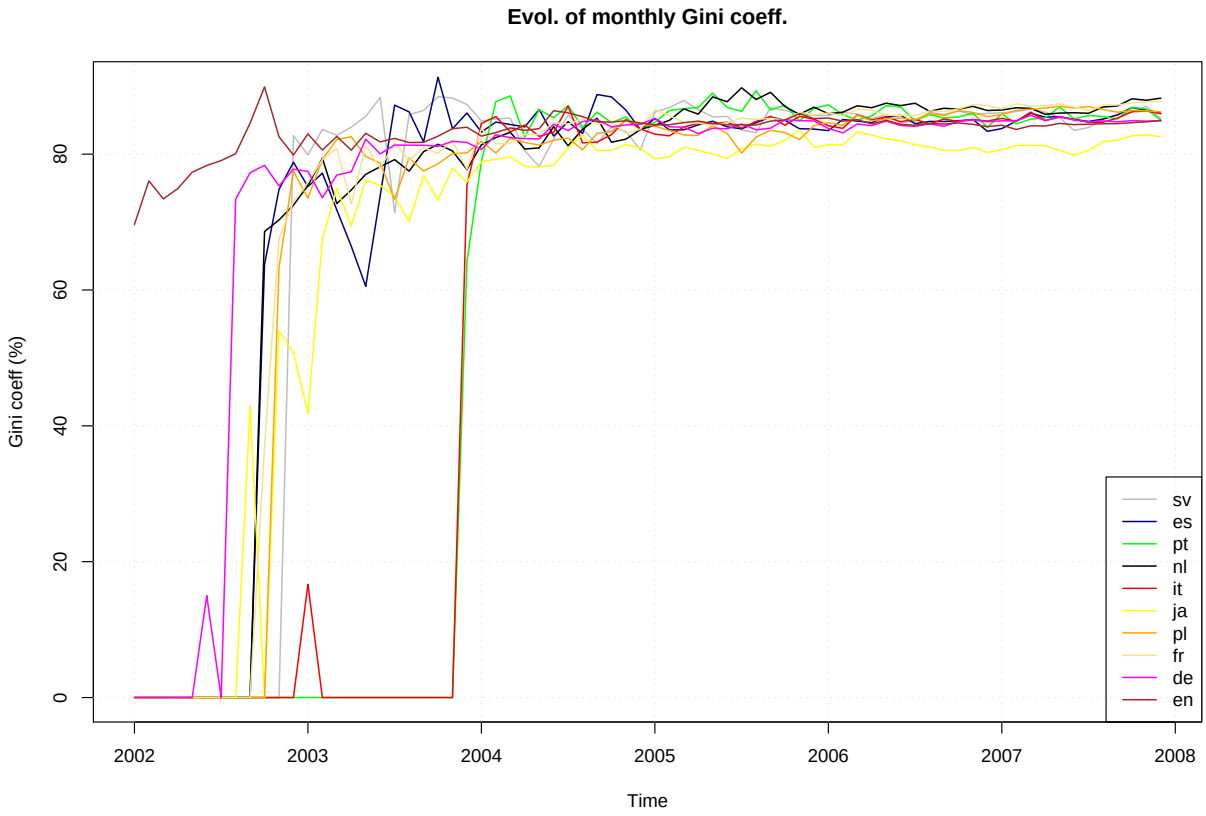

Figure 4.57: Evolution in time of the monthly Gini coefficient for logged authors in the top ten Wikipedias. The plot demonstrates that all language versions reach a self-regulated pattern after their first two years of history. The initial transitory state is debt to the initial influence of anonymous authors, which is subsequently overwhelmed by the number of contributions from logged authors in subsequent years. Interestingly, all language versions get stabilized within a small interval, with monthly Gini coefficients in the range 80-85%

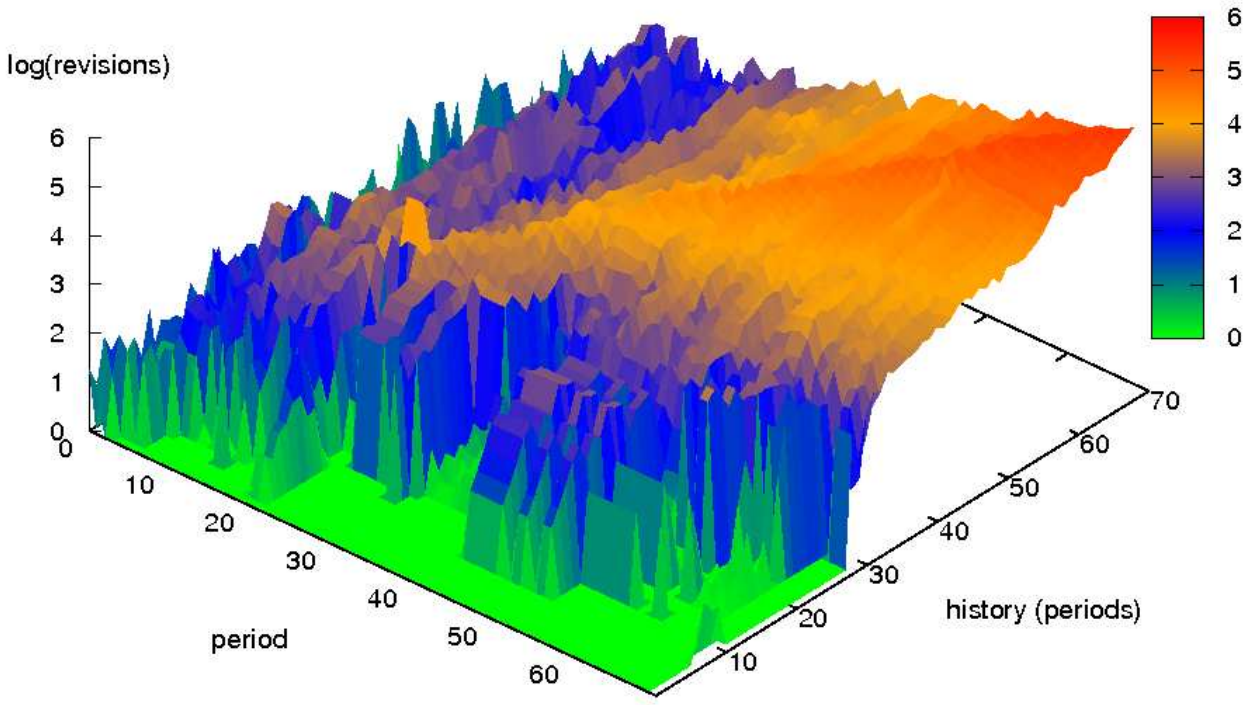

Figure 4.58: Evolution of the number of contributions from the top 10% of most active contributors in each month (x axis), over the remaining months (y axis) in the English Wikipedia. The vertical axis has a logarithmic scale. Core members in early history months continue to contribute to the project, but with lower rates, forcing them to leave the core in favor of new, more active core members in the last months (represented by the red, very active area in the top right corner of the 3D graph)

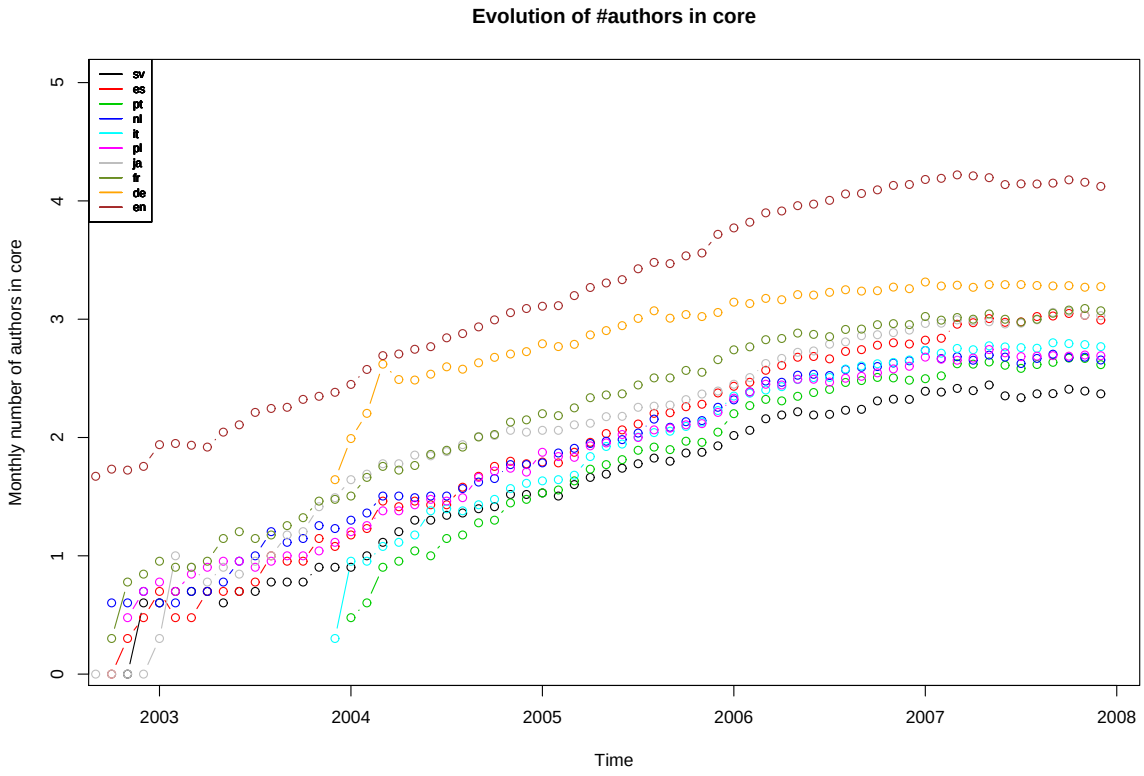

Figure 4.59: Evolution of the number of authors in the core of the top ten Wikipedias. Since the core group is taken as a constant proportion of the 10% most active users, the pattern must follow the shape of the monthly number of active users, with the same leverage effect in the last year. The interesting point here is to quantify how many users do we need in the core to maintain the same production effort as in recent years, since the distribution of revisions among logged authors tend to become more and more unequal as time elapses. We use a log10 scale in the vertical axis to show these results

Chapter 5

Conclusions and Future Research

> "On the contrary, Watson, you can see everything. You fail, however, to reason from what you see. You are too timid in drawing your inferences". *The Adventure of the Blue Carbuncle*, in *The Adventures of Sherlock Holmes*, Arthur Conan Doyle, (1892).

5.1 Featured results

To start with this wrap up chapter, we are going to summarize the main results and conclusions that can be extracted from our empirical analysis of the top ten language versions of Wikipedia. Our presentation is divided in two sections. In the following one, we describe the main conclusions and findings obtained while answering each of the main research questions tackled in this thesis. Then, we discuss the influence that some of these results may exert on the sustainability of the Wikipedia project over the next years, with special attention to the main goal of producing as much quality contents as it were possible.

5.1.1 Research questions tackled in this thesis

In the first chapter of this thesis work, we presented the main research questions that we addressed throughout our quantitative analysis of the top-ten Wikipedias. Chapter 3 covered in detail the steps followed for answering each of these questions. Now, we are going to summarize the most important conclusions that can be extracted from the answers obtained for these questions.

1. **Q1: How does the community of authors in the top-ten Wikipedias evolve over time?**: We found that the monthly number of active logged authors in all language versions under study has reached a steady state from approximately summer 2006, and clearly, over 2007, the last available year in our data samples. As a consequence, the monthly number of revisions performed by logged authors (both considering the whole set of wiki pages in each version, or focusing on the articles population exclusively) has also stabilized over the same period. The same change of the previous steady growing rate was found in the monthly number of revisions performed by anonymous authors. Since we cannot trace each of these authors individually, we cannot formulate any definitive conclusion about this change of trend, though it seems possible that the same cause may apply for this case as well. Finally, looking at the monthly share of contributions from bots in each Wikipedia, we found interesting insights about the precise

strategy followed by each language version to increase the coverage of encyclopaedic terms. For instance, the Polish and Dutch Wikipedias register a very high rate of bot contributions in many periods of their respective histories, something that eventually influences the shares of pages found in each namespace and thus the composition of the whole set of wiki pages in those versions.

2. **Q2: What is the distribution of content and pages in the top-ten Wikipedias?**: The monthly number of active articles in all language versions has experimented the same shift to a steady-state condition that we already found in the monthly number of active logged authors and revisions. As far as redirect pages is concerned, the stabilization in the number of active redirects per month is always clear, also from summer 2006. Moreover, we have found that the typical distribution of the length of pages storing content is bimodal, while discussion pages (like talk pages, category talk or user talk ones) usually present unimodal distributions, thus providing a basic method to distinguish which type of content we may find in a certain population, with reasonable accuracy. In addition to that, the evolution of the distribution of length of articles over time showed that the KDE curves tend to become smoother as time elapses, as a consequence of the natural distribution of different amount of content in each article. The most notable exceptions are the German and Swedish Wikipedias, presenting a smooth curve for all years of their activity history (even the earliest ones). Finally, we remembered that the length of articles presents positive correlation with the number of different authors that contributed to each article, though the relationship is not very tight, since the longest articles in all versions usually present a low number of distinct logged authors.

3. **Q3: How does the coordination among authors in the top-ten Wikipedias evolve over time?**: The monthly number of revisions by logged authors on talk pages has reached also a stabilized trend from 2007, and the same occurred for the number of active logged authors contributing to them (with some decreasing tendencies starting to appear in some versions like the English Wikipedia). However, it is interesting to notice that the number of active talk pages has continued to grow in all versions. In the case of the French Wikipedia, it suffered a leverage effect whose stepper slope let it reach the same activity level found in the German Wikipedia, despite having half the total number of articles than the German version. The analysis of the ratio of talk pages per article is even more interesting. For instance, the Japanese and Dutch Wikipedias exhibit a very low percentage of talk pages per article, revealing a remarkable lack of interest in these communities of authors about discussion on articles content. The Polish Wikipedia is an alarming extreme case, with a number of talk pages one order of magnitude lower than the total number of articles, evidencing the "artificial" method of producing content based on bots work, without considering discussion pages except for fewer number of articles. On the other side, the French and specially the English Wikipedia revealed an extraordinary interest in content discussion. Specially remarkable is the 80.8% found for the English version, which is much larger than any other Wikipedia in the top ten list, and despite this, supports a discussion page for most of its more than 2 million articles. Finally, the same smoothing effect found for the evolution of the length of articles is also identified in the evolution of the length of talk pages over time for all language versions.

4. **Q4: Which are the key parameters defining the social structure and stratification of Wikipedia authors?**: First and foremost, we have identified that the best fit distributions obtained for key productivity parameters affecting authors (like number of revisions per author and number of different articles per author) follow and upper Pareto distribution, suggesting

a scale-free network of authors that shapes the distribution of effort, possibly by means of a preferential attachment process (defined by the distribution of links in pages found in previous research work [106]). An examination of the parameters related to articles reveal that, at present time, the distribution of effort among articles follows a lognormal shape, thus deviating from earlier phases during its evolution, in which they were reported to follow power law patterns [18]. We also confirm that the inequality level of contributions from logged authors is strongly biased towards a small group of very active contributors in each version, which we already named as the *core* of each community. Likewise, we also demonstrate that the inequality level of the distribution of contributions among articles in all versions is also biased (though in a less intense way) towards a group of very popular articles. As well, in subsequent sections, we also find that the group of top quality articles (FAs) in each version belongs to this group of highly popular entries.

5. **Q5: What is the average lifetime of Wikipedia volunteer authors in the project?**: The main conclusion we can infer from our survival analysis performed on the community of authors in the top ten Wikipedias is that there is an extraordinary high mortality rate in all languages. Actually, we show that the monthly number of deaths of logged authors in the top ten language versions surpassed the monthly number of new logged authors coming to contribute for the first time in a certain version. Therefore, the higher mortality rate, since the beginning of 2007, offers a possible explanation for the steady-state reached by the monthly number of contributions and monthly number of active pages in all versions during the same period. A significant proportion of authors (more than 50% in all versions) abandons the project after more than 200 days. Moreover, reaching the core group of very active authors does not ensures that those authors will exhibit better survivability since, in fact, more than 50% of them abandon that core of very active authors after less than 100 days (less than 30 in the case of the Portuguese and English Wikipedias). Complementing this findings, the application of the Cox proportional hazards model let us demonstrate that the participation of logged authors in FAs or talk pages has a significant positive impact to enhance the survivability of such contributors, being the contribution to both key types of pages the one presenting the higher enhancement effect over the average lifetime of authors.

6. **Q6: Can we identify basic quantitative metrics to describe the reputation of Wikipedia authors and the quality of Wikipedia articles?**: We explored very basic metrics to quantify the common patterns found in high quality articles in Wikipedia, those belonging to the group of FAs validated for the respective community of contributors. We found that FAs in Wikipedia consistently present a higher number of different authors and revisions received than average articles in a certain version. In the same way, we also demonstrate that FAs are significantly older than average articles, showing that the refinement process to enhance the content of articles in any of the encyclopaedias actually takes a long time (more than 1,000 days in all cases). Since we have found that the population of authors is not growing any more since 2007, this shift in the demographic patterns of the community of authors may have a direct impact over the ability of the project to produce high quality content, since a stable population of authors is not capable of increasing the number of articles to be reviewed. Finally, as a proof of concept we validate the metrics proposed in an earlier research work by Stein and Hess [108]. We find that those metrics present significant differences in favor of FAs, something that opens the door to the possible application of those metrics, together with some other alternative methods to measure the reputation of Wikipedia authors and articles content [6]. Further research work

must be conducted to explore the adequate combination of parameters that may lead to forecast which articles have the highest potential to become FAs in due course.

7. **Q7: Is it possible to infer, based on previous history data, any sustainability conditions affecting the top-ten Wikipedias in due course?**: As a main conclusion, looking at the evolution of the key parameters already identified as relevant to explain the progress in time of the top ten Wikipedias and their communities, we find that those statistics describing the activity of logged authors tend to follow Pareto-like distributions that become, in general, more and more log-linear as time elapses. On the other hand, metrics describing articles has progressively lost the old Pareto-like shape for their distribution, reaching a lognormal shape during 2007 (probably, as a result of the stabilization of the number of logged authors in all versions, as well). The analysis of the evolution in time of contributions from the core of very active authors identified in each month of history of a certain language version, reveals that former core authors does not provide a comparable amount of effort to the level offered by new, even more active members of the core. Nevertheless, again the evolution parameters point out a somewhat delicate situation, since the monthly inequality level of the contributions from logged still maintains the same values as in previous years. Thus, this indicates that either the inequality of the distribution of revisions maintains the present level (in which case the authors would not be able to address so many articles than in previous years) or else, that the inequality level of this distribution will continue to grow, until core authors begin to find their natural limit in the maximum number of revisions performed and number of different articles reviewed.

As far as we know, this is the first quantitative analysis comparing several language editions of the Wikipedia at the same time, and more precisely, the 10 largest ones. WikiXRay, our tool for automating this analyses, has proven to be effective enough to undertake this endeavor, offering an unparalleled opportunity for other researchers around the world to follow on their own research lines avoiding initial problems preventing them to undertake this studies, due to the complexity of the pre-processing stage after retrieving the corresponding data base dump file. To conclude, our model has proven to be the first one serious attempt to understand and compare some of the largest (if not the largest) communities of volunteers in the Internet, participating in an open content creation process.

5.1.2 Sustainability conditions

The main conclusion that we can infer from the overall results of our quantitative analysis is that there exists a severe risk in the top-ten language versions of Wikipedia, about maintaining 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.

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 revisors 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 encyclopaedias, namely Britannica [44], 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 undoubtly 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, 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.

5.2 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.

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 revisors 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 encyclopaedias, 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 undoubtly 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 creational 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.

5.3 Future research work

All along this thesis, we have presented a number of statistical techniques and different approaches that pave the way for further research works interested in extending and complementing their results. A very valuable contribution in this sense is the set of scripts included in our tool, WikiXRay, aimed to automate quantitative analyses on any language version of Wikipedia. This software may serve as a useful starting point for other researchers who want to implement empirical studies on Wikipedia, since it has already solved some annoying preliminary operations, like retrieving the complete information from database dumps, as well as parsing that files to extract and compute appropriate descriptive information.

Speaking about concrete research lines that we have left still open to further exploration, one of the most promising fields for future research is, without a doubt, analyzing the concrete factors influencing the creation of quality content in Wikipedia articles. As we have seen in section 4.5, some basic metrics can be applied to quantify the reputation of Wikipedia authors based on the quality of content produced by them, and the quality of certain articles standing out from the average corpus of encyclopaedic terms in each language version. Nevertheless, forecasting the evolution of the quality level of content in Wikipedia, and identifying which articles still not included in the group of Featured

Articles have the highest potential to be nominees for joining this selected set of high quality content is still an open question. Most probably, this metrics might have to be refactored or extended to take into account any other factors that may have direct impact in the final quality of Wikipedia content. In spite of this, according to recent research works in this area [3] this will not be an easy task, since we can identify a somewhat complex taxonomy of possible editing strategies from Wikipedia authors that ultimately affect the final quality level of content. In the same way, the complementary information provided by discussion activity in talk pages may also provide additional insights and indicators to help us in the process of identifying potentially new Featured Articles.

Another important aspect that should be taken into account for further research works is looking at the future trends exhibited by key activity parameters (like active authors per month, number of revisions per month, active article/talk pages, and so forth). Very recent reports on this topic has revealed that, within some specific groups like the one corresponding to very active authors, previous stabilized trends has started to shift to decreasing curves, at least some of the language versions for which we have new database dumps available to be analyzed [1]. Analyzing the future trend followed by activity patterns in the Wikipedia community of authors will be crucial to provide adequate assessment for reconducting the content creation process in the project so that the production of quality content and the coverage of encyclopaedic entries do not suffer irreparable damages over the next months. Initiatives like the new project sponsored by Stanton Foundation (USA) to radically improve the usability of MediaWiki interfaces will definitely contribute to mitigate the current negative trend in the top ten language versions, which are losing authors more rapidly than they can attract new volunteers.

Likewise, the evaluation of new emerging tools that facilitate the dynamic visualization of the evolution in time of large data sets, like the ones provided in Wikipedia, definitely worths to be studied in further detail. Inline visual applications like WikiTrust [2] or Wikidashboard [3] offer additional information that may help Wikipedia editors to undertake their tasks. WikiTrust may act as a powerful add-on displaying a graphical, easy-to-read estimation of the level of trustworthiness of articles content, based on the pre-computed reputation metric for the corresponding authors. Wikidashboard is aimed to summarize the evolution in time of the revision activity that took place so far for a certain article, and could also serve as a great inline tool in certain cases in which authors will benefit from rapidly identifying articles in which revision activity has suddenly boost up. On the other side, as far as the off-line statistical analysis of Wikipedia data is concerned, the adoption of state-of-the-art visualization tools like Processing [4], which has already been covered in detail by relevant experts in this field [95], [43], [104] can open the door to completely new ways of interpreting the current state and evolution over time of the Wikipedia project and its associated community of authors.

Finally, further research should be conducted on the analysis of the underlying preferential attachment process that seems to be driving the behavior of Wikipedia authors to create content. A thorough examination of the whole web graph conformed by internal links in each language version will provide definitive conclusions in this respect. On the other hand, the examination of the properties of the web graph conformed by external links included in Wikipedia articles may also provide interesting information complementing other metrics for the quality level of articles.

[1] http://infodisiac.com/blog/2009/01/wikistats-is-back-again/

[2] http://trust.cse.ucsc.edu/

[3] http://wikidashboard.parc.com/

[4]

Bibliography

[1] Wiki research bibliography. consulted on April 13th, 2008.

[2] Wikibibliographie encyclen. consulted on April 13th, 2008.

[3] Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001, 2008.

[4] Inmaculada B. Aban, Mark M. Meerschaert, and Anna K. Panorska. Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, 101(473):270–277, March 2006.

[5] Thomas B. Adler, K. Chatterjee, Luca de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to wikipedia content. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, 2008. ACM Press.

[6] Thomas B. Adler and Luca de Alfaro. 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, 2007. ACM Press.

[7] Thomas B. Adler, Luca de Alfaro, I. Pye, and Raman V. Measuring author contributions to the wikipedia. In *Proceedings of the 4th International Symposium on Wikis*, New York, NY, USA, 2008. ACM Press.

[8] Rodrigo Almeida, Barzan Mozafari, and Junghoo Cho. On the evolution of wikipedia. In *International Conference on Weblogs and Social Media*, Boulder, Colorado, USA, March 2007.

[9] Denise Anthony, Sean Smith, and Tim Williamson. Explaining quality in internet collective goods: Zealots and good samaritans in the case of wikipedia. Electronically.

[10] L. Aronsson. Operation of a large scale general purpose wiki website. Berlin, 2002. Verlag für Wissenschaft und Forschung.

[11] A.B. Atkinson and F. Bourguignon, editors. *Handbook of Income Distribution*, volume 1 of *Handbook of Income Distribution*. Elsevier, 2000.

[12] David Aumüller and Sören Auer. Towards a semantic wiki experience - desktop integration and interactivity in wiksar, 2005.

[13] Angela Beesley. How and why wikipedia works. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 1–2, New York, NY, USA, 2006. ACM Press.

[14] F. Bellomi and R. Bonato. Network analysis for wikipedia. In *Wikimania 2005*. Wikimedia Foundation, 2005.

[15] Yochai Benkler. *The Wealth of Networks : How Social Production Transforms Markets and Freedom*. Yale University Press, May 2006.

[16] Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *GROUP '05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10, New York, NY, USA, 2005. ACM Press.

[17] Michel Buffa, Peter Sander, and Jean-Claude Grattarola. Distant cooperative software development for research and education: three years of experience. In *Proceedings of the International Conference on Computer Aided Learning in Engineering Education (CALIE 04)*, 2004.

[18] Luciana S. Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. Temporal analysis of the wikigraph. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 45–51, Washington, DC, USA, 2006. IEEE Computer Society.

[19] B. Butler, L. Sproull, S. Kiesler, and R. Kraut. Community effort in online groups: Who does the work and why, 2001.

[20] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: the case of wikipedia, Feb 2006.

[21] Carlos Castillo-Salgado et al. Measuring health inequalities: Gini coefficient and concentration index. *Epidemiological Bulletin*, 22(1), March 2001.

[22] Magnus Cedergren. Open content and value creation. *First Monday*, 8(8), August 2003.

[23] Thomas Chesney. An empirical examination of wikipedia's credibility. *First Monday*, 11(11), November 2006.

[24] Andrea Ciffolilli. Phantom authority, self–selective recruitment and retention of members in virtual communities: The case of wikipedia. *First Monday*, 8(12), December 2003.

[25] Aaron Clauset, Cosma R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, Jun 2007.

[26] Dianne Cook and Deborah F. Swayne. *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi (Use R)*. Springer, 1 edition, December 2007.

[27] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1037–1046, New York, NY, USA, 2006. ACM Press.

[28] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[29] Michael J. Crawley. *The R Book*. Wiley, June 2007.

[30] Peter Dalgaard. *Introductory Statistics with R (Statistics and Computing)*, chapter 14. Springer, 2nd edition, August 2008.

[31] Peter Denning, Jim Horning, David Parnas, and Lauren Weinstein. Wikipedia risks. *Commun. ACM*, 48(12):152–152, December 2005.

[32] Ludovic Denoyer and Patrick Gallinari. The wikipedia xml corpus. *SIGIR Forum*, 40(1):64–69, June 2006.

[33] Alain Désilets, Lucas Gonzalez, Sébastien Paquet, and Marta Stojanovic. Translation the wiki way. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 19–32, New York, NY, USA, 2006. ACM Press.

[34] Chris Dibona, Mark Stone, and Danese Cooper. *Open Sources 2.0 : The Continuing Evolution*. O'Reilly Media, Inc., October 2005.

[35] Pierpaolo Dondio, Stephen Barrett, Stefan Weber, and Jean Seigneur. Extracting trust from domain analysis: A case study on the wikipedia project. pages 362–373. 2006.

[36] Robert Dorfman. A formula for the gini coefficient. *The Review of Economics and Statistics*, (61):146–149, March 1979.

[37] Paul Dourish and Victoria Bellotti. Awareness and coordination in shared workspaces. In *CSCW '92: Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 107–114, New York, NY, USA, 1992. ACM Press.

[38] Anja Ebersbach and Markus Glaser. Towards emancipatory use of a medium: The wiki. *International Journal of Information Ethics*, 11, 2004.

[39] Anja Ebersbach, Markus Glaser, and Richard Heigl. *Wiki : Web Collaboration*. Springer, November 2005.

[40] W. Emigh and S. C. Herring. Collaborative authoring on the web: A genre analysis of online encyclopedias. pages 99a–99a, 2005.

[41] Andrea Forte and Amy Bruckman. From wikipedia to the classroom: exploring online publication and learning. In *ICLS '06: Proceedings of the 7th international conference on Learning sciences*, pages 182–188. International Society of the Learning Sciences, 2006.

[42] V. Franco, R. Piirto, H. Y. Hu, B. V. Lewenstein, R. Underwood, and N. K. Vidal. Anatomy of a flame: conflict and community building on the internet. *Technology and Society Magazine, IEEE*, 14(2):12–21, 1995.

[43] Ben Fry. *Visualizing Data*. O'Reilly Media, Inc., December 2007.

[44] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[45] Dan Gillmor. *We the Media*. O'Reilly, August 2004.

[46] Conrando Gini. On the measure of concentration with especial reference to income and wealth. In *Cowless Comission*, 1936.

[47] Michel L. Goldstein, Steven A. Morris, and Gary G. Yen. Fitting to the power-law distribution. Aug 2004.

[48] Michel L. Goldstein, Steven A. Morris, and Gary G. Yen. Problems with fitting to the power-law distribution, August 2004.

[49] Todd Holloway, Miran Bozicevic, and Katy Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors, Dec 2005.

[50] David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis: Regression Modeling of Time to Event Data (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2 edition, March 2008.

[51] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W. Lauw, and Ba-Quy Vuong. Measuring article quality in wikipedia: models and evaluation. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 243–252, New York, NY, USA, 2007. ACM.

[52] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W. Lauw, and Ba-Quy Vuong. On improving wikipedia search using article quality. In *WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 145–152, New York, NY, USA, 2007. ACM.

[53] Aniket Kittur, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proceedings of the 25th Annual ACM Conference on Human Factors in Computing Systems (CHI 2007)*. ACM, April 2007.

[54] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in wikipedia. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, New York, NY, USA, 2007. ACM Press.

[55] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text (Statistics for Biology and Health)*. Springer, 2nd edition, August 2005.

[56] Josef Kolbitsch and Hermann Maurer. Community building around encyclopaedic knowledge. *Journal of Computing and Information Technology*, 13, 2005.

[57] Korfiatis, Nikolaos, Poulos, Marios, Bokos, and George. Evaluating authoritative sources using social networks: an insight from wikipedia. *Online Information Review*, 30(3):252–262, May 2006.

[58] Aaron Krowne and Anil Bazaz. Authority models for collaborative authoring. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*, 2004.

[59] Stacey Kuznetsov. Motivations of contributors to wikipedia. *SIGCAS Comput. Soc.*, 36(2), June 2006.

[60] Bo Leuf and Ward Cunningham. *The Wiki Way: Collaboration and Sharing on the Internet*. Addison-Wesley Professional, April 2001.

[61] Andrew Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *5th International Symposium on Online Journalism*. University of Texas at Austin, April 2004.

[62] Lipczynska and Sonya. Power to the people: the case for wikipedia. *Reference Reviews incorporating ASLIB Book Guide*, 19(2):6–7, February 2005.

[63] Cathy Ma. The social, cultural, economical implications of the wikipedia. *Computers and Writing Online 2005*.

[64] John Maindonald and John Braun. *Data Analysis and Graphics Using R: An Example-based Approach (Cambridge Series in Statistical and Probabilistic Mathematics)*, chapter 8, pages 275–284. Cambridge University Press, December 2006.

[65] Deborah L. Mcguinness, Honglei Zeng, Paulo P. da Silva, Li Ding, Dhyanesh Narayanan, and Mayukh Bhaowal. Investigations into trust for collaborative information repositories: A wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web*, 2006.

[66] Nora Miller. Wikipedia and the disappearing "author". *ETC: A Review of General Semantics*, 62(1):37–40, January 2005.

[67] Jeffrey A. Mills et al. Statistical inference via bootstrapping for measures of inequality. *Epidemiological Bulletin*, 22(12), March/April 1997.

[68] Domas Mituzas. Wikipedia: Site internals, configuration, code examples and management issues. Technical report, April 2007.

[69] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions.

[70] Joseph C. Morris. Distriwiki:: a distributed peer-to-peer wiki network. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 69–74, New York, NY, USA, 2007. ACM.

[71] Claudia Müller, Benedikt Meuthrath, and Anne Baumgrass. Analyzing wiki-based networks to improve knowledge processes in organizations. volume 14, pages 526–545. February 2008.

[72] Christine M. Neuwirth, David S. Kaufer, Ravinder Chandhok, and James H. Morris. Issues in the design of computer support for co-authoring and commenting. In *CSCW '90: Proceedings of the 1990 ACM conference on Computer-supported cooperative work*, pages 183–195, New York, NY, USA, 1990. ACM Press.

[73] G. B. Newby, J. Greenberg, and P. Jones. Open source software development and lotkas law: Bibliometric patterns in programming. *JASIST (Journal of the American Society for Information Science and Technology)*, 54(2):1169–1178, 2003.

[74] M. E. J. Newman. Power laws, pareto distributions and zipf's law, December 2004.

[75] Finn A. Nielsen. Scientific citations in wikipedia, May 2007.

[76] S. Noël and J. M. Robert. How the web is used to support collaborative writing. *Behaviour and Information Technology*, 22(4):245–262, July 2003.

[77] Oded Nov. What motivates wikipedians? *Commun. ACM*, 50(11):60–64, November 2007.

[78] Yann Ollivier and Pierre Senellart. Finding related pages using green measures: An illustration with wikipedia. In *Conference on Artificial Intelligence (AAAI 2007)*. Association for the Advancement of Artificial Intelligence, 2007.

[79] Tim O'Reilly. O'reilly – what is web 2.0, 2005.

[80] S original by Kenneth Hess and R port by R. Gentleman. *muhaz: Hazard Function Estimation in Survival Analysis*. R package version 1.2.3.

[81] Felipe Ortega and Kevin Crowston. Introduction to open movements: Floss, open content and open communities minitrack. In *Proceedings of the 42nd Hawaiian International Conference on System Sciences (HICSS 2009)*, January 2009.

[82] Felipe Ortega and Jesus M. Gonzalez-Barahona. Quantitative analysis of the wikipedia community of users. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 75–86, New York, NY, USA, 2007. ACM.

[83] Felipe Ortega and Jesus M. Gonzalez-Barahona. On the inequality of contributions to wikipedia. In *Proceedings of the 41st Hawaiian International Conference on System Sciences (HICSS 2008)*, January 2008.

[84] Felipe Ortega, Jesus M. Gonzalez-Barahona, and Gregorio Robles. The top ten wikipedias: A quantitative analysis using wikixray. In *Proceedings of the 2nd International Conference on Software and Data Technologies (ICSOFT 2007)*. INSTICC, Springer-Verlag, July 2007.

[85] Sébastien Paquet. Seb's "wikis and knowledge sharing" survey: Results, 2003.

[86] Reid Priedhorsky, Jilin Chen, Shyong, Kathering Panciera, Loren Terveen, and John. Creating, destroying, and restoring value in wikipedia. November 2007.

[87] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[88] Sheizaf Rafaeli, Yaron Ariel, and Tsah Hayat. Wikipedians sense of (virtual) community. In *eighth International Conference General Online Research (GOR06)*, Bielefeld, Germany, 2006.

[89] Sheizaf Rafaeli, Tsahi Hayat, and Yaron Ariel. Wikipedia participants and "ba": Knowledge building and motivations. In *Cyberculture 3rd Global Conference. Prague, Czech Republic*.

[90] Sheizaf Rafaeli, Tsahi Hayat, and Yaron Ariel. Wikipedians' sense of community, motivations, and knowledge building. In *Proceedings of Wikimania 2005 - The First International Wikimedia Conference*, Frankfurt, Germany.

[91] Eric S. Raymond. *The Cathedral & the Bazaar*. O'Reilly, January 2001.

[92] Bendel R.B., Higgins S.S., Teberg J.E., and Pyke D.A. Comparison of skewness coefficient, coefficient of variation, and gini coefficient as inequality measures within populations. *Oecologia*, 78(3):394–400, Mar. 2004.

[93] Joseph M. Reagle. A case of mutual aid: Wikipedia, politeness, and perspective taking. In *Proceedings of Wikimania 2005—The First International Wikimedia Conference*, Frankfurt, Germany.

[94] Joseph M. Reagle. Do as i do: authorial leadership in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 143–156, New York, NY, USA, 2007. ACM.

[95] Casey Reas and Ben Fry. *Processing: A Programming Handbook for Visual Designers and Artists*. The MIT Press, September 2007.

[96] Dirk Riehle. How and why wikipedia works: an interview with angela beesley, elisabeth bauer, and kizu naoko. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, pages 3–8, New York, NY, USA, 2006. ACM Press.

[97] Gregorio Robles. *Software Engineering Research on Libre Software: Data Sources, Methodologies and Results*. PhD in Computer Science, Escuela Superior de Ciencias Experimentales y Tencologia, Universidad Rey Juan Carlos, 2006.

[98] Gregorio Robles and Jesus Gonzalez-Barahona. Contributor turnover in libre software projects. pages 273–286. 2006.

[99] Deepayan Sarkar. *lattice: Lattice Graphics*, 2008. R package version 0.17-15.

[100] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R (Use R)*. Springer, March 2008.

[101] Christoph Sauer, Chuck Smith, and Tomas Benz. Wikicreole: a common wiki markup. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 131–142, New York, NY, USA, 2007. ACM.

[102] Mareike Schoop, Aldo de Moor, and Jan L. G. Dietz. The pragmatic web: a manifesto. *Commun. ACM*, 49(5):75–76, May 2006.

[103] Steve Selvin. *Survival Analysis for Epidemiologic and Medical Research (Practical Guides to Biostatistics and Epidemiology)*. Cambridge University Press, 1 edition, March 2008.

[104] Daniel Shiffman. *Learning Processing: A Beginner's Guide to Programming Images, Animation, and Interaction (Morgan Kaufmann Series in Computer Graphics)*. Morgan Kaufmann, illustrated edition edition, August 2008.

[105] Sander Spek, Eric Postma, and Jaap Herik. Wikipedia: organisation from a bottom-up approach, Nov 2006.

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

[107] A. Spoerri. What is popular on wikipedia and why? *First Monday*, 12, April 2007.

[108] Klaus Stein and Claudia Hess. 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, 2007. ACM.

[109] B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith. Information quality in a community-based encyclopedia. In S. Hawamdeh, editor, *Knowledge Management: Nurturing Culture, Innovation, and Technology - Proceedings of the 2005 International Conference on Knowledge Management*, pages 101–113, Charlotte, NC, 2005. World Scientific Publishing Company.

[110] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality - ICIQ 2005*, pages 442–454, 2005.

[111] Besiki Stvilia and Les Gasser. An activity theoretic model for information quality change. *First Monday*, 13(8), April 2008.

[112] Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. A framework for information quality assessment. *J. Am. Soc. Inf. Sci. Technol.*, 58(12):1720–1733, October 2007.

[113] Besiki Stvilia, Michael B. Twidale, and Linda C. Smith. Information quality: Discussions in wikipedia. 2005.

[114] James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, May 2004.

[115] Aaron Swartz. Who writes wikipedia, September 2006.

[116] Terry Therneau and Thomas Lumley. *survival: Survival analysis, including penalised likelihood.*, 2008. R package version 2.34-1.

[117] Vinod Thomas, Yan Wang, and Xibo Fan. Measuring education inequality - gini coefficients of education. Policy Research Working Paper Series 2525, The World Bank, January 2001.

[118] Guido Urdaneta, Guillaume Pierre, and Maarten van Steen. A decentralized wiki enginge for collaborative wikipedia hosting. In *Proceedings of the 3rd International Conference on Web Information Systems and Technologies*, pages 156–163, March 2007.

[119] Guido Urdaneta, Guillaume Pierre, and Maarten van Steen. Wikipedia workload analysis. September 2007.

[120] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*, chapter 13. Springer, September 2003.

[121] Fernanda B. Viégas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in wikipedia. pages 78–78, 2007.

[122] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with ¡i¿history flow¡/i¿ visualizations. pages 575–582, 2004.

[123] Fernanda B. Viégas, Martin Wattenberg, and Matthew Mckeon. The hidden order of wikipedia. pages 445–454. 2007.

[124] Jakob Voss. Measuring wikipedia. In *International Conference of the International Society for Scientometrics and Informetrics : 10th*. ISSI, July 2005.

[125] Christian Wagner. Wiki: A technology for conversational knowledge management and group collaboration. *Communications of the AIS*, 13:256–289, 2004.

[126] Christian Wagner. Breaking the knowledge acquisition bottleneck through conversational knowledge management. *Information Resources Management Journa*, 2005.

[127] Christian Wagner and Narasimha Bolloju. Supporting knowledge management in organizations with conversational technologies: Discussion forums, weblogs, and wikis. *Journal of Database Management*, 16(2):i–viii, 2005.

[128] A Wagstaff et al. On the measurements of inequalities in health. *Soc. Sci. Med.*, 33(5):545–577, 1991.

[129] Jimmy Wales. Wikipedia in the free culture revolution. In *OOPSLA '05: Companion to the 20th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*, pages 5–5, New York, NY, USA, 2005. ACM.

[130] Neil L. Waters. Why you can't cite wikipedia in my class. *Commun. ACM*, 50(9):15–17, September 2007.

[131] Martin Wattenberg, Fernanda Viégas, and Katherine Hollenbach. Visualizing activity on wikipedia with chromograms. pages 272–287. 2007.

[132] Dennis M. Wilkinson and Bernardo A. Huberman. Assessing the value of coooperation in wikipedia, April 2007.

[133] Dennis M. Wilkinson and Bernardo A. Huberman. Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 157–164, New York, NY, USA, 2007. ACM.

[134] Harris Wu, Mohammad Zubair, and Kurt Maly. Harvesting social knowledge from folksonomies. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 111–114, New York, NY, USA, 2006. ACM Press.

[135] Wenpeng Xiao, Changyan Chi, and Min Yang. On-line collaborative software development via wiki. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 177–183, New York, NY, USA, 2007. ACM.

[136] Achim Zeileis. *ineq: Measuring inequality, concentration and poverty*, 2007. R package version 0.2-8.

[137] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. Mcguinness. Computing trust from revision history. November 2006.

[138] Torsten Zesch and Iryna Gurevych. Analysis of the wikipedia category graph for nlp applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT)*, 2007.

[139] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(1), 2006.

Appendix A

Glossary

7zip : A libre software compression program, originately created for Microsoft Windows platforms, which is now also available for other operating systems like GNU/Linux. It is based on a new compression algorithm, LZMA, developed by Igor Pavlov the creator of this program. Its main features are a very high compression rate (specially for plain text files) and superior performance of its compression algorithm, which can be parallelized on multiple threads.

Age of articles : In the study of the quality of content in Wikipedia, the age of a certain article in a language version is the difference (in days) between the date on which the article was edited for the last time, and the date on which the article was edited for the first time, in that language version.

Age of authors : In the study of the reputation of Wikipedia authors, the age of an author is the difference (in days) between the timestamp of the last revision and the timestamp of the first revision performed by that author in that language version. This is an equivalent definition of that found for the lifetime of authors, in survival analysis.

Anonymous author : Any author who does not create a user account in a certain language edition of Wikipedia, and performs revisions under anonymous identity. Anonymous authors are identified in the database by a common user identifier and the IP address from which the author contacted the system.

Article : Wiki page containing information of encyclopedic articles in a certain language edition of Wikipedia. Articles are stored under the `main` namespace in the MediaWiki database of the corresponding language version.

Author : An individual who belongs to the community of a certain language edition in Wikipedia, and who performed at least one revision in that language edition. Otherwise, it will not appear in the corresponding table of the database, registering each revision performed in the system. For privacy reasons, access to the database table containing the full user list in each language version, along with sensible information like email accounts and so forth, is not allowed. An author is identified by a numeric ID in the database, associated to every revision attributed to her.

Birth : In the survival analysis of the Wikipedia community of logged authors, a new birth has occurred in the community of logged authors in a certain language version when a new author contributes with a revision for the first time in the history of that language version.

Bot : Small software programmes that performs revisions on a certain language edition of Wikipedia in an automated way. Many bots can be uniquely identified due to their special privileged status `'bot'` associated with their correspondent `rev_user` unique identifier. This relationship is reflected in the `user_groups` table in the database. Since bots are not real human users, *they and their revisions are systematically filtered out* in this analysis, unless it is indicated otherwise.

CCDF (Complementary Cumulative Distribution Function) : The opposite of the CDF (and sometimes named using the same term): $1 - F(X) = P(X > x)$

CDF (Cumulative Distribution Function) : A function providing a complete description of the probability distribution of a certain random variable X. It tells us the probability, for each value x, that the random variable X takes a value less than or equal to x: $F(X) = P(X \leq x)$.

Core (of very active authors) : Small group of very active authors providing the majority of the revisions received in a certain language version of Wikipedia.

CPH (Cox Proportional Hazards Model) : A semi-parametric model aimed to estimate the influence of covariates on the hazard function of a certain population under survival analysis. The model assumes that the hazard function can be modeled as a baseline hazard function, on top of which we try to integrate the effects produced by each additional parameter included in the model.

Database dump : Archive containing a record of the data contained in a whole database, or in a single table or a set of tables of a certain database, optionally including a description of the layout of the table/s contained in it. Usually, database dumps are recorded either using SQL syntax or other data representation standards, like CSV (Comma-Separated Values) or XML (Extensible Markup Language). The database dumps for each language version of Wikipedia are offered in SQL format, except for the large file containing the full revision history that is provided in XML format. These dumps are the basic data source from which all quantitative data for this thesis work has been obtained.

Death : In the survival analysis of Wikipedia logged authors, the death of a logged author occurs when this author performed her last revision in the history of a language version, and thus, she never come back again to contribute (as far as the activity registered in the database dump file can show).

Edit-based author reputation : Trustworthiness level of a certain author in a language version of Wikipedia, calculated through the quality level achieved by all revisions performed by that author.in that language version.

Event of interest : An event that may occurred to a subject examined in a survival analysis, which marks the end of the lifetime of that subject in the study. Usually, the event of interest has a negative connotation (like death of a patient, or time until a certain component fails), but generally speaking, any event that can occur to a subject included in the study can be considered as the event of interest

in that analysis. For instance, in our survival analysis of the Wikipedia community of authors, one of the events of interest defined was "time until a certain author reaches the core".

External link : An hyperlink redirecting to a article in a different version of Wikipedia than the one the article belongs to, or alternatively, pointing to any other Internet web page outside the Wikipedia project.

Featured Article (FA) : An article that has deserved to be nominated as one of the top quality articles produced in a certain language edition of Wikipedia. The nomination takes place after an exhaustive reviewing process performed by all interested members of the community, upon an open call issued by the corresponding responsible in that language version. Candidate articles are proposed by community members to enter this reviewing process. The result of a voting process, reflecting the opinions of all reviewers involved, determines whether the article is promoted to this new status or not. The promotion is not permanent, that is, as soon as community members detect that the quality of the article has lowered, they can suggest the article as a candidate for a new reviewing process. In case that the FA does not pass this examination, it can be demoted again to its original non-FA status.

FLOSS (Free, Libre, Open Source Software) : This is a broad term to refer both to *free software* (according to the Free Software Foundation definition) and *open source software* (according to the Open Source Initiative definition).*Libre* is a term well understood in romance languages, such as Spanish, French, Portuguese or Italian, and understandable for speakers of many others, as well. It eliminates the ambiguity of *free* in English (which may also mean *gratuitous*) and is used by some people specially in Europe (although the term is rooted in the early US free software community [1]. In this respect, it is important to notice that although the communities, the motivation and the rationale behind *free* and *open source* are different, the software to which they refer is basically (although not exactly) the same.

Gini coefficient : Numerical value (usually presented in percentage terms) summarizing the inequality level of the distribution of a certain parameter among the individuals of a population under study. In this thesis, it represents the number of revisions accumulated by the lower y% of the authors in a certain version of Wikipedia, or alternatively, the number of revisions received by the lower y% of the total number of articles in a language version.

GNU R : A statistical software package, released under the GPL license, providing a huge set of statistical libraries (accessible on the CRAN website [2]), implementing the most relevant statistical techniques and tools available today. WikiXRay uses the facilities provided by GNU R to implement all the statistical analyses performed on Wikipedia quantitative data.

Hazard function : A curve representing the hazard rate a certain population in a survival analysis for each given point in time. See also *hazard rate*.

Hazard rate : The instantaneous risk of death , at a given point in time, of all individuals in a certain population under survival analysis which are still alive at that point. See also *hazard curve*.

[1] `http://sinetgy.org/jgb/articulos/libre-software-origin/libre-software-origin.html`.
[2] `http://lib.stat.cmu.edu/R/CRAN/`

Internal link : An hyperlink pointing to another article inside the same language version of Wikipedia that the article containing that hyperlink.

Kaplan-Meier estimate : Also known as product-limit estimate, is a method to estimate survival curves from empirical data in survival analysis.

Kernel Density Estimation : Also known as Smooth Density Estimation, it is a non-parametric method to estimate the probability density function of a random variable, from quantitative data extracted in an empirical analysis.

Language version/edition : Each translation of the Wikipedia encyclopaedia, running on independent MediaWiki instances on different web sites. As a result, each version has its own database to store the whole revision history, as well as its own database dumps files.

Lifetime : In the survival analysis of the Wikipedia community of authors, the lifetime of a certain author in a language version is the difference (in days) between the date on which that author performed her last revision and the date on which she made her first revision in that version. See also *age of authors*.

Logged author : Any author registered in a certain language edition of Wikipedia, by creating a user account. Logged authors can be uniquely identified by either their user identifier or their login in the `revision` table of the corresponding MediaWiki database. Therefore, authors must log in the system before performing revisions, to let the database register their identity.

Lognormal distribution : Probability distribution of a random variable whose histogram (in a log-log scale) follows the shape of a normal (or Gaussian) probability distribution.

Lorenz curve : It is a graphical representation of the CDF of a probability distribution for a random variable. In this thesis, it represents the percentage of the total number of revisions assumed by the bottom y% of the authors in a certain Wikipedia community, or else, the total number of revisions received by the bottom y% of the articles in a language version.

Median survival time : The median survival time is the lifetime reached by, at least, 50% of the population under a survival analysis.

MediaWiki : The libre software wiki package, developed in PHP, and originally created to fit the need of supporting the Wikipedia project.

Namespace : Each of the logical areas in which the content of any wiki based on MediaWiki is classified. The database stores, for each wiki page, a numerical identifier indicating which namespace it belongs to.

Page : Any wiki page, disregarding the namespace in which it is stored in the system database, whose information can be edited by a Wikipedia contributor. This includes encyclopedic articles, user pages, discussion pages associated with each article (`talk_pages`), etc. Any page can be ***uniquely*** identified by their corresponding page ID in the database.

Page-based author reputation : Trustworthiness level of a certain author in a language version of Wikipedia, calculated through the quality level achieved by all pages edited by that author.in that language version.

Pareto distribution : A probability distribution of a random variable, characterized by the log-linear behavior of its CDF and CCDF. In other words, if we represent the CCDF of a Pareto distribution in a log-log scale, we will obtain a linear shape with negative slope, starting from a minimum value of the random variable. It is defined by the value acquired by the slope and the minimum threshold value.

Power law : A type of mathematical relationship between two quantitative variables, presenting a log-linear shape in the function depicting the values of the first variable against the other one. It is said that this type of relationships is scale-free (or alternatively, scale-invariant), since a rescaling operation merely shifts the position of the curve, maintaining the shape and the slope unaltered.

Privileged author : Any logged author who received certain special privileges within the system, which are stored in the `user_groups` table of the MediaWiki database of a language version in Wikipedia.

Recentness of authors : It measures, within the study of the reputation of authors, the number of days elapsed between the last revision performed by a certain article in a language version of Wikipedia and the date of the latest revision recorded in the revision history of that language version.

Recentness of articles : In the study of the quality of articles in Wikipedia, it measures the number of days elapsed between the last revision received by a certain article in a language version of Wikipedia, and the date on which the article was created in the system, in that language version.

Redirect : A special type of article, with no content at all, simply providing alternative encyclopedic entries for a certain term.

Reputation (of a Wikipedia author) : Trustworthiness level achieved by a certain author in a language version of Wikipedia, based on the quality level of the whole history of editions performed by that author in that version, or alternatively, calculated using the quality level of the whole set of pages edited by that author along the revision history of that language version.

Restricted mean survival time : If the last observation in a survival analysis is not a death, then the survival function does not reach the zero value at the end, and the mean survival time can not be estimated. To avoid this inconvenient, the restricted mean calculates the mean survival time until the last death registered in the population under analysis, thus ensuring that we will always obtain a finite value.

Revision : Any individual modification on a wiki page in a certain language edition of Wikipedia, that is registered in the database as such and identified by a unique numeric ID.

Revision History : The whole set of revisions performed on a certain language version of Wikipedia throughout its entire history of existence, recorded in the database dump files. We have to point out that this does not include information from deleted wiki pages, which is filtered out from the database dumps.

Stub : An article considered so short as to be considered as a useful encyclopedic article. Stubs are usually new articles recently opened, providing a seed upon which to create a longer, more complete encyclopedic entry. They are usually marked as such using special *templates*, customized in each language edition (sometimes, even for distinct categories in each language version). There is no official policy in Wikipedia regarding the minimum length an article must attain to avoid being considered a stub.

Survival Analysis :Statistical methodology that allows us to build empirical models for data analyses in which the variable of interest can be formulated in terms of *time until an event occurs*.

Survival rate : The percentage of subjects in population examined in a survival analysis that are still alive at a given point in time. See also *survival curve*.

Survival function : A curve depicting the survival rate of a certain population in a survival analysis for each given point in time. See also *survival rate*.

Talk page : Wiki pages containing discussion about the contents of an encyclopedic article in a certain language version. Each talk page is presented next to its corresponding article in the MediaWiki interface. However, newly created articles does not automatically come with a talk page, so it may or may not exist (until a user decides to create it). They are stored under the `talk_page` namespace, with the exception of talk pages associated to *user pages*, which are stored under the `user_talk_page` namespace.

Upper truncated Pareto distribution : A Pareto probability distribution whose CCDF curve suddenly gets to zero (whit a much stepper negative slope), until it reaches a maximum upper value. This one, the minimum threshold value and the slope are the 3 descriptive parameters to define this distribution.

User page : Wiki pages presenting information of a *logged author* in a certain language version. They are stored under the special `user_page` namespace in MediaWiki.

WikiXRay : A libre software Python tool, created to automate the quantitative analysis of the information contained in the database dumps from any language version of Wikipedia. It is the tool used to produce all statistical results and graphs included in this thesis.

Appendix B

Probability distributions

In this appendix, we provide a very basic introduction that should be sufficient for the average reader to get the background needed to understand the statistical analyses undertaken in section 4.3 of this thesis. Our study of the social structure of the Wikipedia community of authors has revealed clear indications that most of the key descriptive parameters found followed Pareto-like probability distributions. Still some other descriptive parameters, focused on per article metrics, have been fitted to lognormal probability distributions. In the following sections, we introduce the application of Pareto distributions in our study, describing the most important properties of the specific subtypes that we have found in our empirical data. Finally, we also provide an introduction to the lognormal distribution and its peculiar properties affecting the analyses performed in this thesis.

B.1 Power laws and Pareto distributions

In this section, we provide a brief introduction to the features and properties of power laws and Pareto distributions, focusing on their application for this thesis work. Readers interested in expanded coverage on this topic are referred to [25], [74] and [69] for a more in-depth treatment of this matter.

Some empirical results obtained in scientific experiments yield measurements varying over a large range of values, typically of several orders of magnitude. We can say that a set of values follows a **power law** if it is drawn from a distribution with a probability function given by:

$$p(x) = Cx^{-\alpha} = e^{c}x^{-\alpha} \leftrightarrow ln[p(x)] = -\alpha \cdot ln(x) + c \tag{B.1}$$

As we can see, the most notable property of power laws is that they have a distinctive linear pattern in a log-log scale graph. The accurate identification of power law distributions in empirical data is a non-trivial task that should be approached with extreme care. The authors of the bibliographic references cited above unanimously state that the safest method to fit power law patterns to empirical data is by using *complementary cumulative distribution functions* (CCDFs) [1]. Instead of plotting a

[1] We have found different information sources using distinct notation to refer to such functions. The majority of authors simply use the term *cumulative distribution function* to refer to the function giving the probability of getting a value of **X** less than or equal to a certain threshold value x (usually represented as $F(x)$ in most textbooks); or instead, to refer to the opposite of this function $1 - F(x)$, following the definition presented in the paragraph above. Nonetheless, other authors reserve the term *complementary* cumulative distribution function to differentiate the second function from the first one. This second approach is the one we have followed in this thesis work, since we felt that for most readers who are not familiar

standard histogram curve, we calculate the probability of obtaining a value of the random variable **X** which is greater than or equal to a specific value x of such random variable:

$$P(x) = \int_x^\infty p(x')dx' \tag{B.2}$$

Thus, this is a function whose slope must be strictly negative (or zero, but never positive), starting at value 1 for $x = 0$ and ending at value 0 for $x = \infty$. If we introduce the formula of the power law, given in equation B.1, inside equation B.2 we obtain:

$$P(x) = \frac{C}{\alpha - 1} x^{-(\alpha-1)} \tag{B.3}$$

Which leads to the relationship between the power law exponent α for our quantity expressed in natural units and the exponent in the CCDF plot, which is precisely $\alpha - 1$. Equation B.3 give us the CCDF for a Pareto probability distribution, which is sometimes also referred to as Zipf's law depending on the way we represent the variables in both coordinated axes [2].

When the problem comes to fit the value of the α coefficient to our data, we must also take into account that most real measurements only follows a Pareto distribution from a certain lower bound onwards. This lower limit must be also taken into account when we report our fits, since the wrong election of this threshold may lead to a completely misleading result. According to Clauset *et al.* in [25], the best way to proceed here is to minimize the maximum distance from the fitted curve to the empirical curve obtained from experimental data. Following the same notation employed by these authors, we have called such distance D in this document, and we always provided it in our tables whenever we had to fit a Pareto distribution to our empirical data, thus offering the adequate assessment information to judge the goodness of our fitting procedures.

We also have to remark here some important properties exhibited by the results obtained in this thesis. According to [74], whenever the fitted coefficient $\alpha < 2$ we can state that the power law distribution has not finite mean. The precise meaning of this property is that, although we can find the finite mean for a real data set drawn from this distribution, this is only due to the fact that the sample has a finite size (like the samples employed in this thesis). However, as the size of the sample grows up, so will do the value of the mean, leveraged by the largest samples drawn from the distribution. Thus the value of the mean will continue to grow without limit. In the same way, it would me useless to offer the estimated variance or the standard deviation for this distribution, since in those cases in which $\alpha\ leq 3$, the second order moment (the mean square) also diverges.

However, for some parameters concerning per author measurements (number of different articles edited per author, number of revisions performed by each author, etc.) we have found a slightly different version of this distribution. In these cases, the power law pattern is maintained until a certain upper value in the CCDF curve, then the function drops off suddenly with a much stepper negative slope. The answer to this challenging deviation from the general trend can be found in [4]. In this paper, the authors suggest to fit a modified version of the Pareto distribution which allows us to define an upper bound limitting the values of our empirical distribution, in addition to the lower bound imposed by the standard Pareto distribution previously presented. The probability density function for an upper truncated Pareto distribution is given in equation B.4

with this issues, the first approach would have added some confusion to our explanations.

[2] Again, we refer to the online paper `www.hpl.hp.com/research/idl/papers/ranking/ranking.html` by Lada A. Adamic for additional useful information on this particular issue.

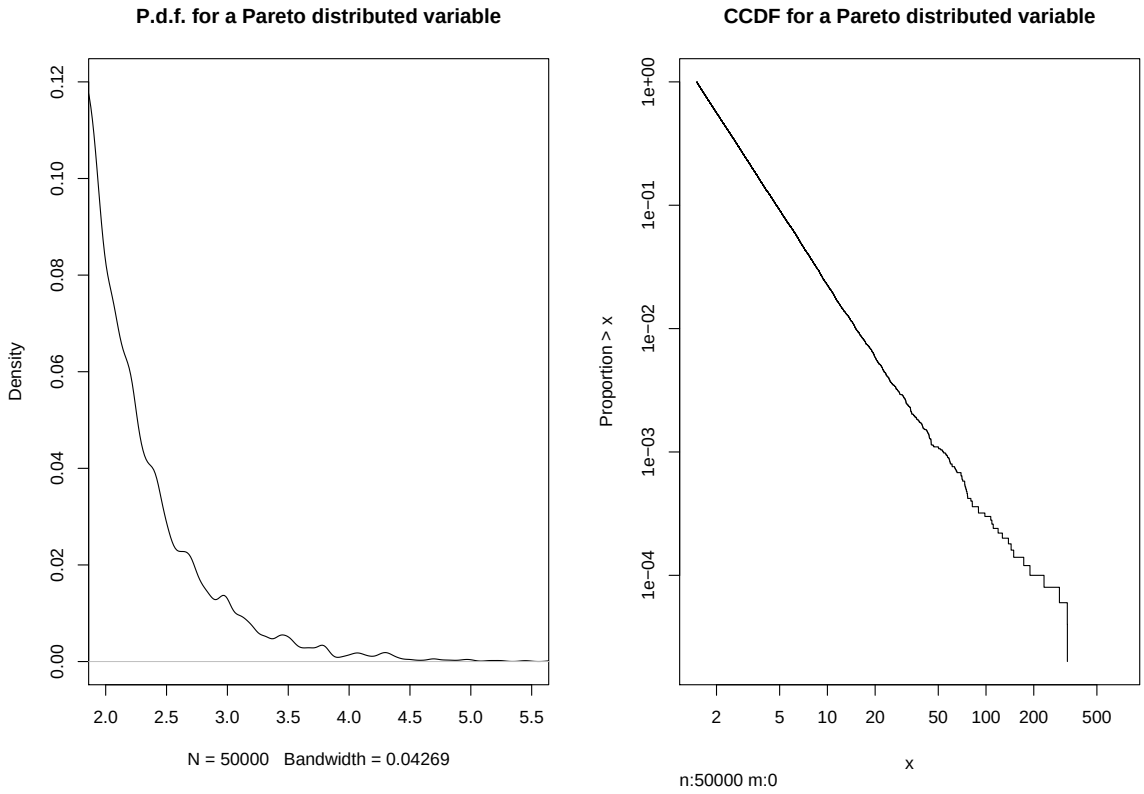

Figure B.1: Example of p.d.f. (left side) and CCDF (right side) for a Pareto probability distribution

$$f(x) = \frac{\alpha \gamma^{\alpha} x^{-\alpha-1}}{1-(\gamma/\nu)^{\alpha}} \tag{B.4}$$

While the expression for the CCDF of this probability distribution is given by equation B.5

$$1 - F(x) = P(X > x) = \frac{\gamma^{\alpha}(x^{-\alpha} - \nu^{-\alpha})}{1-(\gamma/\nu)^{\alpha}} \tag{B.5}$$

Where $0 < \gamma \leq x \leq \nu < \infty$ and $\gamma < \nu$.

In those cases involving the use of upper truncated Pareto distributions, we have included in our results the best fit for the 3 parameters involved in this procedure (the slope, the lower bound and the upper threshold), accompanied by the log likelihood value obtained for assessing the goodness of our fitting procedure. Interested readers should refer to [4] for further information about how to derive the MLE for these parameters.

B.2 The lognormal distribution

Some of the empirical distributions found for key parameters identified in our data samples, specifically those focused on per article metrics (number of different authors per article, number of revisions per article, etc.).

A positive random variable follows a lognormal distribution if $ln(x)$ is normally distributed. Thus, a lognormal distribution presents a probability density function given by equation B.6.

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}x\sigma} exp\left[-\frac{(ln(x)-\mu)^2}{2\sigma^2}\right] \quad x>0, \sigma>0, -\infty<\mu<\infty \tag{B.6}$$

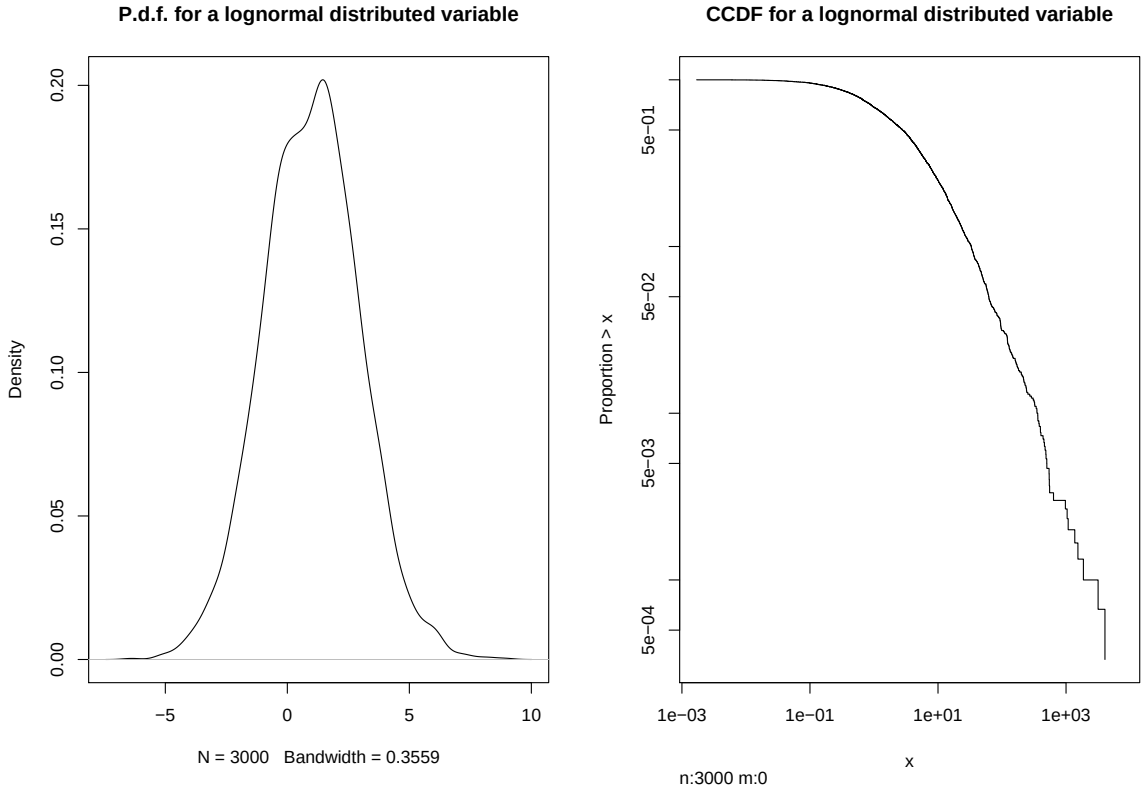

Figure B.2: Example of p.d.f. (left side) and CCDF (right side) for a lognormal probability distribution

Lognormal distributions has several interesting properties. The first one is that, unlike Pareto-like distributions, and although they are heavily right-skewed, they present a modal value, allowing us to make some useful deductions about the commonest cases found in the distribution (like the typical number of revisions received by a certain article). On the other hand, lognormal distributions presents a characteristic curved shape in their CCDF plot (in a log-log scale), which is clearly different from the log-linear pattern presented by Pareto distributions. It is then interesting to notice how some key descriptive parameters (like the number of different revisions per article or the number of distinct logged authors editing a certain article) have seemed to evolve from former Pareto-like distributions to their current lognormal pattern, suggesting that the distribution of these parameters has been reshaped by a change in the inner generational process producing them. Further investigation should be conducted to prove whether the stabilization in the total number of active users per month has generated the drop off in the number of revisions received by each article, thus beginning to accumulate some values around central tendency points.

Appendix C

Introduction to Survival Analysis

In this appendix, we provide some theoretical background about survival analysis and its applications. This material will be suitable for those readers unfamiliar with this set of statistical techniques that played an important role in the analysis of the Wikipedia community of authors included in this thesis.

This chapter is therefore aimed to provide expanded information about the methodology followed to obtain our results, but of course, it can not substitute an in-depth review of appropriate references in scientific literature about this matter. The most accessible monographic publication on survival analysis grounds, and its applications, might be [55]. We can also find shorter introductions in [64] (section 8.7), as well as in chapter 14 of [30]. Another comprehensive coverage of these topics, though focused on epidemiologic research (one of the main application fields of these techniques) is [103]. Interested readers may also find additional references providing a more theoretical approach, explaining the mathematical and probabilistic framework supporting survival analysis. The most remarkable example of this kind is probably the classic reference by Hosmer *et al.* [50].

C.1 Basic concepts in survival analysis

Broadly speaking, **survival analysis** is a set of statistical techniques and procedures aimed to model any kind of problem that can be stated as *"time until a certain event occurs"*. The purpose of the analysis if following a subset of individuals or examples, until the expected event occurs (where elapsed time can be measured in days, months, years, etc.). Then, we try to infer from these data the average behavior of subjects regarding this event, in terms of their expected **survival time**. In many case, the event is directly considered as a failure, because it has some negative connotation (death of patients, failure of engines or devices, etc.).

A key aspect in survival analysis is dealing with *censoring* problems. This situation happens when we started to collect information from a certain subject in the inspected population but, for any reason, at the end of the study we can not know its survival time. This may have been produced by many different causes, like the subject surviving beyond the final date of the study, loosing track of the subject or the subject experiencing another anomalies (different from the event of interest) causing its elimination from the sample. In general terms, the most common cause of censoring in survival studies (and the only one concerning our own study in this thesis) is the case in which subjects survive beyond the limit date of the analysis. This is call *right-censored* data in terms of the survival analysis methodology.

In order to compute survival information, we usually build tables including, for each subject under

study, its unique identifier and the observed survival time. Besides that, to profit from censorship information in our studies we need to include an additional column for each subject indicating whether the survival data is censored or not. Therefore, the advantage of this technique is that we do not lose information from any subject in the study, since both censored and uncensored samples incorporate useful data to the model.

It is usual to denote as T the random variable representing the survival time of subjects under study, whereas t usually represents the individual survival times observed in each subject. The values of the random variable representing the survival time must be contained in the range $T \in (0, \infty)$, and its continuous distribution is specified by its cumulative distribution function $F(t)$, (which offer the probability of any subject of having a lifetime value $T \leq t$), with probability density function $f(t)$. Now we can define the **survival function**.

Survival function: We define the **survival function** $S(t)$ as:

$$S(t) = 1 - F(t) = P(T > t) \tag{C.1}$$

Thus, it expresses the probability for a certain author to stay in the community longer than some specified time t. Theoretically, $S(t)$ is a smooth curve departing from value $S(0) = 1$, until it eventually reach the lowest possible value at $S(\infty) = 0$. However, in practice we use estimation procedures to obtain a step function, approximating the values of $S(t)$ (like we do in this thesis). In the next section, we provide a more detailed explanation on how to proceed to obtain this estimation.

On the other hand we can also work out the **hazard function** of a certain population. While the focus of the survival function is estimating the percentage of the total number of subjects under study who still remain alive after time t, the hazard function provides the opposite point of view. According to Kleinbaum *et al.*, it can be interpreted as *"the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t"*.

Hazard function: We define the **hazard function or force of mortality** $h(t)$ as a function measuring the risk of dying within an infinitesimal time interval Δt, given that the subject is alive at time t. The mathematical expression for this function is:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \ < \ (t + \Delta t) \,|\, T \geq t)}{\Delta t} \tag{C.2}$$

In this thesis work, hazard functions give us the (infinitesimal) risk that a Wikipedia author leaves the project, ceasing to contribute to it. It is important to remark the conditional character of the numerator term in the expression above. The instantaneous risk of death of individuals is always affected by the fact that they still remain alive beyond a certain point in time, t. In fact, there exists a direct relationship between these two functions, given by equations C.3 and C.4.

$$S(t) = exp\left[-\int_0^t h(u)du\right] \tag{C.3}$$

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right] \tag{C.4}$$

And thus, we can see that the hazard function is actually the derivative of the survival function with respect to the time variable. The -1 factor is needed to produce a positive rate for $h(t)$, since $S(t)$

has always a constant or negative slope (but never increasing). It is also interesting to notice that we can calculate directly the mean survival time of all individuals at risk between time t and $t + \delta$ as:

$$\text{mean survival time} = \int_t^{t+\delta} S(u)du \tag{C.5}$$

As we can see, this integral defines the area under $S(t)$ in the interval $[t, t + \delta]$, corresponding to the mean survival time of subjects who died during that interval or lived beyond $t + \delta$.

C.2 Estimation of survival functions

There exist at least a couple of methods to estimate the survival curve $S(t)$ for a population under analysis. The first one is the Kaplan-Meier estimator, which can be calculated from empirical survivability data for observed subjects. The second one is the Altshuler or Fleming-Harrington estimator of the survival curve, for which we need to first obtain an estimator of the hazard function, being this the principal drawback of this method [120]. In this thesis, we use the Kaplan-Meier estimator to obtain the survival curves included in our analysis of Wikipedia authors. The simplest expression for the Kaplan-Meier estimator can be written in this form:

$$\hat{S}(t_j) = \hat{S}(t_{j-1})\hat{P}r(T > t_j | T \geq t_j) \tag{C.6}$$

That is, the survival function at a given point in time is expressed as the product of the probability of surviving beyond the last failure time, multiplied by the probability of the subject surviving beyond the current failure time, given that (or conditioned to) it survived at least until the current failure time. Another possible expression of the same formula, implies representing the estimator as a product limit. In this form, the probability of surviving until the previous failure time is given by the product of all previous conditional probabilities for failure until that time and earlier.

$$\hat{S}(t_{(j-1)}) = \prod_{i=1}^{j-1} \hat{P}r(T > t_{(i)} | T \geq t_{(i)}) \tag{C.7}$$

This will lead us to a step estimator of the survival function, in which the survival probability at a given time will decrease according to the number of observed deaths within survivor subjects until that point. Figure C.1 exhibits an example of estimated survival curve using the Kaplan-Meier method.

In most of the case studies presented in this thesis, the population under analysis included so many individuals that the shape of the step function presents a virtually smoothed aspect, since we have a lot of information points to draw the estimated curve. In addition to that, as we already stated in chapter 4, the confidence intervals for the estimated curves obtained using the Kaplan-Meier method get better as we increase the number of available information points, rendering the very small intervals obtained in the majority of our case studies. In the last section of this appendix we will describe the most common features of representation of survival curves found in *GNU R*.

C.3 The Cox proportional hazards model

We can use conditional covariates in the Kaplan-Meier estimation of the survival curve to test whether survivability of certain groups within our population is significantly different or not. However, this is not the preferred statistic method to test such influence, since we do not obtain precise information

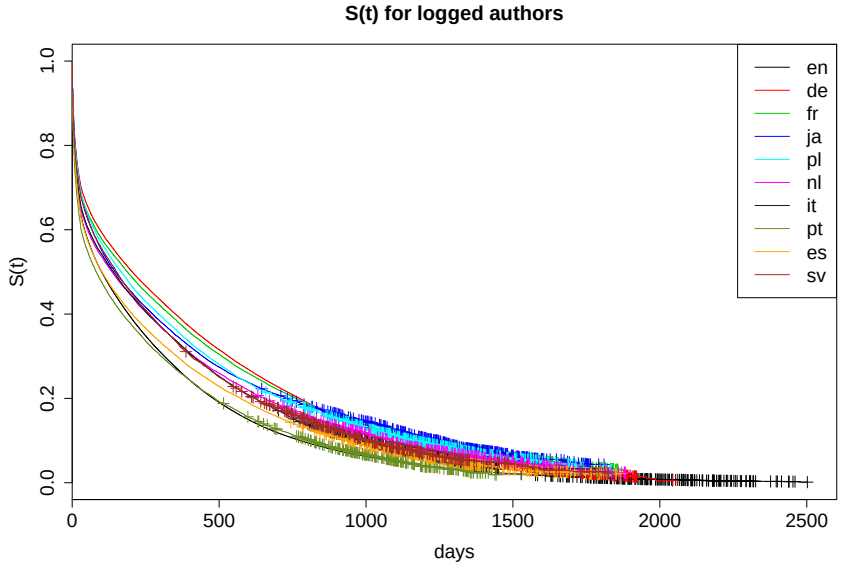

Figure C.1: Example of estimated survival function calculated with the Kaplan-Meier method

about the relationship between different covariates included in our model. There exist a number of *parametric* models (Weibull, exponential, etc.) to obtain estimators for the influence of covariates in the hazard function under analysis, much in the same way that we would proceed in a standard linear regression approach. Nevertheless, we have to be very sure about the validity of the selected parametric model applied to our empirical data. Even though we can test the goodness of fit of selected models, sometimes we could not be completely sure about their correctness.

Fortunately, in 1972 Cox introduced a semiparametric model to allow some flexibility in the adjustment of an appropriate model for hazard functions, based on the influence of covariates [28]. Now, this method is called the *Cox proportional hazards* model, sometimes abbreviated as CPH model. One of the reasons that has made this model so popular is the fact that it offers best fit approximation to the appropriate parametric model in each case. In other words, no matter which is the actual parametric model we should have chosen to fit our empirical data, the CPH will approximate it with reasonable accuracy.

The principles of the CPH model are the following. The central assumption is that we can calculate the hazard functions corresponding to different effects produced by relevant covariates included in the model. To achieve this, we define a common **baseline hazard function**, that will be present in all of them. This baseline hazard function is unspecified, and that is the reason making this method a semiparametric model. Then, each individual hazard function is obtained multiplying the baseline hazard function by the effects introduced by each individual covariate included in the model, weighted with the adequate coefficients (much in the same way we proceed in standard linear regression models). Thus, the hazard model for an individual with a given specification of a set of covariates (represented by $\mathbf{X}$ is:

$$h(t, \mathbf{X}) = h_0(t) exp\left[\sum_{i=1}^{p} \beta_i X_i\right] \tag{C.8}$$

Where:

$$\mathbf{X} = (X_1\, X_2 \ldots X_p) \tag{C.9}$$

Is the set of predictor variables for this individual [55]. Another important assumption here is that, though the baseline hazard function only depends on time, the exponential expression depends only on the covariates, but it should be independent from time. The inclusion of time-dependent variables is allowed in an extended version of the CPH model, which we do not use in this thesis, since we assumed time independence of all covariates considered in our case study.

Using the `survival` package in *GNU R*, we can obtain the results of the coefficients and assessment indicators to test the level of significance for our results. As an example, we offer the output obtained for the fit of the data corresponding to the English Wikipedia, testing the influence of edition of talk pages and FAs on the hazard function experimented by authors:

```
Call:
coxph(formula = survival ~ in_talk + in_FAs, data = target)

  n= 715267
                    coef exp(coef) se(coef)     z p
in_talk           -0.733     0.480  0.00301 -244 0
in_FAs            -0.756     0.469  0.00439 -172 0

                  exp(coef) exp(-coef) lower .95 upper .95
in_talk               0.480       2.08     0.478     0.483
in_FAs                0.469       2.13     0.465     0.473

Rsquare= 0.172   (max possible= 1 )
Likelihood ratio test= 134671  on 2 df,   p=0
Wald test            = 114276  on 2 df,   p=0
Score (logrank) test = 122217  on 2 df,   p=0
```

Table C.1: Output of summary results for the application of CPH model to test the influence of edition in talk pages and FAs on the survivability of logged authors, in the English Wikipedia

The model is fitted by means of maximizing the so-called Cox's likelihood, which can be proved to be used like a real log-likelihood [30]. The first line, as usual, presents the original call provided to GNU R, specifying the type of model we want to use, the covariates involved and the data set on which the model must be applied). The output reports a population size of $n = 715\,267$ logged authors (those who made more than 1 single edit under a logged account in the English Wikipedia). The first column shows the fitted coefficients, corresponding to β_i in equation C.8. For convenience, the actual exponential coefficient and the standard error corresponding to each covariate are also provided in the second and third columns.

As we can see, the p-value for both covariates is practically 0 (we obtain 0 due to precision limits in our computer calculations), demonstrating that both covariates have a high impact on the hazard function of logged authors in the English Wikipedia. As well, we can see that the coefficients are

negative in both cases, indicating that these two covariates have a direct impact on **reducing** the hazard rate of logged authors (alternatively, leveraging their lifetime in the system). With this tool, we were also able to use the *survfit* function to obtain proportional survival functions for each subset of logged authors, showing the influence of each parameter included in our model separately, then for the joint influence of editing FAs and talk pages at the same time (see Figure 4.39).

Finally, the summary outputs the results for three alternative tests of significance (Likelihood ratio, Wald test and a standard logrank test). In all cases, we can see that the significance of our conclusions is undoubtedly, specially thanks to the very high number of subjects followed in our analysis. We can also appreciate that the confidence intervals for both coefficients are very narrow, and they do not contain 0 (as we might expect from a covariate adding relevant information to our model).

Appendix D

Resumen en español

A lo largo de esta década, hemos sido testigos del nacimiento y proliferación de un gran número de proyectos colaborativos que tienen como objetivo común la creación y mejora de contenidos digitales, utilizando Internet como plataforma global de comunicación entre los miembros de la comunidad. De entre estos proyectos, Wikipedia ha logrado erigirse como el epítome de esta nueva tendencia. Un proyecto basado en premisas que, a priori, pueden resultar completamente incompatibles con la creación de material digital coherente y útil, y que sin embargo ha logrado aglutinar la que, con toda probabilidad, es la mayor comunidad virtual de toda la historia que trabaja bajo un proyecto común. El enorme éxito de Wikipedia plantea numerosas preguntas sin resolver para investigadores de muy diversas áreas, en especial relacionadas con la organización interna del proyecto, la distribución del esfuerzo creativo entre los voluntarios que contribuyen, así como la estructuración de los contenidos generados y el papel que juegan los procesos de discusión y coordinación dentro de la dinámica global de Wikipedia.

En esta tesis, abordamos el primer estudio comparativo del que tenemos constancia sobre las 10 mayores versiones de Wikipedia, según el número de artículos creados en cada una de ellas. Este estudio aborda un análisis cuantitativo, basado en rigurosas técnicas estadísticas, que pretende identificar métricas y parámetros descriptivos que influyen decisivamente en el funcionamiento de Wikipedia como comunidad creativa global. Nuestro objetivo, por tanto, ha sido no sólo el de abordar este análisis desde un punto de vista de exploración de datos que describen la actividad del proyecto, sino también el de ajustar modelos estadísticos para estos parámetros y estimadores. Como consecuencia, pretendemos mejorar nuestro nivel de comprensión sobre cómo funciona Wikipedia, cómo ha evolucionado hasta nuestros días, y cuál es el escenario más probable que deberá enfrentar el proyecto en los próximos años, en caso de mantenerse la situación actual. Todos los datos se extraen a partir de ficheros de registro públicamente disponibles en el sistema de descarga oficial mantenido por la Wikimedia Foundation (`http://download.wikimedia.org`). Dichos ficheros contienen, para cada idioma que queremos analizar, un listado detallado de todas las contribuciones que se han registrado en su base de datos, a lo largo de toda su historia de funcionamiento. Por ello, por primera vez nos encontramos en disposición de analizar la que probablemente es la mayor comunidad humana en un proyecto colaborativo, y además, de poder hacerlo con un nivel de detalle excepcional.

Otra de las aportaciones más importantes de esta tesis es la publicación de una herramienta software, WikiXRay, que permite automatizar por completo todo este proceso de análisis. La herramienta se encarga de recuperar del repositorio la información necesaria para cada idioma, descomprimir los ficheros de registro, analizar la información contenida en dichos ficheros para guardarla en una base de datos local, y por último, ejecutar una serie de análisis estadísticos sobre

estos datos para modelar el comportamiento y la evolución de dicho idioma. Esto supone una importante ventaja para otros investigadores interesados en realizar estudios cuantitativos similares sobre Wikipedia, puesto que ahorra un gran trabajo en un apartado especialmente complicado de este proceso como es el tratamiento inicial de los ficheros de registro. Por otro lado, se prevé publicar las bases de datos creadas localmente a partir del análisis de estos ficheros de datos, de forma que otros investigadores puedan tener acceso inmediato a dicha información, lista para ser analizada, sin tan siquiera tener que utilizar esta herramienta. Es de esperar que esta contribución resulte valiosa para la comunidad de investigadores que pretende estudiar Wikipedia desde un punto de vista cuantitativo, en especial gracias a la capacidad de esta herramienta para admitir nuevos módulos desarrollados por terceros que implementen nuevas funcionalidades y análisis que todavía no se han proporcionado.

En las siguientes secciones explicaremos con cierto detalle cuáles han sido los antecedentes que han suscitado la aparición de esta tesis, cuáles son los principales objetivos que pretendía acometer, la metodología que se ha seguido para su implementación y las principales conclusiones que podemos extraer a la vista de los resultados obtenidos.

D.1 Antecedentes

Desde la aparición del primer wiki, "WikiWikiWeb", en el dominio de Internet *c2.com* [1], la popularidad de estas herramientas como soporte para el desarrollo colaborativo de contenidos no ha cesado de crecer. Sin embargo, la idea de utilizar plataformas de creación colaborativa de contenidos se remontan mucho más atrás en el tiempo, puesto que en 1990 ya tenemos constancia de la existencia de un entorno este tipo en la Web, el editor *PREP* descrito por Neuwirth *et al.* en [72]. En este interesante artículo, los autores establecen algunas de las premisas básicas que deberían seguirse en el diseño de sistemas de este tipo:

1. Proporcionar métodos apropiados que den soporte para desarrollo de interacciones sociales entre los participantes en el proceso creativo.

2. Integración adecuada de soporte para aspectos cognitivos del proceso de coautoría, así como para la inclusión de comentarios externos.

3. Implementar herramientas prácticas que proporcionen ambos tipos de interacción dentro de la plataforma.

Más adelante, en 1995, encontramos otro atractivo artículo de Dourish y Bellotti [37], que introduce el concepto de *información de conciencia* para describir los comentarios y datos sobre las actividades que otros autores están desarrollando dentro del documento colaborativo. Los autores también presentan el concepto de *reacción compartida* (*shared feedback*) para referirse a aquellos sistemas de información que ponen a disposición del resto de autores los comentarios y reacciones de otros participantes, presentando estos datos en el espacio de trabajo común. Hoy día, podemos comprobar como muchas de las plataformas wiki a nuestra disposición incorporan esta funcionalidad, presentando de una u otra forma la información sobre los comentarios y anotaciones que otros autores dejan referente al documento en el que se trabaja de forma común. Sin embargo, este tipo de soportes puede dar pie a la aparición de reacciones contraproducentes para el proceso de creación colaborativa, tal y como explican Franco *et al.* en [42] al describir el concepto de *flame*. Aún más cerca en el tiempo, podemos encontrar otro artículo de Noël *et al.* [76] que incluye una comparativa detallada

[1] `http://c2.com/cgi/wiki?WelcomeVisitors`

de 19 plataformas distintas para creación colaborativa, siendo uno de los compendios más completos sobre el estado del arte de este sector.

Por otro lado, otro de los retos para los investigadores es el de comprender las razones que pueden llevar a un gran número de voluntarios a colaborar con un proyecto desinteresadamente (refiriéndonos explícitamente a la participación no remunerada en iniciativas de esta índole). En relación con este tema, Butler *et al.* publicaron en [19] un estudio acerca de los factores que influencian la participación de individuos dentro de una comunidad virtual. Si bien dicho estudio concluye que, como era de esperar, los fundadores y pioneros dentro del proyecto son los que desempeñan la mayor parte de las tareas dentro del proyecto, también se demuestra que otros participantes colaboran de manera muy activa. Es más, estos participantes se concentran en la realización de diferentes tipos de tareas dentro del proyecto, de acuerdo a sus propias iniciativas o afinidades personales, y según los objetivos que consideren más prioritarios desde su punto de vista personal.

Finalmente, otro factor crucial que ha contribuido, de manera determinante, a que los wikis alcancen su nivel de popularidad actual ha sido el concepto de *contenido abierto*, presentado por Cedergren en [22]. La posibilidad de poder mejorar y volver a distribuir el contenido creado de forma colaborativa, de manera que otros puedan importar dicho contenido y continuar trabajando sobre él (aunque dando crédito a los autores originales) supone un paso muy importante para favorecer la aparición de sinergias entre diferentes proyectos. Así pues, los contenidos abiertos permiten establecer las bases de un sistema en el que los contenidos pueden circular libremente entre diferentes comunidades virtuales, sin los problemas que conllevan los derechos de propiedad intelectual restrictivos, respetando al mismo tiempo el crédito que merecen cada uno de los autores que han ayudado a mejorar dichos contenidos. Como consecuencia, este nuevo paradigma operativo abre el camino para la implementación de novedosas formas de colaboración y trabajo creativo mediante wikis, como pueden ser el periodismo participativo, descrito en [61], los sistemas de traducción colaborativa, presentados en [33], o también sistemas para desarrollo de software, presentados en [17] y [135].

Sin embargo, de entre todos los proyectos de desarrollo colaborativo de contenidos que conocemos hoy día, Wikipedia destaca sin duda por ser el más popular, el que ha logrado aglutinar en torno a si a la mayor comunidad de colaboradores hasta la fecha, y uno de los más radicales en cuanto al modelo de admisión de nuevos voluntarios (no se imponen ningún tipo de precondiciones). Este último apartado ha suscitado una gran controversia en muchos sectores (no sólo académicos), sobre la capacidad del proyecto para generar contenidos que sean correctos y precisos, similares a los de una publicación enciclopédica tradicional. Sin embargo, estudios como el conducido por Giles para la revista científica Nature en 2005 [44] ponen de manifiesto que este modelo de desarrollo aparentemente desprovisto de todo control integra, en realidad, complejos mecanismos de coordinación y vigilancia de contenidos que permiten que muchos artículos producidos por Wikipedia puedan competir, prácticamente en igualdad de condiciones, con los generados por otras enciclopedias tradicionales como Britannica. Este hecho, unido al enorme tamaño de la comunidad de usuarios y lectores de Wikipedia (probablemente la mayor comunidad virtual que jamas se haya llegado a reunir en toda la historia de la Humanidad) y la excepcional capacidad de detalle que nos brinda el acceso a los ficheros de registro de toda la actividad llevada a cabo en cada edición de Wikipedia, convierten a este proyecto en uno de los objetivos más atractivos para una gran proporción de la comunidad científica internacional.

En general, podemos clasificar la literatura científica que gira en torno al estudio de Wikipedia en cuatro grandes grupos:

- Análisis cuantitativos: Un primer grupo de publicaciones está compuesto por todos aquellos

estudios enfocados en el análisis de Wikipedia desde un punto de vista puramente cuantitativo. Se trata de estudios que tratan de ofrecer una panorámica amplia del funcionamiento de Wikipedia, la organización de su comunidad de usuarios, y su evolución a lo largo del tiempo. Así pues, usualmente estos análisis suelen integrar diferentes metodologías, técnicas estadísticas y enfoques que se complementan para poder crear modelos que expliquen parte del funcionamiento de Wikipedia, mejorando nuestra nivel de compresión sobre el proyecto. La mayoría de estos estudios se han centrado en la Wikipedia inglesa, al tratarse de la versión de mayor tamaño y que aglutina a la mayor comunidad de colaboradores dentro del proyecto. Por otra parte, también es frecuente encontrar estudios cuantitativos centrados en otras versiones de menor tamaño, puesto que el análisis de la Wikipedia en inglés conlleva en muchos casos problemas técnicos de difícil solución si no se cuenta con la infraestructura adecuada para acometer el procesamiento de ingentes cantidades de información de forma eficiente. Podemos citar entre las publicaciones más destacadas en este ámbito [124], [132], [8] y [84].

- Calidad de contenidos y reputación de los autores: Otro de los aspectos que más despierta la atención de los autores científicos en relación con Wikipedia es la aparentemente inexplicable capacidad del proyecto para crear contenidos de calidad. Intuitivamente podríamos tener la percepción de que un proyecto completamente abierto a las aportaciones de cualquier usuario, sin control alguno sobre su nivel de conocimientos acerca del tema en el que va a contribuir, no sería capaz de generar contenidos enciclopédicos rigurosos y fiables. Muy al contrario, en la práctica existen estudios científicos que han demostrado que, si bien Wikipedia no se caracteriza por su prolífica creación de contenidos de calidad extraordinaria, si que podemos constatar la existencia de una organización rigurosa dentro de la comunidad de usuarios que permite controlar los actos de vandalismo con agilidad, mejorar el contenido de los artículos siguiendo reglas editoriales bien definidas, y en suma, aumentar el nivel de calidad de las entradas enciclopédicas en cada idioma siguiendo un notable sistema de coordinación entre los usuarios. Complementando esta aproximación, otros estudios se han centrado en intentar identificar los parámetros que influyen en la mejora de la reputación de autores en Wikipedia, medida de acuerdo a la calidad de los contenidos aportados por cada autor. Sobresalen en este apartado los trabajos realizados por el equipo del Visual Communication Lab de IBM [123], el excelente trabajo de definición de un entorno teórico para el análisis de calidad de los contenidos en grandes repositorios de datos (en particular en Wikipedia), por Besiki Stvilia, Les Gasser *et al.* [110], [112], [3], así como también la aplicación WikiTrust [2], que evalúa la reputación de autores en Wikipedia y que ha dado pie a un buen número de publicaciones que resumen las contribuciones más relevantes que nos aporta [6], [7], [5].

- Redes sociales, grafos web y modelos de enlaces: El tercer gran apartado de publicaciones científicas sobre Wikipedia analiza las redes de conexiones que podemos encontrar dentro del proyecto, bien tomando como nodos cada uno de los autores miembros de la comunidad, bien considerando las relaciones existentes entre los diferentes artículos enciclopédicos generados, a través de los hiperenlaces que apuntan desde un artículo a otros dentro del mismo idioma o dentro de otras versiones. En la mayoría de artículos y publicaciones dentro de este apartado se intentan aplicar métodos gráficos que permitan visualizar las estructura de conexiones que aparece entre los miembros de la comunidad analizada, así como modelos estadísticos que permitan comprender el proceso generador subyacente que da lugar a la aparición de redes de este tipo. Cabe destacar en este ámbito las contribuciones realizadas por Bellomi *et al.* [14],

[2] `trust.cse.ucsc.edu`

Buriol *et al.* [18],Capocci *et al.* [20] y Holloway *et al.* [49].

- Wikipedia y la Web Semántica: Por último, otro de los apartados que más interés científico ha acaparado en los últimos años es la exploración de nuevas herramientas y estrategias metodológicas que permitan incrementar las funcionalidades de los wikis, añadiendo capacidades de análisis semántico de contenidos, categorización y búsqueda inteligente de información contextual. Wikipedia no ha sido una excepción ha esta norma general, y varias publicaciones han intentado estudiar de qué forma podríamos integrar eficientemente herramientas propias de la Web Semántica dentro del proyecto, bien por medio de sistemas de categorización más eficientes, bien integrando nuevas funcionalidades dentro del motor MediaWiki para proporcionar estas nuevas características en forma de extensiones software. Dentro de este apartado, destacamos los trabajos realizados efectuados por Denoyer y Gallinari [32], Sauer *et al.* [101], ,Ollivier y Senellart [78] y Aumüller y Auer [12].

Tal y como podemos ver, Wikipedia ha recibido la atención de numerosos investigadores hasta el momento. Sin embargo, a pesar de que no son pocas las publicaciones científicas existentes que giran en torno a Wikipedia, todavía hoy somos incapaces de encontrar estudios rigurosos que comparen, a gran escala, las características descriptivas y funcionales más destacadas que podemos identificar en varias Wikipedias de distintos idiomas. La gran cantidad de datos que deben procesarse para culminar este tipo de análisis en un tiempo razonable, unido a la falta de herramientas hasta el momento que permitan acometer estos análisis de manera eficiente, ha supuesto un obstaculo insalvable para efectuar tales tipos de estudios. Es precisamente en este apartado concreto en el que esta tesis pretende rellenar el vacío que hasta ahora existía, proporcionando no solo el primer estudio cuantitativo a gran escala sobre las 10 versiones de Wikiepedia más grandes (en términos de número de artículos en cada una de ellas) sino también una herramienta software (WikiXRay) que facilite la reproducción de los análisis incluidos en esta tesis, así como su extensión para incluir nuevas técnicas de estudio que podamos aplicar sobre estos datos.

D.2 Objetivos

Como ya hemos introducido con anterioridad, el objetivo de esta tesis doctoral es doble. Por un lado, pretendemos ofrecer el que probablemente es el primero estudio cuantitativo a gran escala sobre las diez mayores versiones de Wikipedia, siguiendo una metodología estadística que permita comparar las diferentes versiones con el objeto de identificar parámetros descriptivos característicos que aumenten nuestro conocimiento sobre el funcionamiento de Wikipedia y la organización de su comunidad de usuarios. Por otra parte, esta tesis ofrece como segunda contribución importante el software que ha permitido implementar estos análisis, WikiXRay, desarrollado dentro de nuestro propio grupo de investigación. Este software facilita, por una parte, que otros investigadores puedan replicar sin demasiado esfuerzo los análisis que hemos incluido en esta tesis. Por otro lado, permite que se puedan incluir nuevos análisis al proceso de estudio, aprovechando que el programa resuelve todas las tareas previas de adquisición y almacenamiento local de datos cuantitativos sobre cada idioma, paso previo a la realización de estos análisis.

Centrándonos en los puntos concretos que hemos abordado en esta tesis, la siguiente es la lista de las 7 preguntas que analizamos en nuestro estudio comparativo de las 10 mayores Wikipedias:

1. ¿Cómo evoluciona la comunidad de autores de Wikipedia a lo largo del tiempo?: El tamaño de la comunidad, y el elevado número de modificaciones efectuadas sobre los artículos y otras

páginas wiki dificulta el análisis de la historia completa de los cambios registrados en los archivos de registro de la base de datos. Nuestro propósito en esta caso es el de estudiar la evolución en el tiempo del número de contribuciones recibidas y el número de autores activos para cada mes, en busca de tendencias características que puedan encontrarse en estos gráficos. Estas son métricas básicas, que describirán el nivel de actividad mantenido por los miembros de la comunidad a lo largo del tiempo. También consideraremos la posible influencia de los autores anónimos y los robots (programas que efectúan modificaciones automáticas en los contenidos de las páginas).

2. ¿Cuál es la distribución de las páginas y contenidos en cada idioma?: Cada una de las diferentes versiones de Wikipedia pueden concentrar el esfuerzo de colaboración en distintos tipos de páginas o contenido. Mediante el análisis del porcentaje de cada tipo de página (artículos, redirecciones, páginas de usuario, páginas de discusión...) producidas en las diez mayores Wikipedias, proporcionará valiosos indicios sobre las distintas estrategias seleccionadas por cada comunidad para desarrollar su trabajo. También obtendremos información sobre la importancia de aspectos organizacionales clave para la comunidad (como discusiones o creación de páginas de usuario), así como también sobre categorización de contenidos y extensión del número de términos enciclopédicos incluidos en cada idioma. Del mismo modo, el análisis del tamaño de los artículos de Wikipedia, y su evolución en el tiempo para las diferentes versiones, revelará interesantes rasgos del proceso de creación de contenidos que se lleva a cabo en cada una de las comunidades examinadas.

3. ¿Cómo evoluciona la coordinación entre los distintos autores de cada idioma?: La participación de los autores en las páginas de discusión asociadas a cada artículo resulta un punto crítico para mejorar la calidad de los contenidos de la enciclopedia. Al mismo tiempo, las páginas de discusión proporcionan un foro natural para asegurar la aplicación de políticas editoriales impuestas por la comunidad. El análisis de la evolución en el tiempo de los autores que participan en páginas de discusión, la evolución del número mensual de autores que participan en discusiones y la evolución del tamaño de las páginas de discusión a lo largo del tiempo contribuirá a completar todavía más nuestro análisis de los patrones de comportamiento internos de la comunidad de autores de Wikipedia.

4. ¿Cuáles son los parámetros clave que podemos identificar para estratificar la comunidad de autores de Wikipedia y definir su estructura social?: Para abordar el problema específico de describir en detalle la distribución de esfuerzos entre los miembros de la comunidad de autores de Wikipedia, implementamos un análisis en detalle de la distribución de las revisiones entre los autores y el número de artículos diferentes editados por autor. De igual forma, examinamos el mismo escenario desde una perspectiva distinta, estudiando el reparto de revisiones y autores entre los diferentes artículos de cada idioma. Por último, utilizamos varias medias de desigualdad en distribuciones, bien conocidas dentro del ámbito científico, para estudiar el reparto de revisiones entre autores y artículos, caracterizando así la estratificación de cada comunidad de acuerdo al esfuerzo empleado por cada individuo dentro del grupo.

5. ¿Cuál es la vida media de los autores en Wikipedia?: Un aspecto importante concerniente a la organización y sostenibilidad de cualquier proyecto colaborativo es el cálculo del tiempo medio de vida (medido como tiempo de participación) de los voluntarios que participan en la comunidad. Si el proyecto recibe a más voluntarios de los que pierde, entonces nos encontraremos ante una comunidad creciente, que es capaz de afrontar retos cada vez más

complejos con el transcurso del tiempo. Por el contrario, si el proyecto pierde más miembros de los que es capaz de atraer, este hecho puede imponer condiciones negativas que coarten la sostenibilidad del proyecto en el futuro próximo.

6. ¿Podemos identificar métricas básicas que describan la reputación de los autores en Wikipedia y la calidad de sus artículos? A pesar de que el análisis de la calidad del contenido de Wikipedia, y la reputación de los autores es una tarea compleja, pretendemos identificar métricas básicas que revelen rasgos comunes compartidos por todos los artículos de calidad en Wikipedia. Construiremos nuestras mediciones sobre el trabajo de revisión efectuado por numerosos miembros de la comunidad, que ha seleccionado aquellos artículos que merecen la distinción de *Artículo Destacado* dentro de una edición, debido a su elevada calidad. También comprobaremos la validez de métricas previamente propuestas por Stein y Hess [108], para examinar si pueden aplicarse en medidas de calidad de los artículos y de nivel de reputación de los autores, complementando de esta forma otras iniciativas en marcha en este mismo sentido [6].

7. ¿Podemos deducir, basándonos en los datos previos, algunas condiciones de sostenibilidad que influencien la evolución futura de Wikipedia en los próximos años?: Como conclusión, examinaremos la evolución en el tiempo de algunos de los parámetros y métricas más sobresalientes identificados en los apartados anteriores. El principal objetivo de este análisis será el de deducir implicaciones relevantes para la sostenibilidad de Wikipedia en el futuro, en especial en lo concerniente al número de autores necesarios para soportar su impresionante nivel de crecimiento y el amplio espectro de términos y contenidos cubierto actualmente por el proyecto.

De este modo, nuestro propósito es que las respuestas obtenidas para cada una de las siete preguntas descritas anteriormente proporcionen los elementos necesarios para construir un modelo coherente, que nos permita entender mejor cómo funciona Wikipedia, como se comporta su comunidad de usuarios y cómo es probable que evolucione el proyecto en el futuro inmediato, suponiendo que la situación actual que identifiquemos en nuestro análisis se mantenga durante los próximos meses.

D.3 Metodología

En líneas generales, la metodología que hemos seguido para llevar a cabo esta tesis doctoral, comprende una serie de técnicas estadísticas que nos han permitido obtener resultados significativos para contestar las preguntas que hemos presentado en el apartado anterior. Para llevar a cabo este análisis, hemos utilizado la herramienta que hemos desarrollado a tal efecto, WikiXRay. El proceso de análisis global que hemos llevado a cabo sigue las siguientes fases:

- Obtención de la información de registro de actividad de cada una de las ediciones de Wikipedia analizadas, recuperando los archivos procedentes de la página web en la que Wikimedia Foundation publica esta información [3].

- El siguiente paso es descomprimir cada uno de los archivos recuperados, analizando la información para extraer información cuantitativa que podamos almacenar en una base de datos

[3] `http://download.wikimedia.org`

local para su posterior análisis. Se trata de una de las fases más complicadas, puesto que algunas ediciones analizadas contienen ingentes cantidades de información en los registros de actividad, que debe ser examinada de manera eficiente.

- Una vez extraída la información cuantitativa procedente de los archivos de registro, la herramienta construye una serie de tablas de resultados intermedios, organizando la información de manera apropiada para facilitar los cálculos posteriores dentro de los análisis que hemos implementado en esta tesis.

- Por último, efectuamos todas las operaciones necesarias para implementar cada uno de los análisis que hemos llevado a cabo en esta tesis. Siempre que dichos análisis conlleven la aplicación de técnicas, modelos y herramientas estadísticas, WikiXRay se apoya en la utilización del paquete estadístico *GNU R* [87] para llevarlas a cabo. Se trata de un completo entorno estadístico, liberado como software libre, que en la actualidad acumula más de un millar de paquetes de extensión diferentes, que proporcionan soporte para casi cualquier tipo de análisis o técnica estadística concebible.Los resultados de nuestros análisis, tanto numéricos como en forma de gráficos, se organizan convenientemente en directorios de manera que podamos recuperar esta información ágilmente.

En lo referente a las técnicas estadísticas que hemos empleado en cada uno de nuestros análisis, la siguiente lista proporciona una panorámica completa de todos los procedimientos que se han aplicado en esta tesis:

- Análisis de datos exploratorio: Para comenzar nuestro estudio, hemos aplicado toda una serie de técnicas de descripción de datos estandarizadas en la actualidad para identificar patrones descriptivos característicos en poblaciones y muestras. Estás técnicas incluyen desde gráficas de evolución en dos dimensiones, siguiendo los cambios de los parámetros estudiados a lo largo del tiempo, hasta histogramas y diagramas de densidad de probabilidad que muestran la distribución de lo estadísticos de interés.

- Ajuste de distribuciones de probabilidad: En nuestro análisis de la estructura social de la comunidad de autores de Wikipedia, hemos tenido que ajustar distribuciones teóricas a las distribuciones de probabilidad de los estadísticos más relevantes identificados en nuestro análisis. Las técnicas de ajuste han procurado seguir todas aquellas buenas prácticas que podemos encontrar en la literatura actual, de forma que los resultados obtenidos sean lo más precisos posibles. Esto es especialmente delicado cuando tratamos de ajustar distribuciones de probabilidad a conjuntos de datos de gran tamaño como los que tenemos que manejar en esta tesis. Las distribuciones teóricas que hemos utilizado en nuestros ajustes incluyen la distribución de Pareto, la Pareto truncada y la distribución lognormal.

- Análisis de desigualdad en distribuciones: Para examinar el grado de desigualdad en el reparto del esfuerzo creativo dentro de la comunidad de autores, hemos aplicado una serie de coeficientes estadísticos que permiten medir este grado de desigualdad desde diferentes puntos de vista. El más conocido y útil de estos coeficientes es el índice de Gini, que permite calcular a partir del diagrama de una curva de Lorenz un resumen numérico que da una idea del grado de desigualdad en el reparto de un determinado parámetro entre los miembros de la población.

- Análisis de supervivencia: La más novedosa de las técnicas estadísticas aplicadas en esta tesis es sin duda el análisis de supervivencia. La novedad no radica en el hecho de que sean

técnicas que hayan aparecido hace poco tiempo, puesto que la mayoría de herramientas de análisis de supervivencia llevan aplicándose desde hace varias décadas en campos tan dispares como estudios demográficos, control industrial y especialmente medicina, sobre todo dentro del campo de la epidemiología. Más bien, el aspecto novedoso estriba en que se trata de técnicas relativamente poco conocidas aún en las Ciencias de la Computación. El análisis de supervivencia nos permite abordar problemas que puedan enunciarse como *tiempo transcurrido hasta que un determinado evento ocurre* para cada individuo de la población. Entre los resultados más destacados que podemos obtener con esta técnica, cabe resaltar la vida media de los miembros de la comunidad, y las curvas de supervivencia, que modelan cómo evoluciona la mortalidad entre los individuos de la población a lo largo del tiempo.

Con el objeto de poder realizar un seguimiento preciso de la actividad que ha tenido lugar en cada idioma, hemos restringido nuestro estudio cuantitativo a los llamados autores registrados, aquellos que antes de editar una página se han autenticado en el sistema para realizar dicha edición asociada a su cuenta de usuario dentro de ese idioma. De esta forma, hemos filtrado sistemáticamente de nuestro estudio las contribuciones de los autores anónimos, puesto que el sistema de registro únicamente guarda para identificarlos una dirección IP. Esto supone un problema de cara a poder discernir con claridad cuántos autores están editando a través de una misma dirección IP, debido a que en la actualidad el empleo de sistemas de traducción de direcciones (NAT), *proxies*, cortafuegos y otros dispositivos permiten agrupar las conexiones de múltiples usuarios con la misma IP. Por otro lado, aunque en la primera parte de nuestro análisis estudiamos la influencia de las contribuciones procedentes de programas robot en cada idioma, en todos los epígrafes siguientes hemos filtrado de nuestro análisis dichas contribuciones, para poder concentrarnos exclusivamente en las ediciones procedentes de autores humanos.

En nuestro análisis, consideramos cada edición que ha efectuado un autor registrado en una cierta página de la Wikipedia, dentro de un idioma concreto. Cada una de estas ediciones recibe el nombre de revisión. Así pues, para extraer los datos cuantitativos empleados en esta tesis se han tenido que procesar un total de más de 300 millones de revisiones, efectuadas sobre un total de más de 21 millones de páginas en las 10 mayores Wikipedias. De entre ellas, hemos separado las revisiones procedentes de un total de más de 2 millones y medio de autores registrados. Con el objeto de poder establecer una comparativa equilibrada entre los diferentes idiomas, hemos analizado los datos recogidos por el sistema desde el primer registro almacenado en cada idioma, hasta el 31 de diciembre de 2007. La elección de esta fecha límite ha estado condicionada por la disponibilidad de datos para todos los idiomas. En el caso de la Wikipedia inglesa, no tenemos información de registro válida más allá de esta fecha, por lo que hemos limitado el rango temporal de nuestra comparativa para asegurar que podemos analizar todos los idiomas por igual, ignorando los datos de registro en otras Wikipedias para fechas posteriores.

D.4 Conclusiones

Para concluir este resumen, examinaremos las principales conclusiones que hemos podido extraer en nuestro análisis cuantitativo de las 10 mayores ediciones de Wikipedia, haciendo especial hincapié en aquellos resultados que influencien en mayor medida la posible sostenibilidad del proyecto en el futuro próximo.

El primer resultado que llama poderosamente nuestra atención se refiere a la estabilización que ha empezado a registrar el número de revisiones por mes recibidas en todos los idiomas (tanto

totales como procedentes de usuarios registrados). Si analizamos el número de usuarios activos por mes, observamos le mismo fenómeno que para las gráficas de revisiones: las curvas se estabilizan, aproximadamente, a partir de verano de 2006. Consideramos como usuarios activos aquellos que han realizado al menos una edición sobre una pagina de ese idioma en un mes determinado. Así pues, la razón principal de la estabilización en el esfuerzo editorial registrado en todos los idiomas la encontramos en la deceleración del número de usuarios activos por mes. El siguiente paso de nuestro análisis consiste en estudiar si esta estabilización es aplicable a todos los tipos de páginas en general, o bien afecta a unos tipos concretos de páginas. Los resultados obtenidos indican que la estabilización de la actividad editorial de los autores afecta a todo tipo de páginas por igual, por lo que podemos hablar de un fenómeno generalizado experimentado por todos los idiomas. Este cambio en el patrón evolutivo del esfuerzo desplegado por la comunidad de autores es el primer paso en una cadena de resultados que tendrán consecuencias importantes para la sostenibilidad del proyecto.

No obstante, un resultado que también llama nuestra atención, por ir en contra de la dinámica general identificada, es el crecimiento sostenido del número de páginas de discusión activas (asociadas a cada artículo individual) por mes. Este resultado nos indica que el esfuerzo de los autores puede estar comenzando a derivar hacia la discusión sobre los contenidos de los artículos, en lugar de dedicarse a la edición de contenidos propiamente dicha. Por tanto, los indicios empíricos que encontramos apuntan hacia un cambio en el patrón de comportamiento de los usuarios en todos los idiomas analizados, que comienzan a concentrarse en tareas de coordinación y discusión de contenidos.

Por otra parte, los diferentes patrones de actividad desplegados por los usuarios en cada idioma han dado lugar a distribuciones de contenidos muy diferentes en cada una de las Wikipedias analizadas. Mientras que encontramos algunos idiomas que otorgan una gran importancia a la creación de páginas de discusión e intercambio de información entre miembros de la comunidad (páginas de discusión, páginas de usuario), como el francés, alemán e inglés, otras como el español y el sueco presentan un importante nivel de actividad en la categorización de páginas. Finalmente, idiomas como el polaco demuestran el resultado de un intenso trabajo de creación de contenido por parte de robots, que provoca la aparición de una proporción de páginas de artículo anormalmente alta en comparación con otros idiomas.

En lo que se refiere la estructura social de la comunidad de autores de Wikipedia, la estratificación por número de revisiones realizadas por cada usuario revela una distribución del esfuerzo libre de escala, puesto que podemos ajustar una distribución teórica Pareto truncada a la distribución de los estadísticos de actividad más importantes (como número de revisiones por autor y número de artículos diferentes editados por autor). Por otra parte, los estadísticos por artículo demuestran que la distribución del esfuerzo editorial recibido sigue un patrón lognormal, diferente al patrón log-lineal que se había identificado en estudios previos a este respecto. La principal consecuencia de estas distribuciones de esfuerzo es que un pequeño grupo de autores en cada idioma es el responsable de la mayor parte del esfuerzo editorial que se lleva a cabo. Por otra parte, este patrón de distribución de esfuerzo, unido a recientes estudios sobre el modelo generador de la actividad editorial en Wikipedia [106], apuntan a que un proceso de relación preferencial es el responsable de este patrón de actividad que hemos identificado. Los artículos que registran un mayor número de enlaces en otras páginas que apunten hacia ellos tienen más posibilidades de ser editados por un número mayor de usuarios. Sin embargo, la estabilización en el número de los autores activos por mes ha provocado que la distribución de esfuerzo recibido por los artículos haya evolucionado de un patrón de Pareto a un patrón lognormal. El análisis del nivel de desigualdad en las contribuciones de los usuarios revela que el gran nivel de desigualdad, sesgado hacia el grupo de autores que está desarrollando la mayor parte de la actividad de contribución en cada idioma, no ha cesado de crecer. Este hecho es potencialmente preocupante para la sostenibilidad del proyecto en un futuro próximo, puesto que indica que los

autores están incrementando progresivamente su actividad, hasta que lleguen a un límite que no puedan superar. Por ello, es de esperar que este nivel de estabilización comience a decrecer en un futuro próximo, y empecemos así a ver gráficas de actividad descendentes en las principales Wikipedias.

El análisis demográfico revela parámetros descriptivos todavía más clarificadores. En este análisis, estudiamos en primer lugar el número de nacimientos y muertes registrados por el sistema en cada mes. Consideramos un nacimiento como un nuevo usuario que comienza a editar páginas en Wikipedia por primera vez, mientras que se considera como una muerte a un usuario que abandona el proyecto para no volver a contribuir nunca más [4]. En este contexto, los resultados nos indican que desde que comenzó a producirse la estabilización en el número de usuarios activos por mes, en verano de 2006, el número mensual de muertes ha superado el número mensual de nacimientos registrados en todos los idiomas. Así pues, la estabilización en el número de usuarios activos por mes se debe, en realidad, a que las mayores Wikipedias están perdiendo usuarios, puesto que el número de autores que se marcha del sistema para no volver a contribuir supera el número de nuevos autores que el proyecto es capaz de atraer. Un análisis detallado de los tiempos medios de vida de los autores en todos los idiomas, revela que efectivamente el riesgo de muerte entre los autores más jóvenes es extraordinariamente alto, lo que incide negativamente en el número de voluntarios que finalmente alcanza un grado de actividad significativo dentro del proyecto.

En cuanto al análisis de los patrones cuantitativos comunes en artículos destacados, podemos afirmar a partir de los resultados empíricos obtenidos que existen rasgos comunes compartidos por los contenidos de mayor calidad en todos los idiomas analizados. Los artículos más destacados necesitan, en general, recibir un número de contribuciones mucho mayor que el de los artículos estándar, así como también un mayor periodo de tiempo en el sistema para que los defectos se puedan ir corrigiendo. La diferencia en ambos casos es de un orden de magnitud en comparación con los artículos no destacados, para todos los idiomas. Por otra parte, centrándonos en los autores que editan en artículos de calidad, observamos que se trata de miembros muy experimentados de la comunidad, con más de 1.000 días de actividad en el proyecto (algo si cabe todavía más insólito, habida cuenta de la elevada tasa de mortalidad que vemos en la población de autores más jóvenes). Finalmente, nuestro estudio de validación de las métricas de calidad de artículos y reputación de autores, propuestas por Stein y Hess en [108], revelan que existe una correlación positiva entre las medidas y el nivel de calidad y reputación de los usuarios analizados, por lo que pueden suponer un elemento de utilidad para predecir los artículos candidatos a convertirse en destacados en el futuro próximo, en combinación con otras métricas adicionales.

Para concluir, la revisión de las tendencias evolutivas de algunos de los estadísticos más importantes que hemos identificado en esta tesis confirman la preocupante tendencia identificada en todos los idiomas respecto a la sostenibilidad del proyecto. El nivel de desigualdad mensual de las contribuciones recibidas por todos los idiomas se ha mantenido constante a lo largo del tiempo. Sin embargo, conforme el número de autores registrados ha comenzado a estabilizarse, la cantidad de trabajo editorial que deben llevar a cabo el núcleo de usuarios más activos se hace cada vez mayor. Como recomendación de cara al futuro, la Wikimedia Foundation debería centrarse en el diseño y puesta en práctica de políticas de atracción y fidelización de usuarios más agresivas, puesto que de otro modo, el proyecto comenzará a entrar en un periodo de clara recesión con consecuencias indeterminadas para su continuidad futura. Por otra parte, el extraordinario nivel de crecimiento que siguen experimentando algunos estadísticos de coordinación entre usuarios demuestran que

[4] Esta terminología coincide con la que comúnmente se utiliza en estudios demográficos, así como también en el análisis de sistemas de conmutación y sistemas de comunicación en red, por medio de la Teoría de Colas.

las comunidades de las 10 mayores Wikipedias están empezando a ponderar más intensamente los aspectos organizativos y de discusión y mejora de los contenidos enciclopédicos, lo que supone un punto muy positivo de cara al objetivo del proyecto de mejorar la calidad general de sus artículos de forma progresiva.

Appendix E

License Creative Commons Attribution-ShareAlike 3.0

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE (”CCPL” OR ”LICENSE”). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

   (a) ”Adaptation” means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image (”synching”) will be considered an Adaptation for the purpose of this License.

   (b) ”Collection” means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent works in themselves, which together are

assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined below) for the purposes of this License.

(c) "Creative Commons Compatible License" means a license that is listed at http://creativecommons.org/compatiblelicenses that has been approved by Creative Commons as being essentially equivalent to this License, including, at a minimum, because that license: (i) contains terms that have the same purpose, meaning and effect as the License Elements of this License; and, (ii) explicitly permits the relicensing of adaptations of works made available under that license under this License or a Creative Commons jurisdiction license with the same License Elements as this License.

(d) "Distribute" means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.

(e) "License Elements" means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.

(f) "Licensor" means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.

(g) "Original Author" means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.

(h) "Work" means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

(i) "You" means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

(j) "Publicly Perform" means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place

individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.

(k) "Reproduce" means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.

2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

   (a) to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;

   (b) to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified.";

   (c) to Distribute and Publicly Perform the Work including as incorporated in Collections; and,

   (d) to Distribute and Publicly Perform Adaptations.

   (e) For the avoidance of doubt:

      i. Non-waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

      ii. Waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,

      iii. Voluntary License Schemes. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

   The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

4. Restrictions. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

(a) You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.

(b) You may Distribute or Publicly Perform an Adaptation only under the terms of: (i) this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-ShareAlike 3.0 US)); (iv) a Creative Commons Compatible License. If you license the Adaptation under one of the licenses mentioned in (iv), you must comply with the terms of that license. If you license the Adaptation under the terms of any of the licenses mentioned in (i), (ii) or (iii) (the ”Applicable License”), you must comply with the terms of the Applicable License generally and the following provisions: (I) You must include a copy of, or the URI for, the Applicable License with every copy of each Adaptation You Distribute or Publicly Perform; (II) You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License; (III) You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform; (IV) when You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

(c) If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution (”Attribution Parties”) in Licensor’s copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable,

the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Ssection 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., ”French translation of the Work by Original Author,” or ”Screenplay based on original Work by Original Author”). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

(d) Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author’s honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author’s honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

   UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

(a) This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

(b) Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

(a) Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

(b) Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.

(c) If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

(d) No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

(e) This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

(f) The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License is not intended to restrict the license of any rights under applicable law.

Creative Commons is not a party to this License, and makes no warranty whatsoever in connection with the Work. Creative Commons will not be liable to You or any party on any legal theory for any

damages whatsoever, including without limitation any general, special, incidental or consequential damages arising in connection to this license. Notwithstanding the foregoing two (2) sentences, if Creative Commons has expressly identified itself as the Licensor hereunder, it shall have all rights and obligations of Licensor.

Except for the limited purpose of indicating to the public that the Work is licensed under the CCPL, Creative Commons does not authorize the use by either party of the trademark ”Creative Commons” or any related trademark or logo of Creative Commons without the prior written consent of Creative Commons. Any permitted use will be in compliance with Creative Commons' then-current trademark usage guidelines, as may be published on its website or otherwise made available upon request from time to time. For the avoidance of doubt, this trademark restriction does not form part of the License.

Creative Commons may be contacted at `http://creativecommons.org/`.