Measuring the Wikisphere*

Jeff Stuckman
Department of Computer Science
University of Maryland
College Park, Maryland USA
stuckman@umd.edu

James Purtilo
Department of Computer Science
University of Maryland
College Park, Maryland USA
purtilo@umd.edu

ABSTRACT

Due to the inherent difficulty in obtaining experimental data from wikis, past quantitative wiki research has largely been focused on Wikipedia, limiting the degree that it can be generalized. We developed WikiCrawler, a tool that automatically downloads and analyzes wikis, and studied 151 popular wikis running Mediawiki (none of them Wikipedias). We found that our studied wikis displayed signs of collaborative authorship, validating them as objects of study. We also discovered that, as in Wikipedia, the relative contribution levels of users in the studied wikis were highly unequal, with a small number of users contributing a disproportionate amount of work. In addition, power-law distributions were successfully fitted to the contribution levels of most of the studied wikis, and the parameters of the fitted distributions largely predicted the high inequality that was found. Along with demonstrating our methodology of analyzing wikis from diverse sources, the discovered similarities between wikis suggest that most wikis accumulate edits through a similar underlying mechanism, which could motivate a model of user activity that is applicable to wikis in general.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation*; K.4.3 [**Computers and Society**]: Organizational Impacts—*Computer-supported collaborative work*

General Terms

Measurement

Keywords

wiki, Mediawiki, crawler, gini, power law, distribution, metrics

*Authors are supported in part by Office of Naval Research contract N000140710329 while doing this research.

WikiSym '09, October 25-27, 2009, Orlando, Florida, U.S.A.

1. INTRODUCTION

Mediawiki is an open-source platform used to host wikis (user-editable repositories of content). Wikipedia, the most popular instance of Mediawiki, has popularized wikis, and many webmasters have installed their own wikis using this platform as a result. Many prior studies have analyzed Wikipedia and its user base, but few have examined other wikis. This is because Wikipedia makes database dumps available to researchers for analysis[1], while obtaining dumps of other wikis would require the cooperation of their individual webmasters. The result is a lack of information on how wikis other than Wikipedia are being used, information that could increase the knowledge available to wiki practitioners by making the increasingly sophisticated models and analysis of Wikipedia applicable to wikis in general.

To narrow this knowledge gap, we developed WikiCrawler, a tool that converts wikis running Mediawiki into machine-readable data suitable for research by parsing their generated HTML pages. We then assembled a collection of 151 popular wikis of varying sizes (totaling 132393 wiki pages) and observed that nearly all were authored collaboratively (like Wikipedia). Then, to demonstrate that our methodology can produce useful data for analysis, we analyzed the distributions of activity (across users and articles) for each wiki. We determined that the studied wikis have highly unequal distributions of activity across authors, and that the inequality in these observed distributions can largely be predicted by their underlying power-law forms. We conclude that large-scale quantitative analysis of wikis is practical, and that our methodology can be used to generalize findings for Wikipedia to a broader population of wikis, enabling new applications such as calibration of wiki metrics. Our findings also support the notion that a generative model could be developed that accurately reflects the activity distributions of most wikis.

2. RELATED WORK

The methodologies of many previous studies have included quantitative analysis of Wikipedia. Some of these studies measured the distributions of edits across articles or users, such as [12], which found that article lengths had a log-normal distribution and the number of unique editors per article had a power-law distribution. The mechanisms that generate these distributions were studied by Wilkinson and Huberman [13], who proposed a stochastic model that produces a log-normal distribution for the lengths of articles

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

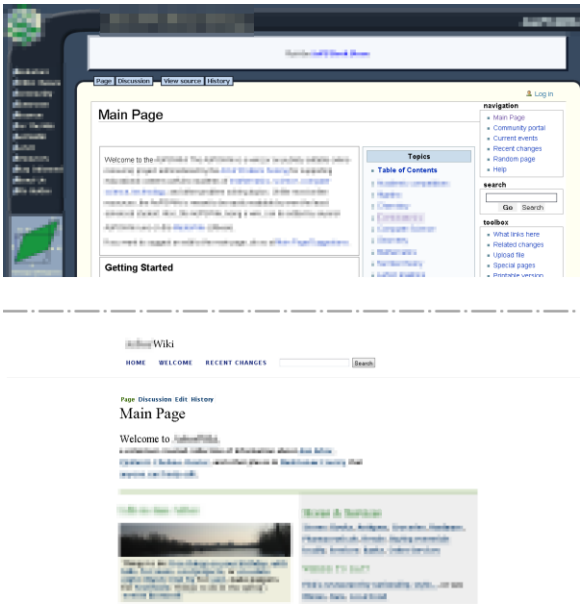

Figure 1: Two examples of wikis hosted by Mediawiki

with a given age.

One actively debated question involves the role of frequent users versus occasional contributers to Wikipedia. Kittur et al. [4] tracked words and revisions of articles to suggest that occasional users are responsible for an increasing amount of Wikipedia's content, while Ortega et al. [8] used a Gini coefficient metric to conclude that few users are responsible for a bulk of the activity. However, the scope of all such studies was limited to Wikipedia.

The S23[2] website, which provides statistics for tens of thousands of public wikis, has been used for research that compared multiple wikis, such as [9] and [10]. We did not use this data source because it only collects a few simple statistics for each wiki, precluding in-depth analysis of distributions of work. Other research involved case studies of individual wikis [1] or analysis of multiple Wikipedia projects [8], but we were not able to find any prior research that analyzed (in depth) multiple wikis with diverse administration, motivating our current work.

3. DATA COLLECTION FROM WIKIS

We developed a Java-based tool which we will call the "WikiCrawler" to collect the data for our analysis. Through the use of webcrawling and screen-scraping techniques, the WikiCrawler converts the HTML served by wikis into data suitable for analysis. Despite the fact that Mediawiki permits extensive user interface customization, most Mediawiki sites use consistent HTML element IDs on their UI elements. For example, the two wikis depicted in Figure 1 look different, but similarities in the underlying HTML allow information to be extracted from both. There are a few variations in the HTML generated by different versions of MediaWiki, but most of the processing rules are identical across versions.

Unlike general-purpose webcrawlers, the WikiCrawler selectively downloads the subset of pages required for the researcher to compute the desired statistics. The researcher accomplishes this by programming rules that indicate which quantities the WikiCrawler should measure, allowing the WikiCrawler to determine which pages to download and parse. The desired quantities are then stored in an SQL database for later analysis.

Lists of pages to download are extracted from the wiki's "all pages" index, and revision histories are obtained by following the appropriate links. User histories are obtained in a similar manner. Only the current revisions of the page text were downloaded, but the complete revision histories were retrieved. To work around server-side URL rewriting rules, the WikiCrawler often harvests links from previously parsed pages instead of generating them, to avoid training the crawler to generate every type of URL needed under a specific rewriting rule.

3.1 Finding wikis

We obtained a list of candidate wikis to analyze by using the Yahoo and Microsoft Live search web services to retrieve the first 1000 results for the string `Main_Page`. Because Mediawiki creates a page called `Main_Page` by default, this allows us to easily find Mediawiki instances. We did not use wiki directory sites such as WikiIndex[3] or the previously mentioned S23 because it was unclear if the wikis in the directories were collected by humans, which could result in a selection bias in our experimental population. Using search engines to form our sample ensures that the only bias is that the wikis were popular enough to appear in a search result.

Filtering non-wikis and duplicates between the two search engines out of our 2000 seed URLs, we found 1445 wikis which could potentially be analyzed. Wikimedia projects were excluded because they can be downloaded and analyzed more easily by using the database dumps mentioned earlier.

The Robot Exclusion Standard[4] allows websites to indicate that robots and crawlers should not visit. (This standard is not enforced by technical means; therefore, compliance is voluntary.) We rejected 77 of these 1445 wikis where our crawling would have violated this protocol.[5]

Because MediaWiki is an open-source product, users often customize it to improve the appearance of the wiki or add new features. These customizations sometimes confound the screen-scraping component of our crawler, which checks for inconsistencies in the downloaded pages to compensate for this. 182 sites were excluded due to such inconsistencies, or because a password was required to access one or more wiki pages, preventing accurate statistics from being compiled. In addition, we manually removed 3 sites because they contained illegal or pornographic content, or because they were duplicates (which happens when multiple virtual hosts are backed by the same wiki.) In the end, 1183 wikis were available for analysis.

3.2 The study population

We estimated the sizes of the 1183 available wikis by ex-

[2] `http://s23.org/wikistats/`

[3] `http://www.wikiindex.org/`

[4] `http://www.robotstxt.org/orig.html`

[5] This is an apparent contradiction, because our list of wikis originally came from search engines. This would mean that major search engines are violating this protocol, or that the wikis can be accessed from an alternate URL that falls outside of the restrictions.

amining the article indices of the main namespace. (The actual number of articles in the wiki is usually lower than this estimate, due to redirection pages that contain no content themselves, among other anomalies.)

The size estimates are depicted in Figure 2. Note that the sizes of wikis with more than 300 pages are distributed with a discrete power law distribution ($p = .284$, $\alpha = 1.77$, $x_{min} = 300$, see Appendix A). This is consistent with the finding by Roth [9] that the sizes of wikis tracked on S23 have a power law distribution.

The number of wikis with less than 300 pages is smaller than what the power law distribution would expect, possibly because the search engines we used were more likely to return larger wikis than smaller ones.

Because it was impractical to download all 1183 wikis in their entirety (totaling over 6.2 million articles) for analysis, we analyzed a random sample of the wikis in our study population.[6] We excluded the smallest wikis (any wiki with fewer than 50 pages), to avoid potentially misleading results from the analysis of very small datasets. We also excluded the wikis with more than 32000 pages, because manually inspecting them revealed that they were large because they were generated by robots which convert large amounts of database data to a wiki format, instead of being organically created by a user base.

We randomly selected 181 of the remaining wikis for our analysis. The processing of 30 wikis failed because persistent PHP or database errors were experienced while crawling, leaving 151 wikis for analysis.

3.3 The analyzed wikis

It is useful to visualize the number of authors, users, and words in the wikis studied to observe the relationships between these basic measures. To do this, we used the WikiCrawler to count the articles, registered users, and words in each wiki chosen for analysis. Figure 3 depicts the number of articles and users in each analyzed wiki, along with the average article sizes. (This final article count is not subject to the inaccuracies in the estimate mentioned previously.) Performing a Spearman's rank correlation test on the two depicted relationships shows a moderate correlation between user and article counts. ($p = 6.93 \cdot 10^{-10}$, $\rho = .474$), while the same test showed the number of users and the average article length to be slightly correlated ($p = .028$, $\rho = .178$). These results indicate that having a large user base tends to increase the number of articles in a wiki, with a less clear effect on the lengths of existing articles. Studying if new users tend to create new articles over adding content to existing ones is a topic for further research.

4. ANALYSIS OF THE WIKIS

When studying a large population of wikis, it is important to determine if the wikis are being used as the Mediawiki authors envisioned (for collaborative content development) or if many of them are being used as simple content-management systems that discourage editing by ordinary users. This was previously seen when excluding wikis with more than 32000 pages from the study sample, and we propose a metric to determine if other wikis are being used in

[6]Despite the small number of wikis analyzed, we still downloaded more than 820,000 HTML pages and the analysis generated a 33 GB database.

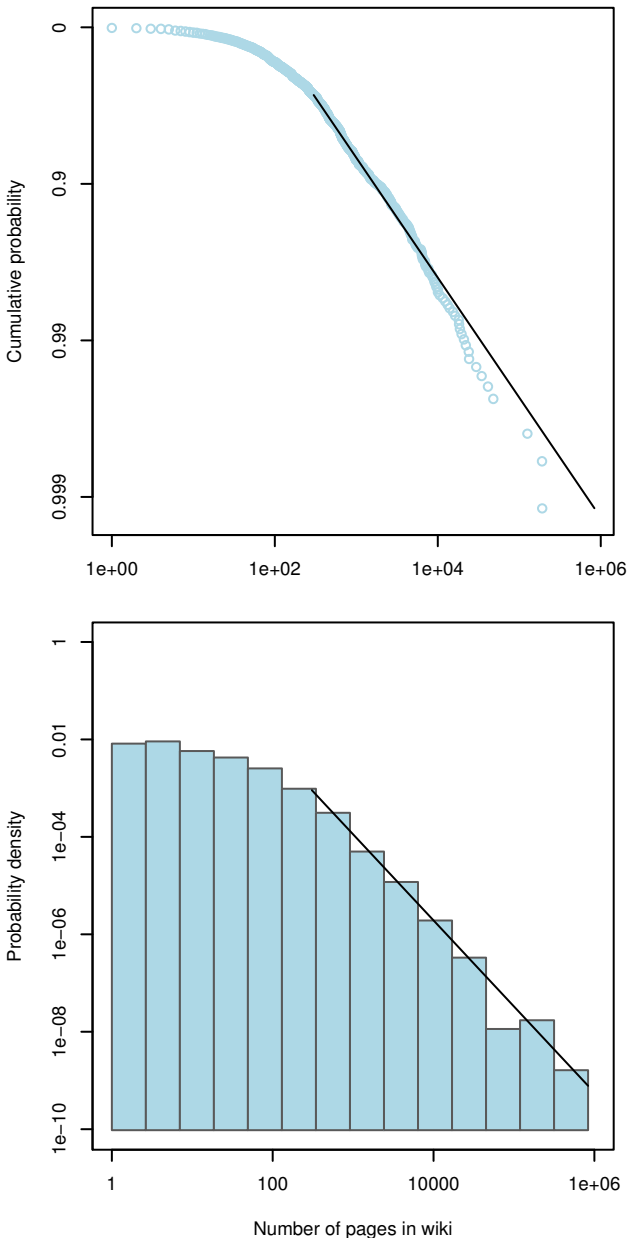

Figure 2: Sizes of wikis in our study population, depicted as a cumulative distribution function (CDF) and a histogram. The goodness of the fit can easily be estimated by comparing the empirical CDF (plotted as dots) with the CDF of the fitted power law function, depicted as a straight line. (The Kolmogorov-Smirnov statistic mentioned in Appendix A can be visualized as the maximum vertical distance between the dots and line.) A histogram is also provided, as it is somewhat more intuitive for visualizing the distribution of the population.

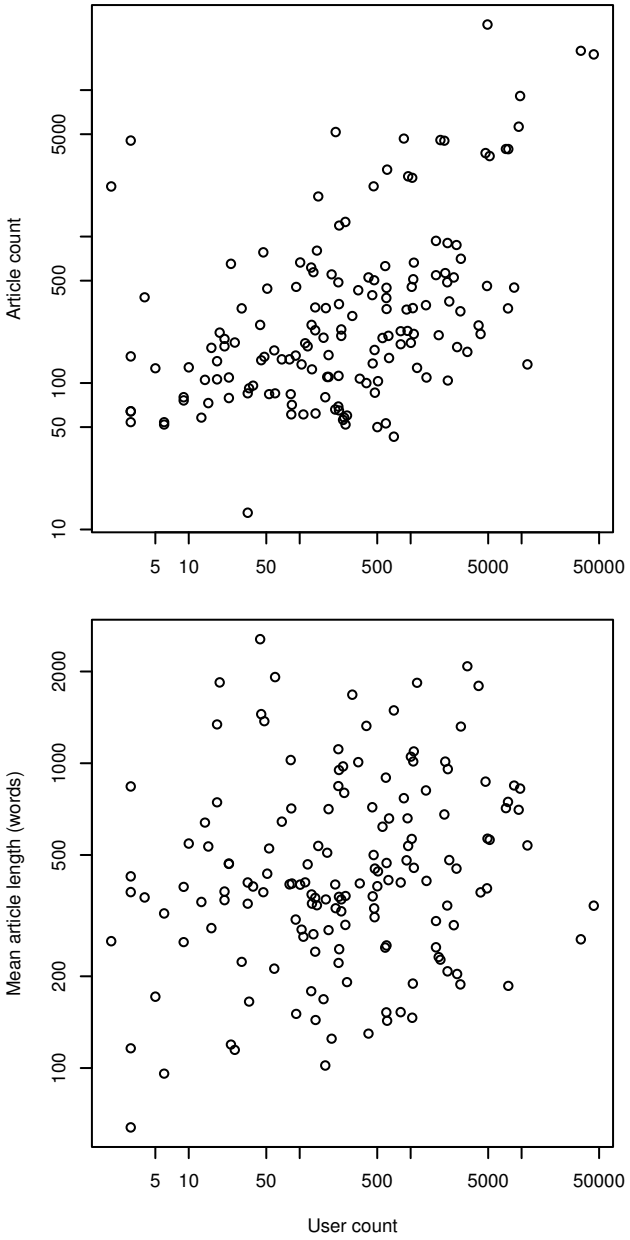

Figure 3: Numbers of articles and average article length of wikis compared with the number of users. Each dot represents one wiki.

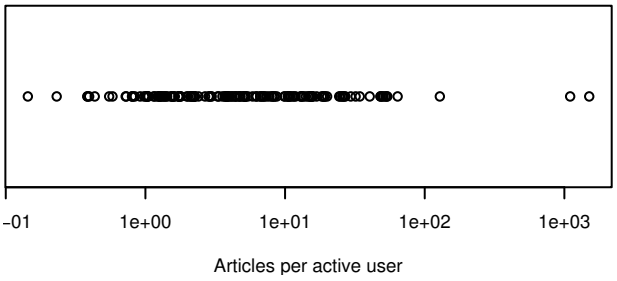

Figure 4: Number of articles per active user. Each dot represents one wiki.

the same way. For each wiki, we divided the number of articles by the number of active users and plotted the results in Figure 4. (Active users are users that edited the wiki at least once. We exclude inactive user accounts because they could have been automatically generated.)

This measure was distributed along a continuum of 0.4-54.0 articles per active user, with four outliers at 64, 128.3, 1100.0, and 1507.0 articles per active user. Inspecting the outliers shows that the third and fourth outlier wikis were automatically generated from a database while the first and second outlier wikis are knowledge bases that users cannot directly edit. Spot checks of other wikis (including the data points closest to the outliers) do not reveal any similar cases (One wiki with 52.3 articles per user was partially generated but had later attracted an active user base.) Therefore, we conclude that measuring the number of articles per active user is an effective way of detecting artificially generated wikis, and that the wikis in our sample are largely being authored collaboratively.

4.1 Concentration of work across wiki users

Past researchers have frequently studied the concentration of work in Wikipedia (the degree that a small number of users is responsible for a large proportion of the content.) In one such study, Kittur et al. [4] examined the claim that Wikipedia reflects "the wisdom of crowds", and concluded that the most active, "elite" users are continually declining in influence. We revisit this question, applied to a larger sample of wikis than Wikipedia alone.

4.1.1 Gini coefficients

Originally developed to measure economic inequality, wiki researchers have used the Gini coefficient (see Appendix B) to determine the degree that few users make a large proportion of the contributions to Wikipedia. Ortega et al. [7], noted that 90% of the revisions were made by 10% of the users of the English Wikipedia, resulting in a high Gini coefficient of .9360. Ortega indicates that this high degree of inequality persists on a monthly basis [8], casting doubt on the theory that wiki content represents "the wisdom of crowds"

To exclude wikis that were too small for a meaningful inequality measurement, we measured inequality in the subset of sampled wikis that had more than 50 active users and 300 pages (totaling 50 wikis). In Table 1, we present the inequality in the contribution levels of all users as measured by the Gini coefficients. A large amount of inequality was found, with most wikis having a Gini coefficient greater than

$(.57, 0.88]$	3
$(0.88, 0.92]$	6
$(0.92, 0.96]$	11
$(0.96, 0.98]$	15
$(0.98, 1]$	15

Table 1: Gini coefficients of sampled wikis, measured across all users (0.0 would indicate that all users contributed equally while 1.0 would represent that a single user contributed everything.)

$(.36, 0.80]$	7
$(0.80, 0.84]$	6
$(0.84, 0.88]$	13
$(0.88, 0.92]$	16
$(0.92, 0.98]$	8

Table 2: Gini coefficients of sampled wikis, measured across active users

.96, which is slightly higher than the Gini coefficients found for Wikipedia editions in [8] (although comparing Gini coefficients of very large and very small wikis can be misleading, as noted in [3]).

In Table 2, we calculate the Gini coefficients of wikis if inactive users (who never edited the wiki) are excluded. (The presence of these users may skew the results because wikis can share a user database with other information systems, resulting in the appearance of many inactive users.) When excluding users who never edited, the inequality is smaller but still significant with a median above .87.

4.1.2 *Probability distributions of activity across editors*

Past research [2] has discovered that several quantities in Wikipedia, such as the number of edits to individual articles and the number of unique contributers to articles, have power-law distributions. Such distributions have also been found in a number of phenomena on the world wide web [5]. Using the fitting algorithm described in Appendix A on the contribution levels of active users, we found that the user contribution distributions in 40 out of the 50 wikis were consistent with power law distributions (at $p \geq .10$ as described in Appendix A).

Because there are some underlying similarities between power law and lognormal distributions [6], and they have been confused for one another in previous research [13], we used the same fitting algorithm to fit the data against a lognormal distribution. Only 7 of the 50 wikis had user contribution distributions that were consistent with a lognormal distribution, suggesting that the power law distribution better matches the underlying model.

Figure 5 presents several examples of wikis that do and do not have clear power law distributions for user contributions. Note that in the lower-right wiki, the number of users making more than 24 edits is well-predicted by the power law fit, while the number of users making less would be overestimated by the fitted distribution. These partial fits (formalized in Appendix A) have also been seen in [12], [2], and other Wikipedia studies that observed power law distributions in various phenomena.

To explore these cases where the empirical distribution

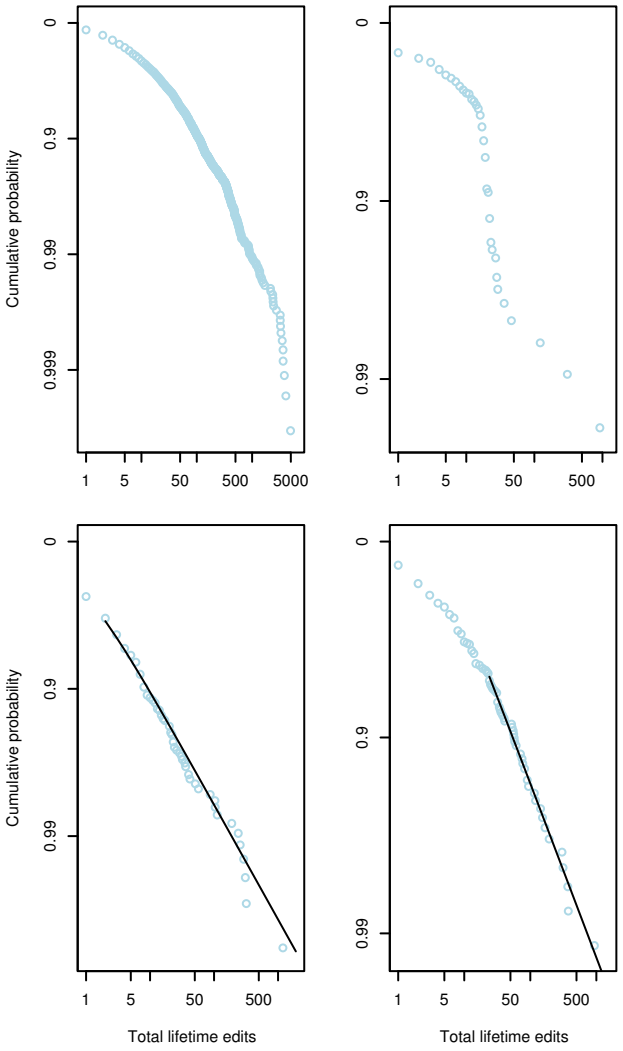

Figure 5: Sample cumulative distributions of user contribution counts. The top two graphs are examples of wikis where user contributions are not consistent with power law distributions, while the bottom two graphs are examples of wikis where they are. Note that the wiki in the lower-right quadrant had a partial fit for the cumulative distribution function. The graphs are to be interpreted as in Figure 2

deviates from power law behavior as the number of contributions shrinks, we compared the number of users represented by the unfitted portion of the empirical distributions with the number of users that would have been expected in that portion if the power law behavior had held. 34 wikis had less probability mass than expected in this portion of the empirical distribution, while 6 wikis had more.

This indicates that most studied wikis had fewer "occasional editors" than would be expected from the fitted distribution. This phenomenon could be a simple consequence of the power law being a poor fit for the underlying distribution; however, we believe that several factors which discourage occasional contributions depress the count of users who made few edits. One such factor is that many wikis allow users to edit wiki pages without a username, which provides a convenience for the occasional contributer, but may be less attractive to frequent contributers who want their work to be recognized and attributed. It is difficult to associate anonymous contributions with an individual, and for this reason, past studies of inequality in Wikipedia have declined to measure the effects of anonymous contributers.

4.1.3 Gini coefficients and probability distributions

Finally, to better understand the underlying reasons for the high Gini coefficients seen in wikis, we explored the relationship between the parameters of the power law distribution and the Gini coefficient of a wiki's concentration of work. For each wiki, we calculated the Gini coefficient of a synthetic power law dataset with the same distribution parameters as the fitted distribution for that wiki (see Appendix A). We found that the Gini coefficients of the synthetic data were very close to those of the actual data, with a median absolute difference of only .017. As seen in Figure 6, a few of the Gini coefficients rose or dropped sharply when switching to the synthetic dataset, but most changed only slightly, and all stayed within a general neighborhood (.13) of the original value.

Ortega [8] noted that the Gini coefficient of work performed within successive time intervals slowly rises and eventually stabilizes. This was cited as evidence that Wikipedia users are not contributing more equally over time, as claimed in [4]. We have demonstrated that the Gini coefficient is mostly determined by the number of users being measured and the parameters of the power law distribution that describes their concentration of work. We are currently investigating if the periodic concentration of work in Wikipedia has a power law distribution, and if the evolution of the Gini coefficient can be explained by a shift in the parameters of the fitted distribution, or a shift in the quantity of contributions each month.

Our discovery that the concentration of work in most wikis is consistent with a power law distribution may provide motivation for a *generative model* of user activity that produces the observed distributions. Because the same family of distribution was found in a majority of the studied wikis, such a model could be generalizable, providing researchers with another tool to characterize the relative importance of "the power of the few" and "the wisdom of crowds" in Wikipedia and other wikis. Similar models include the model developed by Wilkinson and Huberman [13] that explains the concentration of work across Wikipedia articles, and the models that were proposed to explain power law distributions in numerous phenomena on the World Wide Web [5].

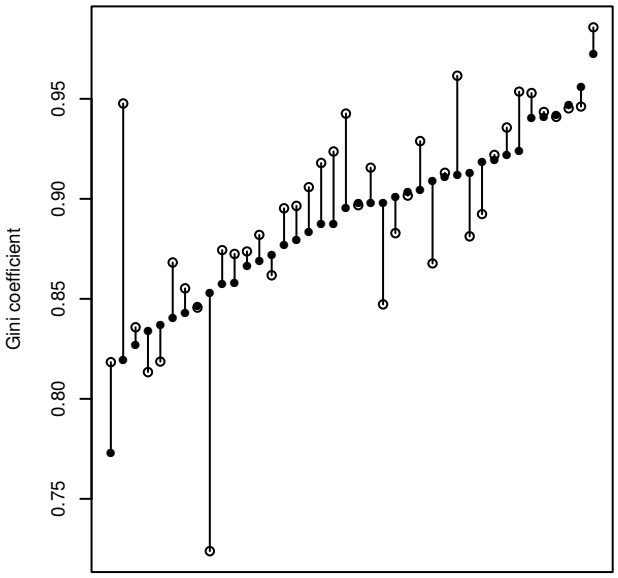

Figure 6: Changes in Gini coefficient. Closed circles represent the actual Gini coefficient of a wiki's user contribution levels, while open circles represent the Gini coefficient that would have been predicted from the parameters of the fitted power law distribution.

For example, one common model that explains how power law distributions develop in graphs is the *preferential attachment* model, which specifies that new nodes tend to attach to nodes with many edges [6]. Regardless of the generative model that is ultimately chosen, the WikiCrawler provides a unique opportunity to verify models, because the timestamps are captured for each wiki contribution. The sequence of timestamps could be checked against what was predicted by the model, allowing for evaluation of the degree that the two are consistent or inconsistent.

4.2 Distributions of wiki article edits

Having fitted power law distributions to the contribution levels of individual users, we then attempted to fit power law distributions to the number of revisions made to each article in the wiki. The number of revisions is a rough measure of the amount of work put into an article, and examining its distribution will help us determine how efforts are spread across the articles in wikis. Measuring wikis with 300 or more articles, 38 out of 66 eligible wikis had distributions of revisions across articles that were consistent with a power law distribution. This is more than would be expected by chance, but the percentage of successful power law fits (58%) was lower than it was when fitting power law distributions to user contribution levels (80%).

This is consistent with the finding [13] that the distribution of edits across Wikipedia articles is log-normal within time slices, tending toward a power law distribution over a long period if certain conditions are met. (Also, in Wikipedia, [2] claimed a power law distribution of edits per article

but did not statistically test the fit.)

5. CONCLUSION

In this research, we introduced a methodology for studying wikis other than Wikipedia and extended some past Wikipedia research to the larger wiki population. We proposed a metric to determine if wikis are the product of collaborative editing, and demonstrated that nearly all wikis in our study sample had this characteristic. We then applied an algorithm for fitting power law distributions to wikis, and concluded that the concentration of work across users in most of the studied wikis was consistent with the power law. This allowed us to demonstrate that the level of inequality in wiki contributions (as measured by the Gini coefficient) was largely determined by the parameters of their underlying power law distributions.

These characteristics (which we found in the majority of the studied wikis) suggest that a mathematical model describing user activity could be developed, opening an avenue for future research. Such a model could help wiki practitioners understand the relative importance of occasional and frequent contributers to a wiki, along with the relative importance of registered and unregistered users. This research also motivates additional study into the inequality of user activity levels in Wikipedia, by relating the shifting levels of inequality to the statistical distribution of work across users. Finally, this work demonstrates that performing large studies with many wikis is practical, which will be useful for applications such as the calibration of wiki metrics and the generalization of research results to a larger population of wikis than Wikipedia alone.

6. REFERENCES

[1] S. Blaschke and K. Stein. Methods and measures for the analysis of corporate wikis: A case study. In *International Communication Association conference*, Washington, DC, USA, 2008. International Communication Association.

[2] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the wikigraph. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 45–51. IEEE Computer Society, 2006.

[3] G. Deltas. The small-sample bias of the gini coefficient: Results and implications for empirical research. *Review of Economics and Statistics*, 85(1):226–234, 2003.

[4] A. Kittur, E. H. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *CHI '07: Proceedings of the Computer/Human Interaction Conference*. ACM, 2007.

[5] A. Medina, I. Matta, and J. Byers. On the origin of power laws in internet topologies. *SIGCOMM Comput. Commun. Rev.*, 30(2):18–28, 2000.

[6] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1:226–251, 2004.

[7] F. Ortega and J. M. Gonzalez-Barahona. Quantitative analysis of the wikipedia community of users. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 75–86. ACM, 2007.

[8] F. Ortega, J. M. Gonzalez-Barahona, and G. Robles. On the inequality of contributions to wikipedia. In *HICSS '08: Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, page 304. IEEE Computer Society, 2008.

[9] C. Roth. Viable wikis: struggle for life in the wikisphere. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 119–124. ACM, 2007.

[10] C. Roth, D. Taraborelli, and N. Gilbert. Measuring wiki viability. In *WikiSym '08: Proceedings of the 2008 international symposium on Wikis*. ACM, 2008.

[11] C. R. Shalizi and M. Newman. Power-law distributions in empirical data. *SIAM Review, to appear*, 2009. arXiv:0706.1062.

[12] J. Voss. Measuring wikipedia. In *International Conference of the International Society for Scientometrics and Informetrics*, Katholieke Universiteit Leuven, Leuven, Belgium, 2005. International Society for Scientometrics and Informetrics.

[13] D. M. Wilkinson and B. A. Huberman. Assessing the value of cooperation in wikipedia. *First Monday*, 12(4), 4 2007.

APPENDIX

A. POWER LAW DISTRIBUTIONS

In its most basic form, a *power law distribution* is a distribution of the form $x^{-\alpha}$ where α is the parameter of the distribution. Such distributions are important because many phenomena related to information systems (such as the Web) are distributed according to a power law distribution [5].

Because the number of edits associated with an article or user is a discrete quantity, we used the discrete power law distribution described by [11] of the form $\frac{x^{-\alpha}}{\zeta(\alpha, x_{min})}$, with ζ being the generalized Zeta function. In practice, power law distributions rarely hold for the entire range of observed values, instead only holding for values greater or equal to a certain x_{min}. For this reason, a parameter x_{min} is estimated from the data, and lesser values are not described by the distribution.

Strictly speaking, we cannot prove that a finite dataset has a power-law distribution (instead of a different distribution that could have generated the same values), so instead we argue that the observed data is *consistent* with a power-law distribution. To do this, we first estimate α and x_{min}, using the maximum likelihood estimation procedure described in [11]. To avoid the case where a very large x_{min} is chosen, leading to a power law distribution that describes just a few data points (which can be trivially true, but uninteresting), we restrict the possible values of x_{min} so the power law distribution describes at least half of the unique values (or "counts") that were observed in the empirical distribution. (In practice, this means that the distribution describes many more than half of the observed counts because the empirical distribution becomes sparse for large values of x.)

It is important to use inferential statistics when claiming that a certain empirical distribution is consistent with a power law distribution, because the traditional method for finding power law distributions (identifying a straight line on a log-log plot) is notoriously unreliable and other common

distributions appear to be power law upon such inspection, as noted in [11]. Using the procedure described in [11] to test the fit, we compute the Kolmogorov-Smirnov goodness-of-fit statistic of the fitted distribution, and use a Monte Carlo procedure to compute the probability (p) that the empirically observed data has a better Kolmogorov-Smirnov statistic than a random dataset drawn from the fitted distribution (corrected for the unfitted values below x_{min}). The power law fit is then accepted if $p > .1$, per the guidelines in [11]. (We do not correct for the experimentwise error rate stemming from testing many distribution fits, because not doing so actually results in a more conservative experiment, due to Type I errors yielding an incorrectly rejected power law fit.)

When generating data with a power law distribution in Section 4.1.3, we generate the same number of data points found in the original distribution, with $x_{min} = 1$ and the same α as in the original distribution. We also constrain the generated values so none of them are greater than the maximum value of the original data, to prevent the generation of extremely large user contribution totals that exceed the amount of work found in the wiki.

B. GINI COEFFICIENTS

The Gini coefficient is a measure of inequality that has been broadly used in economics and information science. It is a unitless quantity that measures how fairly a variable (in our case, wiki edits) is distributed across a population (in our case, users or articles), with 0 signifying perfect equality and 1 signifying the largest possible inequality (which would mean that all edits are concentrated in a single user or article) [3].

To compute the Gini coefficient on a discrete dataset, we use the formula presented in [3], which is a revision of the traditional formula that reduces the bias for small datasets:

$$\frac{n \sum_{i,j} \|y_i - y_j\|}{2\,(n-1)\,n^2 \bar{y}}$$

where y is the sequence of values and n is its length. (Wikipedia research such as [8] did not include the $\frac{n}{n-1}$ correction factor when calculating the Gini coefficient, but its effect is infinitesimal in large datasets such as Wikipedia.)