Participation in wiki communities: A statistical characterization (#63802)

First submission

Guidance from your Editor

Please submit by **7 Sep 2021** for the benefit of the authors (and your $200 publishing discount) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the materials page.

6 Figure file(s)
2 Latex file(s)
3 Raw data file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	***Example***
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Participation in wiki communities: A statistical characterization

Ámbar Tenorio Corresp., 1 , **Javier Arroyo** Corresp., 2 , **Samer Hassan** 2, 3

1 Decentralized Science, Madrid, Spain

2 Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, Madrid, Spain

3 Berkman Klein Center for Internet and Society, Harvard University, Harvard, Massachusetts, United States

Corresponding Authors: Ámbar Tenorio, Javier Arroyo
Email address: atenorio@ucm.es, javier.arroyo@fdi.ucm.es

Peer production online communities are groups of people that collaboratively engage in the building of common resources such as wikis and open source projects. In such communities, participation is highly unequal: few people concentrate the majority of the workload, while the rest provide irregular and sporadic contributions. The distribution of participation is typically characterized as a power-law distribution. However, recent statistical studies on empirical data have challenged the power-law dominance in other domains. This work critically examines the assumption that the distribution of participation in wikis follows such distribution. We use statistical tools to analyse over 6,000 wikis from Fandom/Wikia, the largest wiki repository. We study the empirical distribution of each wiki comparing it with different well-known skewed distributions.

The results show that the power-law performs sensibly poor, surpassed by three others, while the truncated power-law is superior to all others or superior to some and as good as the rest in 99.3% of the cases. Thus, we propose to consider the truncated power law as the distribution to characterize participation distribution in wiki communities. Furthermore, the truncated power-law parameters provide a meaningful interpretation to characterize the community in terms of the frequency of participation of occasional contributors and how unequal are the group of core contributors. Finally, we found a relationship between the parameters and the productivity of the community and its size. These results open research venues for the characterization of communities in wikis and in online peer production.

Participation in wiki communities: A statistical characterization

Ámbar Tenorio[1], Javier Arroyo[2], Samer Hassan[2,3]

[1] Decentralized Science, Madrid, Spain

[2] Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, Spain

[3] Berkman Klein Center for Internet and Society, Harvard University, Harvard Massachusetts, United States

Corresponding Author:

Ambar Tenorio[1]

Email address: atenorio@ucm.es

Participation in wiki communities: A statistical characterization

First Author[1] **and Second Author**[2]

[1]**Address of first author**
[2]**Address of second author**

Corresponding author:
First Author[1]

Email address: f.author@email.com

ABSTRACT

Peer production online communities are groups of people that collaboratively engage in the building of common resources such as wikis and open source projects. In such communities, participation is highly unequal: few people concentrate the majority of the workload, while the rest provide irregular and sporadic contributions. The distribution of participation is typically characterized as a power-law distribution. However, recent statistical studies on empirical data have challenged the power-law dominance in other domains. This work critically examines the assumption that the distribution of participation in wikis follows such distribution. We use statistical tools to analyse over 6,000 wikis from Fandom/Wikia, the largest wiki repository. We study the empirical distribution of each wiki comparing it with different well-known skewed distributions.

The results show that the power-law performs sensibly poor, surpassed by three others, while the truncated power-law is superior to all others or superior to some and as good as the rest in 99.3% of the cases. Thus, we propose to consider the truncated power-law as the distribution to characterize participation distribution in wiki communities. Furthermore, the truncated power law parameters provide a meaningful interpretation to characterize the community in terms of the frequency of participation of occasional contributors and how unequal are the group of core contributors. Finally, we found a relationship between the parameters and the productivity of the community and its size. These results open research venues for the characterization of communities in wikis and in online peer production.

INTRODUCTION

Since the emergence of online communities, one of the major topics of interest is to understand the different levels in which members participate: that is, the distribution of participation, also named distribution of work, or effort. Far from classical organizational structures, and more similar to volunteer-driven social movements, communities show an inherent participation inequality across its participants. Specifically in peer production communities, such as those in wikis and free/open source software, this issue has derived multiple research questions: the concentration of participation in an elite (Shaw and Hill, 2014; Kittur et al., 2007; Priedhorsky et al., 2007), the degree of participation inequality (Fuster Morell, 2010; Ortega et al., 2008; Neis and Zielstra, 2014), the characterization of who participates more (Hill and Shaw, 2013; Reagle, 2012), the process of changing user roles (Arazy et al., 2015; Preece and Shneiderman, 2009), or the evolution of participation depending on multiple factors (Vasilescu et al., 2014; Serrano et al., 2018).

An important bulk of peer production research tends to say that the distribution of participation follows a power-law. Intuitively, this means a very small number of contributors would concentrate most of the participation (or work), highlighting participation inequality. Formally, a power law is a simple relationship between two quantities such that one is proportional to a fixed power of the other.

In the issue at hand, i.e. participation, the two quantified dimensions are the number of contributions, and the share of people in the community that has made such number of contributions. The relationship among them is negative, that is, the higher the number of contributions, the smaller the share of contributors that has made such number of contributions. According to this idea, a small amount of contributions would be common, while larger amounts would be more rare. This fits with the assumption of participation inequality in which most members of the community tend to participate very little (occasional contributors),

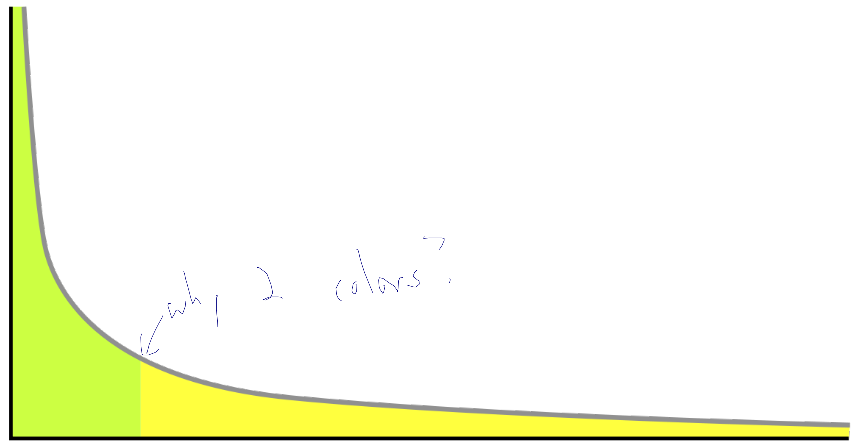

Figure 1. Power law distribution. For participation, the X axis represents the number of contributions made by a person and the Y axis the number of persons that made X contributions. Picture by Hay Kranen PD. available at Wikimedia Commons.

while a few of them account for an enormous amount of contributions (core contributors). In fact, the statement is not ungrounded, since several statistical studies focused on Wikipedia claim that the plot of edits per user follow a power law distribution (Kittur et al., 2007; Stuckman and Purtilo, 2011), and other studies find similar behavior in free/open source communities (Healy and Schussman, 2003; Sowe et al., 2008; Schweik and English, 2012; Cosentino et al., 2017) or other peer production communities (Wu et al., 2009; Wilkinson, 2008).[1]

Figure 1 shows an example of the power law. If we consider it represents a distribution for participation, the distribution models how frequent is to find a person that contributes X times. It can be seen that the frequency quickly declines as X grows, because most of the people only contribute a few times (green area). However, it is possible to find a few contributors with a very high number of contributions (yellow area).

The power law implies an underlying regularity in the behavior of the phenomenon under study. In particular, the power relationship should hold independently of which particular scale we are looking at. This may not be the case in real data. In fact, recent studies in statistics challenge the apparent dominance of power law across multiple fields with the help of modern sophisticated statistical tools (Clauset et al., 2009; Broido and Clauset, 2019). According to these works, power law distributions are complicated to detect because fluctuations occur in the tail of the distribution, and because of the difficulty of identifying the range over which power-law behavior holds.

In the peer production field, the regularity of the power law would imply that the relationship that holds for the occasional contributors would be the same to that for the core members, which may be a strong assumption for a community.

In particular, the tail of the distribution, which represents the activity of core contributors, may not have an extreme behavior as the power law suggests, i.e., the number of extremely active contributors may not be as high. If that is the case, more conservative distributions, such as the the truncated power law, would provide a better fit. In fact, such distribution was found suitable in a comparative analysis of the ten largest Wikipedias (Ortega, 2009).

According to these premises, it seems reasonable to question the characterization of the participation in peer production as a power law, and consider other heavy-tailed distributions. Thus, we will apply the statistical tools proposed by Broido and Clauset (2019) to study peer production distributions, and more precisely participation distributions from wiki communities. The statistical tools proposed in that work provide a test to determine whether a distribution provides a better fit than another with respect to the empirical data provided. Thus, we will use them to analyze whether one candidate distribution consistently provides a better fit than the others. The candidates will be five well-known distributions,

[1]Other studies just mention a highly skewed distribution or similar statements without further specification (Howison et al., 2006; Crowston et al., 2006; Barbrook-Johnson and Tenorio-Fornés, 2017).

namely, the power law, three heavy-tailed distributions with a tail more conservative than the power law (truncated power law, stretched exponential and log-normal) and a non-heavy tailed distribution (exponential), following the example by Broido and Clauset (2019).

In our work, we focus on Wikia, the largest wiki repository which provides a large and diverse sample of peer production communities. Wikia accounts for over 300,000 wikis. However, because of constraints of the statistical methods used, which require a certain minimum of observations, we will use for our analysis the ∼6,000 wikis which have at least 100 users.

The rest of the article proceeds as follows. Section details the process followed to perform the statistical analysis and for the data collection. Section shares the results of the statistical study of user contributions, and discusses its results through the explanation of series of graphs. Afterwards, Section offers an analysis of the winning distribution, i.e. the truncated power law, and proposes an interpretation of its parameters and how they characterize the different wikis under study. The paper closes with some concluding remarks and future work in Section .

METHODOLOGY AND DATA COLLECTION

Methodology

Following Clauset et al. (2009) and Broido and Clauset (2019), our study is divided in two analyses. First, in order to assess if the power law distribution is a plausible model for the given empirical data, we use the authors' goodness of fit test. Then, we perform an exhaustive analysis in order to identify which distribution better describes each wiki within the data set. These two methods are explained in this section.

Goodness of fit

Clauset et al. (2009) propose a statistical test in order to asses if a distribution plausibly follows a power law. First, the power law distribution is used to model the data, finding its slope, or α parameter, and the minimum value from which the power law behavior is observed, or x_{min} parameter.

Afterwards, in order to compare the empirical data to different distributions, we create a set of comparable synthetic data-sets that follow the distribution (i.e. have the same parameters). This allows us to compare the real data with the synthetic data, and see how they deviate from each other. This method is considered more accurate than comparing the deviation with an ideal distribution which real data may never fit. Thus, we artificially create 100 synthetic data-sets per wiki, for each of the five distributions.

Thus, the distance of the real data to its power law model is compared with the distance of the synthetic data sets to their power law models. Note that the synthetic datasets are also fit to power law models to compete in similar conditions These distances are calculated using the Kolmogorov-Smirnov (KS) statistic. The goodness-of-fit test returns a p-value between 0 and 1 representing the number of synthetic dataset fits that outperformed the real data fit. E.g. a p-value of 0.4 represents that the real data fits better the power law than the 40% of the synthetically generated data. This p-value is then used to decide whether to rule out the hypothesis of the data following a power law. In our study, we rule out the power law model hypothesis if the p-value is smaller than 0.1, as Clauset et al. (2009) and Broido and Clauset (2019) do, i.e. if the probability of obtaining a worse fit by chance is smaller than 10%. The number of synthetic data sets used to calculate the p-value determines the accuracy of the result. Following Clauset et al. (2009), for the result to be accurate to within ε, we should generate about $\varepsilon^{-2}/4$ samples. Our study generates 100 synthetic data sets per test, therefore, the results are within an ε of 0.05.

When the number of observations is relatively small, this goodness of fit test cannot rule out a power law model in those cases in which the data follows other distributions such as the log-normal or exponential. For instance, for data following an exponential distribution with $\lambda = 0.125$, at least 100 observations are needed for the average p-value to drop bellow our threshold of 0.1, while for data following a log-normal distribution with $\mu = 0.3$, the average p-value drops below 0.1 from around 300 observations (Clauset et al., 2009). Thus, high p-values in these distributions with small number of observations should not be interpreted as the data following a power law. Moreover, as studied in the following section, even if a distribution plausibly follows a power law, other distributions may fit the data better. This work considers wikis with more than 100 observations (i.e. wikis with over 100 contributors) for the p-value study for two reasons. First, as already mentioned the goodness-of-fit test would not be able to rule out competing distributions. Second, as the wikis with less than 100 contributors represent more than 98% of wikis (See Section), the percentage of wikis passing the test due to the small number of observations may hinder the adequacy of the power-law hypothesis for those wikis with enough data to provide test results significant enough to distinguish from alternative models.

Summarizing, our study considers distributions with more than 100 observations (i.e. wikis with over 100 contributors), performs the goodness-of-fit tests proposed by Clauset et al. (2009) considering those with a p-value greater or equal to 0.1(±0.[illegible]8) to plausibly follow a power law. The results of these tests are presented in Section .

This study was performed using the *poweRlaw* R package (Gillespie, 2014). Besides, the R script source code, required for applying these statistical tests to our data, is available as free/open source software to facilitate replication.[2]

Likelihood-ratio test

The previously described goodness of fit test provides a tool to decide whether to rule out a power law distribution as a good model for the data. However, even if a power law model is not rejected, there may be better alternative distributions. The likelihood-ratio test allows us to compare the likelihood of the empirical data fitting two competing distributions. That is, it establishes which distribution is more likely to fit the data, and whether the difference is significant.

Following the approach described by Clauset et al. (2009), our study compares the likelihood of 5 different skewed distributions. Our hypothesis is that the power law is too ”ambitious” for the observations of the tail. We also expect the distribution to be heavy tailed, i.e. with a decrease of the tail slower than in an *exponential distribution*. In addition to these two distributions that frame the expected tail of our data, our study adds three skewed distributions that would lie in between, presenting a slower decrease in the tail than the exponential but a stronger decrease than the power law: the *truncated power law* (also named power law with exponential cut-off), the *log-normal* and the *stretched exponential*. Both the truncated power law and the log-normal distributions have two terms, while the power law, exponential and stretched exponential have only one. The number of terms of the distributions is relevant, since it is a factor for fitness.

The study exhaustively compares, for each wiki, the fit of the data to those five skewed distributions (power law, truncated power law, log-normal, exponential and stretched exponential), and identifies when the likelihood differences are statistically significant. It uses the Vuong method **(?)**, which considers the variance of the data, and returns a p-value that states if the likelihood differences may be due to the data fluctuations, or are significant in order to favor one distribution over the other.[3] As Clauset et al. (2009), we consider significant the differences with a p-value smaller than 0.1, i.e. those that have less than 10% probabilities of being a result of the data fluctuations. Additionally, in order to avoid over-fitting to the tail of the distribution, we force the method to fit every contributor with at least 10 contributions. If we do not impose this condition, the method could exclude many contributors in order to find a better fit for the most active contributors, for instance a fit for the people with more than 500 contributions.

This study was performed using the *Powerlaw* python package (Alstott et al., 2014). Similar to the previous subsection, the python script source code, required for the performed analysis, is available as free/open source software to facilitate replication.[4]

Data collection

This work investigates the distribution of participation in wikis from Wikia/Fandom studying the number of edits per user. Wikia/Fandom is a suitable research object to draw conclusions about participation in wikis in general. As argued by Shaw and Hill (2014), Wikia is an ideal setting in which to study peer production. Wikia only hosts publicly accessible, openly-licensed, volunteer-produced, peer production projects. To date, it is the largest and more diverse repository of open knowledge peer production, with a rich ecosystem of a broad diversity of topics, languages, community and wiki sizes. Furthermore, Wikia never restricts viewership, nor participation (except that from spammers or vandals). Wikia hosts some of the largest and most successful wikis in multiple topics and languages, such as Marvel or Star Wars fandom wikis, LyricWiki on song lyrics, Proteins scientific wiki, or AmericanFootballDatabase on such sport.

[2] Goodness of fit tests script: ANONYMIZED

[3] The method is adapted by Clauset et al's for nested distributions such as power law and truncated power law, where a family of distributions is a subset of the other. Such modified method, which we use as well, allows to state whether the larger family is indeed needed or both distributions are good models.

[4] Likelihood-ratio test script: ANONYMIZED

To collect our data we used the publicly available Wikia census described by Jiménez-Díaz et al. (2018) and retrieved on the 20th of February 2018.[5] However, as explained in Section , we limit our analysis to wikis with at least 100 registered users which have done at least one edit, and excluding bot users.

Thus, starting from this census data, and complementing it with additional information as explained below, we have created a new dataset to study the distribution of participation, i.e. which is the distribution of edits made by registered users, excluding bots. This dataset is complete, since it includes all the Wikia wikis with at least 100 users which made at least one contribution, resulting in 6,676 wikis, as explained in detail below.

The mentioned Wikia census provides information of ∼300,000 wikis. However, the census does not provide information on the number of edits of each user in each wiki. Thus, such information needs to be generated manually to complement the dataset.

Therefore, in order to retrieve the required data, we need to query the API of each of the wikis hosted in Wikia. Spefically, we need to query the Special:ListUsers API endpoint that every MediaWiki wiki has.[6] Such Special:ListUsers page lists the information of every registered user in a given wiki, e.g. username, number of edits, groups she belongs to, or date of last edit made. A perl script was developed in order to use that endpoint and obtain the number of edits performed by each registered user. In particular, the script queries the endpoint making a request for all users. Afterwards, it filters out the bot users, removing the users belonging to the *bot* and *bot-global* groups. As with the previous scripts, this perl script source code is available as free/open source software to facilitate replication.[7]

The data collection was performed on November 6, 2018 and it is publicly available.[8] It contains information about 295,658 wikis, as 8,433 wikis endpoints were technically unavailable.

This data, i.e. the census wikis with the edits information, was curated to avoid duplicates and to filter out wikis without human participation (i.e. bot only) and without statistical data provided by Wikia. After removing them, the collection contains information about 282,039 wikis.

The reliability of the data collected is considered high. The edit numbers are as reliable as Wikia publicly accessible statistics are (i.e. those from the Special:ListUsers endpoint). We have also done a consistent effort in bot identification in order to filter them out.

For statistical reasons already explained in Section , this work considers only wikis with at least 100 registered (non-bot) users. Thus, the number of considered wikis was further reduced to 6,676. It is important to remark that this is not a sample, but the observed full population of wikis with at least 100 registered users with contributions in Wikia.

RESULTS OF THE STATISTICAL TESTS

According to the goodness of fit test described in Section , the power law is a plausible distribution (i.e. it cannot be ruled out) for the 83% of the 6,676 wikis from Wikia/Fandom with at least 100 registered non-bot users. However, as explained in Section , that does not mean that the power law is the best choice, since other distributions may fit better the empirical data.

Thus, we perform the likelihood-ratio test to compare the pairs of the five candidate distributions as explained in Section . The distributions are: power law, truncated power law, exponential, stretched exponential and log-normal. For each wiki, we perform likelihood-ratio tests comparing all the competing distributions against each other. That is, we perform 10 likelihood-ratio tests for each wiki, since there are 10 possible couples.

Figure 2 summarizes the results of these comparisons. The figure's pentagon apexes shows each of the five considered distributions. An arrow from distribution A to distribution B represents the percentage of wikis in which distribution A was preferred over distribution B in the likelihood-ratio test, while the opposite arrow represents the percentage of wikis where distribution B was superior to distribution A. Note in some cases, the likelihood-ratio test may be inconclusive to determine which of the two distributions is better for a given wiki, and in those cases neither A nor B is superior. It is important to remark that the test being inconclusive means that both distributions fare similarly, which could mean that both are

[5]Wikia census: `https://www.kaggle.com/abeserra/wikia-census`

[6]Note all Wikia wikis use the same wiki software, MediaWiki, maintained by Wikimedia Foundation and used by its projects, including Wikipedia.

[7]Script to retrieve user contributions: ANONYMIZED

[8]ANONYMIZED

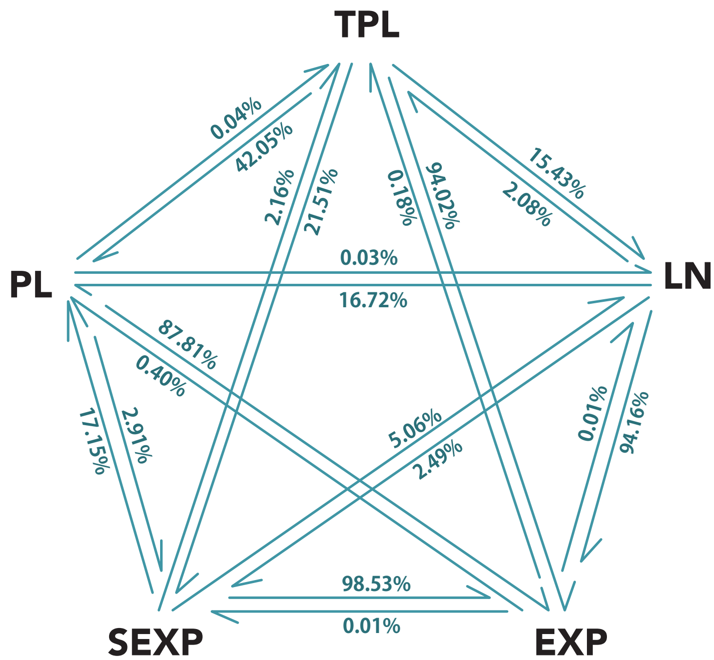

Figure 2. Results of the likelihood-ratio test between the five considered distributions for registered users. The distributions considered are: power law (PL), truncated power law (TPL), log-normal (LN), exponential (EXP) and stretched exponential (SEXP). Each arrow from A to B has the percentage of cases in which A was superior than B.

adequate or even that both are inadequate. For the sake of clarity, the figure omits the complementary percentage where the likelihood-ratio test was inconclusive, although it can be easily calculated.[9]

The analysis of the figure results shows that the power law is not a strong contender, as it is rarely a more likely distribution than any of its competitors, with the exception of the exponential distribution, which is also overwhelmingly defeated by the rest of the candidates.

The defeat of the exponential distribution by all candidates means that a large tail of core users is clearly present in the wiki participation distributions, and thus that an exponential distribution, which is not able to represent heavy tails, is not a good candidate.

However, the power law being defeated by the rest of the heavy-tailed distributions means that the tail is not as heavy or large as a power law would predict. Hence, more moderated heavy-tailed distributions are required. This conclusion is similar to the one drawn in recent works that disprove the supposed prevalence of the power law in other domains (Clauset et al., 2009; Broido and Clauset, 2019).

Thus, a correct characterization of the distributions, in nearly all cases, lies in between the exponential and the power law distributions. Among the rest of the candidates, the truncated power law stands out, since as seen in Figure 2, it is rarely beaten by its competitors: 2.16% against the stretched exponential, 2.08% against the log-normal, 0.18% against the exponential, and 0.04% against the power law distribution. Hence, the likelihood-ratio test clearly supports the truncated power law as the most appropriate distribution to characterize participation.

The appropriateness of the truncated power law is better appreciated when we aggregate the results of the likelihood-ratio tests for each wiki as shown in Table 1. We count the cases where a candidate distribution won all the likelihood-ratio tests for each wiki, which means that that distribution is the right choice for that wiki. In addition, we also counted the times where a candidate distribution lost at least one test, which means that for that wiki the candidate distribution was not the best choice.

It is important to remark that only in 10 wikis (0.15%) no candidate distribution won any likelihood-ratio test which means that they all were equally good (or, more precisely, bad) candidates. We have

[9] In all cases, percentage of A¿B + percentage of A¡B + percentage of inconclusive = 100%

Distribution	Wins all tests	Loses at least one test
Power law	0 (0%)	2816 (42.18%)
Truncated power law	596 (8.93%)	177 (2.65%)
Log-normal	41 (0.61%)	1159 (17.36%)
Stretched exponential	2 (0.03%)	1492 (22.35%)
Exponential	0 (0%)	6578 (98.53%)

Table 1. Aggregated results of the likelihood-ratio tests for each wiki counting the cases where a candidate distribution wins all tests and loses at least one test

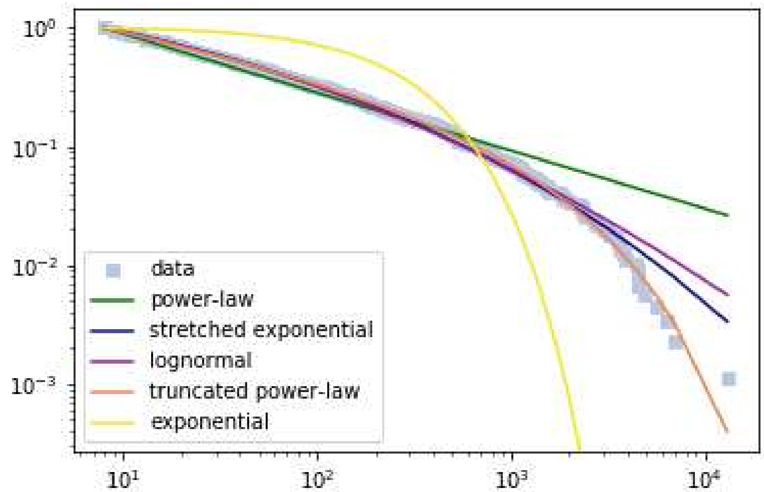

Figure 3. Complementary cumulative distribution function of participation of a wiki and the fitted distributions.The X axis represents the logarithm of number of edits and the Y axis the inverse cumulative relative frequency the percentage of contributors that made at least X edits in the wiki.

inspected these cases and they all exhibit uncommon participation distributions.

According to Table 1, the truncated power law is significantly better than all the candidates in 596 wikis out of the 6,676, i.e. approx. 9% of the wikis considered. While the rest of the distributions fare much worse: only the log-normal and stretched exponential distributions are the best candidates in 41 and 2 wikis, respectively. The power law and the exponential are not the best candidates for any wiki, which reinforces the idea of the suitability of a heavy-tailed distribution but not as heavy as that from the power law.

According to the aggregated results in Table 1, the truncated power law is not the best or among the best candidates for only 177 wikis out of 6,676 wikis (2.65%); more precisely in 67 wikis (1%) looses one test, in 101 wikis (1.51%) looses two tests and in 9 wikis (0.1%) looses three tests. The rest of the distributions fare much worse, e.g. log-normal can be ruled out as the best candidate in the 17.36% of the wikis and the stretched exponential in the 22.73%. This result reinforces the idea of the truncated power law being the *distribution of choice* when trying to characterize the participation distribution in wikis, because it seems difficult to find a better one for most of the cases.

We show an example of participation distribution where the truncated power law won all the tests in Figure 3. The figure shows a log-log plot of the complementary distribution function where the X axis represents the logarithm of the number of edits in the wiki and the Y axis the inverse cumulative relative frequency, i.e. the percentage of contributors that made at least X edits in the wiki. The figure displays the observations (grey squares) and the fitted distributions, i.e. the truncated power law and all the candidate distributions. The observations in the left side of the graph represent the contributors with fewer edits, while those most towards the right are the core contributors that made most edits, i.e., the tail

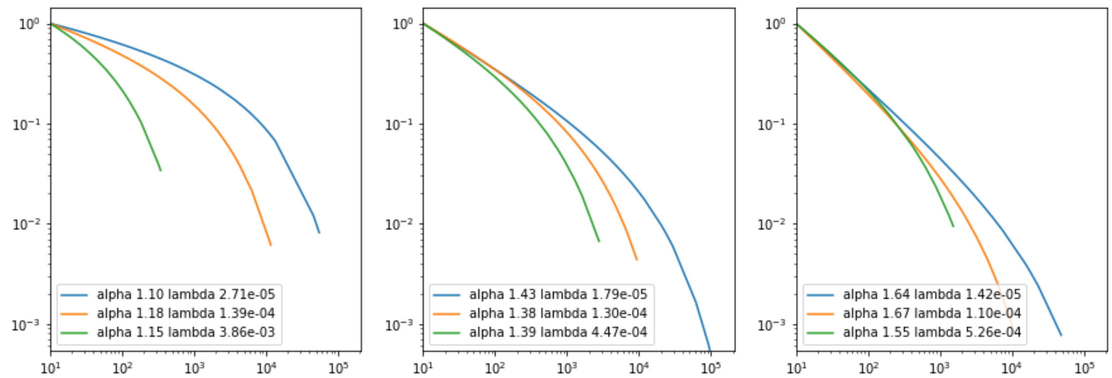

Figure 4. Complementary cumulative distribution functions in logarithmic scales of truncated power laws. Each sub-figure plots three wikis with similar α parameter, adopting smaller values in the left plot, average values in the middle and higher values in the right. The X axis represents the logarithm of number of edits and the Y axis the inverse cumulative relative frequency the percentage of contributors that made at least X edits in the wiki.

of the participation distribution.

In this figure, first we can observe the different tails of the considered distribution. While the exponential has the most conservative tail, the power law is the one that has a heavier tail, while the rest of the distributions have a tail in between them. Regarding the data fitting, the exponential with his bounded tail is not able to model the community behavior at all. The rest of them fit the initial slope, but only the truncated power law is able to successfully grasp the tail behavior, because the others predict a heavier tail.

Note the participation distribution in Figure 3 is one of the 9% examples in which the truncated power law wins all test. Still, as mentioned, in most of the cases ($97,35\%$), the Truncated Power law is not defeated by any other distribution. Such cases typically correspond with participation distributions with tails that can be conveniently fitted by the truncated power law, but also by the log-normal and/or the stretched exponential. So, according to this statistical evidence, the truncated power law is in fact the most adequate distribution for wiki participation.

The statistical analysis carried out shows that the truncated power law is the best distribution to characterize the participation in wikis among those considered, as it is barely rejected and is the only proper fit in 9% of the cases. In the next section, we will interpret the parameters of this distribution in the context of participation and will relate them with the characteristical features of the wiki communities.

ANALYSIS OF THE TRUNCATED POWER LAW FOR CHARACTERIZING PARTICIPATION DISTRIBUTIONS

In this section, we will explore the diversity of participation distributions that are modelled by the truncated power law, but before that, we need to understand better the effect and interpretation of the parameters that define the the truncated power law.

Interpretation of the truncated power law parameters

The truncated power law is defined as a power law multiplied by an exponential: $x^{-\alpha}e^{-\lambda x}$. In the log-log plot, the parameter α is related to the slope of the power law function, while the parameter λ is related to the starting point and/or the steepness of the decay in the tail.

As a result, lower alphas can be associated with a higher frequency of participation of occasional contributors, as their frequency decreases less conspicuously as the number of contributions increase than in the case of higher alphas. In other words, in communities with lower alphas the frequency of contributors with more contributions decreases less significantly.

On the other hand, higher lambdas can be associated with more pronounced deviations from the power law in the tail, which means that more active contributors are less frequent as what the power law would

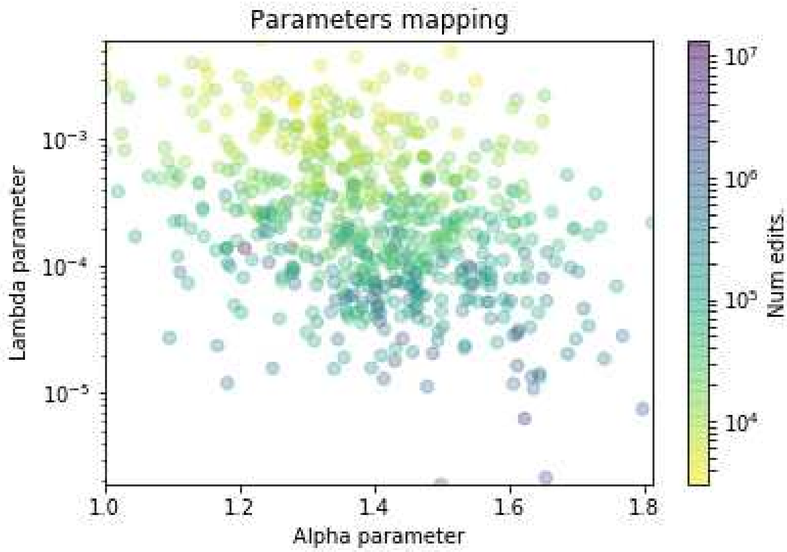

Figure 5. Scatter plot of the TPL-distributed wikis where the color represents the number of edits.

predict. Thus, higher lambdas relate to less inequality among active contributors than predicted by the power law.

In Figure 4, we show the truncated power law of nine wikis with different α and λ parameters that illustrate how diverse may be the participation distributions in wikis. From left to right we show three plots each of them with three participation distributions with roughly similar α values (the alpha values grow from the left to the right plot). In each plot, we show participation distributions with similar α but with different λ values. This figure illustrates the idea that the initial slope of the distributions depends on α values, as it is steeper from the left to the right plots. Besides, in each figure we can appreciate that higher values in the λ parameter are associated with a more pronounced and earlier decay sooner, or, conversely, smaller values allow the power law relationship to prevail longer.

Relationships of the parameters with features from the wiki communities

In this section we explore whether the α and λ parameters are related to some features from wiki communities, namely, the number of edits and the number of participants. We will use scatter plots in which each dot represents a wiki in a 2-dimensional plot. The plot axes represent the values of the α and λ parameters, and the dot is colored according to a color gradient related with the specific wiki feature. More precisely, in Figure 5 the color represents the number of edits, and in Figure 6, it represents the number of users of the wiki. For the sake of clarity, the plot will only display the wikis where the truncated power law distribution won all the likelihood-ratio tests.

The scatter plots show a cloud of dots with no clear relationship among the parameters. The relationship could be inverse, since the cloud rarely includes wikis with large α and λ values or wikis with small α and λ values. However, the variability is very high to see a clear pattern.

When studying the relationship of the parameters with the size of the community in Figure 5, we can observe how the λ parameter seems to be inversely related to the number of edits of the wiki, as the largest wikis are distributed in the lower part of the figure and vice versa. In other words, larger wikis (those with millions of edits) have smaller lambdas, which means that the decay in the tail of their participation distributions is not as significant. It reveals that, given an alpha value, there are more core contributors than in wikis whose participation distributions have higher lambda values, and that results in more productive communities in terms of edits. On the contrary, wikis with higher lambdas have a less populated elite of core contributors which results in smaller wikis in terms of edits.

At Figure 6, we can observe that the number of users of the wiki is related to the combination of both parameters, as we can see that the color gradient evolves from the upper-left towards the bottom-right corner. Participation distributions characterized by high alpha values and low lambda values belong mostly to larger wiki communities (blue dots). Such parameter values determine an extremely sharp decrease in

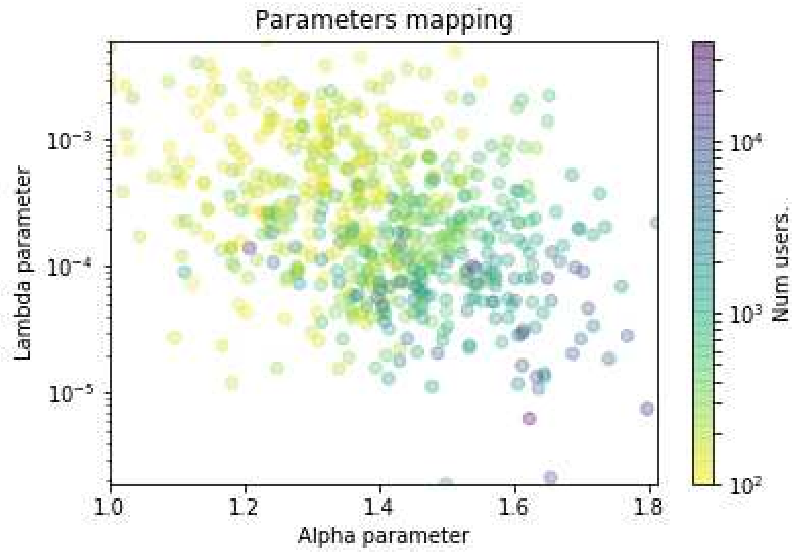

Figure 6. Scatter plot of the TPL-distributed wikis where the color represents the number of contributors.

the (relative) frequency of editors as the number of edits increases, and also a more pronounced decay on the frequency of the most active contributors. In other words, extremely unequal participation distributions can be found mostly in large wiki communities. Conversely, we can find that less unequal distributions of participation (those with low alpha and high lambda values) characterize mostly the distribution of participation of wikis with smaller communities (yellow dots).

We cannot conclude if higher inequality is cause or consequence of larger communities and vice versa. Such confirmation would require further research. However, it seems that there is a clear link between community size and participation distribution.

Furthermore, it is important to bear in mind that we are observing the participation distribution during the whole life of the wiki, that is, the aggregated effect of different communities that interacted in the wiki across time, since new users come and other leave, or contribute in different degrees, throughout their evolution. In fact, larger communities are usually older communities. In this sense, it would be interesting to observe how the yearly participation distribution in these wikis evolved, because the highlighted inequality could potentially be the result of the aggregation throughout the years of more egalitarian distributions of participation.

CONCLUDING REMARKS

In this work, we have critically studied the distribution of participation in wikis. We have analyzed the ~300,000 wikis from Wikia, selecting the 6,676 wikis with at least 100 users to perform our statistical analysis. This is considered an extensive and diverse population, appropriate for an analysis following the approach defined by Clauset et al. (2009). According to our results, the power law is not an appropriate distribution for wiki participation, as it predicts a higher proportion of highly active users than the observed in these communities. This contradicts the bulk of the peer production literature, which refers to the power law as the reference distribution when discussing about contributor participation.

In our statistical analysis we have considered potential alternatives, and from these distributions, the truncated power law gives clearly the best fit with the empirical data. Consequently, it should be considered as the distribution of participation of choice when characterizing wiki communities. Of course, it may not be adequate for some specific communities, and yet it has been able to characterize effectively the vast majority of them, while the other candidates performed significantly worse. In our analysis, we have found that the parameters of the truncated power law distribution (that govern the slope and the decay of the power law relationship in a wiki project) are related with the number of members in the community and the number of edits in the project. However, the reasons behind these findings deserve deeper consideration and are a matter of future research.

The prevalence of the truncated power law as the distribution of choice for characterizing the participation distribution in wikis has several implications:

- The truncated power law implies that the power law behavior holds true only in a limited range of wikis, and that from that point a decay can be observed. In a distribution of participation, it means that the truncated power law fits better not only concerning the frequency of participation of occasional contributors, but also concerning the frequency of the most active ones. The change of slope may also serve to empirically determine a division between core and non-core contributors instead of using arbitrary divisions as in other studies (Kittur et al., 2007). Further research may provide insights on how and why the inner dynamics change, and how we can study better the different emergent roles within peer production communities.
- In a truncated power law, the core contributors, i.e. the highly active members, are rarer than with a power law with the same slope. That means that, when looking at the distribution tail, we can observe a sharper decrease in the frequency of contributors as the edit activity increases. It seems to reinforce the idea that core contributors are somehow special, in the sense that there is a qualitative change in their work and motivations (Burke and Kraut, 2008) and thus higher barriers to join them, and/or the elitization of the core leads to oligarchies (Shaw and Hill, 2014). The reasons behind could be due to community dynamics such as some kind of elitism that prevents more people to be involved as much as those more active in the community, or that many active users experiment a burnout at some point and cease or decrease their activity level (Jiang et al., 2018).

The approach followed by this work has several limitations:

- It is a descriptive quantitative work, and thus it lacks explanatory aspects that further qualitative research could contribute with.
- We are cautious with the generalizability of our findings beyond Wikia to wiki communities and more in general, to other peer production communities. Still, considering the significant size and diversity of the sample used, and similar generalizations performed in the field, for example by Shaw and Hill (2014)), there is good evidence for potential generalizability. In order to support this generalization, these results would need to be validated in other projects, such as the Wikimedia Foundation projects, as well as in other peer production communities such as Free/Open Source Software projects. Thus, we encourage other researchers to replicate our approach with other peer production communities.
- The statistical analysis methods employed require a certain wiki size to have conclusive results, which may constrain their applicability for very small wikis. Despite of having near 300,000 wikis in Wikia, most of them are under 100 users and thus are discarded, using ”only” 6,676 wikis in the analysis.
- We have analyzed the participation in the communities aggregated through time (years), that is, accumulating the participation of all the members from the beginning. However, the members of a wiki community change through time, as change the participation dynamics. The participation distribution could be different when analyzed in a smaller time window, such as a year.

We have already defined several potential lines for future work, but we would like to mention those that we consider more interesting:

- To use a different base population, in order to appropriately generalize for peer production communities and not just wikis. For instance, we could analyze in a similar manner communities from Github, Wikimedia Foundation projects, or Stack Exchange.
- To perform a temporal analysis with a rolling time window, to understand how these distributions evolve over time, especially considering the evolution of the truncated power law parameters and how they relate with participation dynamics and inequality.
- To study the characterization of wikis based on their truncated power law parameters, i.e. clustering similar wikis and explaining the causes or consequences of the different typologies and how they relate with factors such as maturity stage, community dynamics and sustainability.

Our work asserts the truncated power law is probably the most appropriate distribution to represent the distribution of participation in wikis from Wikia. Our results can be better understood if they are observed in the context of a previous study that questioned the prevalence of power law in several fields (Clauset et al., 2009) and the ground-breaking finding that the power-law was indeed rare in real-life networks (Broido and Clauset, 2019). Our finding will thus open new lines of research, revisiting old assumptions in the field, exploring further the causes behind the observed structural change in core contributor participation and the relationships with the sizes of the community and the project and other factors behind the behavior.

ACKNOWLEDGMENTS

ANONYMIZED

REFERENCES

Alstott, J., Bullmore, E., and Plenz, D. (2014). powerlaw: A Python package for analysis of heavy-tailed distributions. *PloS one*, 9(1):e85777.

Arazy, O., Ortega, F., Nov, O., Yeo, L., and Balila, A. (2015). Functional roles and career paths in wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1092–1105. ACM.

Barbrook-Johnson, P. and Tenorio-Fornés, A. (2017). Modelling commons-based peer production: The commoners framework. In *Social Simulation Conference 2017 (SSC2017). Dublin, Ireland*. European Social Simulation Association (ESSA).

Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*.

Burke, M. and Kraut, R. (2008). Taking up the mop: identifying future wikipedia administrators. In *CHI'08 extended abstracts on Human factors in computing systems*, pages 3441–3446.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Cosentino, V., Izquierdo, J. L. C., and Cabot, J. (2017). A systematic mapping study of software development with github. *IEEE Access*, 5:7173–7192.

Crowston, K., Wei, K., Li, Q., and Howison, J. (2006). Core and Periphery in Free/Libre and Open Source Software Team Communications. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006. HICSS '06*, volume 6, pages 118a–118a.

Fuster Morell, M. (2010). Participation in online creation communities: Ecosystemic participation. In *Conference Proceedings of JITP 2010: The Politics of Open Source*, volume 1, pages 270–295.

Gillespie, C. S. (2014). Fitting heavy tailed distributions: the powerlaw package. *arXiv preprint arXiv:1407.3492*.

Healy, K. and Schussman, A. (2003). The ecology of open-source software development. Technical report, Technical report, University of Arizona, USA.

Hill, B. M. and Shaw, A. (2013). The wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one*, 8(6):e65782.

Howison, J., Inoue, K., and Crowston, K. (2006). Social dynamics of free and open source team communications. In Damiani, E., Fitzgerald, B., Scacchi, W., Scotto, M., and Succi, G., editors, *Open Source Systems*, number 203 in IFIP International Federation for Information Processing, pages 319–330. Springer US.

Jiang, L., Mirkovski, K., Wall, J. D., Wagner, C., and Lowry, P. B. (2018). Proposing the core contributor withdrawal theory (ccwt) to understand core contributor withdrawal from online peer-production communities. *Internet Research*.

Jiménez-Díaz, G., Serrano, A., and Arroyo, J. (2018). A wikia census: motives, tools and insights. In *Proceedings of Opensym 2018*. ACM.

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., and Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19.

Neis, P. and Zielstra, D. (2014). Recent developments and future trends in volunteered geographic information research: The case of openstreetmap. *Future Internet*, 6(1):76–106.

Ortega, F. (2009). *Wikipedia: A quantitative analysis*. PhD thesis, PhD thesis. Universidad Rey Juan Carlos, Madrid.

Ortega, F., Gonzalez-Barahona, J. M., and Robles, G. (2008). On the inequality of contributions to wikipedia. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, pages 304–304.

Preece, J. and Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. *AIS transactions on human-computer interaction*, 1(1):5.

Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., and Riedl, J. (2007). Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 259–268. ACM.

Reagle, J. (2012). "free as in sexist?" free culture and the gender gap. *first monday*, 18(1).

Schweik, C. M. and English, R. C. (2012). *Internet success: a study of open-source software commons*. MIT Press.

Serrano, A., Arroyo, J., and Hassan, S. (2018). Webtool for the analysis and visualization of the evolution of wiki online communities. In *Proceedings of the European Conference on Information Systems (ECIS) 2018*. AIS Electronic Library (AISeL).

Shaw, A. and Hill, B. M. (2014). Laboratories of oligarchy? how the iron law extends to peer production. *Journal of Communication*, 64(2):215–238.

Sowe, S. K., Stamelos, I., and Angelis, L. (2008). Understanding knowledge sharing activities in free/open source software projects: An empirical study. *Journal of Systems and Software*, 81(3):431–446.

Stuckman, J. and Purtilo, J. (2011). Analyzing the wikisphere: Methodology and data to support quantitative wiki research. *Journal of the American Society for Information Science and Technology*, 62(8):1564–1576.

Vasilescu, B., Serebrenik, A., Devanbu, P., and Filkov, V. (2014). How social q&a sites are changing knowledge sharing in open source software communities. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 342–354. ACM.

Wilkinson, D. M. (2008). Strong regularities in online peer production. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 302–309. ACM.

Wu, F., Wilkinson, D. M., and Huberman, B. A. (2009). Feedback loops of attention in peer production. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 409–415. IEEE.