Quantitative Analysis of the Wikipedia Community of Users

Felipe Ortega
Universidad Rey Juan Carlos.
Tulipan s/n.
28933, Mostoles. Madrid. SPAIN
jfelipe@gsyc.es

Jesus M. Gonzalez-Barahona
Universidad Rey Juan Carlos.
Tulipan s/n.
28933, Mostoles. Madrid. SPAIN
jgb@gsyc.es

Abstract

Many activities of editors in Wikipedia can be traced using its database dumps, which register detailed information about every single change to every article. Several researchers have used this information to gain knowledge about the production process of articles, and about activity patterns of authors. In this analysis, we have focused on one of those previous works, by Kittur et al. First, we have followed the same methodology with more recent and comprehensive data. Then, we have extended this methodology to precisely identify which fraction of authors are producing most of the changes in Wikipedia's articles, and how the behaviour of these authors evolves over time. This enabled us not only to validate some of the previous results, but also to find new interesting evidences. We have found that the analysis of sysops is not a good method for estimating different levels of contributions, since it is dependent on the policy for electing them (which changes over time and for each language). Moreover, we have found new activity patterns classifying authors by their contributions during specific periods of time, instead of using their total number of contributions over the whole life of Wikipedia. Finally, we present a tool that automates this extended methodology, implementing a quick and complete quantitative analysis of every language edition in Wikipedia.

Categories and Subject Descriptors H.3.7 [*Information Storage and Retrieval*]: Digital Libraries—system issues

General Terms Performance

Keywords quantitative analysis, methodology, Wikipedia, WikiXRay

WikiSym'07, October 21–23, 2007, Montréal, Québec, Canada.

1. Introduction

Wikis present a new paradigm of website with dynamic contents created by its own users. When we refer to content development systems, *wiki* is now synonymous of *collaborative*, *agile*, *powerful* and even *easy*. They are now a core component of what is presented as the Web 2.0, providing tools to create contents through collaboration and interaction of users.

Wikipedia is, by far, the most successful example of this new paradigm of web services. With more than 200 different language editions, and one of the biggest communities of users of the Internet, Wikipedia has showed us the power of collaborative content development. As of April 26th, 2007, a total of 1,755,932 articles are already available in the English language edition, and the top 10 language editions (English, German, French, Japanese, Polish, Dutch, Italian, Portuguese, Spanish and Swedish) accumulate a total sum of 4,852,810 articles. Many scientific works are increasingly referencing Wikipedia, and many learning systems employ it as a primary source of information.

One of the most relevant contributions that Wikipedia has made to the wiki community is the MediaWiki software, a libre (free, open source) software that facilitates the task of creating, configuring, maintining and using a wiki. MediaWiki provides some very useful tools for wiki users and administrators, for instance:

- *Easy-to-use editing interface*: A toolbar provides easy access to the most common editing functions. This reduces the learning curve of new users, who can get up to speed in editing contents fastly.
- *Content classification*: Contents can be classified attending to their topic into several categories, presented in special pages, thus helping users searching for a certain topic.
- *Discussion pages for every article*: Each article in MediaWiki is accompanied by its talk page, a special page allowing users interested in editing that article to exchange their impressions, and collaborate towards obtaining consensus about the article's contents and presentation.

- *Contents organization tweaks*: Some automatic features, like the creation of a table of contents for long articles (those with at least 4 main epigraphs) and automatic archive indexation and linking (for example, to archive previous discussions in talk pages), foster correct content arrangement and clarity.
- *Automatic back-up and recovery tools*: MediaWiki also provides automatic tools for administrative tasks like back-up and recovery from previous crashes, as well as updating the wiki contents with new dump versions (for example, in wikis whose contents are mirrors of another wiki).

Wikipedia sets up a content development philosophy very similar to the one we find in libre software projects. The edition of articles is completely open to any user who feels like doing so, no matter its level of knowledge about the topics included in the article. This rises some important issues regarding content quality and preservation:

- *Content quality and accuracy*: As we have already mentioned, Wikipedia classifies its articles in categories, to enhance content arrangement and clearness, making the search for a certain topic easier. In addition to this, a short time ago the English version of Wikipedia began the Wikipedia 1.0 initiative. A group of volunteers searches through the encyclopedia for articles with a high quality level, both in terms of accurate contents and correct presentation. The goal of this initiative is to retrieve enough articles to construct a *stable* version of the English Wikipedia, with trusted contents, that could be distributed in CD or DVD media.
- *Preservation of contents*: As Wikipedia is completely open to anyone interested in editting its contents, it must support many acts of vandalism from people whose main purpose is just to damage Wikipedia's contents or reputation. Despite these attacks, Wikipedia has proved a great resiliency against these issues. A fundamental point to fight against them is the role of *sysops*, Wikipedia users that have received special privileges to block certain users or IP addresses that have attempted to damage contents, among other important tasks that we will summarize shortly.

Another similarity that Wikipedia shares with libre software projects is the promotion of certain users to a privileged status, like *sysops*, in order to look after the good health of the system infrastructure. Special types of Wikipedia power-users include:

- *Administrator (Sysop)*: Administrators, usually also known as admins, or *sysops*, are special users with certain privileges that enable them to help with maintenance tasks. Among their main attributions, as we mentioned above, is the power of protecting pages from anonymous edits, deleting pages, blocking other editors, as well as undoing these actions. It is a common practice within admins to be elected by other users in the community according to their reputation, knowledge about Wikipedia infrastructure and policies, and proved contributions to the contents of the encyclopedia. In the English language version, admins acquire their new status in a permanent way, unless abuse actions recommend revocation of their privileges. We will see that this policy may vary among different language editions.
- *Bureaucrat*: Bureaucrats implement administrative decisions like promoting other users to administrator or *bureaucrat* status, granting or deleting a user's *bot* status, (*bots* are special users for automated programs that implement repetitive tasks), or renaming a user's account.
- *Steward*: Users with even higher privileges. They can grant or revoke virtually all kind of user access levels.
- *Oversight*: This users have the power of removing the history of revisions for a certain article. This revisions can only be restored by system developers.
- *Checkuser*: They can retrieve the IP address employed by a certain username, as well as all edits made by users with a certain IP, or within a certain IP range. Logs of the checkuser's activities are always available to the rest of the community.

As a consequence, we end up with a system with many of the ingredients to become a success: an easy-to-use interface, strategies for content preservation and efficient classification, and power-users that distribute privileges and watch for system and contents' health. The question now is, where do contributions to Wikipedia come from? Do most of the contributions come from power-users, with a deep and strong implication level in the project? Do they come from the group of average users, with just a few contributions per month?

In this paper, we look back to previous methodologies presented to answer these questions. We evaluate their results, which were obtained only for the English version of Wikipedia, and we point out some important parameters that may have been overlooked. We then extend this methodology to other language versions, showing why we need to consider new parameters to fully explain behavioral patterns in different communities of users. Finally, we present an enhanced methodology that shows us a more complete picture of contributions in Wikipedia, and we apply it to analyse the community of users in the English edition, as well as the Swedish and Norwegian editions for comparison purposes.

2. METHODOLOGIES FOR QUANTITATIVE ANALYSIS OF WIKIPEDIA

Although there is still very little research presenting quantitative analyses of Wikipedia, there are some previous works

we need to mention in order to give an accurate picture of the current state-of-the-art in this interesting field. In the following paragraphs, we offer a brief presentation of these previous investigations. Then, in the next section we focus on proposed methodologies to answer the question: where do contributions to Wikipedia come from? This is followed by a discussion presenting additional parameters that we should take into account. Finally, we present our own proposal for a improved methodology to undertake this challenging question.

2.1 Previous Methodologies

One of the first research works presenting quantitative analysis results about Wikipedia was conducted by Jakob Voss [7]. In this paper, he presents some interesting preliminary results about the evolution of contents and authors, mainly focusing on the German version of the Wikipedia: the number of distinct authors per article follows a power-law, while the number of distinct articles per author follows Lotka's Law. Buriol et al. [2] showed that the growth in the number of articles and users in the English edition were consistent with Voss' results. The authors find many similarities among several language versions of Wikipedia, as well as with the structure of the World Wide Web. This should be no surprise, because in some way, wikis are simply another flavor of websites where contents may be linked from other contents (using HTML hyperlinks).

Viegas et al. [5] found an alternative approach for studying contribution patterns to Wikipedia articles. They have developed a software tool, *History Flow*, that can navigate through the complete history of any article. This way, it is possible to identify periods of intense growth in the content of articles, acts of vandalism and other interesting patterns in users' contributions. In a more recent paper [6], Viegas et al. use their software tool to describe the collaborative process behind articles creation, how users employ Wikipedia's talk pages for reaching consensus and demanding additional contents and the periodic process of archiving previous discussions. In this paper, we can also get a graphical demonstration of the impressive growth rate of Wikipedia, looking at the history graphs for some popular articles.

The first study that tried to answer the question about what is the major source of contributions in Wikipedia was [1]. In this paper, the authors measured the quality of contributions by the percentage of aggregated contents that remains in the subsequent revisions of a certain article. They concluded that two main groups of contributors are responsible for most of the high quality contributions. On one hand, high quality contents come from *zealots*, registered users with a strong interest in reputation and a high level of participation. On the other hand, we have *good Samaritans*, that is, anonymous users with a low level of participation in the project. They also show that there is a strong correlation between the quality of contributed contents and the level of contributions made by individual authors. In the group of registered users, the greater the number of contributions per user is, the better the quality of those contents is too. If we turn to the group of anonymous users, quality contents come from users with a smaller number of contributions, and that quality decreases as the number of contributions per user raises.

Finally, we find a more serious attempt to establish a formal methodology for measuring these parameters in a recent paper by Kittur et al. [4]. In this work, the authors try to confirm the theory that, after an initial period in which contributions usually came from registered users with a high level of participation, these days Wikipedia receives the majority of its contributions from those users with a very low level of participation. Therefore, these authors argue that the reason for the current growth of Wikipedia is what they call the *wisdom of the crowd*, or in other words, *the rise of the bourgeoisie*.

In the next section, we will revisit these results, creating similar graphs to those appearing in that paper, but with a new set of data. This will lead us to some interesting conclusions about who are really making the majority of contributions to Wikipedia.

3. REVISITING THE RISE OF THE BOURGEOISIE

In their previous research paper, Kittur et al. propose a simple method for classifying Wikipedia's users, applying it to the English edition of the encyclopedia. Results were generated using the history dump created on July 2nd, 2006, with 4.7 million wiki pages, and 2.4 million article related entries. They classify Wikipedia's users into 5 different categories, according to the total number of contributions each user made to the encyclopedia up to that point: more than 10,000 edits (10k+); between 5,001 and 10,000 edits (5-10k); between 1,001 and 5,000 edits (1-5k); between 101 and 1,000 edits (100-1k); and 100 or fewer edits (< 100). They also generate some graphics isolating contributions that come from administrators of the English Wikipedia (that is, *sysops*), comparing these data with the contributions generated by the 5 categories we mentioned above.

Before extending the methodology, we decided to reproduce these results, though using a more recent dump (November 4th, 2006) of the English edition. For this reason, our graphics show some additional data that could not be considered in this previous research. We also retrieved a more recent dump (April 4th, 2007) with the database table containing users' privileges (table *user_groups* in MediaWiki), to recreate graphics concerning *sysops*. We present the evolution in time of these results in monthly periods, starting from the first month of existence of that language edition (that is, in the English edition period 0 corresponds to January 2001, period 1 corresponds to February 2001, and so on). As the database dumps did not include the complete history of the last period (period 70, corresponding to

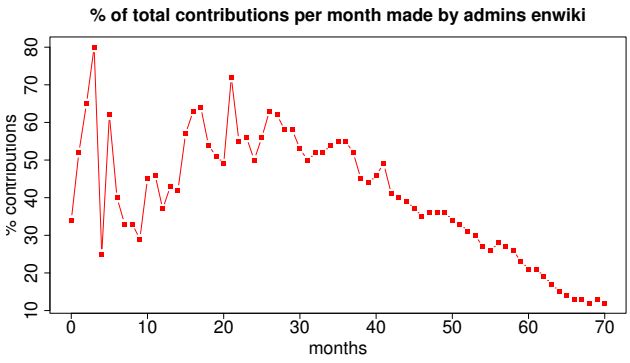

Figure 1. Percentage of total number of contributions per month made by *sysops* in the English edition of Wikipedia.

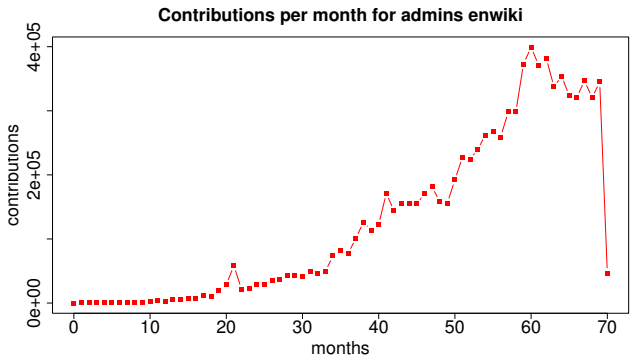

Figure 2. Total number of contributions per month made by *sysops* in the English edition of Wikipedia.

November 2006), the graphics we show in this article present a strong dropoff in this period. So that, results within period 70 should not be taken into account. It was not possible to manually eliminate these remaining data without affecting the automatic tool we have developed for this quantitative analysis, which we will describe shortly in the final section.

For this new sets of graphics, we have filtered contributions made by *bots*, special programs created to automatically load new contents retrieved from publicly available information sources, and to implement some maintenance tasks as controlling vandalism and spam. We only take into account contributions made to Wikipedia's pages that fall in the *main* namespace category, corresponding to articles, including *redirects* (articles linking to the main denomination of a certain topic) and *stubs* (articles with few or very poor contents that still need to be improved). We should also precise that, for each month, we only count those users (and admins) who made at least one edit in that month.

Figure 1 reproduces the percentage of total edits per month made by *sysops*. We should remark that, in the English edition of Wikipedia, *sysops* never leave their special status once they have reached it. At first sight, we can detect a new peak at the beginning of the graph (period 3) that did not appear in the reproduced study. The reason for this is that we are considering a broader population, since we are using a more up-to-date version of the list of *sysops*, and specifically some new users who made almost the 80% of the total number of contributions during that month. We can see that the percentage of contributions made by *sysops* still tends to decrease, from period 36 (January, 2004) to present time. We can also see an overall increment in the percentage of total edits per month made by admins with respect to the results presented in the previous work, due to the higher number of *sysops* we consider this time.

Even more interesting is the recreation in Figure 2, showing the total number of contributions per month made by admins. In the reproduced study, the authors pointed out a drop in the number of edits made by admins, starting from period 60 (January, 2006), that seemed to be merely produced due to some new admins that were not included in the list yet. We need to point out that in the English Wikipedia, admins are selected through a peer review process. Some users are proposed to become *sysops*, and then the community makes positive, negative or neutral votes in response to each proposal. For this reason, there is an inherent start-up time associated with the process of becoming admin, as Kittur et al. already explained.

Nevertheless, the new graph shows that maybe that is not the motive for this behaviour. In the additional periods that we analyse now, there is a clear trend towards stabilization of the number of edits made by admins, and now we can be sure that we are considering all the possible users that reached the *sysop* status. The most reasonable hypothesis now is a switch in the behaviour of admins, that cease to contribute to articles at the same growing rate than before.

We now turn to the classification of users in the English Wikipedia based on the total number of contributions they made so far, where Kittur et al. presented the so called *rise of the bourgeoisie*. They showed that the mass of users with a lower number of contributions per month is responsible for most of the contributions received by the English Wikipedia in the last months.

In fact, this seems to be the case looking at the graphs we show in the following figures (created from our own data, but applying the methodology from the reproduced study). In Figure 3 we can see that the percentage of the total number of contributions per month made by users with less than 100 edits is increasing steadily. Figure 4 presents the number of contributions per month for each editing level. In this Figure, as well as in Figures 5 and 6, we take the log10 of the total number of contributions. This is perfectly consistent with the reproduced study, since the number of contributions per month made by users with less than 100 edits is growing at a much faster rate than any other group. At the same time, the average number of contributions per user in each editing level, depicted in Figure 5, shows that the values per month for each group have remained very stable over time. Finally, Figure 6 depicts the evolution in time of the population in

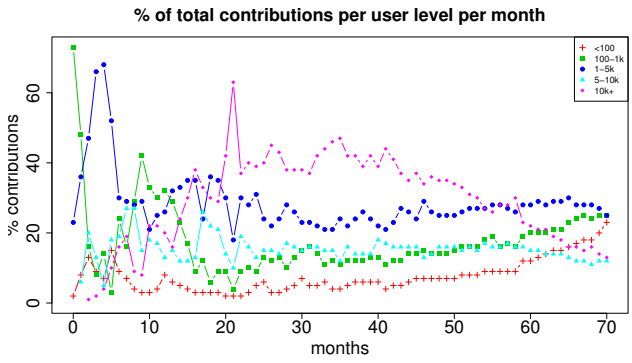

Figure 3. Percentage of the total number contributions per month made by users with distinct editing levels in the English edition of Wikipedia.

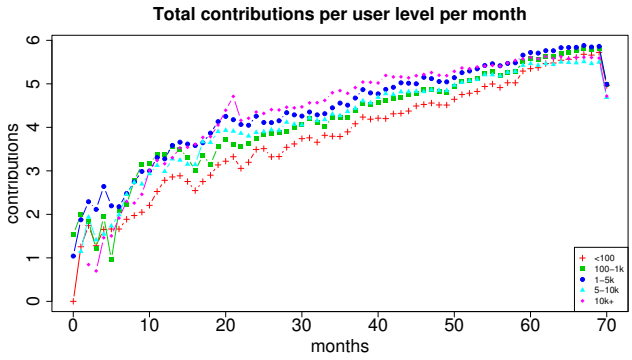

Figure 4. Number of contributions per month made by users with distinct editing levels in the English edition of Wikipedia. (log-scale)

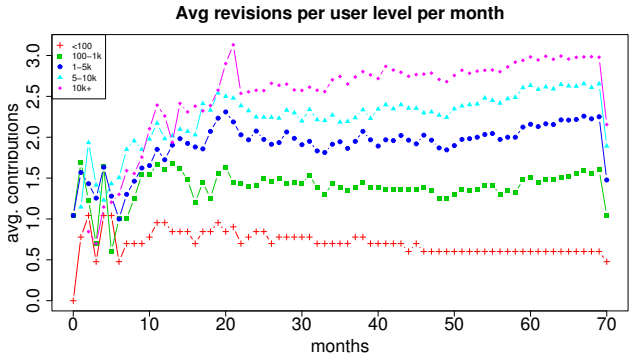

Figure 5. Average number of contributions per user in each editing level per month in the English edition of Wikipedia. (log-scale)

each user group. As we can see, most of the groups shows an exponential growing rate, except for the group of users with a total of more than 10,000 edits. Again, the authors explained this leverage due to the fact that new admins were not taken into account. But the reproduced figures, that consider the whole population of *sysops*, reflects that this tendency is not a mere artifact, but a behavioural change. Figure 7 shows that the percentage of users in the group with less than 100 edits is steadily rising, a result that is consistent with the previous data presented up to this point.

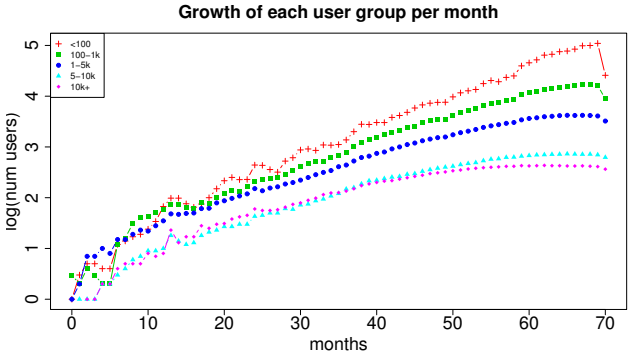

Figure 6. Evolution in time of the population in each user group in the English edition of Wikipedia. (log-scale)

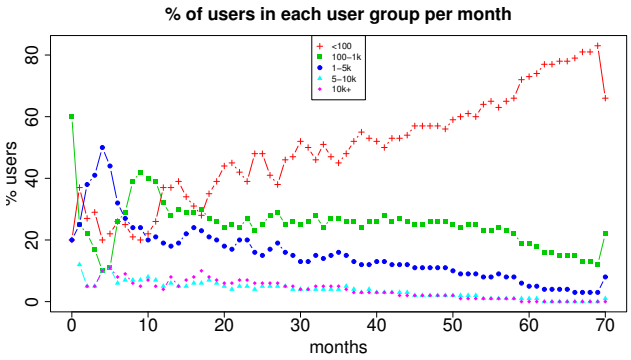

Figure 7. Percentage of users in each user group per month in the English edition of Wikipedia.

4. EXTENDING PREVIOUS METHODOLOGIES

So far, we have reproduced some of the most relevant results found in previous methodologies, proposed for the quantitative analysis of Wikipedia. In this section, we will extend these previous methodologies to the quantitative analysis of other language editions of Wikipedia. We will also enhance these previous methodologies, quantifying per period parameters. That will lead us to get a more accurate picture of the evolution in time of the Wikipedia's community of users, as well as to show some interesting conclusions that can be extracted from this extended analysis.

4.1 Extending the previous methodology to other language versions

As we have seen in the previous section, some interesting behaviors were hidden in the previous versions of these graphs, due to the fact that they could not consider data we have included in this recreation. But there are some additional points that worth a more in-depth discussion.

First of all, we consider the analysis of data from the *sysops* population. The graphics above have shown that, in the English edition of Wikipedia, contributions that come from admins are steadily decreasing, supposedly in favor of the contributions from the group of users with less edits per month.

However, it would be interesting to know how many of the current admins were included in each user group per month. This information is presented in Figure 8, where we show the number of current admins that falls in each user group per month. Hence, we can see in this graphic the evolution of the current admins population through the different classes of users. Clearly, the three most contributing groups provide the majority of *sysops* in the English Wikipedia. However, we should metion that the number of contributions to articles in the main namespace is not the unique measure considered for electing *sysops*, though it is very common to require that the *sysop* candidate have at least a few hundred edits.

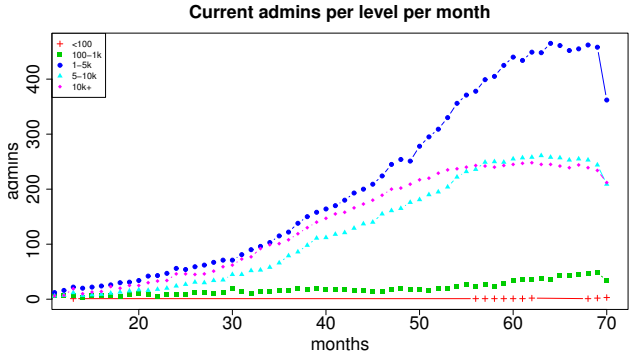

Figure 8. Number of *sysops* in each user group per month in the English edition of Wikipedia.

So, we could ask the following question: how many users that we now include in the group with less edits will become admins in the near future? We cannot give a definite answer for that question, because nobody can predict future, but from historic results we can definitely figure out that many of them will not reach that status. Furthermore, if we compare Figure 8 and Figure 6, we can see that many authors falling in the three most active groups of users have not been selected as admins yet. For this reason, we cannot consider that a dropoff in the percentage of contributions per month made by admins (presented in Figure 1) reflects the behavioral pattern of the whole population of most active users, since many of them are not *sysops*.

Another aspect we should consider is how extensible this methodology is when we apply it to the study of other language editions of Wikipedia. If we take, for instance, the Swedish edition, we can see a different behaviour for *sysops*, in Figure 9 and Figure 10. While the total number of contributions per month made by *sysops* follows a very similar pattern to that of the English Wikipedia, the percentage of the total number of contributions per month from admins is quite distinct. Figure 9 shows that this percentage has remained very stable, around the 40% of the total number of contributions, from period 17 (May, 2004) onwards, in contrast to the steadily decreasing trend exhibited by the English Wikipedia's *sysops* in Figure 1.

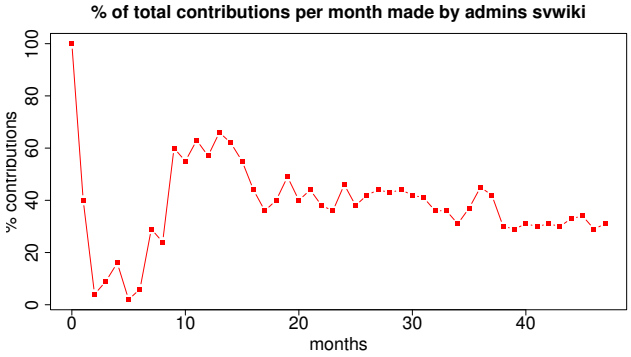

Figure 9. Percentage of total number of contributions per month made by *sysops* in the Swedish edition of Wikipedia.

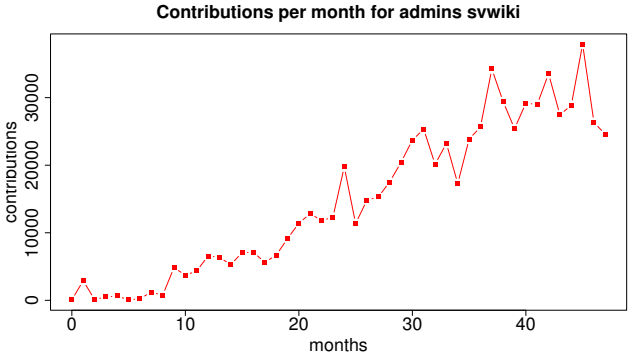

Figure 10. Total number of contributions per month made by *sysops* in the Swedish edition of Wikipedia.

This behavioural change may be due to a radically different approach for selecting admins [1]. In this language edition, admins do not acquire a permanent privilege, but they rather must be re-elected every year, according to the work they have developed. Therefore, the figure shows us a somewhat constant level of effort maintained by this group, in order to retain the admin status. That creates a framework that favors a higher number of contributions per user than in any other

[1] http://en.wikipedia.org/wiki/Swedish_Wikipedia#History

language edition. In the Swedish Wikipedia there is no *rise of the bourgeoisie*, or at least not yet.

As we can see in Figure 11, the percentage of total contributions per month made by users in the most active groups is still above the level exhibited by users with fewer total contributions, although it seems to be a slight trend towards changing since period 30. As a result, regarding admins anal-

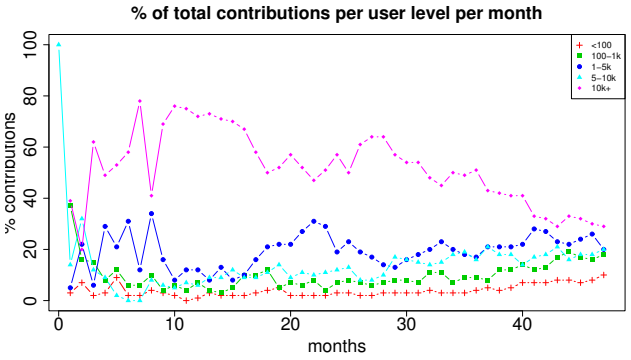

Figure 11. Percentage of total contributions per month made by users with distinct editing levels in the Swedish edition of Wikipedia.

ysis we cannot make a general assumption about their impact in the contributions made by a certain community. On the contrary, we should take into account special conditions like, for example, the way that community of users selects admins and whether or not these special users maintain their status of privilege indefinitely. We should carefully revise these parameters when we turn to the analysis of a different language version, because they may even affect behavioural patterns of the whole community of users.

Another factor we should remark is a generational change that may have been overlooked in the previous research; many language editions presents, like English, a tendency in which the population of users between 5,000 and 10,000 edits in total has surpassed the population with more than 10,000 edits, as we saw in Figure 6. This presents us the hypothesis that, in short time, these users will leave this group to be included in the more than 10,000 edits class, since the average number of edits per user per month is approximately constant.

This evolution in the composition of the community of users may remain hidden in subsequent presentations of this methodology, as long as we use the aggregate number of edits to classify users. The total number of contributions made so far by a certain user through the whole history of a language edition may shadow very active and young users, with a higher number of contributions per month in the recent history, that get lost in the mass of users with less overall number of edits. Besides that, following the previous methodology we do not know, for example, how many of the less active users made one or two edits to Wikipedia in a certain month, and then never came back again to contribute. The effect of the group of users with less than 100 total edits is worth to be analysed, but we should also take into account period by period measurements, to be sure that we can get a precise picture of the whole edition process. That leads us to our proposal for an enhanced methodology to automate the quantitative analysis of all Wikipedia language editions with our own software tool, *WikiXRay*.

4.2 Per Period Analysis of the English Wikipedia's Community of Users

Trying to expand our knowledge about how the community of most active editors evolves over time, we have extended the previous methodology analysing users according to their activity level during specific periods of time (instead of considering the whole history of a certain edition of Wikipedia). This avoids the influence of duration, and location, of the active period of time for a contributor in the way she is classified. As we discussed in the previous subsection, the consequence of using the previous methodology is that some new editors are classified in low-contributing populations despite being very active, just because they had not enough time to contribute significantly yet.

As a result, following the same presentation approach we have split the history of edits in the English Wikipedia in monthly periods, considering the first one (January, 2001, for the English edition) as period 0. Then, for each language edition we have analysed the number of contributions per month from the most active editors, with an special interest in how stable is the group of top contributors. We have considered two criteria for identifying the most active editors:

- We sorted editors according to their number of contributions per month. Then, we identified the 5% of editors with the highest number of contributions in each month.
- Later, we considered again the list of editors sorted by their number of contributions per month, but this time isolating the most active editors responsible for the 10% of the total number of changes in each month.

We should recall at this point that these data has been obtained from the database dumps containing the complete edit history of each Wikipedia page, filtering out any page that does not belong to the *main* namespace. Therefore, we only consider to compute these results those users that have made at least one contribution in each month.

In the rest of this section, we show how both populations have a similar behaviour, and how, after an initial period of instability, authors in both populations remain active during long periods of time, suggesting that the core of Wikipedia's top contributors is stable over time.

Figure 12 plots the edit history of the 5% of authors with the highest number of edits in each month in the English Wikipedia. This 3D graph has been created as follows: the x axis indicates the period in which we identify the group of the 5% of editors with the highest number of edits in

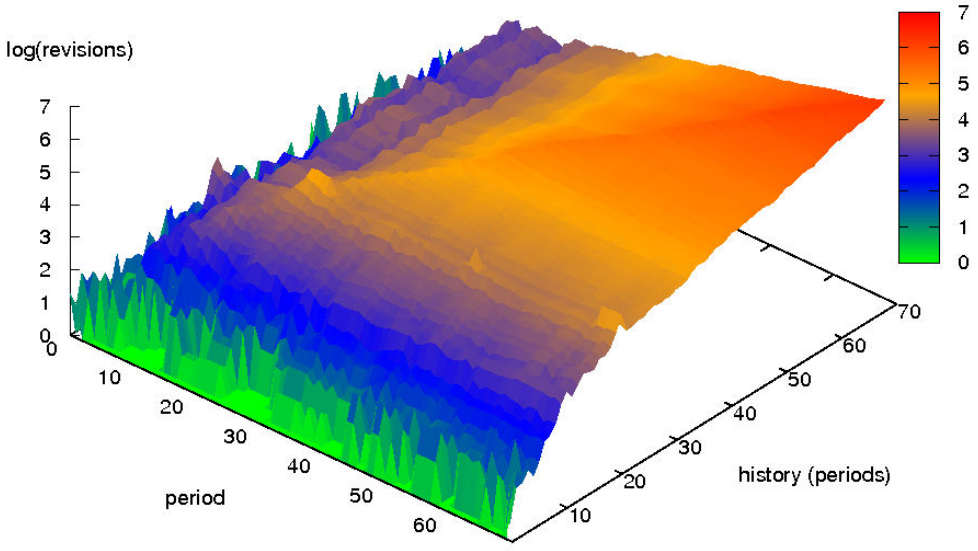

Figure 12. History of the activity in the remaining periods of the 5% of editors with the highest number of contributions in each period (English Wikipedia).

that month. Then, we plot on the y axis the edit history of that group of editors in the remaining periods of the English Wikipedia. On the z axis, we plot the total number of contributions quantified for each group in every month. For instance, for the group of 5% of authors with highest number of edits in period 40 (x axis) we trace the complete history of their edits in the remaining periods (y axis). We have taken the log10 of the total number of edits (z axis) to ensure that recent periods, with a higher number of edits, do not hide results for previous months. As we can see, there is a group of very active editors for each month that has also been very active in other months. The graph does not show any significant valleys, demonstrating that the community of users with the higher number of edits in each month presents a quite constant behaviour over time for the English Wikipedia. One of the parameters we are interested in is the level of inequality that can be found for contributions. Analyzing inequality will allow us to see if both phenomena present similar patterns.

We will measure inequality by means of the Gini coefficient. This coefficient, introduced by Conrado Gini [3] to measure income inequality in economics, shows how unequal something is distributed among a group of people. To calculate the Gini coefficient, first we have to obtain the Lorenz curve, a graphical representation of the cumulative distribution function of a probability distribution. Perfect distribution among authors is hence given by a 45 degree line. The Gini coefficient is given by the area between the two curves, providing how far the actual distribution is from perfect equality. Figure 13 presents the Lorenz curve for the English edition of Wikipedia, considering editors that made at least one edit in that language edition. As we can see, approximately 90% of the active editors is responsible altogether for less than 10% of the total number of contributions, (Gini coefficient of 0.9360). Hence, we find a small group of very active editors maintaining a high activity level throughout the whole history of the English Wikipedia.

We show in Figure 14 another interesting view quantifying, for the English Wikipedia, the edit history in the remaining months of the most active editors in each month who accumulate the 10% of the total number of edits in that period. The graph has been created following the same presentation method explained above. We also take the log10 of the total number of contributions to plot values in the z axis. Again, we can see that the behavior of this group of users is fairly similar in periods other than the one in which they reached the group of top editors with the 10% of the total number of contributions in that month. Nevertheless, in this graphic we appreciate some valleys towards the early edit history of these users, between periods 20 to 40. This fact indicates that those users did not participate with such a high editing level in the early history of the English Wikipedia.

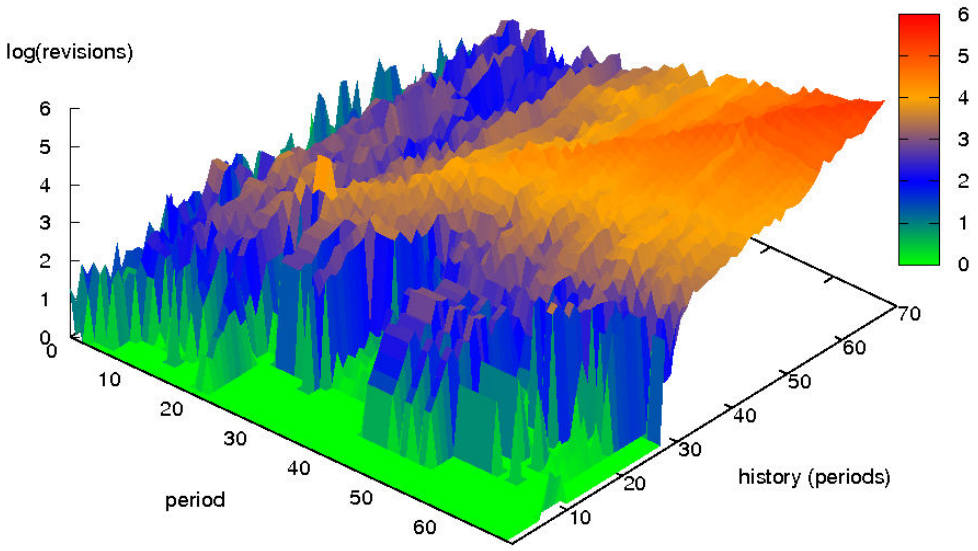

Figure 14. History of the activity in the remaining periods of the most active contributors in each period accumulating the 10% of the total number of edits in that period (English Wikipedia).

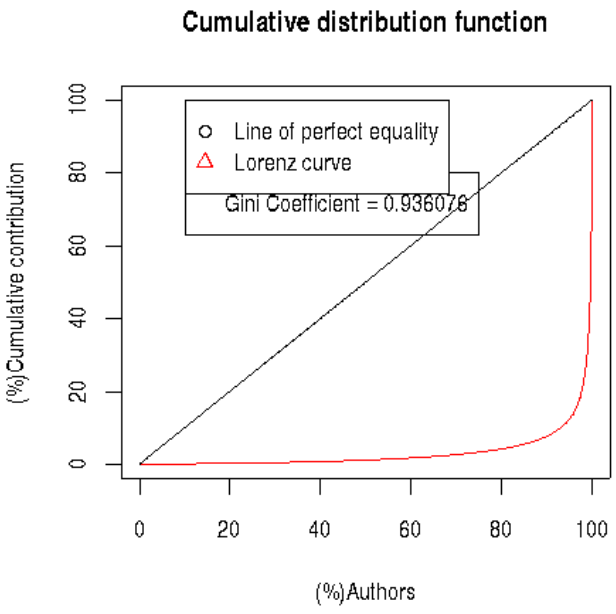

Figure 13. Gini graph for cumulative contributions from registered editors with at least one contribution per month (English Wikipedia).

To evaluate the average profile that a certain user must present to be considered under this analysis, we summarize in Figures 15 and 16 the number of authors in this group for each period, and the minimum number of edits that a certain user must reach to be included in the analysis for each period, respectively. As we can see from these graphs, though the number of most active users in each month accumulating the 10% of the total number of contributions in that month is steadily rising, (most notably from period 51 onwards), the evolution of these groups of users presents periodic falls that worth a more in deep analysis. The number of contributions that a certain user must reach in each period to get into this group presents a disproportionated maximum value in period 21 that also deserves further analysis. We suspect that this outlier is due to a *bot*, unregistered as such in the database dump.

Finally, in Figure 17 we show that this methodology can be extended to other language versions of Wikipedia as well, depicting the same graph as in Figure 14, but this time for the Norwegian edition of Wikipedia. This is an example of a medium size Wikipedia, edited in a language that not many users employ outside its own country. It is just the opposite situation of the English edition, which can receive contributions from a broader community of users all over the world. Here, we take the natural logarithm to plot the number of contributions in the z axis, instead of the log10 previously

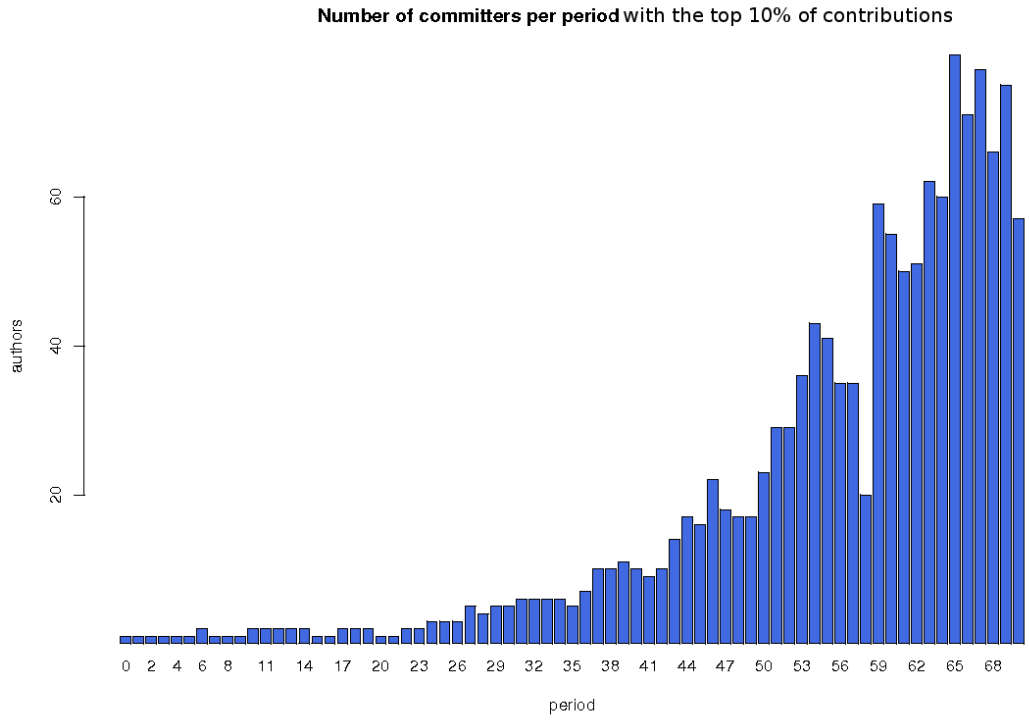

Figure 15. Number of most active editors in each month who concentrate the 10% of the total number of edits in that month (English Wikipedia).

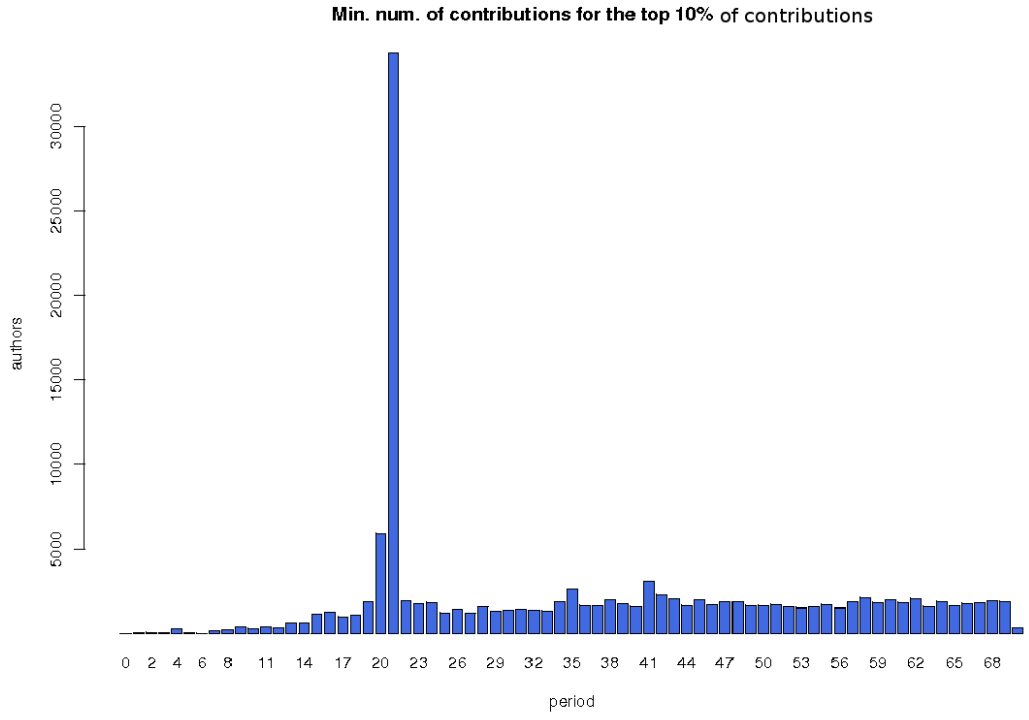

Figure 16. Number of revisions to enter the group of most active editors in each month who concentrate the 10% of the total number of edits in that month (English Wikipedia).

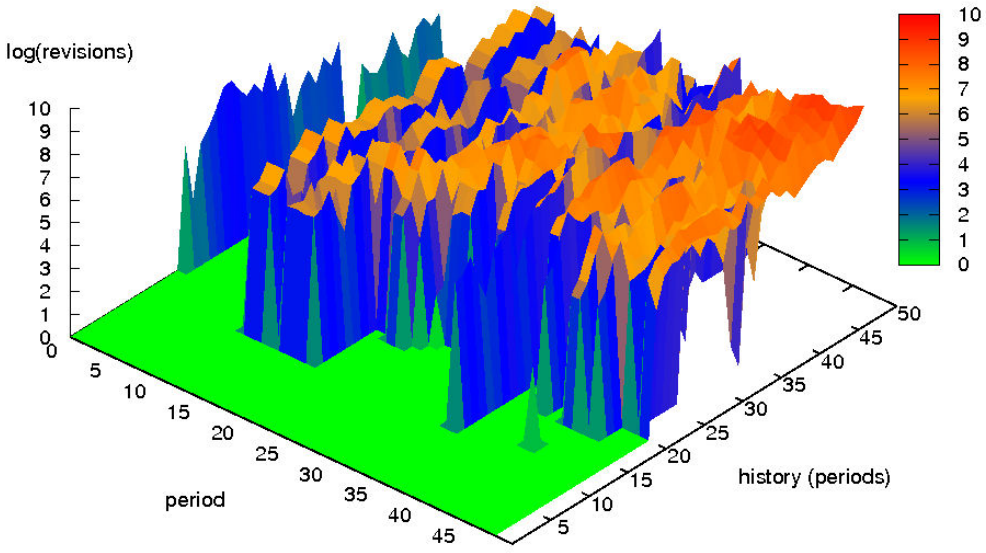

Figure 17. History of the activity in the remaining periods of the most active contributors in each period accumulating the 10% of the total number of edits in that period (Norwegian Wikipedia).

employed. This time, we can see remarkable valleys in the recent history, showing periods of clear inactivity of the most active users in each month responsible for the 10% of the total number of contributions in that month. Thus we can infer that these are younger users that did not contribute to this language edition during its early history. Most notably, we detect for those editors an interval of almost complete inactivity from periods 2 to 10. These young editors could not have been identified following the previous methodology proposed by Kittur et al., as they would have been included in the group of less active editors, since we were classifying contributors regarding their total number of edits troughout the whole history of a certain language edition.

4.3 Automating Quantitative Analyses

WikiXRay[2] is a Python software package licensed under the GNU GPL, whose main goal is to provide a set of tools for automating the quantitative analysis of all Wikipedia language editions. In fact, this tool could be used in any wiki based on MediaWiki.

Some features of *WikiXRay* can be applied to the quantitative analysis of the Wikipedia's community of users, in particular:

1. *WikiXRay* has reproduced previous methodologies for the quantitative analysis of the different language editions of Wikipedia. We have use this capability to create the graphics presented in this research paper.
2. As well as reproducing previous methodologies that use aggregated contributions to classify users, *WikiXRay* is capable of classifying users by their number of edits per period. Currently, we consider monthly periods, but it is fairly straightforward to switch *WikiXRay*, so we can take periods by quarters, semesters, etc. This way, we can have a more accurate picture about what is really happening in each month, rather than trusting solely in aggregated data to show behavioral patterns.
3. *WikiXRay* can depict 3D graphs showing the evolution for other different periods of relevant groups of users within an individual period. We can show, for example, the historical behavior in the remaining periods of the most active editors in period 40, (measured from the first month of history), concentrating the 10% of the total number of edits in that period. This feature is very useful to verify if certain behavioral patterns are just local or they can be extrapolated to the whole history of that language edition.
4. *WikiXRay* allows us to easily extend this methodology to other relevant parameters like articles. For example,

[2] http://meta.wikimedia.org/wiki/WikiXRay

instead of focusing on the most active contributors in period 40 concentrating the 10% of the total number of edits in that period, we can analyse the most popular articles in period 40, (those obtaining the highest number of contributions), that receive the 10% of the total number of edits in that period. We could then use those data to correlate them with the number of distinct authors per article, the length of those articles, etc.

We have applied our tool to the quantitative analysis of the English edition of Wikipedia, trying to validate some of the results and previous methodologies we have presented so far. We have also used it to extend those previous methodologies to other language editions of Wikipedia and to include per period parameters in the quantitative analysis process.

5. CONCLUSIONS AND FUTURE WORK

In this research paper, we revised previous methodologies proposed for the quantitative analysis of Wikipedia. We showed that, although these methodologies reveal some interesting behavioral patterns, they can also hide another important phenomena that could lead us to acquire a more complete and detailed picture of the Wikipedia's community of users.

Later, we presented our own proposal for enhancing these previous methodologies to undertake the quantitative analysis of Wikipedia. This extended methodology can be applied to further language editions other than English, and it focuses on the study of per period parameters to precisely follow the evolution in time of the Wikipedia community of users.

We finally showed some graphs summarizing the most relevant parameters that we can study with this new methodology using *WikiXRay*, a Python software tool that automates the quantitative analysis of all Wikipedia language editions. Further development of *WikiXRay* will include extending this per period analysis method to Wikipedia articles, and reflecting possible correlations between featured parameters that will let us better explain the behavior of the communities of users in all language editions of Wikipedia.

References

[1] D. Anthony, S. W. Smith, and T. Williamson. Explaining quality in internet collective goods: Zealots and good samaritans. the case of wikipedia. November 2005.

[2] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal evolution of the wikigraph. In *Proceedings of the Web Intelligence Conference*, December 2006.

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

[4] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. 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)*, April-May 2007.

[5] F. B. Viegas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582, April 2004.

[6] F. B. Viegas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, page 78a. Computer Society Press, January 2007.

[7] J. Voss. Measuring wikipedia. In *Proceedings of the ISSI 2005*, July 2005.