Community Management through Meta-Data in Wikipedia

Matthijs den Besten
Ecole Polytechnique
Centre de Recherche en Gestion
Chaire Innovation et Régulation des services numériques
32 boulevard Victor
75739 Paris Cedex 15, France
+33 1 45 52 64 27
matthijs.den-besten@polytechnique.edu

Loris Gaio
University of Trento
Department of Computer and Management Science
Via Inama 5
I-38100, Trento, Italy
+39 0461 882142
lgaio@cs.unitn.it

Alessandro Rossi
University of Trento
Department of Computer and Management Science
Via Inama 5
I-38100, Trento, Italy
+39 0461 882101
arossi@cs.unitn.it

Jean-Michel Dalle
Université Pierre et Marie Curie
4, Place Jussieu
75252 Paris cedex 05
France
+33 1 44 18 07 15
jean-michel.dalle@upmc.fr

DRAFT

ABSTRACT

We argue that templates in Wikipedia represent a variety of meta-data which are used as a means to coordinate collaborative work. We suggest that such “management through meta-data” might be an important coordination mechanism for online communities.

KEYWORDS

Peer production, coordination, Wikipedia, template messages, bugs, exploration, exploitation.

INTRODUCTION

The dominant trend among organizations in the past two decades or so has been towards specialization and disintegration (Langlois, 2001). The advent of the Internet and other information and communication technologies has helped to make this shift possible (Malone and Laubacher, 1998), but the hardware in itself is not enough. In addition, we can observe a growing reliance within firms on textual media as a means of coordination (Greenman, 2005). Consider the enormous success enjoyed by post-it notes in the office environment. First introduced in 1978, post-it notes have become one of the most successful and widespread products for the office in recent history. Post-it notes serve as temporary bookmarks, as password reminders stuck on the monitor, and also as placeholders for short memos attached to dossiers which are passed around in the office. In fact, it should come as no surprise that looser organization coincides with denser documentation as text is the main form of indirect communication. The challenge for firms is to find the correct mechanisms for indirect communication, because their ability to disintegrate and open up ultimately depends on the coordination power afforded through such communication.

Meta-data are small relatively standardized pieces of information which are added on top of other pieces of information. In the context of distributed problem solving, one could interpret meta-data as the coordination layer on top of the object that is being collectively manipulated. For instance, if we consider the development of software, a feature is the object that is being produced collaboratively, and a feature request is a collection of meta-data describing that feature. Another instance is represented by software bugs which are accompanied by a bug reports. Many relevant open source software projects invest heavily in capturing these meta-data and the management of development activities consists to a large extent in manipulating these meta-data. Consider, for instance, the case of Firefox, the open source Internet browser whose bug tracking systems prominently present bugs with a high number of duplicates while ignoring bugs with a high number of votes. This example show that it can be quite tricky to manage efficiently meta-data. A similar example is offered by a study of determinants of the acceptance of requests for comments (RFC) by the Internet Engineering Task Force (IETF), which found that RFCs whose authors were well-known had a higher chance to be accepted than RFCs with relatively obscure authors (Simcoe et al. 2008). Moreover, the study found that it was not the actual authorship that mattered but rather whether the author had been mentioned in the summary information on the RFC that was distributed by email (alternatively the author could disappear under “et al.”). Hence, this study shows that in some distributed settings it should be expected that the participants’ attention span is quite limited. Where the attention span of participants is longer, such as in the case of the innovation challenges set by the Innocentive Web-based community, the formulation of the problem and by extension the format and quality of the meta-data are so critical that a business could be created around the expertise in such activities of problem formulation (Sieg et al., 2009).

Wikipedia, the large online encyclopedia that anyone can edit, makes use of something very similar to post-it notes: on many of its pages one can find elements known as “templates” which serve to mark-up the page with short pieces of information. Occasionally these meta-data are purely informative, but in many cases they note a particular defect in the page and make a general request to redress it. Often, the templates contain a standard infobox, with minor variations meant for addressing specific issues. We analyzed in some detail the editing activities associated with two frequent template instantiations: the “unsimple” (recently relabeled as “complex”) template, which indicates readability concerns in articles of Simple Wikipedia (a special Wikipedia targeting non native English reader); and the “NPOV” template, which indicate a perception of neutrality bias in articles of the main English Wikipedia. Even though for a sizeable minority of the articles templates seem to be ignored, usually their effect is palpable. One could say that in response to the notification of a defect the community of editors working on the article change their editing activities and only revert to normal after the defect notification has been removed. It is almost as observing a bunch of honeybees, whose flying patterns become far more directed after they have notices one of their colleagues perform a waggle dance pointing them to a new food source. Like bees, the editors of Wikipedia are autonomous agents who have limited resources for coordination and communication but are nevertheless quite responsive to unambiguous and short requests and, while the waggle dance focuses the efforts of bees on a particular food source whose whereabouts are indicated by the shape of this dance, the template focuses efforts of editors on the particular type of editing which is suggested by the template.

The waggle dance of bees is an example of insect organization that can be classified as stigmergic. We think that in general many insights about self-organization can be gained from entomology. Like the keepers of a beehive, firms who engage in network innovation typically will have to manage relations with a large number of patrons in their network. Yet, unlike agents of the in-house innovation model, the patrons in a network are hard to control. Moreover, the firm cannot assume that its patrons invest many resources to understand what it is exactly that the firm wants. So, this is where simple devices like the Wikipedia templates have an important role to play. For the firm engaged in networked innovation in general, the implication of our study is that relatively low-cost mechanisms for coordination and communication like virtual post-it notes can be very important for the effectiveness of the network and that by facilitating and monitoring such devices the firm can have a better idea where its innovation is heading.

Below we first devote a few more words to the theoretical lens that we borrow from entomology. Next we introduce Wikipedia. In separate seconds we describe our studies of the “unsimple” template and the “NPOV” template, respectively, before we finish with a discussion and conclusion.

1. BACKGROUND

1.1 Stigmergy & Organization

The emergence of online communities has fostered a new interest in distributed problem solving. Most of this interest has been focused on how to extract information or solutions from various and notably peripheral problem-solvers (Lakhani et al., 2007). This approach, if very promising, partly leaves aside the enormous coordination efforts that are crucial to all production activities undertaken by online communities. Indeed, what is new and amazing about online peer production (Benkler, 2006) is not only that it attracts motivated participants who together provide solutions to distributed problems, but also that the more or less “spontaneous” coordination (Crowston, et al., 2005; Malone and Crowston, 1994) of those efforts gives birth to efficient software products composed of millions of lines of code or to a widely used encyclopedia with millions of pages such as Wikipedia. In this respect, recent works have suggested that some characteristic features of ongoing collective endeavors and of their organization could serve as *coordination signals* (Dalle and David, 2007; Den Besten et al., 2008), which would foster and trigger the allocation of decentralized efforts.

Moving one step further along this road, we are led to consider how online communities can *consciously* make use of coordination mechanisms in order to manage the editing activities. Wikipedia has precisely implemented such a mechanism through the use of template messages which appear as labels or tags on wiki-pages. By studying how these tags influence the coordination of work activities within online communities, we can gain further insights about how distributed problem solving works and about how it can be made more efficient within online communities by *signaling* problems in a way that is adapted to the self-organized nature of those communities, and more precisely to its stigmergic aspects. Essentially, tags can attract the attention of potential contributors, whose problem-solving efforts are therefore *oriented* in a direction that is particularly appropriate to collective peer production. For instance at times it can be important to switch from exploration of an environment to its exploitation and in the context of Wikipedia sometimes it might be necessary to switch from pure editing to conflict resolution.

Three examples of self-organization among social insects are worth to mention as far as stigmergy is concerned: the construction of termite-hills, path-finding by ants, and the honeybee waggle dance as a guide to sources of food. Termites and ants are guided by pheromones, i.e. chemicals excreted by other termites and ants of their colony. Termites' marginal choice where to add to the termite hill is guided by pheromone and ants' choice of which path to follow is equally determined by the intensity of pheromone. Termites' behavior inspired Dalle et al. (2009) to model the construction of open source software as the result of many developers' choices where to invest effort based on characteristics of the existing code-tree. The ant's behavior is in a way reminiscent of Lih's (2004) finding that topics on Wikipedia typically improve a lot after they have been in the news. Also Schroeder and den Besten's (2008) study of collective annotation of a novel suggest that collaboration works better if everyone is on the same page. While termites can be said to be guided by signals they find on the object under construction, and ants by signals left by other subjects, the waggle dance of bees introduces an element of intentionality. As such can be said to represent the form of stigmergy with the highest expressive power. The aim of our study of Wikipedia is to determine the extent of managerial leverage which this extra expressive power affords.

1.2 Wikipedia

Wikipedia is a collective endeavor that has managed to engage a very large group of people in order to construct an encyclopedia. Wikipedia is far from being the first attempt to rely on extensive external interactions. For instance, the Oxford English dictionary made heavy use of volunteers, as did the French government after the revolution when it enlisted hairdressers to compile logarithmic tables. Where Wikipedia is innovative, and where firms can learn, is in the structures and techniques that it has devised to steer the community of reader-editors that constitute its patrons.

Issues pertaining to the reliability of open content collections are at the core of the agenda of both scholars and practitioners interested in commons-based peer production. As put forward by Larry Sanger, Wikipedia co-founder:

> “It's fun, first of all. But it can be fun for intellectually serious people only if we know that we're creating something of quality. And how do we know that? The basic outlines of the answer ought to be fairly obvious to anyone who has read Eric S. Raymond's famous essay on the open source movement, ‘The Cathedral and the Bazaar’. Remember, if we can edit any page, then we can edit *each other's work*. Given enough eyeballs, all errors are shallow. We catch each other's mistakes and enjoy correcting them.” (Sanger, 2001)

Others have been more agnostic regarding the possibility of large mass peer screening to act as a substitute for source authoritativeness as a means for assuring quality (Den Besten and Dalle, 2008). Obviously, as far as trustworthiness is concerned, content peer production has also its share of skeptics in the scientific literature (Denning et al., 2005), in practitioners’ view (Keen, 2007) and in popular media (Cuozzo, 2008). Despite some exceptions (Giles, 2005), this lively debate has mostly being fueled by claims that have still to move towards the stage of sound empirical validation.

We build from previous empirical research in the field that has started to shed light on the role of institutions and organizational practices in channeling the largely unstructured efforts of voluntary contributors (Den Besten and Dalle, 2008; Den Besten et al. 2008; Kittur and Kraut, 2008). According to this line of research, peer production within wiki platforms makes extensive use of template messages – standard info-boxes placed on top of a given page – as coordination tool which ease the contribution to the production process of the various participants. In Wikipedia, for instance, there is an overwhelming number of templates, a.k.a. tags, which are used as a means to facilitate various goals and activities, such as to flag particular anomalies and dysfunctions of pages (e.g., violations of common policies or guidelines), and to call for specific actions for contributors (e.g., cleaning up, improvements in the organization of the text, and so on).

Simple Wikipedia

The fundamental reason for the choice of Simple Wikipedia over various other publicly available wiki–based collections lies in the strong commitment by the active participants in the activities of collaborative editing to a writing style which poses a strong emphasis on simplicity and readability (Wikipedia, 2008a-b). Accordingly, the template “complex” (which in the early days of the collection was labeled “unsimple”) is used by editors in order to signal that a particular article is unsatisfactory as far as readability is concerned. An additional rationale pushing further the argument for choosing “complex” from Simple Wikipedia over other alternatives is represented by the possibility of computing measures of simplicity/readability, derived from the computational linguistic tradition, which can be used as objective appraisal of the gravity of a bug and of the effectiveness of the work done to fix it.

NPOV

The second case study on template is drawn from the larger English Wikipedia, where we focus on the template “NPOV”, which signals breaches of the neutrality principle. All Wikipedia articles are meant to be written from a neutral point of view and along with "Verifiability" and "No original research”, “Neutral point of view” is one of Wikipedia’s three main policies regarding content editing. There are several ways an article could violate this principle and they may range from single biased statements to more general imbalances in the article structure or undue weights given to various aspects pertaining to the subject. Neutrality concerns may give way to heated disputes and "NPOV" page on Wikipedia lists several good editing practices meant to minimize their emergence or to confine their scope.

2. METHODOLOGY

In the previous section we argued that the use of meta-data is crucials as a means to coordinate collaborative work in distributed efforts such as Wikipedia. In what follows we will show how meta-data can be effective. Our approach is to study, as examples of meta-data, templates that appear to voice important concerns to a considerable part of the editors of Wikipedia. By exploiting publicly available archives, we retrieve the edit history of the articles where these templates have appeared and we compare the evolution of edits before and after the appearance of the template.

Meta-data can be said to be effective if the signal they send is accurate and trusted, if they provoke the appropriate reaction – i.e. result in a treatment of the defect that has been signaled – and if the reaction provoked is reasonably fast. Accuracy of a template can be determined by comparing the state of an article when the template first appears against a benchmark such as the average state of articles in the collection. In many cases there are textual indicators that correlate with the defect that is singled out by the template. For instance, articles that attract the `"unsimple"` template can be expected to score low in metrics of readability. Besides, to evaluate the appropriate reaction to a template we can compute, for instance, the proportion of articles where templates are not ignored completely or reverted immediately. In order to compare the state of articles before and after the appearance of a template we can again rely on the differences in characteristics of texts such as the number of words or references in the text and indicators like readability. In extension to this, we can compute measures of textual similarity as a more generic metric of the differences between texts. Finally, using survival analysis, we can ask the following question – given the accuracy of the signal, what does it take to obtain a treatment with the desired effect? We consider that a treatment has obtained the desired effect either through self-reporting or by means of measurement. That is, we consider that a treatment has been completed when:

1) The template signaling the defect has been removed;

2) The state of the article has changed so that the defect is no longer there.

In our model we hypothesize that the duration of the treatment depends on the following elements:

- The severity (and accuracy) of the defect as reflected by the state of the article when the tag first appeared;
- The attention paid to the article:
    - `reactime`: reaction time (days) measured as time between tagging and the first subsequent revision of the article;
    - tag/untag: a dummy variable showing that the user who put the tag also removed it.

- The division of labor within the community – involvement of zealots, Samaritans, etc.
    - shareAdm{R1,R2}: share of administrators with respect to registered contributors in pre-tagging regime and 2, respectively;
    - shareAno{R1,R2}: share of anonymous users anonymous versus registered contributors in pre-tagging regime and 2, respectively;
- And effort exerted:
    - revsReg{R1,R2}: number of edits by registered contributors only in pre-tagging regime and 2, respectively;
    - uniqueReg{R1,R2}: total number of distinct registered contributors in pre-tagging regime and 2, respectively;

In the subsections below we explain in a bit more detail several novel elements in our methodology.

2.1 Readability

The readability of an article is determined by computing the Flesch readability score of the article's text with help of the GNU Style package. This score is a function of the number of syllables per word and the number of words per sentence (Flesch, 1979). More precisely, the formula 'score = 206.835 – 84.6*syllables/words – 1.1015*words/sentences' yields a number that is usually between 0 and 100 and between 60 and 70 for standard English texts. This Flesch reading easy formula, which has been elaborated on the basis of school texts by Flesch in 1948, has been very popular, especially in the US, as a measure of plain English. Its popularity rests on the fact that the formula is easy to compute, yet often accurate. Even word processing applications, such as Microsoft Word, often provide this score as part of their statistics.

2.2 Textual similarities

The approach proposed here stems from the analysis of text similarities, which is used to compare documents in large text corpora, in order to assess the repetitions of patterns. In this respect, similarity is thus considered a measurable property assessing the degree of relation between two or more information artifacts.

There is a plethora of similarity measures (for an extensive review see Lee, 2008) to evaluate this feature; in this study we will take into account two particular vector-based metrics that have been selected for their particular properties.

A measure often used for comparing different documents is represented by the Jaccard coefficient (*J*), a distance metric defined as follows: given two documents *A* and *B*, let *a* and *b* the sets of terms occurring in *A* and *B* respectively. Define *I* as the intersection of *a* and *b*, and *K* as their union. Then the Jaccard similarity is the number of elements (cardinality) of *I* divided by the cardinality of *K*, thus

$J = |I| / |K|$.

Conversely, the Cosine similarity (*C*) is computed in the following way: let A_s and B_s be sets of terms occurring in *A* and *B*, as in the previous measure; define *K* as the union of A_s and B_s, and let k_i be the i-th element in K. Then the vector terms in A and B are:

$$a = [nA(k_1), nA(k_2),\ldots, nA(k_n)]$$

$$b = [nB(k_1), nB(k_2),\ldots, nB(k_n)]$$

where $nA(k_i)$ is the number of occurrences of term k_i in A, and $nB(k_i)$ is the same for B. In this respect, Cosine similarity between two original document sets is defined as

$$C = (a \times b) / ||a||\ ||b||.$$

that is the ratio between the scalar product of vectors a and b and their Euclidean norm.

2.3 Survival Analysis

We can consider the act of tagging the page with a template as a signal of dysfunction, where editing out the template might correspond to an indefinite remission (a cure), or to a temporary remission until the illness shows up again. In this respect, survival analysis seems to be an appropriate set of tools to be used to address and analyze the usage of Wikipedia tags as means of coordination. In this respect, the dynamic of tagging can be analyzed as a survival process, linking the probability of entry/exit of a page into a “pathological state” with regard to various explanatory variables. According to this framework, we perform survival both on the durations before and during the pathological state, exploring how different variables affect inception, treatment, and persistence of such condition.

Survival analysis is a collection of statistical methods and approaches used to describe and analyze time-to-event information. In survival analysis, it is mainstream the notion of ‘failure’ to define the occurrence of the event of interest (without attaching a value judgment to such manifestation, for example when the event might be a ‘success’, such as recovery from illness). The term ‘survival time’ specifies the length of time taken for failure to occur. When the variable under consideration is the length of time taken for an event to occur (e.g. death), the count of events as a function of time can be used to build a cumulative density function $F(t)$, which represents the proportion of individuals who have died as a function of t, and is known as the cumulative death distribution function. The inverse of such function is the proportion of individuals in the population who have survived to time t and is denoted as $S(t)$. A rather well-known method to derive the survival function from empirical data is the Kaplan-Meyer method; because counts of individuals at discrete time points are usually used, survival curves are normally presented in step format.

Moreover, in longitudinal studies exact survival time is only known for those individuals who show the event of interest during the observation period. For those who are disease free at the end of the follow-up period, all we can say is that they did not show the event of interest during the observation time. These individuals are called censored data. An attractive feature of survival analysis is that we are able to include the information contributed by censored observations right up until they are removed from the set. In this instance, particular importance bears the notion of right-censored data, that is subjects for whom is known that failure will occur some time after the recorded follow-up interval.

When no theoretical distribution adequately fits the data, then non-parametric methods are used in order to efficiently describe the survival pattern of the observed phenomenon, as in the case of Cox-PH (proportional hazard) regression model, which can be used with particular assumptions, the most important is the independence of the hazard function respect to time.

If survival data are consistent with a parametric distribution, then parameters can be estimated in order to stylize the survival, and statistical inference can be based on the chosen distribution. This is the case for exponential, Weibull, and gamma regression models. An extension of these models is the Accelerated Failure Time model: an AFT model assumes that the effect of a covariate is to multiply the predicted event time by some constant, acting multiplicatively on the failure time by a scale factor. The effect of a predictor (covariate) is to alter the rate at which a subject proceeds along the time axis (i.e., to accelerate the time to event); this family of models allowed to identify possible variable effects over time.

3. Data

3.1 Access

3.1.1 Simple Wikipedia

We used the July 2008 archive of Simple Wikipedia, available at: http://downloads.wikimedia.org, which for every revision made on an article page lists the following data: the user–id of the editor (IP address in case of anonymous edits), date and time of the edit, comments made by the editor and the full (wiki markup) text of the revision. We selected from the archive all the revisions corresponding to article pages which had been tagged at least once with the "unsimple"|"complex" templates. In order to avoid biases due to very short series for some datapoints in the survival analysis, we restricted the analysis to article pages which had been revised at least 15 times. For each article page we limited our extraction to all revisions belonging to the interval spanning from the first edit to the revision antecedent to the removal of the "complex" template[1]. After this selection, we ended up with 378 article pages for the analysis.

3.1.2 English Wikipedia

We retrieved the November 2006 .xml meta-history dump of the English version of Wikipedia, available at: http://downloads.wikimedia.org. We subsequently produced an .xml sub-archive made from all article pages tagged at least once in their lifetime with the "NPOV" template. There is a large family of template messages used to signal the breach of the neutrality policy in Wikipedia. Table 1 shows the frequencies of the various existing NPOV template messages. For a data consistency rationale we limited the analysis to strict "NPOV" templates (which accounts for around 80 per cent of all instances), while disregarding all remaining NPOV template messages

[1] In the case of pages in which the "complex" tag has never been removed (a.k.a. censored pages) we took all the available revisions. Also, we did not consider instances of repeated flagging of one page, where one page, after returning in the "simple" regime, is flagged once again as complex.

(around 14 percent is related to NPOV messages place at section level and 6 per cent are represented of a large number of variations of marginal use).

In order to avoid some inconsistencies on the original .xml archive of Wikipedia (due to some older conversion scripts which have been in place until February 2002, some older articles have incomplete histories where the initial revisions are missing), we filtered out around 700 articles with starting date older than March, 1st 2002.

After this filtering, we ended up with a selection of 6042 article pages for the analysis. While some studies on the English Wikipedia have shown that actual changes in a given article page are sometimes the result of longer discussions occurring at the level of the corresponding talk page (Kittur and Kraut, 2008; Viegas et al., 2007), the use of talk pages as a means to anticipate and discuss actual changes is not investigated here and our analysis relies solely on data collected from article pages.

Table 1. The NPOV template message family.

tag	# article pages	# article revisions
[“NPOV”]	6815	160772
[“NPOV-section”]	941	37452
[“msg:NPOV”]	196	3700
[“Long NPOV”]	143	5672
[“SectNPOV”]	134	8404
[“sectNPOV”]	106	6302
[other 129 tags]	260	37745
TOTAL	8595	260047

3.2 Pre-processing

De-wikification of the text and categorization of registered users (in terms of administrators, bots, registered and anonymous users) have been performed according to previous literature (Den Besten and Dalle, 2008; Den Besten et al., 2008). The distinction between registered and anonymous users is based on the author identification in the edit meta-data: where a user-id is provided, we attributed edits to “known” users, while where an IP-address is given, edits are considered by anonymous users. The group of “known” users is further split in administrators, bots, and regular users based the user-groups attribution given as a separate table in the Wikipedia archive (it has to be noted that this method assumes that users do not change sub-groups – an assumption which is probably not always warranted in the case of administrators). Finally, readability and similarity metrics were computed according to Den Besten et al. (2008).

3.3 Regime definition

Our main purpose is to characterize the existing differences in the production process of an article with respect to the presence of absence of a specific template. Accordingly, we analyze the dynamics surrounding the birth of an article page, the emergence of readability (or neutrality) concerns and their resolution.

We designate the period which goes from the article page inception to the appearance of the “unsimple” (or “NPOV”) template as “pre-tagging regime” and we label as “post-tagging regime” the subsequent period which subsists until the template is removed. As a matter of fact if one tracks down the appearance of a template in the revision history of an article one frequently observes repeated cycles of appearance-disappearance.

This dynamic can be interpreted by considering that it is not uncommon for an article page to develop the readability or neutrality concerns at different periods over time. Using a medical analogy, if we consider the act of placing a template on the page as a marker of a pathology, removing the template might correspond to an indefinite recovery or a temporary remission if the illness shows up again. In our analysis, for articles presenting repeated illnesses, we restrict our focus to the first occurrence and treatment.

3.4 Filtering vandal activity

Previous work has highlighted the short life span for vandal edits in wiki-collections (Viegas et al, 2004). While this generally reassures us that the impact of these malicious activities on the quality of the whole archive is limited, at the same time we still feel that when studying the process of development of articles one has to carefully evaluate whether vandal edits might introduce distortions in the interpretation of the data.

In our particular case, vandal edits replacing a non-negligible part of the article with other text, might wipe out also the wiki-code present in the preamble of the article (where template messages are placed). In this respect both those vandal edits and the corresponding revert edits aimed at cleaning from vandalizations induce cycling of the template on the article page that is repeatedly placed and removed in subsequent revisions of the article page.

The above-mentioned elements suggest that while filtering data for vandalism is not a major concern for the purpose of our analysis, yet a careful operationalization of the "post-tagging regime" is essential in order to inform the subsequent analyses. In this respect, taking as post-tagging regime the period which goes from the first appearance to the first removal of the template, might introduce a "shortening bias" due to vandal editing. Consequently, we employed different methods to filter vandal bias from the datasets, which were tailored to the complexity of the corresponding archive.

Simple Wikipedia

In the case of Simple Wikipedia, given the relatively limited number of revision involved, we decided to manually clean the dataset from vandalisms. This fixed both the issue of anticipated termination of post-tagging regime and also allowed to obtain unbiased measures of work activity related to article pages (e.g., number of revisions, number of unique contributors, etc.). We performed this activity both using comment analysis (in order to single out reverts which were explicitly accounted by editors as fixes to vandal edits) and MD5 hash (computed over the full text of a revision) comparisons across subsequent revisions of an article page. Overall, we filtered out from the dataset around 11 per cent of the revisions which were vandal or revert edits.

English Wikipedia

In the case of English Wikipedia, we decided to employ an automatized procedure to filter vandal edits. While we were aware of the existence of algorithms for the automatic detection of vandalisms (Potthast et al., 2008) we decided to employ a simpler heuristic meant only to fix the issue of anticipated termination of post-tagging regime. In case of sequence of placement/removal of the NPOV template, we assumed that post-tagging regime ended only when the removal lasted at least one full day, while removals lasting shorter than that were considered as due to the effect of fast-paced vandalisms (or disputes over the NPOV status of the page).[2]

4. Findings

In the following we report on three series of analysis performed on our two case studies: some descriptive statistics comparing the regime before and the one after the emergence of a template: an analysis of the speed of textual changes occurring across time and across revisions before and after the emergence of a template and a survival analysis on the duration of those regimes, describing the covariates accounting for the emergence and resolution of readability or neutrality concerns.

4.1 The Unsimple Tag in Simple Wikipedia

4.1.1 Accuracy: Descriptive statistics

Table 2 presents some descriptive statistics on duration, number of revisions and number of editors involved, computed over the complete life of the 378 articles tagged at least once with the “unsimple|complex” template. Table 3 compares durations of pre-tagging regime and post-tagging regime. The comparison shows similar distributions for the right side of pre-tagging regime and post-tagging regime, while the first quartile of pre-tagging regime shows a considerable share of articles tagged just right after their inception (this is consistent with a popular practice in SimpleWiki of using the current English Wikipedia entry as a the initial revision of a new SimpleWiki entry). For post-tagging regime, Table 3 also distinguishes between uncensored articles (*n*=255, where the template has been removed at some point) and censored articles (*n*=123, still tagged as complex at the time of the dataset collection). The latter ones show relatively longer durations, suggesting that for some articles the treatment of readability issues might be very critical.

Table 2. Summary statistics for the “complex” articles

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
duration (days)	273.7	823.5	1082	1125	1480	2481
revisions	5	16	29	53.73	59	559
editors	2	10	17	30	33.5	222

[2] The performance of this heuristics was also tested against the Simple Wikipedia database, where it proved to be able to detect the correct ending of the post-tagging regime for a large majority of the cases.

Table 3. Summary statistics for pre-tagging regime and post-tagging regime durations (days)

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
R1	0.0003	3	260	356	622	1765
R2 (all)	0.0023	66	269	356	625	1401
R2 (closed disputes)	0.0023	24	122	178	270	962
R2 (ongoing disputes)	273	614	653	727	811	1401

4.1.2 Effect: Similarity

Figure 1 shows the average magnitude of textual changes experienced for revisions of articles before (left side) and after tagging (right side). This measure of magnitude is computed using the Cosine similarity index for each revision with respect to the revision corresponding to the first appearance of the tag “unsimple” (which corresponds to time=0 in the *x*-axis). Values corresponding to the same time lag category (unit = 1 day) are then averaged. The red line corresponds to the average Cosine similarity index computed over the whole population of “unsimple” articles (treatment group), while the black line corresponds to the same measure computer over a control set made by articles that were never tagged with the “unsimple” template. Two different control sets were constructed taken semi-random samples which were corrected in order to produce a matching set of articles with lengths (measured in terms of number of revisions) similar to the articles belonging to the treatment group. For each article belonging to the control set, revision no. 0 was assumed to be its *n*th revision, where *n* represented the number of revisions needed for the “unsimple” tag to appear in the matched article of the treatment group. Results shown are robust for both random samples. While the pace of textual changes for the treatment and control groups are comparable before the appearance of the “unsimple” tag, they diverge considerably after tagging. In particular, it is possible to observe a sharp decrease in the rate of textual changes for the treatment group before and after tagging, and with respect to the control group. We consider this as a clear indication that the editing style changes after the application of the template: where standard editing (before tagging) may consist of adding large chunks of text, after the “unsimple” template appears, editing seems to be dominated by relatively smaller changes in wording, which may aim at finding a solution to readability issues by way of fine tuning and an incremental editing strategy, in other words, solving the issue while preserving as much as possible of the content.

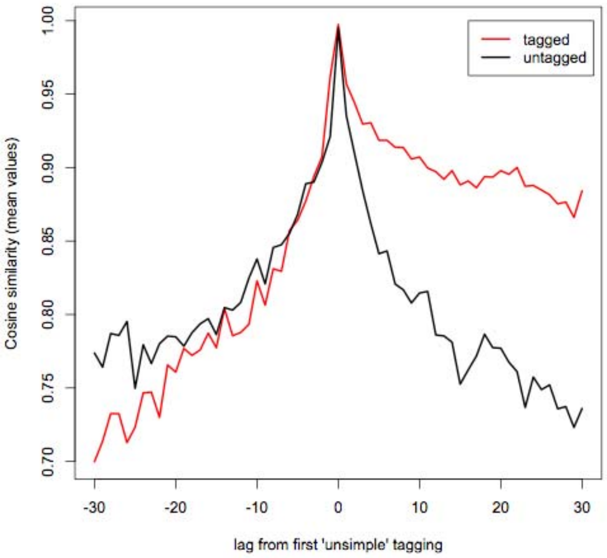

Figure 1 Average similarity plot (Cosine index) computed over time with respect to the first revision in which the tag "unsimple" appears (time = 0)

4.1.3 Persistence: Survival Analysis

We applied survival analysis to study two different albeit intertwined phenomena: (*i*) transition of article pages from the initial "simple" phase (from now on: pre-tagging regime) to the subsequent "unsimple" phase (from now on: post-tagging regime) and (*ii*) exit from post-tagging regime. The observation periods are, respectively, from the very first version of an article page to the revision in which the template "complex" appears, and from the latter to the revision in which the template is edited out. By definition of the sample, for the first event (exit from pre-tagging regime) all observations are uncensored, while for the second event some observation are censored, meaning that in some cases the template has never been removed from the article page.

4.1.3.1 Evidence from pre-tagging regime

According to a Kaplan Meyer estimate, pre-tagging regime seems to fit quite well to a Cox Proportionality Hazard class model.In order to assess the different effects of covariates in the termination of pre-tagging regime, we start considering the impact of division of labor, and in particular the incidence of efforts by different kind of users towards duration of pre-tagging regime. For this purpose we need preliminary to screen for the possible existence of multicollinearity issues between the various variables.

Table 4 summarizes the correlation between the duration of pre-tagging regime and two families of covariates: variables related to efforts (edits) exerted by different categories of participants and variables measuring the number of participants (for such categories). The table shows the existence of strong correlation between participants and edits for each category considered: this suggests to avoid to use both families together in survival estimation.

Consequently, a CoxPH model has been fitted in order to explain the impact of the three families of covariates earlier described. Overall, only the variables pertaining to intensity and division of labor seem to have a significant effect in explaining the length of pre-tagging regime, while variables regarding other features of the pages, such as size, readability, similarity have no explanatory power. For sake of compactness we present the final models only, which are summarized in Table 5.

Table 4. Correlation matrix for pre-tagging regime covariates

	duration1	regrevs1	admrevs1	anonrevs1	botrevs1	reg1	adm1	anon1	bot1
duration1	1								
regrevs1	.401	1							
admrevs1	.398	.296	1						
anonrevs1	.511	.459	.525	1					
botrevs1	.615	.304	.387	.447	1				
reg1	.710	.651	.467	.652	.525	1			
adm1	.575	.364	.807	.605	.573	.588	1		
anon1	.618	.465	.590	.912	.506	.755	.673	1	
bot1	.640	.324	.367	.406	.895	.531	.537	.441	1

In Model 1 the duration of pre-tagging regime is negatively affected by the number of revisions by all categories of contributors. Similarly, there is a negative impact on duration when considering the number of different contributors per category (Model 2). The latter model seems to have a higher descriptive power as far as Rsquare and model tests are concerned.

Overall the two models seem to suggest that both the level of effort on a page (in terms of revisions) and the number of participants in the editing process seem to anticipate the emergence of readability concerns. At this point of the analysis it is still difficult to judge whether this shortening is more due to a variant of the Linus' law (more eyeballs resulting in the anticipatory detection of a defect) or rather due to diminishing returns related with increases in the number of contributors. While the second model seems to be more ambiguous in this respect, the first one seems more clearly to suggest a connection between increases in work intensities and the emergence of a bug as the result of coordination conflicts. Nevertheless this issue seems to be worth of further scrutiny.

Table 5. Survival Analysis on Post-tagging regime Inception

Variable	Model 1	Model 2
regrevs1	-0.039*** (0.0129)	_
admrevs1	-0.037* (0.0203)	_
anonrevs1	-0.027** (0.0089)	_
botrevs1	-0.105*** (0.0150)	_
reg1	_	-0.147*** (0.0267)
adm1	_	-0.135*** (0.0510)
anon1	_	-0.025* (0.0157)
bot1	_	-0.254*** (0.0375)
Rsquare	0.370	0.492
L ratio	175	256
Wald	122	177
logrank	120	185

p-values significance: *<0.1, **<0.05, ***<0.01

4.1.3.2 Evidence from post-tagging regime

Similarly to the previous regime, for post-tagging regime durations a Kaplan Meyer estimate has been computed and the model seem again to fit quite well a Cox Proportionality Hazard class model.

Table 6 confirms the existence of a strong correlation between the number of revisions made by different classes of participants and the number of participants (for the same classes), again suggesting to avoid the use of both families in the same model estimation in order to avoid for multicollinearity problems.

Similarly to what has been done for pre-tagging regime, we test for the same hypotheses related to efforts and division of labor; we look whether the total number of revisions and the number of different contributors in the various classes do play a significant role in exiting from post-tagging regime.

Results are reported in Table 7. Here the variables related to the intensity of efforts (number of revisions) are not significant with the exception of revision made by bots (Model 1). On the contrary, all classes of users are significant when considering the number of different contributors per category (Model 2). In particular the shortening of the pathological regime seems to be affected by the presence of administrators, registered and bot users, while the presence of anonymous users seems to delay the fixing process.

Table 6. Correlation matrix for post-tagging regime covariates

	duration2	duration1	regrevs2	admrevs2	anonrevs2	botrevs2	reg2	adm2	anon2	bot2	react
duration2	1										.199
duration1	.000	1									-.066
regrevs2	.307	.182	1								-.031
admrevs2	.376	.190	.816	1							-.038
anonrevs2	.327	.197	.792	.929	1						-.038
botrevs2	.599	.178	.467	.618	.571	1					-.061
reg2	.438	.220	.887	.904	.888	.635	1				-.028
adm2	.543	.226	.710	.897	.842	.645	.849	1			.006
anon2	.355	.204	.803	.931	.988	.583	.908	.866	1		-.028
bot2	.706	.159	.358	.490	.436	.922	.525	.578	.453	1	-.034
react	.199	-.067	-.031	-.038	-.038	-.061	-.028	.006	-.028	-.034	1

Similarly to the previous Subsection, other covariates (in particular the textual-related covariate) have no incidence on the survival process. In particular, the reaction time to flagging has a negligible impact on post-tagging regime duration (for simplicity the model is not reported). Model 3 allows to introduce in the survival the duration of pre-tagging regime (that can be also thought as the overall life of the page at starting of post-tagging regime) as a covariate. This variable is significant and affects positively the duration of post-tagging regime. A possible interpretation is that the older the page at time of flagging, the more difficult is to solve successfully readability issues.

A final remark is worth on the variable measuring the efforts made by users which originally tagged the page. This variable is not significant, thus hinting to a quite different story with respect to open source development as far as to bug fixing is concerned, and reinforcing a view of open content creation communities as made more by “passers-by” users, rather than by contributors which commit themselves to a particular artifact on a long term perspective.

As far as pre-tagging regime is concerned, we showed that entry in the pathological regime is affected both by the number of users and their efforts, and the former model seems to be relatively more robust. Conversely, no structural feature of pages like size, readability, similarity, and so on are helpful in explaining the “complex” tagging. Overall, survival findings might highlight the existence of competing explanations regarding the shortening of pre-tagging regime duration (complexity/coordination issues vs. “eyeballs” hypothesis), which call for further scrutiny.

Table 7. Survival Analysis on Post-tagging regime Termination

Variable	Model 1	Model 2	Model 3
regrevs2	-0.014 (0.0098)	_	_
admrevs2	0.015 (0.0196)	_	_
anonrevs2	0.007 (0.0064)	_	_
botrevs2	-0.091*** (0.0103)	_	_
reg2	_	-0.060* (0.0373)	-0.084** (0.0379)
adm2	_	-0.233*** (0.0516)	-0.262*** (0.0519)
anon2	_	0.074*** (0.0131)	0.081*** (0.0135)
bot2	_	-0.224*** (0.0222)	-0.228*** (0.0220)
duration1	_	_	.0004** (0.0001)
Rsquare	0.320	0.499	0.537
L ratio	146	262	291
Wald	83.4	152	174
logrank	89	189	213

Regarding post-tagging regime, exit from the pathological state seems to depend on factors related on the number of participants only. In particular, while anonymous users have detrimental effects, all three categories of registered users seem to help in sorting the readability issue. during the regime shortens its duration.

Finally, we mentioned that both entry and exit cannot be traced back neither to reaction time measures, nor other structural features of pages, such as readability, similarity, and so on. In this respect we think that, other statistical models, i.e. event analysis, might represent a more suitable way to study in a more dynamic way their effect on pages being tagged.

4.2 The NPOV tag in English Wikipedia

4.2.1 Accuracy: Descriptive Statistics

Table 8 details the summary statistics for the duration of the articles in the sample computed on the complete life of articles and within regimes 1 and 2 (ended disputes only).

Table 8. Summary statistics for durations (days).

	Min.	1st Q.	Median	Mean	3rd Q.	Max
complete life (*N*=6042)	2.19	397.80	700.40	747.30	1055.00	1710.00
r1 (*N*=6042)	0.00	63.84	304.06	421.16	691.45	1691.72
r2 (*N*=5315)	0.00	0.31	6.21	37.57	38.47	832.97

Table 8 shows that articles usually develop neutrality issues during their maturity (the median duration of pre-tagging regime is 304 days) while the resolution is a relatively faster process.

Table 9 offers some summary statistics computed on the number of revisions performed by human editors only (the activity of bots is not considered). Considerations similar to the one given for durations still hold here. In particular relatively few revision are usually required to fix the neutrality concerns for the majority of articles, while the right tail also suggest that for a minority of them the process can take up a very high number of interventions.

Table 9. Summary statistics for revisions (human editors only).

	Min.	1st Q.	Median	Mean	3rd Q.	Max
complete life (*N*=6042)	2.00	32.00	90.00	292.50	275.00	15120.00
r1 (*N*=6042)	0.00	10.00	31.00	109.90	97.00	7107.00
r2 (*N*=5315)	1.00	1.00	3.00	13.78	11.00	977.00

In a similar vein, Table 10 offers some summary statistics on the number of human editors.

Table 10. Summary statistics for number of unique editors (human editors only).

	Min.	1st Q.	Median	Mean	3rd Q.	Max
complete life (*N*=6042)	1.00	16.00	39.00	108.00	106.00	4411.00
r1 (*N*=6042)	1.00	5.00	14.00	43.97	40.00	2418.00
r2 (*N*=5315)	1.00	1.00	2.00	5.76	5.00	281.00

A more interesting statistic is offered in Table 11, which is computed taking the revisions/editor ratio. The sensible difference between pre-tagging regime and post-tagging regime is here represented by relative increase of participation compared to contribution in post-tagging regime.

All statistics from the Tables 8-11 are computed using ended disputes only. By contrast, Table 12 collects the same statistics for ongoing disputes only. One reason for an article to be still disputed at the date of the dataset collection could be its relatively young NPOV debate (e.g.: the article might have been tagged as NPOV just a few days before the dataset dump). Another interpretation might be that the set of ongoing NPOV controversies is made by articles with very long debates over the neutrality issue. The comparison between Table 12 and Tables 8-11 suggests that for the majority of articles the latter interpretation might be at work: in particular, larger duration values suggest that these articles can be viewed as completely neglected "open issues" at the time of the database collection.

Table 11. Summary statistics for the human revisions/human unique editors ratio.

	Min.	1st Q.	Median	Mean	3rd Q.	Max
complete life (*N*=6042)	1.00	1.63	2.10	2.62	2.86	82.60
r1 (*N*=6042)	1.00	1.48	2.00	2.68	2.83	120.70
r2 (*N*=5315)	1.00	1.00	1.25	1.88	2.00	28.33

Table 12. Summary statistics for ongoing NPOV disputes only, post-tagging regime, *N*=628.

	Min.	1st Q.	Median	Mean	3rd Q.	Max
duration	0.22	29.67	71.64	122.47	158.47	855.28
revisions	1.00	2.00	6.00	15.53	16.00	555.00
editors	1.00	2.00	4.00	7.55	9.00	163.00
revs/editors	1.00	1.00	1.33	1.69	1.85	17.34

Table 13 and 14 present per time unit statistics obtained dividing, respectively, revisions and number of editors by the corresponding duration of the regime.

Table 13. Summary statistics for revisions/days ratio.

	Min.	1st Q.	Median	Mean	3rd Q.	Max
lifetime (*N*=6042)	0.00	0.06	0.15	0.45	0.38	130.40
r1(*N*=6042)	0.00	0.07	0.16	31.81	0.58	2979.00
ended r2 (*N*=5414)	0.00	0.19	1.02	149.90	13.24	43480.00
ongoing r2 (*N*=628)	0.00	0.04	0.09	0.31	0.24	10.31

Both metrics shows a considerable increase in participation per time unit to the editing process from pre-tagging regime to post-tagging regime (for ended disputes), while the statistics of post-tagging regime for ongoing disputes show a decline in work intensity. This evidence suggests that there are pages for which the effect of tagging does not seem to trigger any attention, thus confirming the idea that most of the ongoing disputes are articles which failed to attract the attention needed to address the neutrality concerns.

Table 14. Summary statistics for editors/days ratio.

	Min.	1st Q.	Median	Mean	3rd Q.	Max
complete life (*N*=6042)	0.01	0.05	0.08	0.16	0.15	23.36
pre-tagging regime (*N*=6042)	0.01	0.05	0.08	21.19	0.21	2882.01
ended r2 (*N*=5414)	0.01	0.13	0.62	139.41	9.12	43480.01
ongoing r2 (*N*=628)	0.01	0.04	0.07	0.18	0.15	4.50

4.2.2 Effect: Similarity analysis

Figure 4 makes use of some metrics taken from computational linguistic in order to measure to what extent the text of articles is updated over time before and after the emergence of the NPOV template. In the plots, the *x*-axis is centered over the revision corresponding to the beginning of post-tagging regime. Cosine and Jaccard similarity measures are then computed for all previous/subsequent revisions (upper side), or in the time domain (lower side), with respect to the first revision of post-tagging regime. Values are then averaged.

Similarly to the case of Simple Wikipedia, the plots show asymmetric speeds of change for the text of articles in pre-tagging regime and post-tagging regime. In particular in the latter it is possible to observe a considerable decrease in the rate of textual change starting from around the 10th revision or after around 10 days. These, in turn, correspond respectively to about the 3rd quartile of the number of revisions and about the 65th percentile of the durations in post-tagging regime , respectively. Recalling that we consider only articles with ended disputes, the plot seems to suggest a counterintuitive stylized fact: the longer the dispute the slower the relative pace of textual change which the article is subject.

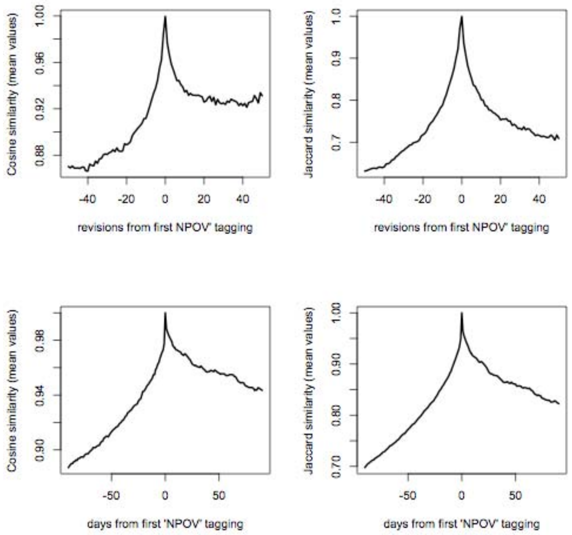

Figure 2. Similarity plots (left side=Cosine metric, right side=Jaccard metric) centered over the first emergence of the NPOV template, computed over the revision domain (upper side) and the time domain (lower side), English Wikipedia.

Furthermore, the two similarity metrics seems to behave differently. In particular, the Cosine index changes at a lesser degree and seems to reach a plateau, while the Jaccard index is characterized by a higher pace of change and does not seem to reach a plateau. Overall, based on the fact that these two metrics are built differently [11], the plot suggests that the typical style of editing for NPOV articles is characterized both by preserving of a relatively large body of redundant "lemmas" (words) between the revisions (or over time), while deletions and introductions are mainly of non-redundant lemmas – that is, of new words.

We consider these findings as preliminary evidence of the existence of different coordination regimes at work in controversial pages: while “easy to solve” disputes are characterized by a pace of textual change similar to the period before tagging occurs, “hard to solve” ones call for a slowdown of the textual changes while the solution appears to be worked our via fine tuned edits, in an incremental way. This might be regarded as a different coordination regime on the page, in which case the effect of NPOV tagging would be partly similar to the use of Simple Wikipedia “unsimple” tagging shown in the previous Section, in that it would trigger the nature of decentralized coordination.

4.2.3 Persistence: Survival analysis

We now apply survival analysis to study the dynamics of NPOV post-tagging regime and to identify variables directly influencing the resolution of neutrality issues. Based on previous findings about the peculiar nature of ongoing NPOV disputes, we restrict our analysis to ended disputes. More specifically, out of a sample of 6042 pages, a subset of 823 articles with open disputes at the time of the dataset collection have been dropped from the samples.

A first inspection of survival durations allows us to speculate around the nature of distribution of regime times; in this respect, Figure 3, which depicts the empirical cumulate density function for durations of post-tagging regime, confirms the clue for a parametric distribution of regime durations. In order to find out the most appropriate model for a parametric survival analysis, we tested a selection of distribution function of durations as candidate for maximum likelihood (MLE) fitting and parameter estimation.

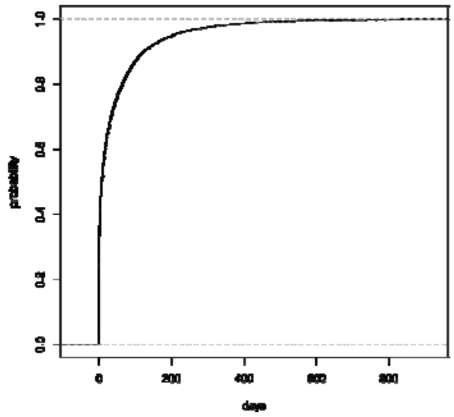

Figure 3. ECDF of durations for post-tagging regime, English Wikipedia.

In order to identify the most promising candidate, alternative distributions have been sorted according to AIC measure (Aikake, 1974; Collett, 2003) and the X_L statistic, that is $2^*(LL_i\text{-}LL_{i+1})$, measuring the increase in Log likelihood between model *I* and model *i*+1, have been computed to test the null hypothesis *H*0 that data follows a given distribution, versus the alternative *H*1 that the underlining distribution follows the next candidate; then *HL* is compared to *p*-value for the usual chi-square. Results of this procedure are summarized in Table 15.

The table summarizes the fitting values for four distribution candidates (exponential, gamma, lognormal and Weibull) and identifies the Weibull (extreme value) distribution as the best choice.

Table 15. Distribution fitting for post-tagging regime.

	estimates					
Model	***a***	***b***	***LL***	X_L	***p*-value**	**AIC**
Exponential	59.06	—	-11473.3	3207.6	<.0001	11474.
Gamma	0.447	59.06	-9869.5	350.4	<.0001	9871.5
Lognormal	2.635	2.999	-9694.3	330.8	<.0001	9696.3
Weibull	0.574	34.26	-9528.9	—	—	9530.9

Subsequently, to fit survival models a set of accelerated failure time models (AFT) has been used, in order to appraise the effect of covariates on survival time. An AFT model assumes that the effect of a covariate is to multiply the predicted event time by some constant, acting multiplicatively on the failure time by a scale factor. The effect of a predictor (covariate) is to alter the rate at which a page proceeds along the time axis (i.e., to accelerate the time to failure); this family of models allowed to identify possible variable effects over time.

In the survival analysis the following variables have been used as covariates in parametric regression:

1. revsAdm{R1,R2}: number of edits by administrator in pre-tagging regime and 2, respectively;
2. revsAno{R1,R2}: number of edits by anonymous contributors in pre-tagging regime and 2, respectively;
3. revsReg{R1,R2}: number of edits by registered contributors only in pre-tagging regime and 2, respectively;
4. uniqueReg{R1,R2}: total number of distinct registered contributors in pre-tagging regime and 2, respectively;
5. shareAdm{R1,R2}: share of administrators with respect to registered contributors in pre-tagging regime and 2, respectively;
6. shareAno{R1,R2}: share of anonymous users versus registered contributors in pre-tagging regime and 2, respectively;
7. deltaReadability: difference in readability measure between the first revision tagged NPOV and the first succeeding edit untagged;
8. tag/untag: a dummy variable showing that the user who put the tag also removed it.

In order to explain the impact of the covariates depicted above, for post-tagging regime we fitted a set of parametric accelerated failure time models. For sake of compactness we present the estimates of the coefficients and their standard errors in parentheses, along with some diagnostics, which are summarized in Table 16.

Three different regression models have been estimated, each using distinct covariate sets, in order to explain the post-tagging regime duration: first in terms of total effort before a tag is set (model 1), then using both effort and division of labor variables (model 2), and eventually considering the difference in readability of the page (model 3).

The overall model diagnostics reported in Table 16 show that model with effort, division of labor and readability covariates best explains regime length; moreover, the scale effect is quite unimportant in all variants, being its log near to one; this provides evidence to the hypothesis that covariate effects do not change much in time.

The largest majority of covariates are significant at the 1% level, for all regressions. As for casual relations, the hypothesis of direct relationship between effort in post-tagging regime and remission time is confirmed, since all parameters regarding counting revisions and unique contributors are positive.

Effort (in terms of no. of revisions) and participation (in terms of no. of unique editors) before NPOV tagging (pre-tagging regime) are associated with a shortening of post-tagging regime, apart from administrators who seem to play a quite different role. Overall, the models seem to suggest that both effort on the page and the number of participants in the editing process seem to render less probable the emergence of neutrality concerns. At this point of the analysis, it is still difficult to judge whether this shortening is more due to circumstances similar to “Linus’ law” (Raymond, 1999)

– more eyeballs resulting in an improved detection of concerns – or to the fact that a more limited number of editors might be associated with “stronger points of views”, in the sense of (Suh et al., 2007), and therefore with a sense of “ownership” vis-à-vis a given topic.

Table 16. Survival analysis – NPOV – English Wikipedia.

Model	**1**	**2**	**3**
(intercept)	2.6831*** (0.04653)	1.5114*** (0.0673)	1.5517*** (0.0773)
revsAdmR1	-0.0435*** (0.01124)	-0.0752*** (0.0092)	-0.0747*** (0.0092)
revsAnoR1	-0.0103** (0.00550)	-0.0097*** (0.0038)	-0.0099*** (0.0037)
revsRegR1	-0.0237*** (0.00626)	-0.0338*** (0.0049)	-0.0328*** (0.0049)
uniqueRegR1	—	-0.0140*** (0.0005)	-0.0139*** (0.0005)
shareAdmR1	—	0.1974*** (0.0750)	0.2122*** (0.0754)
shareAnoR1	—	-0.1997*** (0.0394)	-0.1959*** (0.0393)
revsAdmR2	—	0.0835*** (0.0215)	0.0826*** (0.0213)
revsAnoR2	—	0.0741*** (0.0254)	0.0752*** (0.0254)
revsRegR2	—	0.0304** (0.0139)	0.0336** (0.0141)
uniqueRegR2	—	0.2670*** (0.0119)	0.2694*** (0.0120)
shareAdmR2	—	0.2670*** (0.0119)	0.8101*** (0.0884)
shareAnoR2	—	0.5284*** (0.0568)	0.5162*** (0.0570)
tag/untag	—	-0.1581** (0.0651)	-0.1502** (0.0651)
deltaReadability	—	—	-0.010*** (0.0023)
Log(scale)	0.9662 (0.0111)	0.8284 (0.0111)	0.8265 (0.0111)
Log likelihood	-18613	-17865	-17852
Likelihood Ratio	30.53	1495.6	1521.8

Significance levels: *** = 0.01; ** = 0.05; * = 0.1

Compared to previous findings (Kittur et al., 2007) according to which the number of unique editors involved in an article would negatively correlate with conflict, our findings suggest that this phenomenon would rather be verified for the pre-tagging involvement of editors (in pre-tagging regime), while more contributors to explicitly controversial articles would rather tend to lengthen the controversy (post-tagging regime). In addition, the positive effect of the relative number of edits by administrators during pre-tagging regime upon post-tagging regime duration could suggest that a higher involvement of this class of user could signal a problem in the neutrality of the page before it is tagged and create friction in the process of regime termination.

Both the difference in readability measures along the regime, and the dummy variable related to tag marking/removing are negatively correlated with the duration of the regime. The first covariate might hint towards a way of resolving some disputes: pages characterized by difficulties related to linguistic or composition issues, such as ambiguities or other misfits, can be quickly solved via edits resulting in an improved readability. The second covariate underlines the fact that some contributors use the tagging/untagging mechanism in an intentional way.

Finally, several controls were included both for size and age of pages, in order to test the robustness of the models towards page mass and lifespan effects: all controls were not significant. In this respect, a drawback of our results at this stage is that we do not control for the level of exposure of the article, but we also intend to do so in later studies, in addition to taking “revert” edits explicitly into account[3].

Other tests, not reported here for sake of brevity, were also run while introducing a conflict measure (“*conflictuality ratio*”), defined as the ratio between the number of times the NPOV tag was placed or removed from the article page over the whole duration of the page. Preliminary results suggest that this control would be significant in all regressions, with a negative coefficient, but would not change the sign or significance of any other of the dependent variables, which also suggests that a higher conflictuality ratio would result in controversies being solved on average more rapidly. We also ran various other controls for distribution subsets only, sorting out cases which lie on the extreme 10th and 25th quantiles of the conflictuality ratio, which overall confirmed the robustness of our findings. Again, preliminary results for post-tagging regime in the upper decile of conflictuality suggested a change in the sign of revsAdmR2, i.e. that contributions by administrators shorten the duration of post-tagging regime, instead of lengthening it, as they do in the whole sample. This finding could be easily interpreted in line with previous research suggesting that administrators could play a distinct coordination role in mediating conflicts for the most disputed pages (Suh et al., 2007). Taken together, these preliminary findings might be consistent with the idea that higher conflictuality could also be associated with different coordination regimes.

5. CONCLUSIONS

Adopting ideas from entomology and inspired by examples of stigmergic self-organization among insects, we have argued that the success of post-it notes can be attributed to their role as enablers of lean management. We have asserted that

[3] We also plan to introduce controls for the complexity of pages in future work.

template tags in Wikipedia are virtually akin to post-it notes and are crucial to the success of Wikipedia as a distributed problem solving organization. In order to illustrate our argument, we have selected articles in Wikipedia that feature widely used tags. In particular, we have looked at the use of unsimple/complex in Simple Wikipedia and NPOV in main Wikipedia; retrieved all revisions of articles where the tag appeared; and analyzed the evolution of the article on the basis of textual similarity and the survival of the tags.

With regards to these two tags we have found for NPOV that resolution is quicker when many participate in the editing of the page before tagging and a few concentrate on fixing afterwards and for the unsimple/complex pair that the appearance of a tag is associated with a clear shift in editing behavior. These findings reinforce the idea that simple coordination devices like post-it notes can sometimes play a very important role in collective problem solving within and among organizations. Yet, our analysis has clear limits. To begin with, we still have very little intuition as to the kind of template tags and by extension post-it notes that are most effective. Nor do we have a clear idea why the tags that we have investigated seem to have the effect they have. Do they work only when people are paying sufficient attention? Do they reflect the fact that someone takes ownership of the problem? Is there something else at play? Further research will be needed to sort out these issues.

6. ACKNOWLEDGMENTS

Loris Gaio and Alessandro Rossi received partial funding from MIUR (projects FIRB03 and PRIN05).

7. REFERENCES

H. Akaike, “A New Look at Statistical Model Identification”, IEEE Transactions on Automatic control, 19, pp. 716-723, 1974.

Y. Benkler, The Wealth of Networks: How Social Production Transforms Markets and Freedom. Yale University Press, New Haven, CT, 2006.

D. Collett, Modelling Survival Data in Medical Research (2nd ed.), Chapman & Hall/CRC, Boca Raton, FL, 2003.

K. Crowston, and B. Scozzi, “Coordination practices for bug fixing within FLOSS development teams”, In: Proceedings of 1st CSAC. Porto, Portugal, 2004.

K. Crowston, K. Wei, Q. Li, U.Y. Eseryel, and J. Howison, “Coordination of Free / Libre Open Source Software Development”, in: Twenty-Sixth International Conference on Information Systems, 2005.

S. Cuozzo, “Urban Myths: What Wikipedia gets wrong about NYC”, New York Post. http://www.nypost.com/seven/08242008/postopinion/opedcolumnists/urban_myths_125773.htm?page=0, 2008.

J.-M. Dalle, and P. David, “Simulating Code Growth in Libre (Open-Source) Mode”, in N. Curien and E. Brousseau, eds., The Economics of the Internet, Cambridge University Press, Cambridge, UK, 2007.

J.-M. Dalle, P.A. David, F. Rullani, “Linking coordination, motivations and code structure in successful open source projects: A ‘stigmergic’ approach”, presented at the Academy of Management (AoM), August 7 – 11, 2009, Chicago, IL, USA, 2009.

J.-M. Dalle, and M. den Besten, “Different bug fixing regimes? A preliminary case for superbugs”, In: Proceedings of the 3rd International Conference on Open Source Systems. Limerick, Ireland, 2007.

J.-M. Dalle, M. den Besten, and H. Masmoudi, “Channelling Firefox developers: Mom and dad aren't happy yet”, In: Open Source Development, Communities and Quality, B. Russo, E. Damiani, S. Hissam, B. Lundell, G. Succi, Eds., IFIP International Federation for Information Processing;275. Boston, Springer:265–271, 2008.

M. den Besten, and J.-M. Dalle, Keep it Simple: A Companion for Simple Wikipedia? Industry & Innovation. 15(2):169–178, 2008.

M. den Besten, J.-M. Dalle, and F. Galia, “The Allocation of Collaborative Efforts in Open-Source Software”, Information Economics and Policy, 20(4), pp. 316-322, 2008.

M. den Besten, A. Rossi, L. Gaio, M. Loubser, and J.-M. Dalle, “Mining for practices in community collections: finds from Simple Wikipedia”, In: Open Source Development, Communities and Quality, B. Russo, E. Damiani, S. Hissam, B. Lundell, G. Succi, Eds., IFIP International Federation for Information Processing;275. Boston, Springer:105–120, 2008.

P. Denning, J. Horning, D. Parnas. and L. Weinstein, “Wikipedia risks”, Communications of the ACM 12(48): 152, 2005.

R. Flesch. How to Write Plain English. Harper and Row, New York, NY, 1979.

J. Giles, “Internet encyclopaedias go head to head”, Nature 438(7070):900–901, 2005.

J. Ito, “Wikipedia attacked by ignorant reporter”, Joi Ito's Web. http://joi.ito.com/weblog/2004/08/29/wikipedia-attac.html, 2004.

K.R. Lakhani, L.B. Jeppesen, P.A. Lohse, and J.A. Panetta, “The Value of Openness in Scientific Problem Solving”, Harvard Business School Working Paper, number 07–050, 2007.

A. Keen, The Cult of the Amateur: How Today's Internet is Killing Our Culture. Doubleday Business, New York, 2007.

A. Kittur, and R.E. Kraut, “Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination”, CSCW 2008: Proceedings of the ACM Conference on Computer-Supported Cooperative Work. New York: ACM Press, 2008.

A. Kittur, B. Suh, B.A. Pendleton, and E.H. Chi, “He Says, She Says: Conflict and Coordination in Wikipedia”, Proceedings of the 25th International Conference on Human factors in computing systems - CHI '07, ACM Press, pp. 453-462, 2007.

L. Lee, Measures of distributional similarity, Proceedings of the 37th conference on Association for Computational Linguistics, Montreal, Association for Computational Linguistics, Morristown, pp. 25–32, 1998.

M. Loubser, and M. den Besten, “Wikipedia Admins and Templates: The Organizational Capabilities of a Peer Production Effort”, SSRN. http://ssrn.com/abstract=1116171, 2008.

T.W. Malone, and K. Crowston, “The Interdisciplinary Study of Coordination”, ACM Computing Surveys, 26(1), pp. 87-119, 1994.

M. Potthast, B. Stein and R. Gerling, “Automatic Vandalism Detection in Wikipedia”, In: Advances in Information Retrieval, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, R.W. White, Eds., Boston, Springer:663–668, 2008.

E. Raymond, The Cathedral and the Bazaar, O'Reilly, Sebastopol, CA, 1999.

A. Rossi, L. Gaio, J.-M. Dalle, and M. den Besten, "Novelty Creation within Organizations: The Case of Collaborative/User Generated Content", unpublished manuscript prepared for the Special Panel Session on "Novelty in Organizational Adaptation", 2008 Academy of Management Annual Meeting, Anaheim, CA, 2008.

L. Sanger, "Wikipedia is wide open. Why is it growing so fast? Why isn't it full of nonsense?", Kuro5hin. http://www.kuro5hin.org/story/2001/9/24/43858/2479, 2001.

J. H. Sieg, M. W. Wallin, and G. von Krogh. Managerial challenges in open innovation: a study of innovation intermediation in the chemical industry. In *Proceedings of the 2009 EURAM Conference on Renaissance and Renewal in Management Studies; Track 14: Innovation - continuing the journey*, 2009.

T. S. Simcoe, D. Waguespack, and L. Fleming. What's in a (missing) name? status and signaling in open standards development. NET Institute Working Paper 08-31, SSRN, 2008

B. Suh, E. Chi, B.A. Pendleton, and A. Kittur, "Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations", VAST 2007: IEEE Symposium on Visual Analytics Science and Technology, 2007.

Wikipedia:Simple English Wikipedia. 2008. In: Simple Wikipedia, the free encyclopedia. http://simple.wikipedia.org/wiki/Wikipedia:Simple_English_Wikipedia. Accessed 10 November 2008.

Wikipedia:How to write Simple English articles. 2008. In: Simple Wikipedia, the free encyclopedia. http://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_articles. Accessed 10 November 2008.

F. Viégas, 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 2004, New York, ACM Press: 575–582, 2004.

F. Viégas, M. Wattenberg, F. Kriss, and F. van Ham, "Talk before you type: Coordination in Wikipedia", in: 40th Annual Hawaii International Conference on System Sciences (HICSS'07), 2007.