Measuring Author Contributions to the Wikipedia*

B. Thomas Adler[1] Luca de Alfaro[2] Ian Pye[1] Vishwanath Raman[1]

[1]Computer Science Dept.
UC Santa Cruz, CA, USA
{thumper, vishwa, ipye}@soe.ucsc.edu

[2]Computer Engineering Dept.
UC Santa Cruz, CA, USA
luca@soe.ucsc.edu

ABSTRACT

We consider the problem of measuring user contributions to versioned, collaborative bodies of information, such as wikis. Measuring the contributions of individual authors can be used to divide revenue, to recognize merit, to award status promotions, and to choose the order of authors when citing the content. In the context of the Wikipedia, previous works on author contribution estimation have focused on two criteria: the total text created, and the total number of edits performed. We show that neither of these criteria work well: both techniques are vulnerable to manipulation, and the total-text criterion fails to reward people who polish or re-arrange the content.

We consider and compare various alternative criteria that take into account the *quality* of a contribution, in addition to the quantity, and we analyze how the criteria differ in the way they rank authors according to their contributions. As an outcome of this study, we propose to adopt *total edit longevity* as a measure of author contribution. Edit longevity is resistant to simple attacks, since edits are counted towards an author's contribution only if other authors accept the contribution. Edit longevity equally rewards people who create content, and people who rearrange or polish the content. Finally, edit longevity distinguishes the people who contribute little (who have contribution close to zero) from spammers or vandals, whose contribution quickly grows negative.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Computer-supported cooperative work, Web-based interaction*; K.4.3 [**Computers and Society**]: Organizational Impacts—*Computer-supported collaborative work*; J.4 [**Social and Behavioral Sciences**]: Miscellaneous

*This research has been partially supported by the CITRIS: Center for Information Technology Research in the Interest of Society.

WikiSym '08, September 8–10, Porto, Portugal.

1. INTRODUCTION

On-line collaboration is fast becoming one of the primary ways in which content, and information, is created and shared. From open-source software projects, to on-line encyclopedias such as the Wikipedia, open on-line collaboration taps into the knowledge, time, and resources of millions of people, and it enables content creation on a speed and scale never seen before.

In such collaborative systems for content creation, a relevant problem is how to measure the contributions of individual authors. This problem arises in several contexts. In the Wikipedia, it may be desirable to have a measure of how much work various users have performed, as a help to decide how to distribute promotions and honors (such as Barnstars). In a corporate setting, a measure of user contribution can be used to promote the use of collaboration tools such as wikis, while ensuring that the contributions of the individual employees can be duly recognized. In wikis that generate revenue, measures of author contribution can be used as a basis for deciding revenue sharing among authors. In this paper, we propose and analyze several quantitative measures of author contribution for versioned bodies of textual information that evolve via successive modifications, or *edits*, perfomed by individual authors. We focus on wikis, and in particular on the Wikipedia, an on-line encyclopedia built on wiki technology [5].

On the Wikipedia, the problem of measuring author contributions has so far been considered mainly in the context of gaining a better understanding of the dynamics of how authors contribute to the Wikipedia. In particular, measures of author contribution have been used to discuss the issue of whether it is a large group of novice users, or a small group of experienced editors, who contributes most of the content of the Wikipedia [23, 19, 8]. In these discussions, author contribution was measured via the number of edits performed by authors (edit count) [23, 24, 8, 18, 12, 17], or by the total amount of text the authors introduced (text count) [19]. We argue that neither of these two measures is robust, or fully informative. Both edit count and text count can easily be gamed. In the case of edit count, an author can increase her edit count simply by doing a small modification, then undoing it (perhaps with an accompanying message of apology). These small changes can be spread across the millions of pages of the English Wikipedia, to make detection harder. Fighting this kind of abuse requires time-consuming human intervention, as it requires the examination of the edit log performed by individual users. In addition, adopting edit count as a measure of contribution, and basing important decisions on such a measure, would create a strong incentive towards this kind of tampering, with negative consequences for the stability and quality of Wikipedia content. Text count can be gamed in a similar, if not easier, fashion, as it is possible to introduce in a single edit a very large amount of text, which is then promptly removed in a subsequent edit. These measures

also under-estimate the contributions that particular groups of authors give to the Wikipedia. Edit count fails to recognize adequately, authors who sporadically provide large amounts of bulk text, thus contributing to the basic information build-up. Text count fails to properly recognize the contributions of those users who mostly rearrange existing content, remove vandalism, and perform other fundamental maintenance duties that involve little new text creation.

To provide a more precise, and more robust, measure of author contribution, we introduce several measures that capture not only the *quantity* of contributions (how many edits, how much text), but also the *quality*, analyzing how long the contributions last as content evolves. We analyze the trade-offs involved in the different measures, and the differences in the author rankings they produce. We emphasize that we make a distinction between contribution and reputation; the first is meant to measure productivity, and the second, reliability. A reputation system generally incorporates some measure of productivity, as we do in [2]. This work explores different notions of productivity, particularly as a tool for analyzing user behavior, but does not evaluate the effectiveness of any measure as a reputation system.

Of the measures we explore, our favorite measure is *edit longevity*. The measure combines the amount of change performed by an author, with how long the change lasts. To compute the edit longevity of an author, we consider all the edits performed by the author. For each edit, we compute both the *edit size* and the *edit quality*. The edit size is measured as the edit distance between the article revision produced by the author, and the previous revision; the problem of computing edit distances has been studied in [22, 20, 4]. The edit quality is in the interval $[-1, 1]$, and measures how long the change lasts in the system: it is close to 1 for edits that are preserved fully in subsequent revisions, and it is close to -1 for edits that are reverted [2]. The edit longevity of an author combines these quantities, and is computed as the sum, over all the edits performed by the author, of the edit size multiplied by the edit quality. We show that edit longevity has several desirable properties. It cannot be easily gamed, since the quality of an edit by author A depends on how long authors *different from* A preserve the edit done by A. Edit longevity is sensitive to the size of contributions, giving appropriate reward to authors who sporadically contribute large amounts of text. Edit longevity also rewards the authors who mostly engage in improvement and maintenance work, as edit distance, measures not only new text, but also text deletions and displacements [2]. Furthermore, edit longevity successfully prevents vandals and spammers from accruing contributions.

In addition to edit longevity, some other measures we introduce may serve a purpose in specific contexts.

Text longevity is a measure of how much text has been introduced by an author, in which each text contribution is weighed according to how long the text lasts in subsequent revisions. This measure has the advantage of a very obvious definition: while the estimation of edit quality requires agreement on a particular formula based on edit distances, the estimation of text longevity requires simply, the ability of tracking text through revisions. In contexts where revenue must be divided, a simpler definition has the appeal of being easier to define; for instance, in a legally binding contract. Text longevity, however, has two drawbacks. The first is that text longevity fails to adequately reward authors who mainly engage in maintenance edits, inserting little new text. The second is that text longevity fails to penalize spammers and vandals, assigning to them the same low amount of positive contribution reserved for novices.

Text longevity with penalty rewards authors in proportion to the text they insert, but in addition to this reward, authors of reverted contributions (that is, edits with negative quality) also accrue large negative "fines". Text longevity with penalty is particularly effective in distinguishing productive members of the author community from vandals and spammers. The drawback of text longevity with penalty, compared with edit longevity, is that it fails to adequately reward authors who perform large amounts of maintenance work.

Related Work

Measuring contributions of individuals in group collaborative efforts have been studied for several decades, in the context of software project management. Most programmers are familiar with the KLOC (thousand lines of code) measurement [15, 6] — it counts how many lines of software a programmer writes per week, and is analogous to text count. Although intended to measure the progress of a project, it also implicitly measures each programmers contribution to the project.

The problem of measuring author contributions to the Wikipedia first arose during a debate on which group of authors was chiefly responsible for Wikipedia content. A count of edits indicated that a small group of dedicated authors gave the largest contribution [23]; this result was later disputed in [19], where it was argued that the majority of the content was due to the large number of authors who perform only occasional contributions to the Wikipedia. Kittur et al., [8] discovered that the percentage of edits made by the masses is larger and growing when compared to the authors who are either sysops or have a very large number of edits to their credit. Burke and Kraut use a wide variety of edit counts to predict the authors to be promoted to the status of Wikipedia administrators [3]. There are many works that measure user contributions by the number of edits which authors make to an article [23, 24, 8, 18, 12, 17]. In [8], they consider the number of edits as well as the change in edits and state that their conclusion remains unaffected with either of those measures. Wilkinson and Huberman show that the quality of an article improves with the number of edits and the number of distinct authors that revise that article [24].

Another approach consists in measuring contributions indirectly, by noting citations [13, 7, 11]. While this approach enables one to judge the relevance of a complete article, it cannot easily be used to ascribe the merit of the article to its individual contributors. In [9], the authors examine the sequence of edits to an article and build a social network of authors based on the amount of citation (how much text is removed or preserved) by later authors; this social network is then analyzed to derive authority values for authors.

In [16], the total amount of work that went into the making of the complete Wikipedia is estimated to be about 100 million hours of human thought. This metric is very different from the metrics we propose in this paper, as it is based on the effect on the author (the amount of time required of the author to contribute), rather than on the effect on the wiki (how large the contributions are, and how long they last).

2. DEFINITIONS

The following notation will be used throughout the paper. We consider the set $\mathbb{P}$ of all articles in the main English Wikipedia. We denote the set of authors of pages on the Wikipedia by $\mathbb{A}$. We assume that we have $n > 0$ versions $v_0, v_1, v_2, \ldots, v_n$ of a page p; version v_0 is empty, and version v_i, for $1 \leq i \leq n$, is obtained by an author performing a revision $r_i : v_{i-1} \rightsquigarrow v_i$. Since each revision r_i is performed by one author, we refer to the author who edited revision r_i as a_i. We denote the set of all revisions of a page by $\mathbb{R}$. We refer to the change set corresponding to $r_i : v_{i-1} \rightsquigarrow v_i$ as the edit performed at r_i : the edit consists of the text insertions, deletions, displacements, and replacements that led from v_{i-1} to v_i. We define the map $E : \mathbb{A} \times \mathbb{P} \rightarrow 2^{\mathbb{R}}$, which given an author $a \in \mathbb{A}$ and a

page $p \in \mathbb{P}$, returns a set of revisions that were created by author a for page p. When editing a versioned document, authors commonly save several versions in a short time frame. We filter the versions, keeping only the last of consecutive versions by the same author; we assume thus that for $1 \leq i \leq n$ we have $a_{i-1} \neq a_i$. Every version v_i for $0 \leq i \leq n$, consists of a sequence $[w_1^i, \ldots, w_{m_i}^i]$ of words, where m_i is the number of words in v_i; version v_0 consists of the empty sequence.

Quantity Measures.

Given a series of versions $v_0, \ldots, v_n$ of a page p, we assume that we can compute the following quantity measures:

- $txt(v_i, v_j)$, for $0 < i \leq j \leq n$, is the amount of text (measured in number of words) that is introduced by r_i in v_i, and that is still present (and due to the same author a_i) in v_j. $txt(v_i, v_i)$ is the amount of new text added by a_i through r_i. We define $txt(r_i) = txt(v_i, v_i)$, and refer to this as the *text contribution* of r_i.
- $d(v_i, v_j)$, for $0 \leq i < j \leq n$, is the *edit distance* between v_i and v_j, and measures how much change (word additions, deletions, replacements, displacements, etc.) there has been in going from v_i to v_j. We define $d(r_i) = d(v_{i-1}, v_i)$, for the *edit contribution* made in a revision r_i.

There are several ways to compute edit distance [10, 20], usually based on insertions and deletions of characters. Our formulation is instead based on words as the fundamental unit, to more closely approximate how people perceive edits. We define the edit distance in terms of the following quantities: $I(v_i, v_j)$ is the number of words that are inserted; $D(v_i, v_j)$ is the number of words that are deleted; $M(v_i, v_j)$ is the number of words that are moved, times the fraction of the document that they move across [2]. The edit distance between two versions, v_i and v_j, is then given by:

$$d(v_i, v_j) = \max(I, D) - \frac{1}{2}\min(I, D) + M$$

A more precise treatment of this definition is available in [2], along with reasoning for this particular choice of edit distance and a discussion of text tracking for authorship.

Quality Measures.

In addition to the quantity measures defined above, we define the following quality measures. We first consider the edits in a revision r_i made by author a_i. Given $d(v_{i-1}, v_i)$, the edit distance between versions v_{i-1} and v_i, we would like to associate a higher edit quality to revision r_i if the edits made take the page closer to subsequent versions of the page. For example, if none of the edits made in r_i are reverted in subsequent revisions, then the edits made in r_i have taken the page in the same *direction* as subsequent versions of the page and hence merit the highest quality measure. We define $\alpha_{edit}(v_i, v_j)$, the quality of the edits performed in revision r_i of a page (with respect to v_j) as follows:

$$\alpha_{edit}(v_i, v_j) = \frac{d(v_{i-1}, v_j) - d(v_i, v_j)}{d(v_{i-1}, v_i)}$$

The triangle inequality generally holds, so that $\alpha_{edit}(v_i, v_j)$ typically varies from -1 for revisions which are completely reverted, to $+1$ for revisions which are completely preserved; when the value falls outside this range, we cap it to one of these two values.

Due to the occasional vandalism that happens, we prefer to judge quality using several succeeding versions. We define a map $J : \mathbb{R} \rightarrow 2^{\mathbb{R}}$, such that if r_i is a revision of a page p, $J(r_i)$ consists of the first ten revisions after r_i that have author different from that of r_i. If there are fewer than ten revisions after r_i that have author different from that of r_i, then $J(r_i)$ returns all such revisions. We use the versions in $J(r_i)$ as judges of the quality of r_i. We define the *average edit quality* $\overline{\alpha}_{edit}(r_i)$ of a revision r_i with $|J(r_i)| \neq \emptyset$ as follows:

$$\overline{\alpha}_{edit}(r_i) = \frac{1}{|J(r_i)|} \cdot \left(\sum_{r_j \in J(r_i)} \alpha_{edit}(v_i, v_j) \right)$$

Thus, $\overline{\alpha}_{edit}(r_i)$ is the average of the α_{edit}values determined by each of the judges of r_i.

The quality of a text contribution to a page is a function of how much of the original text was edited out in subsequent revisions of the page. If none of the text was removed, then the text quality of revision r_i should be 1. One model for how text behaves over time is that it decays in an approximately geometric fasion: the largest chunk that is deleted is right after text is added, and subsequent revisions remove smaller and smaller chunks until what is left is the stable core which is generally accepted by other authors. We define the text quality measure $\alpha_{text}(r_i)$ as the solution to the following equation that expresses text decay in this manner:

$$\sum_{r_j \in J(r_i)} txt(i, j) = txt(i, i) \cdot \left(1 + \sum_{r_j \in J(r_i)} (\alpha_{text}(r_i))^{j-i} \right)$$

The resulting value for $\alpha_{text}(r_i)$ ranges from 0 for text which is completely removed, to $+1$ for text which is completely preserved.

A different way to measure the quality of a text contribution is to simply sum the amount of text that remains, over the succeeding revisions. For a revision r_i, by considering the amount of original text introduced in r_i that survives in the next ten revisions, we define the following additional quality measure for text contributions,

$$\beta_{text}(r_i) = \frac{1}{txt(i, i)} \cdot \left(\sum_{r_j \in J(r_i)} txt(i, j) \right)$$

This value generally ranges from 0 for text which is immediately removed, to $+10$ for text which completely survives all ten revisions. (Due to how text tracking works, if a piece of text is copied within an article, the original author might receive credit for more words than she originally wrote.)

3. CONTRIBUTION MEASURES

We would like to define quantitative measures of author contributions to the Wikipedia. Typically, contribution metrics measure only the *quantity* of a contribution, which implicitly assumes that all authors are working towards a common goal. We propose that collaborations including anonymous authors might be better served by factoring *quality* with quantity. Each of the metrics we define below is formulated as *quality* · *quantity*, to make explicit what each value is.

For every page p in the Wikipedia, we consider a revision r_i performed by the author a_i for some $0 < i \leq n$. Each of the subsequent authors $a_{i+1}, a_{i+2}, \ldots$ can either retain, or remove, the edits performed by a_i in revision r_i. These authors who subsequently edit version v_i of the article are considered *judges* of r_i and hence of the contribution made by author a_i. We define various measures of author contribution, taking into account the amount of text added or edits performed by the author and the quality of those changes. We would like to measure contributions both in absolute terms, as the amount of text that was added by an author or the amount of edits made by an author, and in relative terms, where we take into account the quality of the edits. The contributions of all authors is

cumulative over the entire revision history of the Wikipedia; for our experiments, we picked revisions of all articles previous to October 1, 2006.

3.1 Number of Edits

The simplest quantitative measure of contribution for authors is to compute the number of revisions they authored. In previous works, this is referred to as the *number of edits* made by an author [23, 24, 8, 17]. We follow the tradition, and define this precisely as:

$$\forall a \in \mathbb{A}.\mathsf{NumEdits}(a) = \sum_{p \in \mathbb{P}} \sum_{r \in E(a,p)} 1 \cdot 1.$$

3.2 Text Only

Another very natural measure of author contribution is to count up how many words were added by each author, during the course of all their revisions. Since there is no quality measure involved, we refer to this measure as TextOnly, and define it as:

$$\forall a \in \mathbb{A}.\mathsf{TextOnly}(a) = \sum_{p \in \mathbb{P}} \sum_{r \in E(a,p)} 1 \cdot txt(r).$$

We refer to this measure as the *absolute* text contribution measure.

3.3 Edit Only

Correcting grammar, polishing the article structure, and reverting vandalism are all chores [3] which must be done to keep the Wikipedia presentable. Counting the number of words added can partially account for these chores if there is no text authorship tracking done; without text authorship tracking, however, vandals could easily subvert any system for measuring contributions. Instead, we note that measuring the size of the change in each revision is able to reward both authors who write new text, as well as authors who polish existing text. More formally, we measure the edit distance between the the version of a page that was generated by revision r_i and the version that immediately preceded it. The EditOnly measure is thus defined as:

$$\forall a \in \mathbb{A}.\mathsf{EditOnly}(a) = \sum_{p \in \mathbb{P}} \sum_{r \in E(a,p)} 1 \cdot d(r).$$

We refer to this measure as the *absolute* edit contribution measure.

3.4 Text Longevity

The next level of sophistication is to incorporate non-constant quality measures into the calculation of contribution. We desire the text longevity of a revision to be the amount of original text that was added by the author a_i for a revision r_i, discounted by the text quality measure $\alpha_{text}(r_i)$, which describes how the text decays over the next several revision.

$$\forall a \in \mathbb{A}.\mathsf{TextLongevity}(a) = \sum_{p \in \mathbb{P}} \sum_{r \in E(a,p)} \alpha_{text}(r) \cdot txt(r)$$

3.5 Edit Longevity

Similar to the text longevity measure, we define the edit longevity of a revision r_i as the edit contribution, discounted by the average edit quality measure $\overline{\alpha}_{edit}(r_i)$. As with all the measures, we accumulate contributions based on edit longevity over all revisions edited by an author:

$$\forall a \in \mathbb{A}.\mathsf{EditLongevity}(a) = \sum_{p \in \mathbb{P}} \sum_{r \in E(a,p)} \overline{\alpha}_{edit}(r) \cdot d(r)$$

3.6 Ten Revisions

A simpler method for measuring how useful newly inserted text is, is to simply add up how many words survive over the next ten revisions. Large contributions are thus richly rewarded, if they survive; smaller contributions have a slightly better chance of surviving for the entire ten revisions, thus encouraging change — but not too much change.

We consider the ten revisions that follow any revision r_i of an article, and accumulate the amount of text contribution that was made in r_i that remained in each of those ten subsequent revisions of the article. We call this measure TenRevisions and define it as follows:

$$\forall a \in \mathbb{A}.\mathsf{TenRevisions}(a) = \sum_{p \in \mathbb{P}} \sum_{r \in E(a,p)} \beta_{text}(r) \cdot txt(r).$$

Note that our definition of $\beta_{text}(r)$ incorporates the text survival for the next ten revisions, thus simplifying the definition here into the *quality* · *quantity* presentation we are favoring.

3.7 Text Longevity with Penalty

A last variation that we propose is to combine text longevity with edit longevity in such a way that authors of new content are rewarded, but vandals are actively punished for both inserting and deleting text. Text longevity, as we have defined it, already does not reward vandals — vandals either insert no text, or the text they insert is immediately removed; both cases result in a text longevity of zero for the revision. Vandals are still able to accumulate positive contributions from other revisions, however, while disrupting other authors with their vandalism. By only counting edit longevity when it is negative, we are able to punish vandals for any kind of vandalism which is reverted. This leads to the following definition of our punishing measure:

$$\forall a \in \mathbb{A}.\mathsf{TextLongevityWithPenalty}(a) =$$
$$\mathsf{TextLongevity}(a) + \sum_{p \in \mathbb{P}} \sum_{r \in E(a,p)} \min(0, \overline{\alpha}_{edit}(r)) \cdot d(r).$$

4. IMPLEMENTATION

As part of our previous research into author reputation and text trust [2, 1], we have created a modular tool for processing XML dumps from the Wikipedia. It analyzes all the revisions of a page, filtering down the revisions to remove consecutive edits by the same author, and computing differences between revisions to track the author of each word and measure how the author might have rearranged the page. These results can be passed to any of several modules to do additional processing; we use the tool to reduce the enormous collection of data down to a much smaller *statistics file*. We process the statistics file with a second tool, which we instrumented to calculate the various contribution measures we have defined. The original tools are open-source, and can be downloaded from our project page, at `http://trust.cse.ucsc.edu/`. More precise details about text tracking and edit distance are available in [2].

Our analysis is based on main namespace article revisions from the Wikipedia dump of February 6, 2007, which we process to create a reduced statistics file. The statistics file contains information about every version, including the amount of text added, the edit distance from the previous version, and information about how the edit persists for ten revisions into the future. To ensure that each version we considered had revisions after it, we consider only versions before October 1, 2006. After further processing on the file, we used R[14] to analyze the resulting data.

Bots. During the course of our analysis, we found that some authors were extraordinary outliers for multiple measures. Some

investigation into the most extreme cases revealed that bots were making automated edits to the Wikipedia, and that a few bots dwarfed manual labor in the edit based measures EditLongevity and EditOnly. We also found that there are bots that improve content, and bots that vandalize it. We chose to identify bots as those with a username which ends in the string "bot;" While this does not include every bot (especially the ones that vandalize), it is a useful first approximation. We found 614 bots in total as of October 1, 2006.

Vandals. There is a similar problem in trying to define vandals, since such authors don't register themselves as such. For our purposes, we decided to define a vandal as someone who, on average, makes an edit which is completely reverted. Precisely, we define a vandal who meets one of two criteria: $\alpha_{text} < 0.05$, or $\overline{\alpha}_{edit} < -0.9$. We justify this choice in the next section.

5. ANALYSIS

We begin our analysis with some information about the data we are analyzing. Our reduced statistics file includes over 25 million revision records. Figures 1 and 2 were created by drawing a random sample of 5 million records, due to memory limitations of the the software package.

In Figure 1, we show the frequency distribution of the two quality measures α_{text} and $\overline{\alpha}_{edit}$ over the revisions we sampled. We see both measures are heavily biased towards $+1$, indicating that most revisions to the Wikipedia are generally considered useful by succeeding authors. This confirms the intuition that more "good people" than "bad people" must contribute, otherwise the Wikipedia would have a difficult time maintaining the community which continues to extend the online encyclopedia in a useful way.

Delving directly into the data for text quality, we observe that 10% of the revisions made had $\alpha_{text} \leq 0.05$ while 66.67% of the revisions had $\alpha_{text} > 0.95$. When $\alpha_{text} = 0$, the text is immediately deleted in the next revision, so we can infer that these revisions are the work of vandals. When we look at the size of contributions made, we noticed that 6% of the amount of new text added had $\alpha_{text} = 0$, whereas 76.21% of the new text added had $\alpha_{text} > 0.95$. From this we conclude that authors mostly add good new text.

The data is less stark for edit quality. When we looked at revisions, we saw that 1.9% of the revisions had $\overline{\alpha}_{edit} \leq -0.9$, whereas 51.12% had $\overline{\alpha}_{edit} > 0.9$. In fact, 84.71% had positive edit quality. In terms of edit contributions, we noticed that 7.5% of the edit contributions had $\overline{\alpha}_{edit} \leq -0.9$, whereas 61.39% had $\overline{\alpha}_{edit} > 0.9$. Moreover, 1.6% of the edit contributions were immediately reverted. From these statistics, we conclude that authors mostly do good edits.

Figure 2 shows the absolute text and edit contributions, $txt(r_i)$ and $d(r_i)$, for the sets of sampled revisions. It is important to note that these two graphs are using the logarithm of the size of contribution, along the x-axis; edit sizes can fall below $+1$, due to the way we compute edit distance for moved words as a fraction of how much of the document they move across. Thus, the frequency count for edit sizes between 0 and 1 suggests that a good fraction of revisions involve rearranging of text. Beyond that, we can conclude that contributions, as measured by text added or by edit distance, are predominantly under 100 words.

In Figure 3 we show the average edit quality and average text quality for all non-anonymous authors. In order to compute this, we took all revisions created by each author and took an average of the text and edit qualities of those revisions. We notice that 15.9% of authors had $\alpha_{text} \leq 0.05$ and 6.3% of authors had $\overline{\alpha}_{edit} \leq -0.9$. These are shown by the bars on the left extreme of the histograms

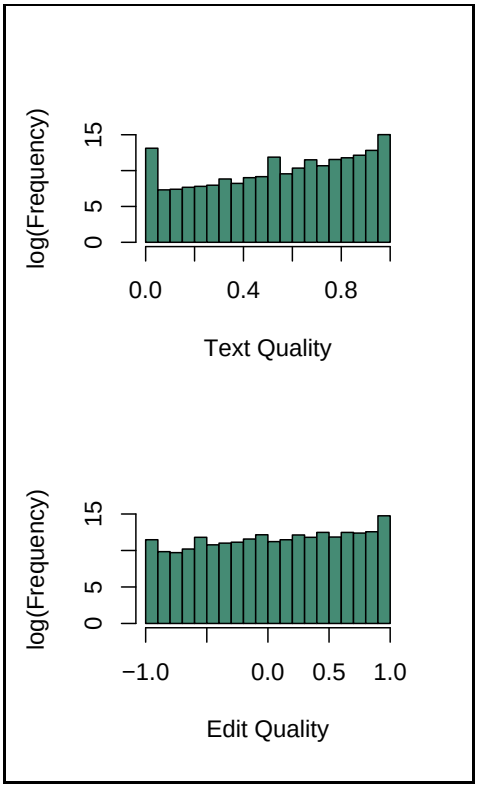

Figure 1: This graph shows the text quality α_{text} and edit quality measure $\overline{\alpha}_{edit}$ for 5 million randomly selected records of each type.

in Figure 3. This sharp increase in the number of authors at the lowest end of our quality measures, combined with our previous analysis of revisions and contributions with respect to quality, gives us some justification to define vandals as those authors who have either $\alpha_{text} \leq 0.05$ or $\overline{\alpha}_{edit} \leq -0.9$ on average. We state that the identification of vandals can be made more precise using more sophisticated analyses of our data, but we don't deal with that in this paper.

During our investigations comparing the proposed measures, we found an unusually large fraction of non-anonymous authors having scores relatively close to zero. This suggested that many users had made a relatively small number of revisions, and that the absolute text and edit contributions of the revisions tended to be small, or that the quality tended towards zero. This is consistent with the power law distribution for edits per author (Lotka's law) detected by [21]; we confirmed the distribution for our data and observed that 362,461 authors made only one edit: over 46% of the total 777,223 authors we tracked. In Figure 4 we show the edit quality measure for these authors. In contrast to the edit quality distribution over all authors from Figure 1, we notice that the edit quality for these authors are almost evenly distributed across the entire quality range (except for the two extreme values).

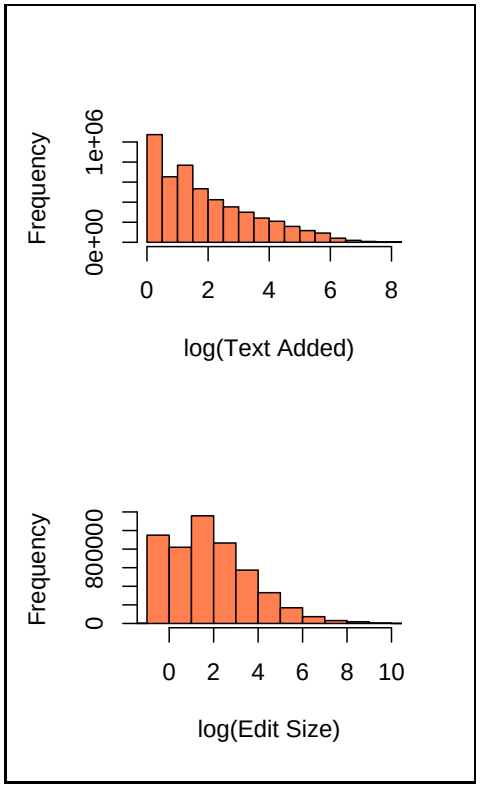

Figure 2: This graph shows the absolute text and edit contributions on a log scale, for 5 million randomly selected records of each type.

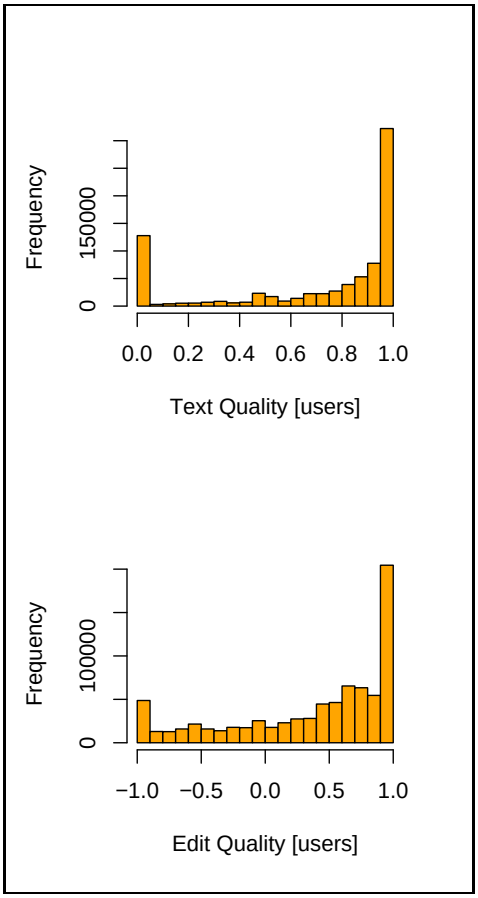

Figure 3: This graph shows the average text quality α_{text} and the average edit quality measure $\overline{\alpha}_{edit}$ over all non-anonymous authors.

5.1 Comparing Measures

We next present the correlations between the various measures in Table 1. These are correlations with respect to the amount of contributions made by all non-anonymous authors, excluding those we've classified as vandals. From the correlation table, we notice that text based measures are better positively correlated with each other. Similarly, the edit based measures are better positively correlated with each other as we expected. The measures EditLongevity and EditOnly are highly correlated as borne out by the fact that a large percentage of the edits are of good quality. We notice that the same is true for TextLongevity and TextOnly. The correlation between TextLongevityWithPenalty and the absolute measure EditOnly is low, demonstrating that TextLongevityWithPenalty penalizes authors for bad edits, gives no credit to good edits, and accumulates the quality discounted text contribution measure TextLongevity. Therefore, authors need to contribute high quality text, while ensuring that they have no bad edits to get a high score on TextLongevityWithPenalty. TenRevisions being a text contribution measure, is highly correlated with the other text contribution measures TextOnly and TextLongevity. NumEdits is positively correlated with all measures as we would expect, since the majority of contributions are deemed good by each of the quality measures.

While TextOnly and EditOnly appear to be reasonable measures of author contribution, we have found evidence that vandals accrue large contributions against these measures. For instance, we found that author 1065172 is in the $99th$ percentile when measured using TextOnly, but is nearly at the bottom of the ranks, at 0.000001 quantile when we look at his TextLongevityWithPenalty measure. We found five revisions in which this author added new text, but four of those were immediately reverted. The only revision that was kept around was a one word addition to a page! From the edits made by this author, we saw that he is a spammer. On the other hand, using TextLongevity instead of TextOnly we noticed that the author was below the $25th$ percentile. On the EditLongevity measure, this author was below the 0.001 quantile; among the lowest in rank. Therefore, we argue that the measures that discount TextOnly and EditOnly by a text or edit quality measure are more indicative of the "useful" work added to the Wikipedia. We argue that NumEdits is not as good a measure, since vandals and bots can easily make large numbers of bad edits.

We present two figures, Figure 5 and Figure 6, which have been restricted to a region containing the bulk of the data points. In Figure 5, we see a vee shape, which separates the authors into two groups: those that have positive edit quality and those that have negative edit quality, as measured by $\overline{\alpha}_{edit}$. The worse the quality of edits made by authors the less they accumulate of the EditLongevity measure, whereas the EditOnly measure, being oblivious to edit quality, attributes the same contribution to an author whose contributions persists as it does to an author whose contributions do not. On the negative side of EditLongevity, there are points that represent vandals, who edit large sections of existing pages, which are then immediately reverted. Clearly, EditOnly ranks some of these authors very highly, whereas EditLongevity is able to distinguish

$Measures$	$EditLong$	$EditOnly$	$NumEdits$	$TenRevs$	$TextLong$	$TextOnly$	$TextWPen$
$EditLong$	1.000	0.999	0.28	0.070	0.075	0.16	-0.32
$EditOnly$	0.999	1.000	0.29	0.071	0.077	0.16	-0.33
$NumEdits$	0.283	0.286	1.00	0.361	0.417	0.45	0.27
$TenRevs$	0.070	0.071	0.36	1.000	0.983	0.96	0.89
$TextLong$	0.075	0.077	0.42	0.983	1.000	0.98	0.90
$TextOnly$	0.158	0.164	0.45	0.963	0.983	1.00	0.82
$TextWPen$	-0.320	-0.326	0.27	0.886	0.897	0.82	1.00

Table 1: This table gives the pairwise correlations of the different measures we have defined in this paper.

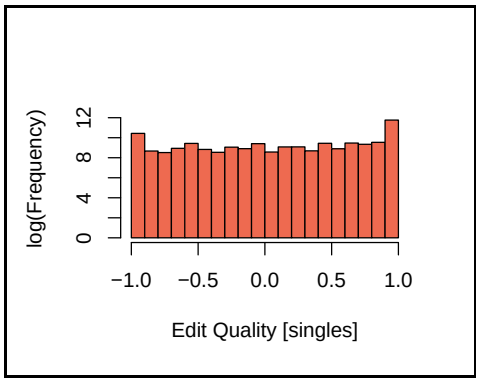

Figure 4: This plot shows the $\overline{\alpha}_{edit}$ of the non-anonymous authors who made a single edit contribution.

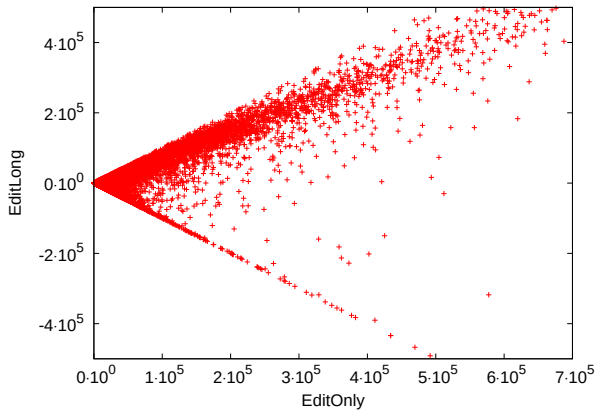

Figure 5: Comparing the absolute edit contribution of a user with the edit longevity. Notice that authors who are "all bad" are easily identifiable – and sometimes quite prolific.

them and rank them very low.

In Figure 6, we see a similar vee shape; in this case, TextLongevity cannot go below zero as the text quality measure is always non-negative, so vandals, by our definition, receive no contribution. As before, the measure that incorporates quality can distinguish vandals from non-vandals and attribute a contribution measure to authors that is proportional to the merit of their contribution.

Of the various measures we introduced, TextLongevityWithPenalty is perhaps the one with the least tolerance, since by this measure, the only way an author can accumulate contribution is by adding new text that persists and by making edits that are judged to be of good quality. Further, this measure does not reward authors for good edits, but penalizes them for bad edits. In Figure 7, we plot TextOnly against TextLongevityWithPenalty. We see the vee shape, with vandals falling on a noticeable line in the fourth quadrant, that has no TextOnly contribution. Since almost all new text added by vandals is immediate reverted, and their edits always have low quality, we notice that they get low negative TextLongevityWithPenalty contributions. In fact, we noticed that the bottom ten authors by rank when measured according to TextLongevityWithPenalty were all vandals with the exception of $AntiVandalBot$. We explain this in the subsection on bots.

5.2 Ranking Authors

A different direction we explored was how these different measures end up ranking different authors. Since the contribution measures varied over such a wide range of values, with most people within a smaller region around zero, we hoped that ranking the authors would give us better insight into how the measures differed.

To this end, we computed the percentile rank (rounded up to the next even value for clarity in the image) of all non-anonymous authors, including those that we had classified as vandals, and then plotted them in 3-dimensional histograms; see Figures 9 and 10. The correlation structure implied by Table 1 becomes apparent. An important point to remember about Figures 9 and 10 is that the low-lying regions of the graph are rarely zero — there are roughly between one and ten authors at each intersection, but this is so small compared to the areas that correlate that we cannot see it on the graph.

We also include a 3-dimensional histogram comparing the percentile rankings as determined by EditLongevity and NumEdits, in Figure 11. The "rows of fences" we see in Figure 11 are due to the large number of authors who make only a handful of edits; the NumEdits measure neither distinguishes them from each other, nor is it capable of distinguishing good contributions from bad contributions. This last point is important, that even users in the lowest percentile of EditLongevity can be rated very highly by NumEdits—demonstrating that it is much easier to game the NumEdits measure to achieve a high rank, while doing bad work.

5.3 Bot Behavior

There are several bots operating on the contents of the Wikipedia. Many bots are sanctioned by the community, and do useful chores such as automatically removing text which is likely to be vandalism, correcting spelling, and adding geographical data. There are also bots which are created to vandalize pages, and sometimes well-intentioned bots run amock and accidentally vandalize pages as well. During the course of comparing the various contribution measures with each other, we found several bots (both good and

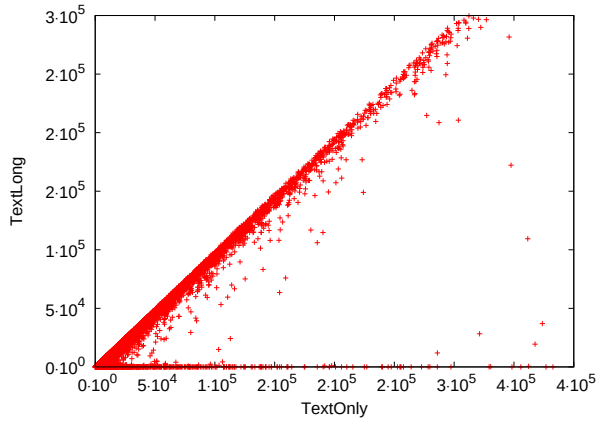

Figure 6: Comparing the absolute text contribution with the contribution as measured by text longevity. We see that large contributors are either "all bad" or nearly "all good."

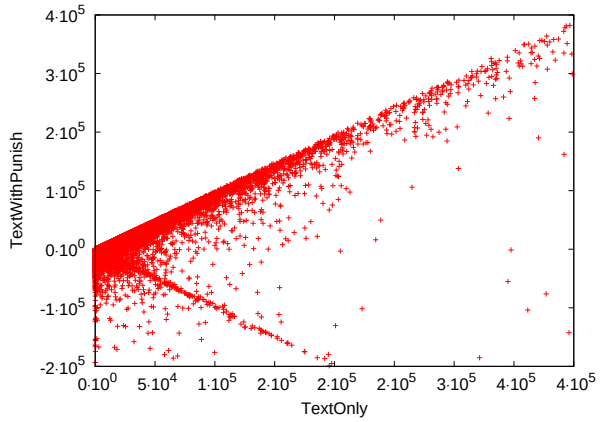

Figure 7: Comparing the absolute text contribution of an author with their contribution as measured by TextLongevityWithPenalty**.**

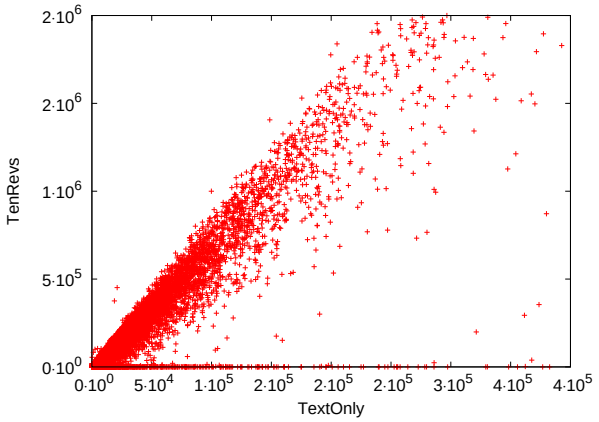

Figure 8: This graph compares how much text is initially added by a user (along the x-axis), with how much of the text survives over the next ten filtered revisions (along the y-axis). The higher up the y-axis a point is, the more text that survived all ten revisions. Most authors add under 100,000 words, and about half of what they add survives.

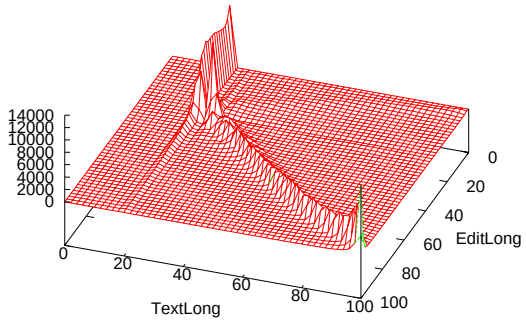

Figure 9: EditLongevity **vs** TextLongevity

bad) which were obvious outliers in the data. To analyze bots as a group, we selected all users which included the "Bot" moniker in their username; this self-identification does include some malicious bots, but obviously favors selection of good bots.

The edit and text quality measures for all bots are similar to that of all authors shown in Figure 1. We noticed that bots create a large number of revisions with high quality. We found that 69.56% of the revisions made by bots have a text quality measure of $\alpha_{text} > 0.95$. The percentage of revisions made by bots with $\alpha_{text} \leq 0.05$ was 9.2%. We found that 66.92% of the new text added by bots were with $\alpha_{text} > 0.95$ and 14.14% of the new text added by bots were with $\alpha_{text} = 0$, which means they were immediately reverted. Similarly, on the edit contributions of bots we found that 54.42% of the revisions with edits made by bots were of high edit quality, with $\overline{\alpha}_{edit} > 0.9$. The number of revisions having $\overline{\alpha}_{edit} < -0.9$ being negligible; 1% from our analysis. When we counted all edit revisions that had a negative edit quality we saw that 12.73% of the revisions were judged to be of poor quality with $\overline{\alpha}_{edit} < 0$. We found that 93.3% of the edit contributions made by bots had positive edit quality and the remaining 6.4% had negative edit quality. More interestingly, 65.20% of the edit contributions made by bots had $\overline{\alpha}_{edit} > 0.9$, which means they were not edited out in subsequent revisions and represent the sheer amount of work done by bots that is of very high quality. The contributions with $\overline{\alpha}_{edit} < -0.9$ are 1.8%. This indicates that a large part of the text additions made by bots and a large part of the edit contributions made by bots survive indefinitely.

Furthermore, our analysis indicates that bots make large amounts of edit contributions compared to text contributions; the ratio of the size of edits EditOnly to the size of new text TextOnly for all bots is 11.61. Since the penalizing measure TextLongevityWithPenalty does not credit authors for good edits but reduces their TextLongevity contributions, by the amount of their bad edits as measured by EditLongevity, we notice that edits judged as being of poor quality overwhelm the smaller text contributions of bots in general, and $AntiVandalBot$ in particular, resulting in a small overall contribution. We also note here that $SmackBot$ did much better on this measure. $SmackBot$ contributes more text than $AntiVandalBot$. Most of its edits are

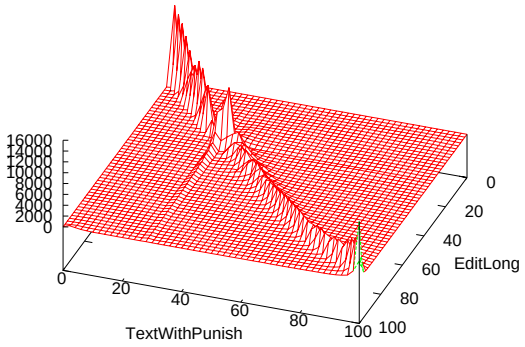

Figure 10: EditLongevity **vs** TextLongevityWithPenalty

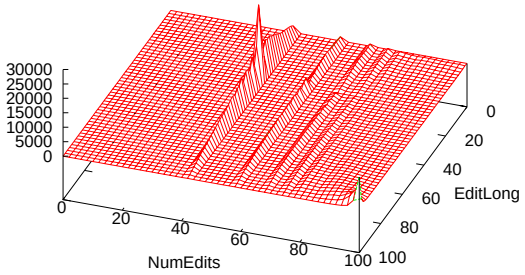

Figure 11: EditLongevity **vs** NumEdits

of smaller size than $AntiVandalBot$. Since they have similar quality measures, $AntiVandalBot$ ends up with a lower score on TextLongevityWithPenalty when compared to $SmackBot$.

5.4 Sources of Error

Since we use filtered revisions, namely we collapse all consecutive revisions by the same author, and since we treat all anonymous authors identically, consecutive edits made by anonymous authors cannot be distinguished. We therefore discard all anonymous authors from our analysis: in any case, we are not measuring their contributions, as they cannot be individually attributed. We have noticed that there are anonymous authors who do good work on the Wikipedia, but at this point we have not implemented a mechanism to attribute them a contribution measure.

We ignore the time difference between edits. When pages receive many views with little editing, it suggests that the article is substantially correct; perhaps later edits are due to changing facts, and not because of poor quality. Articles which are the subject of current events are particularly likely to have their edit quality misjudged. Relatedly, grouping revisions by author ignores the fact that edits separated by days or months are less related and have most likely been reviewed by others.

5.5 Comparing Contributions

Defining multiple contribution measures affords us the opportunity to examine and quantify the user behaviors over the large scale of edits performed. We looked at the list of all blocked authors. [1] We separated them from the others with the objective of determining how many of these authors met our definition of vandals. We were surprised to note that over 51% of authors had $\alpha_{text} > 0.95$ and 39% of authors had $\overline{\alpha}_{edit} > 0.9$. In fact, over 47% of the blocked authors make text contributions that have an average text quality over 0.95. Similarly, over 32% of the these authors make edit contributions that have an average edit quality over 0.9. We note that 11.2% of these authors qualify as vandals by our measure, based on their average edit quality and 24.9% qualify as vandals based on their average text quality. But a large percentage of the authors in the blocked authors list are not vandals, as determined by our definition.

A couple of cases in point are those of authors 3362 and 10784. They are both blocked, but are over the $99th$ percentile on EditLongevity, TextLongevity and TextLongevityWithPenalty. One was blocked by Jimbo Wales and the other was blocked as he was suspected of using multiple accounts.

We end this subsection, by mentioning the top rankers against all measures. The highest ranks across all contributions were secured by authors 3903 and $AntiVandalBot$. Author 3903 had the top rank with respect to measures TextOnly, TextLongevity, TextLongevityWithPenalty and TenRevisions. $AntiVandalBot$ had the top rank with respect to the measures EditLongevity and EditOnly. Interestingly, $SmackBot$ was the second highest scorer after author 3903 on measures TextLongevity and TextLongevityWithPenalty.

6. CONCLUSIONS

As group collaboration becomes more prevalent, the problem of how to compute author contributions becomes increasingly relevant. Our motivation was to explore simple models of user behavior that can be incorporated into reputation systems (e.g., [2]), but we feel that factoring in a notion of *quality* alongside *quantity* can also be revealing in studies about user behavior and the amount of useful information added to the Wikipedia because it cancels out the work of vandals and the work of those who fix the vandalism. We have presented and compared several possible ways to measure author contribution, including two measures popularized by previous works. What we discovered is that there is substantial agreement between the measures for clear cases of valuable contributions, and varying results for authors making questionable contributions.

There are several measures we have defined that have a desirable property, namely, giving credit where it is due and making sure that authors who make short-lived contributions get a low score. We believe that TextLongevity or EditLongevity are equally viable as contribution measures.

The EditLongevity measure is a very interesting measure in our opinion. This measure uses edit distance (as counted in words) to measure the size of the contribution while taking into account the longevity of that contribution, quantified using the edit quality measure $\overline{\alpha}_{edit}$. Since the edit quality measures how much an edit takes a page towards a future version of that page, we find this a good way of measuring contribution. The TextLongevityWithPenalty measure is good at identifying vandals, but fails as a good contribution measure as it does not reward good edits.

As a side effect of our analysis and comparison, we were able

[1] Retrieved on May 8, 2008, directly from the Wikipedia database. It corresponds to the data available at `http://en.wikipedia.org/wiki/Wikipedia:List_of_banned_users`.

to identify some unusual author behaviors. We discovered that the highest contributor by our edit measures was a bot, the second highest contributor by TextLongevity and TextLongevityWithPenalty was again a bot, and that there are evil bots which create a significant amount of vandalism. We also discovered that making large and good text and edit contributions are not always sufficient to be in good standing on the Wikipedia.

There are several directions for future work on measuring author contributions. Our approach has been to consider *content-driven* quality metrics, where no human judgements are necessary, focusing on various measures of *longevity*. Other quality measures are equally viable, such as a "thumbs-up or thumbs-down" rating system for contributions, and the challenge is in both defining them and interpreting the results within context. For example, we have described long-lived content as "good," but might have also described the content as having reached a group consensus. Factoring quality measures into contribution measures can be useful in other collaborative endeavors such as source code archives, or even forum postings. Again, interpretation should be approached with care; for example, a wiki on current events might value short-lived content. Finally, although we have observed that there is general agreement between the metrics we have examined, the differences between them highlight groups of users who behave unusually. We have tried to explain a few of the prominent groups, but there is still much to understand about various behaviors that users exhibit.

7. REFERENCES

[1] B.T. Adler, J. Benterou, K. Chatterjee, L. de Alfaro, I. Pye, and V. Raman. Assigning trust to wikipedia content. Technical Report UCSC-CRL-07-09, School of Engineering, University of California, Santa Cruz, CA, USA, 2007.

[2] B.T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. of the 16th Intl. World Wide Web Conf. (WWW 2007)*. ACM Press, 2007.

[3] M. Burke and R. Kraut. Taking up the mop: identifying future Wikipedia administrators. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 3441–3446, New York, NY, USA, 2008. ACM.

[4] G. Cormode and S. Muthukrishnan. The string edit distance matching problem with moves. *ACM Trans. Algorithms*, 3(1):2, 2007.

[5] W. Cunningham and B. Leuf. *The Wiki Way. Quick Collaboration on the Web*. Addison-Wesley, 2001.

[6] R. E. Park et al. Software size measurement: A framework for counting source statements. Technical Report CMU/SEI-92-TR-020, Carnegie Mellon University, September 1992.

[7] C. L. Giles and I. G. Councill. Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing. *Proc. of the National Academy of Sciences*, 101(51):17599 – 17604, 2004.

[8] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the Few vs. Wisdom of the Crowd: Wikipedia and the rise of the Bourgeoisie. *Alt.CHI*, 2007.

[9] N. Korfiatis, M. Poulos, and G. Bokos. Evaluating authoritative source using social networks: an insight from Wikipedia. *Online Information Review*, 30(3):252–262, 2006.

[10] V.I. Levenshtein. Binary codes capable of correcting insertions and reversals. *Sov. Phys. Dokl.*, 10:707–710, 1966.

[11] D.L. McGuinness, H. Zeng, P.P. da Silva, L. Ding, D. Narayanan, and M. Bhaowal. Investigation into trust for collaborative information repositories: A Wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web*, 2006.

[12] F. Ortega and J. M. G. Barahona. Quantitative analysis of the Wikipedia community of users. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 75–86, New York, NY, USA, 2007. ACM.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[14] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[15] H. P. Schultz. Software management metrics. Technical Report AD-A196 916, MITRE, May 1988.

[16] C. Shirky. Gin, television, and social surplus. `http://www.herecomeseverybody.org/2008/04/looking-for-the-mouse.html`, April 2008. (Retrieved on 9-May-2008.).

[17] K. Stein and C. Hess. Does it matter who contributes: a study on featured articles in the german wikipedia. In *HT '07: Proceedings of the 18th conference on Hypertext and hypermedia*, pages 171–174, New York, NY, USA, 2007. ACM.

[18] B. Suh, E. H. Chi, A. Kittur, and B. A. Pendleton. Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1037–1040, New York, NY, USA, 2008. ACM.

[19] A. Swartz. Who writes Wikipedia? `http://www.aaronsw.com/weblog/whowriteswikipedia`, September 2006. (Retrieved on 9-May-2008.).

[20] W.F. Tichy. The string-to-string correction problem with block move. *ACM Trans. on Computer Systems*, 2(4), 1984.

[21] J. Voß. Measuring wikipedia. In *Proc. of the 10th Intl. Conf. of the ISSI*, 2005.

[22] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, 1974.

[23] J. Wales. Wikipedia, emergence, and the wisdom of crowds. `http://lists.wikimedia.org/pipermail/wikipedia-l/2005-May/021764.html`, May 2005. (Retrieved 9-May-2008.).

[24] D. M. Wilkinson and B. A. Huberman. Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 157–164, New York, NY, USA, 2007. ACM.