Characterizing online communities with their “signals”

Hang M. Ung
Hewlett Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304 USA and Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex FRANCE hang.ung@hp.com

Jean-Michel Dalle
Université Pierre et Marie Curie, 4 Place Jussieu, 75005 Paris FRANCE, jean-michel.dalle@upmc.fr

In order to analyze the heterogeneity of individual and collective online behaviors, we develop a methodology based on online signals. Modeling as a power law the distribution of time intervals between successive revisions of the online encyclopedia Wikipedia, either by an individual user or on an article, we show that part of the heterogeneity of the characteristic parameters of these distributions can be explained: for individual users, by their focus and multitasking behaviors, and, for articles, by different coordination regimes.

Key words: online communities, heterogeneity, coordination, Wikipedia, online signals, signal processing

1. Introduction

With the growing availability of digital records of human online activity, researchers are rapidly developing new methodologies to address the challenge of understanding human behavior, both at the individual level and collectively. In particular, large communities of interacting users, such as members of collaborative projects, social networks or participatory websites, have drawn much interest from researchers. Advanced visualization tools have for instance been developed to unveil conflicts in Wikipedia (Viégas et al. 2004); social network analysis is being applied to open-source communities to understand its social structure (Crowston and Howison 2005); and natural-language processing techniques are refined to mine opinions on Web discussion forums (Pang and Lee 2008).

However, these methodologies are not necessarily well-adapted to address a fundamental yet simple issue, which has many names but which we could call the “aggregation” issue: how do many individual actions result in aggregate outputs whose properties, if difficult to measure, still render them highly relevant socially and economically? Users are heterogeneous within online communities; they coordinate almost only through the Internet and yet, the aggregate result of their actions can be a fully functional software used by millions or an encyclopedia read by millions. This is done without a price system or even contracts, and we don’t know much yet about how heterogeneous users contribute to the production of online goods and how they coordinate their action.

Our assumption in this paper is that how users contribute, how they allocate their efforts and how they coordinate, is partly reflected by their “online signal”. Since the vast majority of online actions are logged with their timestamp, we indeed create artificial signals corresponding to the activity of an individual user or of a group of users interacting on a given topic (e.g. a forum thread, a wiki page). By applying signal processing techniques to the analysis of traces of human online activity thus viewed from a time perspective, as if they were “online signals”, and using Wikipedia as a test-bed, we describe a simple but general method to characterize activity patterns both at the single user level and for a group of people interacting on an article. We show that, taken together, some of the characteristics of users and of coordination regimes suggest a more complex view of the aggregation of efforts in Wikipedia than is perhaps currently thought.
The paper is organized as follows. In Section 2, we review related work on Wikipedia and also the literature describing and modeling the distribution of inter-event intervals (IEI). After describing our dataset in Section 3, we confirm the power law-like nature of IEI distribution of individual users in Section 4, and we then attempt to explain the heterogeneity of the slope of the power law across users by characteristics of their behavior, such as their focus, the average size of contributions or their tendency to multitask. In Section 5, we perform a similar analysis with the IEI distributions of articles. In this case, we introduce variables related to interaction modes and coordination regimes. We conclude and discuss future work in Section 6.

2. Related Work

Wikipedia is an online encyclopedia to which anyone can contribute, mainly by creating new articles or by modifying existing one. Wikipedia is widely recognized as a successful online collaborative project and has been studied by researchers and practitioners from various disciplines since its inception in 2001.

This research has notably focused on understanding the motivations for contributing and the causes for drop-out (e.g. Wilkinson 2008, Nov 2007). According to this literature, a main driver of the heterogeneity of users is their level of contribution. Anthony et al. (2005) distinguish “Zealots” who are committed users, from “Good Samaritans” who contribute infrequently, and show that both types are relevant for Wikipedia. Similarly, Panciera et al. (2009) emphasize the “power editors” or wikipedians, and shows the role of the initial experience in the retention of users.

But Wikipedia has also been widely studied in relation to its apparently successful collaboration model. Referring to the more general literature on collaborative problem-solving, Wilkinson and Huberman (2007) question whether the quality of an article is related to the number of distinct editors, while den Besten et al. (2008b) highlight the role of “tags” in coordinating efforts. Other aspects such as vandalism and conflicts (Viégas et al. 2004, Kittur et al. 2007), as well as the importance of talk pages in the collaboration process (Viegas et al. 2007), have also drawn much interest. In many respects, these studies tend to suggest the existence of different “coordination regimes” or “modes” of collaboration, yet without explicitly characterizing them.

A limited number of studies exploiting the timing of events can be found in the literature on Internet user behaviors. In particular, researchers interested in repetition in user search or browsing behavior (e.g. Sanderson and Dumais 2007, Adar et al. 2008) have identified periodicity patterns. Spectral analysis has also been explored as a tool to unveil rhythmic patterns in contributions of users in open-source software projects (Dalle et al. 2006). However, to our knowledge, there has been no attempt to investigate the distribution of the time intervals between successive actions (inter-event intervals, here simply IEI) from a given user or on a given page and interpret it as partly reflecting characteristics of individual or collective behavior.

However, a well established result in the physics literature is that time intervals between successive events produced by a single individual (e.g., emails sent or web sites visited by any given user) follow a long-tailed distribution (Barabási 2005). This relatively striking characteristic has never been exploited as a method to study individual or collective online activity. Instead, the physics literature has focused on suggesting theoretical explanations and understanding the precise mathematical characteristics of these long-tailed distributions. Furthermore, this regularity has been viewed by physicists as a universal feature of human dynamics, thus emphasizing the existence of two “universality classes” and accordingly of a very limited number of power law exponents (Vázquez et al. 2006), which implies that variations with respect to individual and collective heterogeneity have been neglected.
Characterizing online communities with their "signals"
European Academy of Management Annual Conference 2010, Rome, Italy.

3. Data Collection and Processing
We retrieved the archive of the French Wikipedia as of March 2009\(^1\), consisting in 3 million pages and in 37 million versions, or revisions of these pages. In addition to a unique identifier, pages are associated with a title and a namespace which defines whether a page is an article, a talk page or another type of Wikipedia page. From revisions, we built a list of users, among which 196 000 are registered users and over 1.6 million are “anonymous” users identified by an IP address. All data was stored in a MySQL database and processed with Perl and R.

For IEI analysis, we considered the timestamps of all revisions of the 3,848 registered non-bot users who contributed at least 500 times. To correct timestamps from the common user behavior consisting in regularly saving their work (as we usually do when editing a text), revisions were marked as \(p\)-intermediate when they were followed, on the page history, by another revision from the same user within a period of 30 minutes. This resulted in removing 29.6% of all contributions. Therefore, variables and IEI distributions presented in this paper were calculated ignoring \(p\)-intermediate revisions.

Similarly, we considered the timestamps of revisions for 3000 articles selected randomly amongst those that had received between 500 and 2000 contributions. As for users, \(p\)-intermediate revisions were ignored in this sample.

4. User signals
4.1. Pooled user signal
To get an overview of the IEI distribution, we first compute IEI individually for each user and then we pool those values together to obtain a single set of 8.6 million IEI: \(ALL-U\). We assess the empirical distribution of \(ALL-U\) by plotting a log-log histogram as shown on Fig. 1. We observe that the distribution is long-tailed and approximately linear in the range [1 min, 8 hours], which is a characteristic of power law distributions.

A striking feature, not reported in the physics literature on IEI distributions (most probably because of the limited size of the datasets used) has to do with the right part of the histogram, which exhibits distinct daily cyclic variations, as shown by dashed vertical lines at 1, 2 and 3 days on Fig. 1. This is consistent with the literature on “inter-arrival time” on websites (Adar et al. 2008). Similar circadian patterns for user data from the websites Digg and Twitter were also found (results not reported here), suggesting an understandable yet very general phenomenon.

\(^1\)http://download.wikimedia.org/frwiki/
Characterizing online communities with their “signals”

European Academy of Management Annual Conference 2010, Rome, Italy.

0.5 1.0 1.5 2.0

0 1 2 3 4 5

Figure 2  Histogram of slope parameter $\alpha$ obtained from fitting users IEI to a power-law distribution on the range $[x_{min}, +\infty]$ (NOMAX, grey) or $[x_{min}, 1 \text{ day}]$ (MAX1D, black)

Figure 3  Empirical IEI distributions for 4 individual users, with their MAX1D power law fit (grey line). Axes are the same as on Fig. 1.

We model the IEI distribution by a power law distribution, therefore described by two parameters: a scaling exponent $\alpha$, corresponding to the slope of the line in log-log plots, and a cut-off $x_{min}$. Using the procedure described in Clauset et al. (2007), we use the analytical expression of the Maximum-Likelihood Estimator to first calculate the slope $\alpha$ as a function of a given cut-off value $x_{min}$. This procedure allowed us to fit $x_{min} = 0.05$ hours (we call this value $x_{min-all}$) and $\alpha = 1.11$, which is in the range of what has been previously observed in communication activity (Barabási 2005) and in line with the plateau, below approx. 1 min, observed on Fig. 1.

4.2. Individual user signals

So far, we have been ignoring differences between users, although different users are likely to exhibit different online behaviors and thus different IEI distributions. Indeed, using a two-sided Kolmogorov-Smirnov (K-S) test, we easily verify that, for over 97% of the pairs of users $(i, j)$, the null hypothesis “IEIs distribution of user $i$ and $j$ are the same” can be rejected at level 1%. Stochastic variations alone most probably do not account for differences between individual IEI distributions.

To characterize heterogeneous users and their differences, we thus model individual users’ IEI distributions as power laws, using for the sake of simplicity a common cut-off value $x_{min} = x_{min-all}$ for all users (more refined techniques led to similar results). We estimate $\alpha$ assuming a power law either in the whole range $[x_{min}, +\infty]$ (NOMAX) or only in $[x_{min}, 1 \text{ day}]$ (MAX1D). Values obtained are presented on Fig. 2, and Fig. 3 shows 4 users and the fitted power law with MAX1D. MAX1D values of $\alpha$ spread over a wider range (approx. [0.5, 2.0]) than NOMAX values (approx. [1.1, 1.8]), while the former presents a more standard Gaussian shape.

In contrast with Vazquez et al. (2006), where it has been suggested that for email communication, individual users’ IEI distributions are power laws with a common “universal” slope $\alpha = 1$, we do
not find a common parameter $\alpha$ for all users. Indeed, our results rather tend to highlight the heterogeneity of users’ IEI distributions.

Note also that we do not try here to assess the goodness-of-fit of a power law model\(^2\). The methodology we suggest here consists in using a simple model such as the power law one to extract a synthetic parameter from individual users’ IEI distributions – the slope $\alpha$ of the power law-like domain – in order to characterize the heterogeneity and the differences in user behavior.

To analyze the heterogeneity of user IEI, we now compute several variables to characterize each user:

- Frequency ($freq$) is the total number of revisions divided by the time elapsed between the first and the last revision.
- Average contribution size ($size$) is obtained by averaging over all the revisions from the user, the absolute variation in number of words, relatively to their respective previous revisions.
- Number of revisions per page ($revs/page$) is a simple measure of the tendency for a user to make revisions to pages she or he has already been contributed to. It is the ratio between the total number of revisions of the user and the number of distinct pages (including non-article pages) these revisions affect.
- Multitasking ($mtask$): this variable captures the tendency of users to work simultaneously on several pages within a single work session. Schematically, a user would modify page $P_1$, save it, modify $P_2$, save it and then switch back to $P_1$ (sequence A: $P_1 P_2 P_1$). Assuming this sequence is done within a 30 minutes interval, and that no one else modified $P_1$, then, as described previously, the first revision on $P_1$ is marked as p-intermediate. Thus, the proportion of p-intermediate revisions reflects the multitasking behavior. However, to account also for sequences such as $P_1 P_2 P_1$ (sequence B), we mark as $u$-intermediate the first revision on page $P_1$, i.e. revisions that are followed in the user history by a revision on the same page, within a 30 minutes interval. $mtask$ is then defined as the proportion of p-intermediate amongst all the non $u$-intermediate: values for sequence A and B are 0.33 and zero, respectively. A high $mtask$ value corresponds to a highly multitasking user.

We then conduct a standard OLS regression in order to explain the slope which characterizes each user based on these variables, whose results are presented in Table 1 ($R^2 = 0.31$). Note that after checking the distribution of each variable, we removed 3 outliers from our database of 3848 users, according to $revs/page < 25$ and $size < 1000$. Note also that we further conducted PLS regressions (not reported here) to verify the signs of the parameter estimates, considering that there are partial colinearities between the variables described above. Finally, it should be kept in mind that a higher value of $freq$ does not imply a higher slope $\alpha$ in the context of our power law model: for instance, dividing all intervals by 2 doubles $freq$ while leaving $\alpha$ unchanged.

All variables are significant at the 0.0001 level and explain together 31% of the variance of the slope. Therefore, the heterogeneity of user IEI distribution is partly explained by their behavioral characteristics. More specifically, both $mtask$ and $freq$ have a positive impact on the slope of the power law fit, which is related to a higher proportion of small inter-event pauses compared to longer

\(^2\) In fact, the higher $\alpha$ values obtained with NOMAX vs. MAXID suggest a log-normal distribution could also have been a relevant approximation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Param. estimate</th>
<th>Standard error</th>
<th>t-test</th>
<th>$Prob &gt; t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>1.24498</td>
<td>0.00591</td>
<td>210.71</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$freq$</td>
<td>0.15589</td>
<td>0.00762</td>
<td>20.45</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$revs/page$</td>
<td>-0.05341</td>
<td>0.00186</td>
<td>-28.67</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$mtask$</td>
<td>0.58278</td>
<td>0.05181</td>
<td>11.25</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$size$</td>
<td>0.00051675</td>
<td>0.00005217</td>
<td>-9.91</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

Table 1
ones. By contrast, \textit{revs/page} and \textit{size} have a negative impact associated with a higher proportion of longer inter-event pauses. Even controlling by \textit{freq}, \textit{size} is still significant, confirming the intuition that users making larger contributions have less small intervals between their contributions.

Controlling by \textit{freq}, these results suggest that how much users are focused (higher \textit{revs/page} and \textit{size}) or on contrary more multitasking (high \textit{mtask}), are relevant characteristics of their behavior as measured by IEI distributions. More focused users tend to make more longer pauses and less shorter pauses, for instance by updating a small number of articles every now and then. In constrast, more dispersed users, such as admins, might be executing rapidly and successively small tasks during “work sessions”. To support this interpretation, we also measured that the average value of \textit{revs/page} for the 66 admins is 1.78 and is substantially smaller than for the value of 2.35 for non-admins.

This interpretation could be either consistent with an online division of labor or with the existence of online roles, or just with the fact that online users are simply more or less focused in the way they work. As a consequence, it is intuitively clear that both focused and less-focused users could be both necessary and complementary for an encyclopedia such as Wikipedia, but also that their relative ratio in the community of Wikipedians might result in very different aggregate outputs. To further address this issue, we now turn to studying online page signals.

5. Page signals

We now apply a similar methodology to our 3000 Wikipedia articles. First, we find that IEI distributions for articles (see Fig. 4) exhibit a pattern similar to user IEI distributions, with a left-hand side plateau, right-hand side circadian fluctuations and an intermediate linear, power law-like domain.

As far as we know, this is a new finding for inter-event interval on pages, even with respect to the physics literature, which is only focused on user IEI distributions. Using the same two procedures as for users, we calculate slopes \( \alpha \) for the IEI distributions of articles and find values in the range \([0.3,1.2]\) for \textit{MAX1D} and \([1.1,1.4]\) for \textit{NOMAX}.

Still following a similar methodology as for individual users, we define seven variables in order to describe the nature of interactions on a given article:

- Frequency (\textit{freq}) is obtained by dividing the number of revisions by the age of the article, and serves as a control.
- Featured articles are associated with \textit{adqnum} = 1 and non-featured with \textit{adqnum} = 0.
- \textit{length} refers to the length of the last revision of the article, in number of words.
- Talk Intensity (\textit{tkintensity}) is the ratio between the number of revisions on the related talk page and the number of revisions on the article. Over 99% of our 3000 articles have related talk pages.
Characterizing online communities with their “signals”
European Academy of Management Annual Conference 2010, Rome, Italy.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Param. estimate</th>
<th>Standard error</th>
<th>t-test</th>
<th>Prob &gt; t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.64914</td>
<td>0.00887</td>
<td>73.17</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>freq</td>
<td>1.71127</td>
<td>0.22182</td>
<td>7.71</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>adqnum</td>
<td>0.02394</td>
<td>0.00699</td>
<td>3.42</td>
<td>0.0006</td>
</tr>
<tr>
<td>length</td>
<td>-1.62486E-7</td>
<td>5.076523E-8</td>
<td>-3.20</td>
<td>0.0014</td>
</tr>
<tr>
<td>tkintensity</td>
<td>0.20860</td>
<td>0.01567</td>
<td>13.31</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>share rv</td>
<td>0.69383</td>
<td>0.03344</td>
<td>20.75</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>share anon</td>
<td>-0.18811</td>
<td>0.01621</td>
<td>-11.61</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

Table 2

pages, which are used as a coordination mechanism within Wikipedia since they host discussions related to the content of an article.

- Share of reverts (share rv) is the proportion “reverts”, i.e. revisions that purely revert the article to a previous version. This happens most likely during conflicts and “edit wars”.
- Share of anonymous (share anon) is the proportion of contributions from anonymous users.
- Average revisions per user (revs/user) is calculated by dividing the number of revision by the number of users who contributed on that article.

Based on these variables, the next step is to conduct a standard OLS regression in order to explain slope $\alpha$ (obtained with MAX1D), which characterizes each article. Results are presented in Table 2 ($R^2 = 0.23$). We removed 19 outliers from our 3000 articles after checking the distribution of each variable (according to $slope < 1.2$, $tkintensity < 1$, $revs/user < 20$ and $freq < 0.15$) and conducted PLS regressions to confirm the signs of parameter estimates. Note also that $revs/user$ was eventually removed from the regressions due to collinearities between dependent variables.

All variables are significant at either the 0.001 or even the 0.0001 level, and collectively explain 23% of the variance. Variables freq, adqnum, tkintensity and share rv are positively correlated to the slope of the power law fit (higher proportion of smaller inter-event pauses compared to longer ones). In contrast, length and share anon have a negative impact on the slope (higher proportion of longer inter-event pauses).

Controlling by freq, adqnum and length, Our results suggest that more interactive and “hotter” pages, as measured by the intensive use of talk pages (higher tkintensity) or by a higher proportion of reverts (share rv), are associated with higher slopes and a higher proportion of small intervals. This finding is not completely in line with Viegas et al. (2007), where talk pages are reported as a coordination mechanism which could allow users to “talk before they type”. Here, talking and typing seem to occur simultaneously, and the use of talk pages within Wikipedia could characterize shorter-term reactions and rapid successions of contributions – activity bursts – rather than reflection and longer-term planning. Conversely, it seems that, for articles that attract a larger share of anonymous contributions (higher share anon), longer pauses are comparatively more frequent.

These results could hint towards different coordination regime, with either “looser” interaction between editors, mediated by the article itself, or more interactive articles where coordination would be mediated by discussions on talk pages. This finding could also be consistent with the idea suggested by Suh et al. (2007), according to which some editors could devote themselves to some topics up to the point of acting as if they “owned” them. Of course, anonymous contributions on an article could also prompt “response” contributions from “owner(s)”, resulting in intermediary coordination regimes.

6. Conclusion and future works

We have suggested in this paper that by processing online activity as “online signals”, researchers could identify relevant characteristics of the individual and collective behavior of online users.
Taking Wikipedia as an example, and applying this signal processing methodology to inter-event distributions, we have suggested that the focus of users and the interactivity of coordination could belong to such relevant characteristics.

Combined together, these preliminary findings suggest a more complex vision of aggregation phenomena within online communities, where collective outputs would significantly depend both on how differently focused contributors collaborate and how they interact together. This result would also be consistent with the results of Dalle and David (2007) on open-source software projects, where both stigmergy (indirect interactions mediated by the code itself) and heterogeneously motivated users have been found to explain the nature of aggregate outputs.

Indeed, a very different article would probably result from the collaboration of weakly focused agents who do not talk together, than it would from the collaboration of highly focused agents who do talk a lot together. It is not obvious, however, whether these different coordination regimes could result in higher quality articles, since both regimes could be associated with potential drawbacks or "bugs", which would result in coordination failures, such as insufficient commitment for the former regime and conflictuality for the latter.

Therefore, an implied and more general issue which these results then raise is to determine in what respect online communities need and are able to develop rules, conventions, mechanisms or institutions which allow them to address the many bugs that result from the potential imperfections of online coordination as they result from self-organized aggregation and allocation of efforts. In the case of Wikipedia, den Besten et al. (2008b) suggest that the use of tags and templates, and thus more generally metadata, tends to play such a role; while den Besten et al. (2008a) suggest that, in the case of open-source software, the complexity of the code might contribute to address directly this issue without the need for metadata.

Needless to say, we believe that a several other characteristics of both users and pages would be worth analyzing in future works. Another extension would be to analyze the stationarity of online signals, as we would expect some users at least to change their behavior through time, a change that should be thus reflected on their online signals as well. More generally, we believe that it would be also valuable to perform similar analyses on other types of online signals, and notably on signals extracted from the activity of users who belong to online social networks.

References


den Besten, Matthijs, Max Loubser, Jean-Michel Dalle. 2008b. Wikipedia as a distributed Problem-Solving network. SSRN eLibrary.


