Cumulative Growth in User-Generated Content Production: Evidence from Wikipedia*

Aleksi Aaltonen
Warwick Business School
The University of Warwick

Stephan Seiler
Stanford University
Centre for Economic Performance

This draft: May 18, 2015

Open content production platforms typically allow users to gradually create content and react to previous contributions. Using detailed edit-level data across a large number of Wikipedia articles, we investigate how past edits shape current editing activity. We find that cumulative past contributions, embodied by the current article length, lead to significantly more editing activity, while controlling for a host of factors such as popularity of the topic and platform-level growth trends. The magnitude of the effect is large; content growth over an eight-year period would have been 45% lower in its absence. Our findings suggest other open content production environments are likely to also benefit from similar cumulative growth effects. In the presence of such effects, managerial interventions that increase content are amplified because they trigger further contributions.

JEL Classification: D24, L23, L86, M11, O31

Keywords: Wikipedia, Open Source, User-generated Content, Knowledge Accumulation

*We thank conference participants at Marketing Dynamics (UNC Chapel Hill) and NBER digitization for feedback. We also benefitted greatly from discussions with Kate Casey, Ben Faber, Avi Goldfarb, Andreea Gorbatai, Shane Greenstein, Anders Jensen, Michael Kummer, Christos Makridis, Petra Moser, Navdeep Sahni, and Felix Weinhardt. All errors are our own.

1 Introduction

The recent proliferation of user-generated content marks the emergence of a new kind of production. Rather than using managerial procedures to arrive at a pre-specified, proprietary output, Benkler (2006) characterizes the new form of content creation as *commons-based peer production*, a process that is "decentralized, collaborative, and nonproprietary; based on sharing [...] outputs among widely distributed, loosely connected individuals." One of the most successful examples of this new form of production is Wikipedia. Since its inception in 2001, the online encyclopedia has grown to 4.6 million articles and 23 million registered users in its English version alone.[1] A distinctive characteristic of Wikipedia is the cumulative process by which individual contributors provide small fragments of content that gradually add up to an encyclopedia article. Such granular division of labor differs considerably from a traditional editorial process in which separate authors are contracted to deliver complete, authoritative articles (Aaltonen and Kallinikos (2013)). Another characteristic that distinguishes Wikipedia from traditional firm-based production is the absence of monetary incentives, which has led to an interest in understanding the motivational basis of Wikipedia. Studies have identified a range of motivations that drive contributions (Nov (2007)), analyzed specific contribution mechanisms in detail (Gorbatai (2011) and Hansen, Berente, and Lyytinen (2009)), discussed the implications of social structure on contributor behavior (Zhang and Zhu (2011)), and identified disincentives to contributing (Halfaker, Geiger, Morgan, and Riedl (2013)).

In this paper, we posit that the way in which the new form of content production is organized is *inherently* motivating. We argue that the gradual nature of content development encourages and inspires users to contribute more when some amount of content already exists. This motivational mechanism emerges from the fact that subsequent contributors are able to build on already existing content rather than having to contribute an entire article. The existence of content thus lowers the cost of editing and makes incremental edits a useful contribution to the cumulative effort. Furthermore, existing content can influence users by providing new information about a topic or by making potential areas for further contributions salient to them (Olivera, Goodman, and Tan (2008)). As a consequence, articles that are edited more heavily and therefore grow in length will continue to be edited more. In this paper, we study whether such a *cumulative growth effect* exists and whether it is a quantitatively important driver of content growth on Wikipedia.

To study this phenomenon, we rely on a comprehensive data set of editing activity on Wikipedia that contains the full text of every version of each article.[2] The data allow us to examine editing behavior at a great level of detail, making the online encyclopedia an ideal testbed for studying the new form of content production. Using data on a large set of articles over an eight-year period, we find evidence for a cumulative growth effect. More specifically, controlling for article fixed effects and a platform-level time trend, the current length of the article has a positive impact on the amount of editing activity it receives. Based on a battery

[1] See http://en.wikipedia.org/ wiki/Wikipedia:Size_of_Wikipedia, retrieved 12/23/2014

[2] Even a casual user can easily access an article's edit history by clicking on the "View history" tab in the top-right corner of an article page.

of sensitivity checks, we argue the identified effect of article length on editing activity is causal. The effect is quantitatively important; growth in editing activity during our sample period from 2002 to 2009 would have been 45% lower in the absence of the effect. Finally, we show that editing activity triggered by changes in article length leads into improvements in content quality.

The main managerial lesson that we draw from the findings is that, in the presence of a positive effect of current content on editing activity, any action that increases content can trigger further contributions. The effect thus results in path dependence in editing activity in the sense that content additions lead to *permanently* higher editing activity. Therefore, even small additions early in an article's life can lead to substantial differences in its growth trajectory. Two ways to achieve content increases are, for instance, to incentivize users to contribute content or, even more directly, to pre-populate articles with content. Both managerial interventions lead to a magnified effect due to the knock-on effect on future edits. Because many other platforms mimic Wikipedia, often using the same software and a similar page layout, these findings are likely to carry over to those related platforms. For instance, the for-profit platform Wikia hosts a wide range of wikis on topics relating to popular culture that attracts over 100 million monthly visitors.[3] Second, many non-profit projects, such as Wiktionary (a dictionary) and Wikiversity (a collection of open source teaching materials), use the same user interface as Wikipedia.[4] Finally, many prominent companies such as Sony, Xerox, Disney, Microsoft, and Intel use internal wikis to create, store, and share knowledge within the company.[5] They intend to harness the same principles as Wikipedia, and many, such as Intelpedia, use the same open-source software that underpins Wikipedia.[6] The lessons we can learn from studying content evolution on Wikipedia can therefore inform the design of these related platforms that mimic it.

This paper contributes to the literature on content growth in Wikipedia and to the literature on user interaction in open content production. In contrast to the predominantly descriptive papers documenting the growth in content production on Wikipedia, such as Almeida, Mozafar, and Cho (2007), Suh, Convertino, Chi, and Pirolli (2009) or Voss (2005), our aim is to understand a particular driver of the growth process. We share this goal with several studies that investigate other determinants of editing behavior and content growth. Zhang and Zhu (2011) show that the number of other users on the platform, namely, audience size, positively influences the amount of editing. Ransbotham and Kane (2011) investigate the effect of contributor turnover on article quality and find a curvi-linear relationship with an intermediate level of turnover being optimal. Kittur and Kraut (2008) and Arazy, Nov, Patterson, and Yeo (2011) analyze the effect of coordination between contributors and user composition on article quality, respectively. We add to this literature by identifying an additional driver of content growth (and quality improvement): the cumulative effect of current content on editing intensity. This effect is not mutually exclusive from the other mechanisms

[3] See http://www.wikia.com/Wikia, retrieved 9/4/2014.

[4] Wikipedia itself documents a large number of Wikis: http://en.wikipedia.org/wiki/List_of_wikis.

[5] See Bloomberg Business Week, "No Rest for the Wiki," accessed 9/4/2014

[6] See Socialmedia.biz, "The story of Intelpedia: A model corporate wiki," accessed 9/4/2014.

analyzed in the literature, but turns out to be a quantitatively particularly important one. Our paper is also closely related to Gorbatai (2011), who shows that expert editors become more active when observing prior edits by novice users that signal an interest in the topic. Finally, Kummer (2013) studies how exogenous shocks in readership spill over to neighboring articles and lead to increased readership and editing behavior on those articles. His study differs from our paper by investigating the effect of linkages between articles on readership and edits, whereas our focus is on editing dynamics within an article.[7]

The structure of the paper is as follows: In the next section, we provide a description of the data as well as descriptive statistics. Section (3) presents the main empirical results as well as robustness checks. In sections (4) and (5), we analyze the effect of additional activity on article quality and characterize the changing nature of edits as an article grows in length. Finally, we assess the quantitative importance of the estimated effect and provide some concluding remarks.

2 Data and Descriptive Statistics

We use the English Wikipedia XML database dump extracted on January 30, 2010, that has been made freely available by the Wikimedia Foundation.[8] The data contains the full text of every version of all articles from the beginning of the online encyclopedia in January 2001 to January 2010, allowing us to track the evolution of content across edits for each article. We preprocessed the XML records in the raw data using Python scripts into a tabular data set representing 19,376,577 articles and 306,829,058 edits. Our analysis focuses on a subset of articles that belong to one particular category: the "Roman Empire." We choose this category, which comprises 1,310 unique articles, because knowledge on the topic is presumably undergoing relatively little change during our sample period. This focus removes an additional layer of complexity, which is the incorporation of new information into Wikipedia. In the appendix, we provide details on how we selected the set of articles.

We transform the XML records into a numerical format and compute the length of the article at each version as well as the amount of change in content, measured by the number of characters a particular edit of the article changed. More precisely, for two consecutive versions of the same article, we compute the number of characters that need to be added, deleted, or changed (each of these actions is counted equally) in order to convert one version of the article into the next. For ease of exposition, we will refer to this metric simply as "edit distance" in the remainder of the paper. In order to compute this measure, we use an algorithm known as Levenshtein edit distance (Levenshtein (1966)), which is used in areas such as signal processing, information retrieval, and computational biology (Myers (1986), Navarro (2001), Spiliopoulos and Sofianopoulou (2007)), to quantify the degree of (dis-)similarity between strings. We provide a more detailed description of the procedure and its implementation in

[7] Other papers that also analyze data from Wikipedia, but look at questions less related to our analysis, include Greenstein and Zhu (2012a) and Greenstein and Zhu (2012b), who document the extent of political slant on Wikipedia. Nagaraj (2013) uses Wikipedia data to assess the effect of copyright on creative reuse.

[8] enwiki-20100130-articles-meta-history.xml.7z (size: 5.9 Terabytes)

the appendix. The calculations are computationally heavy but offer an intuitive definition of string difference, that is, the amount of change in content induced by an edit. We are also able to track users across multiple edits by tracking their username or IP address.

2.1 Data Selection

To study the dynamics of editing activity over time, we further cut our sample along several dimensions. First, we exclude edits performed by bots, that is, non-human user accounts that implement automated edits. Second, we remove edits that constitute acts of vandalism. And finally, we do not consider edits that restore a previous version of the same article. In this way, we limit our analysis to "productive" editing activity by human users.

All three types of edits are relatively frequent (see Halfaker, Kittur, Kraut, and Riedl (2009), Vigas, Wattenberg, and Dave (2004), and Piskorski and Gorbatai (2013) for details on reversions) on Wikipedia; therefore, defining them correctly is important. We relegate the description of how we identify bot-edits to the appendix, but outline briefly how we deal with reversions and vandalism. To deal with both issues, we first need to define when an edit constitutes a reversion, which we do by using an assessment of string (dis-)similarity similar in spirit to the edit-distance computation. Specifically, we compare every version of a particular article with the previous 100 versions (if that many exist) and assess whether the current version is identical to any of the previous ones.[9] If we find such an instance, we label as a reversion the edit that restores a previous version; all edits that are undone by the reversion we refer to as reverted edits.

We define acts of vandalism as edits that involve *only* deletion of content and that are subsequently (without any other edits in between) reverted. In other words, we remove any deletion of content that is of a temporary nature. Using this definition, we classify a little over 2% of all edits as vandalism and remove them from the sample.[10] Furthermore, we remove all reversions but maintain the reverted edits in order not to overstate editing activity. For instance, consider the unsuccessful attempt to add 1,000 characters' worth of content. In the data, this attempt will be recorded as two edits (the addition of content and a subsequent revert action), both with an edit distance of 1,000 characters. Such a sequence of edits leads to a seemingly large amount of editing activity while actually leaving the article unchanged. We do keep *reverted* edits in our sample because they constitute legitimate editing activity despite the fact that they do not have a lasting impact on the article. However, we keep track of the reverted edits and later investigate whether they occur disproportionately on longer articles. About 14% of edits constitute reversions.

[9] Note that previous research used other, usually less conservative, definitions. For instance, Suh, Convertino, Chi, and Pirolli (2009) classify edits that have certain keywords (e.g., "revert") in the comment provided by the editing user as reverting edits. Instead of relying on "self-declared" reverts, we compare the actual content by classifying as a reverting edit every instance that returns the article content to a previous version of the article. Relative to Suh, Convertino, Chi, and Pirolli (2009), we find a substantially larger fraction of reverts, presumably because of these classification differences.

[10] This definition has some limitations. First, for our definition, we rely on the fact that the vandalism actually has been detected and subsequently reverted. Second, our definition is not able to capture more "subtle" vandalism that involves factually incorrect additions or changes to the content. Nevertheless, we believe we are able to purge a relatively large set of vandalizing edits with this procedure.

In the descriptive statistics below and in the main empirical analysis, we remove these three types of edits when measuring editing activity. However, we do keep track of the aggregate article length at every point in time regardless of the kind of edits that led to an article reaching a particular length. In other words, we want the current content stock captured by the articles' length to reflect all past activity. When considering whether past activity triggers more contributions; however, we confine ourselves to non-automated "productive" editing activity. Note also that, because longer articles are more likely to be edited by bots and attract more vandalism and reversions, we would estimate a larger effect of article length on editing activity when we retain the set of edits described above.

2.2 Editing Behavior

In this section, we provide some basic descriptive statistics on the key variables used in the estimation. We start by describing the magnitude as well as the nature of edits for our final sample of 62,925 edits across all 1,310 Roman Empire articles. The first line of Table (1) reports the amount of content change induced by individual edits measured by the edit-distance metric defined above. We find an average edit distance of 630 and a median edit distance of 37 characters (about half a sentence in the English language). We find a large degree of heterogeneity in the length of edits, with some very large edits in the right tail of the distribution. For instance, the 99th percentile of the distribution takes a value of almost 9,000 characters, which is orders of magnitude larger than the median edit.

The edit-distance metric is arguably the most direct way to measure the amount of change induced by an individual edit.[11] However, it does not allow us to explore in more detail the nature of the content change. To dig deeper into the nature of edits, we focus on one dimension of particular relevance for our study: the degree of content addition versus deletion induced by an edit. To capture the extent of addition and/or deletion of content, we use a simple metric that combines information from edit distances and length changes. In particular, it has to hold that $|\Delta Length| <= EditDistance$. At the extremes, an edit that only adds new content will have $\Delta Length = EditDistance$, whereas for a deletion of content, it holds that $-\Delta Length = EditDistance$. Based on the relationship between the two variables, we compute $\Delta Length/EditDistance \in [-1, 1]$. We find that 43% of edits are pure additions of content (i.e., $\Delta Length/EditDistance = 1$), whereas 8% are pure deletions. The remaining edits are intermediate cases in which some existing content is deleted but new content is also added. Edits within the intermediate range are roughly uniformly distributed over the range of our metric. Next, we report the number of reverted edits according to our definition provided above (remember we exclude reversions from the sample) and find that 14% of edits within the Roman Empire category are reverted. In other words, the content that such edits provide is later removed and these edits have no lasting impact on an article's content. We later characterize the edits triggered by past contributions along the two dimensions just described

[11] Consider, for instance, the case of an edit that *replaces* large parts of an article with new content and might entail little change in article length despite substantial content changes. Our edit-distance metric is able to capture such changes, which one would miss when using article-length changes as a measure of content change.

in order to assess the longevity of triggered edits and the extent to which they provide new content.

For most of our empirical analysis, we aggregate editing activity at the article/week level and measure aggregate editing activity over a fixed weekly time window. Importantly, individual articles often have spells of inactivity, something the summary statistics at the edit level do not capture. We document the distribution of three key variables that measure editing activity in the lower panel of Table (1): the number of edits, number of users, and cumulative edit distance per week (added up across individual edits if multiple edits occur within a week). The unit of observation is an article/week combination, of which we have a total of 252,427 across the 1,310 articles and up to 433 weeks per article. Defining the number of edits involves some judgement calls because in the raw data, every time a new version of the article is saved, a separate entry is recorded. Sometimes users save an article multiple times in a short time interval, and considering all consecutive saved versions by the same user as a single edit is therefore reasonable. We therefore aggregate any edits by the same user within an eight-hour window (without any other user editing the article within the same time window) into a single edit. Because of the arbitrary nature of the edit aggregation, we prefer to work with the number of users per week for most of the analysis. The variable is defined as the number of *distinct* users that edited the articles in a given week regardless of the number of edits per user or sequence of edits.[12] In about 86% of article-weeks, we observe no editing activity. The average number of users is equal to 0.224, and rarely is more than one user editing an article in any given week. The average weekly edit distance is equal to 157 characters.

2.3 Content Growth Patterns

As a backdrop to our empirical analysis, we provide some key empirical facts on the content growth process to which, as we argue later, the cumulative growth effect contributed considerably. Table (2) reports the number of articles created each year and the amount of editing activity on those articles. We find the number of new articles created increases almost monotonically until 2005 and decreases afterward. The next two columns report the total number of users active each year and the number of edits on any article within the category. For both measures, we see a substantial increase in activity peaking in 2007. Finally, we look at the amount of editing captured by the cumulative annual edit distance across all articles. The pattern for this variable is similar to the other measures of editing activity: a strong increase early on and a slight decrease in the later years. In the case of all three metrics, the eventual slowdown and decrease is substantially smaller than the initial "ramp-up," which is consistent with findings elsewhere, such as Suh, Convertino, Chi, and Pirolli (2009).

Similar to Almeida, Mozafar, and Cho (2007), we find the ratio of edits per user as well as the edit distance per edit is stable over time. Therefore, an increase in the user pool rather than changes in users' editing intensity drives most of the growth process on Wikipedia. This

[12]Note the number of edits (as defined above) is highly correlated (correlation coefficient of 0.95) with the number of users per week (which is not affected by multiple saved versions).

pattern is of particular relevance because we later find the cumulative growth effect also operates on this dimension: longer articles have more users editing them, but the amount of editing activity per user is unchanged. We also report how the types of edits being made change over time and find that edits in later years tend to involve more deletion of content. Also, the fraction of edits being reverted increases from close to zero in the early years to about 15% toward the end of our sample period. This pattern is consistent with an increase in reverted edits over time that Kittur, Suh, Pendleton, and Chi (2007) and Halfaker, Geiger, Morgan, and Riedl (2013) document. Finally, we report the mean and median article length in each year in the final two columns of the table.

3 Cumulative Advantage in Content Growth

To estimate the effect of the current content stock on editing activity, we regress the number of weekly users on the length of the article (in units of 10,000 characters) at the beginning of the respective week. Leaving out articles that were started in 2009 or later, because of a short time series, we have 1,191 articles and up to 432 weeks of data for the earliest article, yielding a total of 247,002 observations.[13] To control for the general appeal and popularity of each article, we include a set of article fixed effects in the model. We also control for a general time trend in editing behavior within Wikipedia as a whole by including a set of weekly dummies. We cluster standard errors at the article level.[14] Formally, we run the regression

$$UserNum_{jt} = \beta ArticleLength_{jt} + \theta_j + \psi_t + \varepsilon_{jt}, \tag{1}$$

where j denotes a specific article and t denotes a week. θ_j and ψ_t are a set of article and week fixed effects, respectively. ε_{jt} denotes the error term.

We now turn to discussing under which assumptions the estimated coefficient on article length can be interpreted as a causal effect. The hypothetical experiment that we would like to run is one in which content is randomly added to certain articles but not others. Such an intervention would allow us to estimate β by comparing the editing activity on "treated" articles with activity on the remaining set of control group articles. Clearly, in our setting, article length does not vary randomly across articles but is a function of the general appeal, popularity, and potentially controversial nature of the topic the article covers. Our conjecture is that whereas factors such as article popularity systematically affect article length, a large random component exists concerning when a knowledgable user comes across an article and provides content, thus increasing its length. We argue that, after controlling for the article-

[13] We drop the first week for each article because by construction, the founding week contains at least one edit.

[14] This level of clustering assumes articles can be treated as independent observations with their own process of content generation. This assumption might be violated if users edit multiple articles on related topics. Content on one article might therefore influence editing activity on another. We cannot directly test for such dependencies between articles, but believe they are not quantitatively important in our context. First, very few users actually edit multiple articles. About 85% of users provide content to only one article. Second, Kummer (2013), who studies spillovers between articles, finds that shocks to readership on one article do generate higher readership on related articles, but the effect on editing activity on related articles is very small.

specific average editing behavior via article fixed effects and a general growth trend across all articles (and article length), the specific timing of when an article experiences changes in length can be treated as exogenous, particularly in the Roman Empire category, for which specific events outside of Wikipedia are unlikely to trigger interest and therefore edits. This type of variation substitutes for the experimental variation in article length outlined above and allows us to recover a causal effect.

To illustrate the source of relevant variation more concretely, consider two articles that have a similar average editing frequency due to their inherent popularity. However, for idiosyncratic reasons, one article is edited more heavily early in its lifespan and therefore grows in length earlier.[15] The effect of such a difference in article length on the number of weekly users identifies our coefficient of interest. Put differently, after controlling for differences in average edit intensity via article fixed effects, we treat longer articles in any given week as a valid counterfactual for articles of shorter length in the same week. Moreover, our context contains ample variation in article length even after we control for article and week fixed effects due to large variation in edit distance. The first and last row of Table (1) illustrate this point: the distribution for the edit-distance variable is highly skewed with a number of very large edits in the right tail of the distribution. The typical pattern of article-length growth therefore looks like the patterns reported in Figure (1) for two exemplary articles from the Roman Empire category: a smooth growth path with the exception of occasional large jumps in length. These discrete changes provide a major source of variation, and in a later robustness check, we focus specifically on those changes.

More formally, we argue that, after controlling for article fixed effects and a general growth trend, we can treat article length as uncorrelated with the regression error. The key identifying assumption is therefore that any factor that might correlate with both article length and the amount of editing activity - such as the popularity of the topic - does not vary differentially over time across articles.[16]

3.1 Estimation Results

Based on the specification presented above, the first column of Table (3) reports the coefficient on article length, which is equal to 0.204 and highly significant. In other words, about 50,000 additional characters (700 sentences) of article length are associated with one more active user per week. To get a sense of the magnitude of the effect, note the median article in 2009 is about 4,200 characters long. The article will therefore be edited by about 0.08 additional users *per week* compared to when it first appeared. The median article that was created in 2002, the first year in our data, was 15,100 characters longer in 2009. This length change leads to an additional 0.30 users each week because of the estimated cumulative growth effect.

[15] In terms of the regression equation above, one can think of this mechanism as capturing article-length differences that are caused by differences in past realization of the error term ε_{jt} rather than any systematic factors such as article popularity.

[16] In the online appendix, we formulate a simple theoretical model of editing behavior. Within the framework of the model, we formalize the necessary identification assumptions for the two-way fixed-effect specification to yield a causal estimate.

Given an average of 0.22 weekly users and a standard deviation of 0.85 in 2009, this effect is large in magnitude. We also note that because of the skewed length distribution, which we document in the final two columns of Table (2), the effect magnitude for the mean (rather than median) article in terms of length is even larger.

Second, we use the cumulative weekly edit distance as the dependent variable instead of the number of users. For this specification, we find a significant coefficient of 245, which can be interpreted as 10,000 characters of article length (about 140 sentences), leading to about 3.5 sentences of additional weekly editing activity. For the median 2002 vintage article in 2009, this increase entails an additional 370 characters, or 5 sentences, being contributed each week. This effect is large relative to the mean weekly edit distance, which is equal to 140 characters, yet the effect might seem small relative to the large standard deviation of the edit-distance variable, which is equal to 8,500 characters. However, the distribution of weekly total edit distance is highly skewed. Therefore, whether its standard deviation is a good benchmark for the effect size is unclear. For this reason and to test whether large outlier values drive our results, we rerun the regression using a version of the edit-distance variable that caps individual edits at 10,000 characters (roughly the 98th percentile of the edit-distance distribution) before aggregating them at the weekly level. When we switch to the capped edit distance as the dependent variable, we obtain a positive and significant coefficient, but of smaller magnitude than for our baseline case. We find that 10,000 characters of additional article length lead to 91 characters of additional edits rather than 245. Note, however, that in terms of standard deviations of the underlying variable (reported in the first row of Table (3)), the effect is actually stronger for the capped edit-distance measure. Note also that the large edits are legitimate data points, and in terms of effect size, one should not exclude them, because those edits have a strong impact on the respective article. The regressions based on the capped measure simply provide evidence that large edits are not the main driver of the results.

For the remainder of the paper, we will use the number of weekly users as our main measure of editing activity. Edit distance is arguably the most direct measure of the extent of change on an article; however, it has high variance due to the existence of very heavy edits in the right tail of its distribution. We therefore prefer to work with the number of users as the main dependent variable, which is much less affected by outliers. Furthermore, the growth patterns presented in Table (2) show that average edit distance per user is fairly stable over time, and an increase in the number of users drives growth in editing activity. Later, we also test explicitly whether increases in article length lead to relatively longer or shorter edits, and find they do not. We are therefore able to focus on the number of users as our main measure of editing activity, without missing other important dimensions of the growth process.

3.2 Robustness Checks: Article-specific Growth Trends

The main threat to identification in our context is the possible presence of article-specific time trends in editing activity. The identifying assumption underpinning our analysis so far has been that after controlling for article and week fixed effects, we can attribute any systematic

differences in growth trajectories to differences in article length. However, as activity on Wikipedia is growing, some articles might be benefitting more from the increase in the user pool, for instance, if new users join and disproportionately start editing popular articles that already have a high level of editing activity. In this case, article fixed effects are not able to fully capture the differences in editing activity due to popularity differences.

As we show in detail below, article-specific time trends do not in fact play an important role in our context. We run a set of three different robustness tests to control for such article-level time trends and do not find a significant change in our coefficient of interest on article length. First, we re-estimate our baseline regression using the most straightforward and "brute-force" way to control for article-specific growth trends: on top of article fixed effects, we allow for a linear and square effect of article age on editing activity,[17]; that is, we estimate

$$UserNum_{jt} = \beta ArticleLength_{jt} + \theta_j + \psi_t + \gamma_j * Age_{jt} + \delta_j * Age_{jt}^2 + \nu_{jt}.$$

Note that this specification includes a set of article and week fixed effects as well as two additional coefficients *per article* that capture article-specific trends. Given the shape of the aggregate and article-level growth patterns, which are characterized by an initial increase and later slowdown, we believe the linear and quadratic article-specific age controls do a good job of capturing growth dynamics at the article level.[18] We report results from this regression in column (4) of Table (3). The coefficient on article length in this specification is equal to 0.130 (0.041) and not significantly different from our baseline regression.

This robustness check relies (even more than our baseline regression) on the presence of discrete and large jumps in article length due to occasional large edits. We document the presence of such edits in Table (1), which shows a highly skewed distribution of edit distances with a long right tail, and we discussed their significance in the context of identification in the previous section. When including article-specific time trends we control for the "smooth part" of the article-level growth trend in editing activity, but identify our coefficient of interest from large jumps in article length. As before, we argue that (even if articles had their own time trends due to a difference in popularity between topics) the specific timing of large edits is driven by the random arrival of knowledgable users that can add a large amount of content, and can be treated as exogenous.

To take advantage of the variation induced by large edits even more directly, we run a second test for which we select weeks with changes in article length of more than 1,000 characters. For each instance of a large change in length, we compute the number of users in the week preceding the change as well as the week following the length increase. We then regress the change in the number of users on the change in article length. Formally, we run a differenced version our original regression:

[17] Note that article-specific age controls are statistically identical to including interactions of a time trend with article dummies.

[18] We also estimated the specification using cubic trends as well as only linear ones. For the linear case, we find a coefficient (standard error) of 0.130 (0.037). When using cubic trends, we obtain 0.077 (0.022).

$$UserNum_{jt+1} - UserNum_{jt-1} \quad = \quad \beta(ArticleLength_{jt+1} - ArticleLength_{jt-1}) + (\psi_{t+1} - \psi_{t-1}) + \mu_{jt}$$

Note that we omit the week that contains the large edit itself in order to compare time periods that are strictly before or after the jump in article length. When estimating the regression, we treat $(\psi_{t+1} - \psi_{t-1})$, which captures the aggregate growth trend as part of the error term. Similar to a regression discontinuity design, we rely on the fact that other than the article-length increase, nothing changed that could have an effect on editing activity. We find a positive and significant coefficient that we report in column (5) of Table (3). In terms of magnitude, the estimated of coefficient of 0.219 is similar to the baseline coefficient of 0.204.[19]

As a third and final test for dealing with article-level growth trends, we run a placebo test. The idea for this test is the following: if some articles experience more editing activity and grow faster because of their inherent popularity, we should see a correlation between current and *all* past editing activity. Instead, if we are correctly identifying a cumulative growth effect, current activity should only respond to past editing activity that is still embodied in the current content of the article. Put differently, content that once existed in the article but was later deleted should not trigger current users to contribute. We should therefore only see a response of editing behavior to *surviving* edits rather than all past editing activity. The fact that cumulative past edits, captured by the cumulative edit-distance measure, and article length differ substantially for many articles because of deletion and replacement of content allows us to run a regression in which we include both variables.[20] We report the results from such a regression in the final column of Table (3). We find that after controlling for article length, the cumulative edit distance has no additional explanatory power. The estimate is not only statistically insignificant, but the magnitude is also very small (note the different units used for article length and edit distance). Furthermore, the coefficient estimate on article length of 0.184 is similar to the baseline specification. [21] The results from this regression provide evidence that editing activity is correlated with current content stock but not the amount of all past contributions including non-surviving edits, lending further support to the notion that we are correctly identifying a causal effect of article length on editing activity.

In summary, our findings are robust to including flexible article-specific age trends, to analyzing changes in editing behavior around dramatic length changes, and to the inclusion of cumulative edit distance on top of article length in the regression. Taken together, these tests suggest article-specific growth trends in editing activity are not likely to be a confounding

[19] To further probe the robustness of this result, we also analyze the change in the number of users for a longer time window around the length change. Specifically, we compute the change in the number of weekly users between $t-1$ and $t+\tau$ for values of τ between 1 (the result reported above) and 5 and find the estimated coefficients are are not significantly different from each other. Also, all five coefficient estimates are significantly different from zero.

[20] Note that if an article experiences no deletion or replacement of content, the two measures would be identical. For most articles, the metrics diverge at some point in their lifetime.

[21] We also run a further set of robustness checks in which we cap the cumulative edit distance at a certain percentile of its distribution in order to ensure outlier values are not the main driver of the null results. When capping the cumulative edit measure at the 90th or 80th percentile of its distribution, we get quantitatively similar results.

factor in our baseline regression.

3.3 Robustness Checks: Information Shocks

A secondary threat to a causal interpretation lies in the presence of information shocks that are persistent over time. New information could become available to users from outside of Wikipedia at a particular point in time, but all users might not respond to the information immediately. Instead, different users might incorporate the new information into the article over an extended period of time. This response would lead to a burst of editing activity over a period of time, and we might incorrectly infer that the later edits within that time window are happening in response to the earlier ones. To avoid this issue, we explicitly chose a set of articles that we presume new information did not particularly affect. Most likely, the stock of knowledge regarding historic topics such as the Roman Empire among the user pool changes little over time. Therefore, we think information shocks are less likely to be present for the set of articles considered, yet we also test whether our estimates are robust to an IV strategy in which we instrument the current length of the article with lagged article length. The idea is to use article length from a time period far enough away that the effect of any information shock on lagged article length will have no longer affect current editing. Apart from instrumenting article length, we run the same specification with article and week fixed effects as in our baseline case. Controlling for these fixed effects is important because the lagged-length instrument does not deal with across-article differences in editing behavior due to difference in the popularity of the topic. Instead, the instrument deals with a separate issue, which is correlation in editing activity within an article over time due to information shocks. We report results using various lags in Table (B1) in the appendix and find the results are robust.

4 Cumulative Growth and Article Quality

So far, we have shown that increases in article length lead to more editing activity. However, we have not yet characterized the nature of this additional activity. In this section, we analyze one particulary important dimension: whether the additional activity translates into improvements in article quality. To quantify article quality at a given point in time, we employ two different measures. First, we compute the number of references an article contains (relative to article length) at any point in time. Second, we use Wikipedia's internal quality categorization scheme,[22] which assigns articles to a set of seven distinct categories ranging from a low-quality "Stub" to a high-quality "Featured Article".[23] One downside of the second metric is the fact that we only observe quality for a subset of article/week combinations. In particular, we observe no quality information prior to about mid-2004. In later years, we

[22] See http://en.wikipedia.org/wiki/Template:Grading_scheme

[23] The seven categories are Stub, Start, C, B, Good Article, A, and Featured Article. The quality information is included either on an article page itself or (more often) on the talk page belonging to the article. We compute our quality measure by scraping this information from both sources. In the appendix, we provide more details on how the quality information is extracted.

observe quality information for roughly 60% of article-weeks. The number of references is available for the entire sample. We also note that the two measures are closely related. A prominent reason for an article being categorized as low quality is a lack of sufficient references.[24] In the appendix, we provide more details on how the quality measures were extracted from the data.

We think of an article's quality as being determined by all the editoral input on the article up to the point at which we measure quality, and therefore regress quality on the *cumulative* number of weekly users up to this point. In other words, we use the number-of-weekly-users variable used in our baseline regression, but for each week, we compute the cumulative value of the variable up to that point in time. Similar to our baseline regression, we control for article and week fixed effects and cluster standard errors at the article level. When regressing the number of references per 10,000 characters of article length on the cumulative user count, we find a positive and significant effect. The results from this regression are reported in column (1) of Table (4). This regression shows that editing activity increases article quality by increasing the ratio of references relative to text length. However, we are more specifically interested in whether editing activity *that is caused by the current content level* also improves quality. We therefore want to isolate the part of the cumulative editing activity that is caused by article-length variation. Based on the relationship estimated in the previous section, we know the cumulative number of users depends on the entire evolution of article length up to the particular point in time. More specifically, article length in a given week affects the number of users in that same week. Therefore, when we analyze the cumulative number of users, the logical analogue is to consider the part of the *cumulative* user count that *cumulative* article length predicts. Based on this reasoning, we implement an IV regression of quality on the cumulative number of users, where the latter is instrumented with cumulative article length. Unsurprisingly, and in line with our baseline regression, cumulative article length is highly predictive of the cumulative number of users with an F-stat on the excluded instrument of 23.19. The second-stage coefficient is equal to 6.163 and statistically significant, which shows that editing activity triggered by the current content stock does lead to improvements in article quality. Furthermore, the comparison with the OLS estimate shows us that the rate of quality improvement entailed by the triggered edits is similar to the effect of *any* editing activity, if not slightly larger.

As a second measure of quality, we use the information from Wikipedia's internal quality categorization described above. This measure is available for fewer article/week observations, and the frequency of each category occurring is uneven. For roughly 75% of article-weeks for which we observe the quality category, the assigned category is a stub. We therefore use a dummy for whether the article is a stub as our second measure of article quality. We regress this dummy variable on the cumulative number of users using both an OLS regression, as well as a version in which we instrument the cumulative number of users with cumulative article length. Results from the two regressions are reported in columns (3) and (4) of Table

[24]For example, Wikipedia's quality guidelines cite as a condition for a B-class article that "[t]he article is suitably referenced [...]. It has reliable sources, and any important or controversial material which is likely to be challenged is cited." (see http://en.wikipedia.org/wiki/Template:Grading_scheme)

(4). We find an estimate of -0.245 for the OLS case that is not statistically significant. From the IV regression, we obtain a significant estimate of -0.720, indicating that triggered edits increase the probability of an article ceasing to be a stub. Similar to the regressions based on the number of references, the results here indicate the triggered edits do improve quality and possibly have a stronger effect on quality than the average edit. Relative to the analysis based on references, the difference between the OLS and IV coefficient is even more pronounced when we use the stub dummy.

In summary, we find evidence that edits triggered by current content, captured by the article's length, do lead to quality improvements. Moreover, we find weak evidence that the triggered edits have a stronger effect on quality relative to the average edit in our sample.

5 Other Dimensions of Editing Behavior

To further explore the nature of edits triggered by the current content stock, we analyze both edit intensity, that is, the amount of change induced by the edit, as well as the type of edits being made. To quantify the latter, we look at the extent of deletion/replacement of content versus addition of new content. This approach allows us to assess whether longer articles experience more conflict and controversy between users, and hence some of the triggered edits do not add content but rather remove or replace content provided by previous edits. We also analyze the extent to which the triggered edits are later reverted, that is, whether article length increases lead to edits that have a lasting impact on the article or only editing activity that does not ultimately survive. We report results regarding these different dimensions of editing behavior in Table (5).

We first test whether the length of edits changes as a function of article length. This test is particularly important for our purpose, because we focused on the number of users as our main measure of editing activity. Although we found that longer articles are edited by more users, users making shorter edits might counteract this effect. First, remember that in columns (2) and (3) of Table (3), we show that longer articles experience more editing activity as measured by the total weekly edit distance as well as capped edit distance. Second, to more explicitly relate edit distance with the number of users, we analyze whether *editing activity per user* reacts to changes in article length and find it does not. We use the same setup as our baseline regression, except that we are only able to use article-week pairs that contain at least one edit. For these weeks, we compute edit distance per user and regress it on article length. Doing so, we obtain a coefficient that is insignificant and small in magnitude compared to the mean and standard deviation of the edit-distance measure. We report the results in column (1) of the top panel of Table (5). To be sure the noisiness induced by outlier values is not the only reason for not finding an effect, we also rerun the regression using capped edit distance per user as the dependent variable in column (2). Again, we find no significant effect.

Next, we analyze whether edits that contain relatively more or fewer additions versus deletion of content characterize longer articles. For this purpose, we use our measure of content addition / deletion introduced earlier ($\Delta Length/EditDistance \in [-1, 1]$) as the dependent

variable in the regression. We find a negative and significant coefficient of -0.010, which implies that edits on longer articles are more likely to delete a larger portion of the previous content. However, the magnitude of the effect is small compared to the mean (standard deviation) of the variable, which is 0.460 (0.629). As a further point of reference, note that the metric falls by about 0.11 between 2002 and 2009, as shown in Table (2). This decrease is an order of magnitude larger than the -0.010 change induced by an increase of 10,000 characters in article length.

Finally, we also use the fraction of reverted edits as the dependent variable. We find a positive and significant effect of 0.011, which shows that edits on longer articles are more likely to be overturned by subsequent edits of other users. However, the magnitude is again quite small compared to the variable's mean (0.083) and standard deviation (0.248) as well as to the increase in the metric over time reported in Table (2).[25]

6 Quantification of the Cumulative Growth Effect

To assess the overall importance of the cumulative growth effect, we use the regression relationship between article length and the number of weekly users to simulate article growth trajectories. This approach helps us quantify the longer-term impact of the cumulative growth effect. Although the regression predictions allow us to assess the immediate effect of article length on editing activity, we cannot directly use them to assess the long-term effect. In the presence of cumulative growth effects, any length increase will lead to more editing, which entails a further length increase, which in turn increases editing activity, and so forth. In this way, changes early in an article's lifespan can influence the entire trajectory of the article, thus leading to path-dependent content growth. To implement a simulation of content growth trajectories, we rely on several pieces of data. Because of the absence of a structural model that gives us guidance on how to combine the different variables involved, the simulation has a back-of-the-envelope character.

We first implement the simulation for the case in which no cumulative growth occurs and editing activity increases only because of the general growth trend of the platform. This implementation serves as a benchmark against which we compare the typical trajectory when the cumulative effect is present. We initiate an article in the first week of our sample in late 2001 to trace out a growth trajectory over the entire sample period. In a given week, the expected number of users is given by article and week fixed effects. We fix the article fixed effect across all simulations such that the average editing intensity is close to the one in our full sample, and use the estimated time-period fixed effect for each of the weeks within our sample.[26] Given the expected number of users, we simulate the actual realization of the

[25] In the online appendix, we further explore changes in the types of edits being made. In particular, we show how the share of different types of users (new vs. returning users that previously edited the same article) as well as the type of edits the different types are making (in terms of length, addition/deletion of content, etc.) change with article length. However, the fact that correctly defining user types is difficult is a disadvantage of this additional analysis and we hence relegate it to the online appendix.

[26] We simplify the simulation at this point. Rather than simulating how many users edit in any given week, we simulate whether any edit happens in that week. In other words, we are treating the number of users as

variable in a given week. We then take a random draw from the empirical distribution of length changes for each edit[27] and compute the new article length at the beginning of the next week. If no edit occurs, article length remains unchanged. Starting from the first week, we simulate edits and length changes in each week and update article length accordingly. Doing so allows us to trace out the growth trajectory for the simulated article. We repeat the procedure for a large set of 10,000 simulations and calculate the average length across simulated articles at different points in time. The results from this simulation are reported in column (1) of Table (6).

The most relevant metric for our purposes is the comparison with the trajectories that include the cumulative growth effect. The simulations for this case are implemented in the same way as described above, but for the fact that the number of weekly users is determined by the two sets of fixed effect *plus* the cumulative growth term. Results are reported in the second column of Table (6) and reveal some interesting differences relative to the reference case without cumulative growth in column (1). We find that articles under both scenarios are initially similar in length, but diverge more and more over time. After an eight-year period, articles that did not benefit from the cumulative effect are on average 45% shorter in length. To illustrate the mechanics of this divergence more concretely, we plot the length evolution as well as the expected number of weekly users (as predicted by the regression relationship) for one specific article from our simulation in Figure (2) with and without the cumulative growth effect. Taken together, the evolution of article length and the number of users illustrate the self-reinforcing nature of the mechanism. Initially, both article length and the number of users are quite similar in both scenarios. However, in later years, as the article grows in length, the number of users increases more in the cumulative growth scenario, which leads to faster growth in length, which further increases the number of users. At the end of the sample period, after eight years, article length for this particular article is about 80% longer and the number of weekly users is four times larger in the presence of the cumulative growth effect. The differences are similar in magnitude to the mean differences in both variables across all simulated articles.

7 Conclusion

In this article, we showed that on Wikipedia, the current body of content tends to motivate further contributions. We quantify the importance of such a *cumulative growth effect* as a driver of content growth and find the magnitude of the effect is economically important: it accounts for almost half of the content growth between 2002 and 2010 for a typical article in the Roman Empire category. Importantly, we also find the additional activity induced by the

a binary outcome and ignore that the variable (although very rarely) takes on values larger than one. This approach turns the simulation into a simple Bernoulli draw for which the success probability is given by the probability of an edit occurring.

[27] We take draws from the unconditional distribution of length changes across all edits. Ideally, one would want to condition on the particular week of the sample in which the edit occurs. As reported earlier in the paper, we do, however, find little evidence that the amount of editing activity per edit changes over time. Therefore, the unconditional distribution should constitute a reasonable approximation.

cumulative growth effect leads to an increase in content quality. These findings are robust to a whole battery of checks, suggesting we have identified a causal mechanism.

Wikipedia is one of the most prominent examples of a new form of content production, and numerous public and private content production environments attempt to mimic its success. These examples include internal wikis of private companies as well as commercial (Wikia) and non-profit (Wikiversity) projects, many of which are based on the same or similar technological platform and content production process. Our findings regarding the quantitative importance of the cumulative growth effect are therefore likely to be relevant for the design of other open content production platforms as well. Importantly, the platform provider has some degree of control over the source of the motivational mechanism we identify, that is, the current stock of content. This leads to important managerial implications regarding how to leverage the cumulative growth effect. Specifically, any action that increases the current content stock will trigger further contributions. A content production platform can thus benefit from pre-populating articles with content to trigger further edits via the effect. Such triggers may, for instance, involve transferring an existing stock of content to an open platform, or incentivizing users to provide initial content via monetary rewards.

References

Aaltonen, A., and J. Kallinikos (2013): "Coordination and Learning in Wikipedia: Revisiting the Dynamics of Exploitation and Exploration (Research in the Sociology of Organizations)," in *Managing "Human Resources" by Exploiting and Exploring Peoples Potentials*, ed. by M. Holmqvist, and A. Spicer, pp. 161–192. Emerald.

Almeida, R., B. Mozafar, and J. Cho (2007): "On the Evolution of Wikipedia," in *Proceedings of the ICWSM*, Boulder, Co.

Arazy, O., O. Nov, R. Patterson, and L. Yeo (2011): "Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict," *Journal of Management Information Systems*, 27, 71–98.

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

Gorbatai, A. (2011): "Aligning Collective Production with Social Needs: Evidence from Wikipedia," unpublished manuscript.

Greenstein, S., and F. Zhu (2012a): "Collective Intelligence and Neutral Point of View: The Case of Wikipedia," *NBER working paper 18167*.

——— (2012b): "Is Wikipedia biased?," *American Economic Review, Papers and Proceedings*, 102(3), 343–348.

Halfaker, A., R. S. Geiger, J. Morgan, and J. Riedl (2013): "The Rise and Decline of an Open Collaboration System: How Wikipedia's reaction to sudden popularity is causing its decline," *American Behavioral Scientist*, 57(5), 664–688.

Halfaker, A., A. Kittur, R. Kraut, and J. Riedl (2009): "A Jury of Your Peers: Quality, Experience and Ownership in Wikipedia," in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, Orlando, Florida.

Hansen, S., N. Berente, and K. Lyytinen (2009): "Wikipedia, Critical Social Theory, and the Possibility of Rational Discourse," *The Information Society*, 25(1), 38–59.

Kittur, A., and R. E. Kraut (2008): "Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination," in *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pp. 37–46, New York.

Kittur, A., B. Suh, B. A. Pendleton, and E. H. Chi (2007): "He Says, She Says: Conflict and Coordination in Wikipedia," in *Proceedings of the 2007 SIGCHI Conference on Human Factors in Computing Systems*, pp. 453–462, New York.

Kummer, M. (2013): "Spillovers in Networks of User Generated Content Evidence from 23 Natural Experiments on Wikipedia," ZEW Discussion Paper No. 13-098.

Levenshtein, V. I. (1966): "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Cybernetics and Control Theory*, 10(8), 707–710.

Myers, E. W. (1986): "An O(ND) Difference Algorithm and Its Variations," *Algorithmica*, 1(2), 251–266.

Nagaraj, A. (2013): "Does Copyright Affect Creative Reuse? Evidence from the Digitization of Baseball Digest," unpublished manuscript.

Navarro, G. (2001): "A Guided Tour to Approximate String Matching," *ACM Computing Surveys*, 33(1), 31–88.

Nov, O. (2007): "What Motivates Wikipedians?," *Communications of the ACM*, 50(11), 60–64.

Olivera, F., P. S. Goodman, and S. S. Tan (2008): "Contribution Behaviors in Distributed Environments," *MIS Quarterly*, 32(1), 23–42.

Piskorski, M. J., and A. Gorbatai (2013): "Testing Coleman's Social-Norm Enforcement Mechanism: Evidence from Wikipedia," Harvard Business School Working Paper.

Ransbotham, S., and G. C. Kane (2011): "Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia," *MISQuarterly*, 35(3), 613–627.

Spiliopoulos, K., and S. Sofianopoulou (2007): "Calculating Distances for Dissimilar Strings: The Shortest Path Formulation Revisited," *European Journal of Operational Research*, 177, 525–539.

Suh, B., G. Convertino, E. H. Chi, and P. Pirolli (2009): "The Singularity is not Near: Slowing Growth of Wikipedia," in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, Orlando, Florida.

Vigas, F. B., M. Wattenberg, and K. Dave (2004): "Studying Cooperation and Conflict between Authors with History Flow Visualizations," *CHI Letters*, 6(1), 575–582.

Voss, J. (2005): "Measuring Wikipedia," in *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*.

Zhang, M., and F. Zhu (2011): "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia," *American Economic Review*, 101(4), 1601–1615.

EDIT LEVEL		Fraction	Mean	S.D.	Median	75th	90th	95th	99th
Edit-Distance			632	13409	37	146	752	1863	8872
Adding / Deletion Measure	Addition	42.86							
	Deletion	7.81							
	Mix	49.33	0.19	0.58	0.17	0.73	0.94	0.98	0.99
Reverted Edits		14.22							

WEEK LEVEL	Weeks with no Edit	Mean	S.D.	Median	75th	90th	95th	99th
Number of Edits	85.88	0.249	1.065	0	0	1	1	4
Number of Users	85.88	0.224	0.853	0	0	1	1	3
Edit-Distance	85.88	157	8500	0	0	23	118	2104

Table 1: **Descriptive Statistics.** The top panel reports descriptive statistics on measures of edit length as well as type of edit across all 62,925 edits in the sample. The bottom panel reports measures of editing activity at the article-week level (including article-week pairs without any edit). The sample contains 252,427 article-week observations. Edit distance is defined as the number of characters that are added, deleted, or replaced by the edit. The addition/deletion measure varies from -1 (pure deletion) to 1 (pure addition) with the intermediate values representing edits that involve both addition and deletion of content. Reverted edits are defined as edits that are overturned; that is, a prior version of the article is reinstated.

Year	Number of Pages Created	Number of Users	Number of Edits	Cumulative Edit Distance (Unit: Characters)	Add/ Delete Metric	Fraction of Reverted Edits	Average Article Length	Median Article Length
2002	84	180	550	369,500	0.53	0.01	2,373	1,385
2003	71	413	969	521,304	0.60	0.02	3,601	2,057
2004	120	1,239	2,681	1,167,481	0.53	0.04	3,806	2,033
2005	326	3,185	7,295	4,479,649	0.47	0.06	3,726	1,727
2006	211	6,033	12,397	9,367,665	0.45	0.14	4,586	2,294
2007	184	7,019	13,556	8,148,506	0.44	0.20	5,825	3,111
2008	195	6,067	12,555	8,149,164	0.41	0.17	7,035	3,511
2009	119	5,667	12,922	7,533,972	0.42	0.14	8,522	4,194

Table 2: **Content Evolution at the Category Level.** The table reports metrics of editing activity aggregated across all articles in the Roman Empire category on a yearly basis.

	Baseline Specification	Alternative Measures of Editing Activity		Robustness Checks		
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Number of Users	Edit-Distance	Capped Edit-Dist.	Number of Users	Δ Number of Users	Number of Users
Mean of the DV	0.218	140	81	0.218	0.218	0.218
S.D. of the DV	0.852	8567	979	0.852	0.852	0.852
Sample	Full Sample	Full Sample	Full Sample	Full Sample	Large Edits Only	Full Sample
Article Length (Unit: 10,000 Characters)	0.204*** (0.054)	245.3*** (91.9)	90.7** (39.2)	0.130*** (0.041)		0.184*** (0.048)
Δ Article Length (Unit: 10,000 Characters)					0.219*** (0.057)	
Cumulative Edit-Distance (Unit: 100,000 characters)						0.008 (0.005)
Article FEs	Yes	Yes	Yes	Yes	No	Yes
Week FEs	Yes	Yes	Yes	Yes	No	Yes
Article Age-Trends	No	No	No	Yes	No	No
Observations	247,002	247,002	247,002	247,002	3,329	247,002
Articles	1,191	1,191	1,191	1,191	1,191	1,191
Weeks	432	432	432	432	334	432

Table 3: **The Effect of Article Length on Editing Activity.** The unit of observation is an article-week pair. Standard errors are clustered at the article level. Column (5) uses only the subsample of large (more than 1,000 characters) edits. Column (4) includes a linear and squared age trend for each article.

	(1)	(2)	(3)	(4)
Dependent Variable	References Per 10,000 Characters	References Per 10,000 Characters	Stub Dummy	Stub Dummy
Estimation Method	OLS	IV	OLS	IV
Article Length (Unit: 10,000 Characters)	5.133***	6.163***	-0.245	-0.720**
	(1.392)	(1.804)	(0.162)	(0.341)
Article FEs	Yes	Yes	Yes	Yes
Week FEs	Yes	Yes	Yes	Yes
Excluded Instrument F-stat		23.19		20.31
Observations	247,002	247,002	120,555	120,555
Articles	1,191	1,191	891	891
Weeks	432	432	293	293

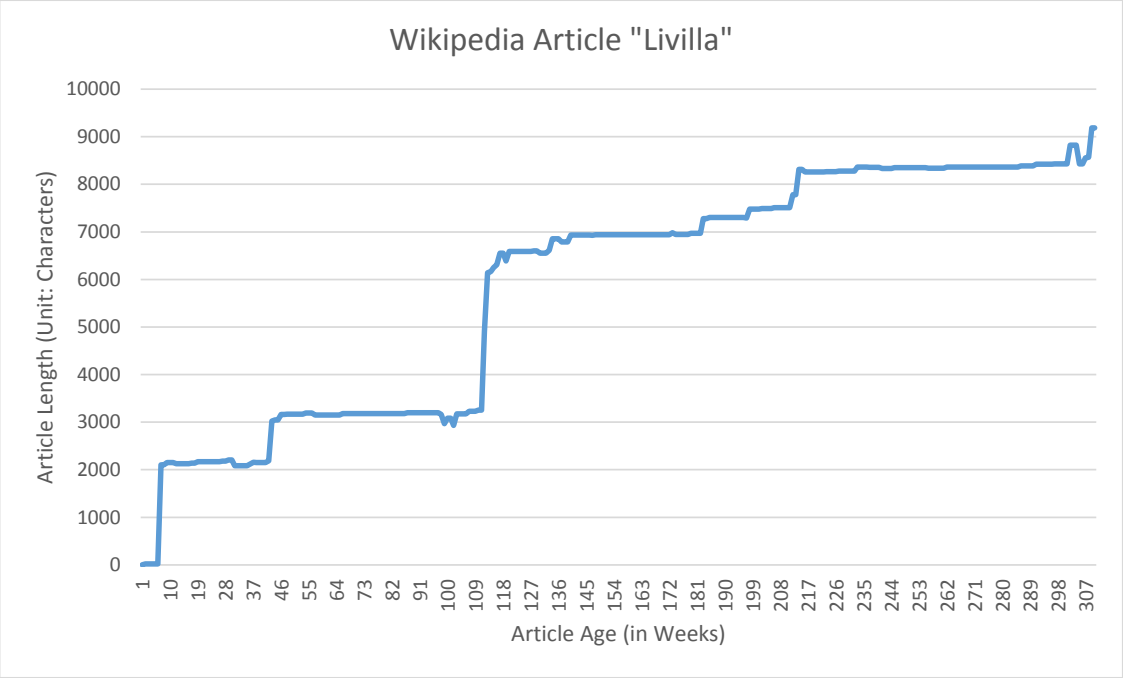
Table 4: **The Effect of Article Length on Content Quality.** The unit of observation is a week-article pair. Standard errors are clustered at the article level. The stub dummy is only available for a subset of observations.

	(1)	(2)	(3)	(4)
Dependent Variable	Edit-Distance Per User	Capped Edit-Dist. Per User	Addition/ Deletion Metric	Fraction of Reverted Edits
Mean of the DV	413	317	0.460	0.083
S.D. of the DV	2812	1213	0.629	0.248
Article Length (Unit: 10,000 Characters)	-55.957	-73.383	-0.010**	0.011**
	(134.237)	(66.667)	(0.005)	(0.005)
Article FEs	Yes	Yes	Yes	Yes
Week FEs	Yes	Yes	Yes	Yes
Observations	33,953	33,953	33,953	33,953
Articles	1,186	1,186	1,186	1,186
Weeks	414	414	414	414

Table 5: **Change in Editing Behavior as a Function of Article Length.** The unit of observation is a week-article pair. Standard errors are clustered at the article level. The dependent variable is defined only for article-week combinations with at least one edit in all regressions. The number of observations is accordingly smaller than in our baseline regression.

		Trajectory WITHOUT Cumulative Growth	Trajectory WITH Cumulative Growth
Article Length (Units: 10,000 Characters)	After 1 Year	0.043	0.044
	After 2 Years	0.079	0.092
	After 3 Years	0.376	0.435
	After 4 Years	0.906	1.117
	After 5 Years	1.434	1.899
	...	...	...
	After 8 Years	2.528	4.366

Table 6: **Article Length Trajectories with and without the Cumulative Growth Effect.** The table reports mean article length at different points in time for a set of simulated article trajectories based on the regression estimates.



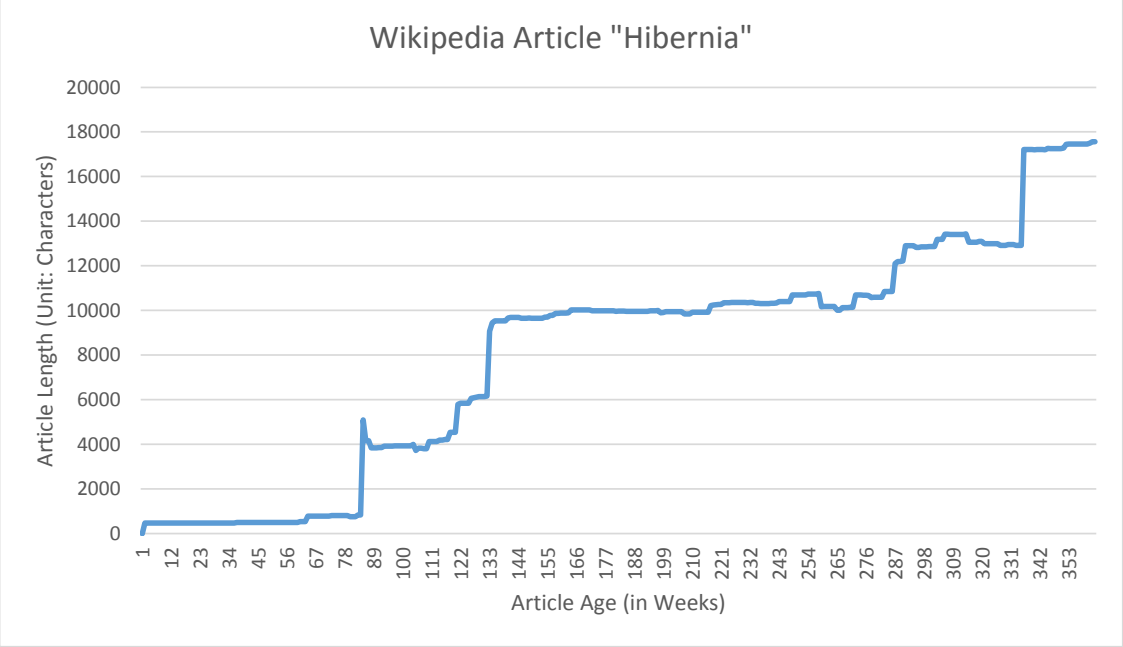


Figure 1: **Evolution of Article Length.** The graph plots the evolution of article length for two articles over their respective lifespan.

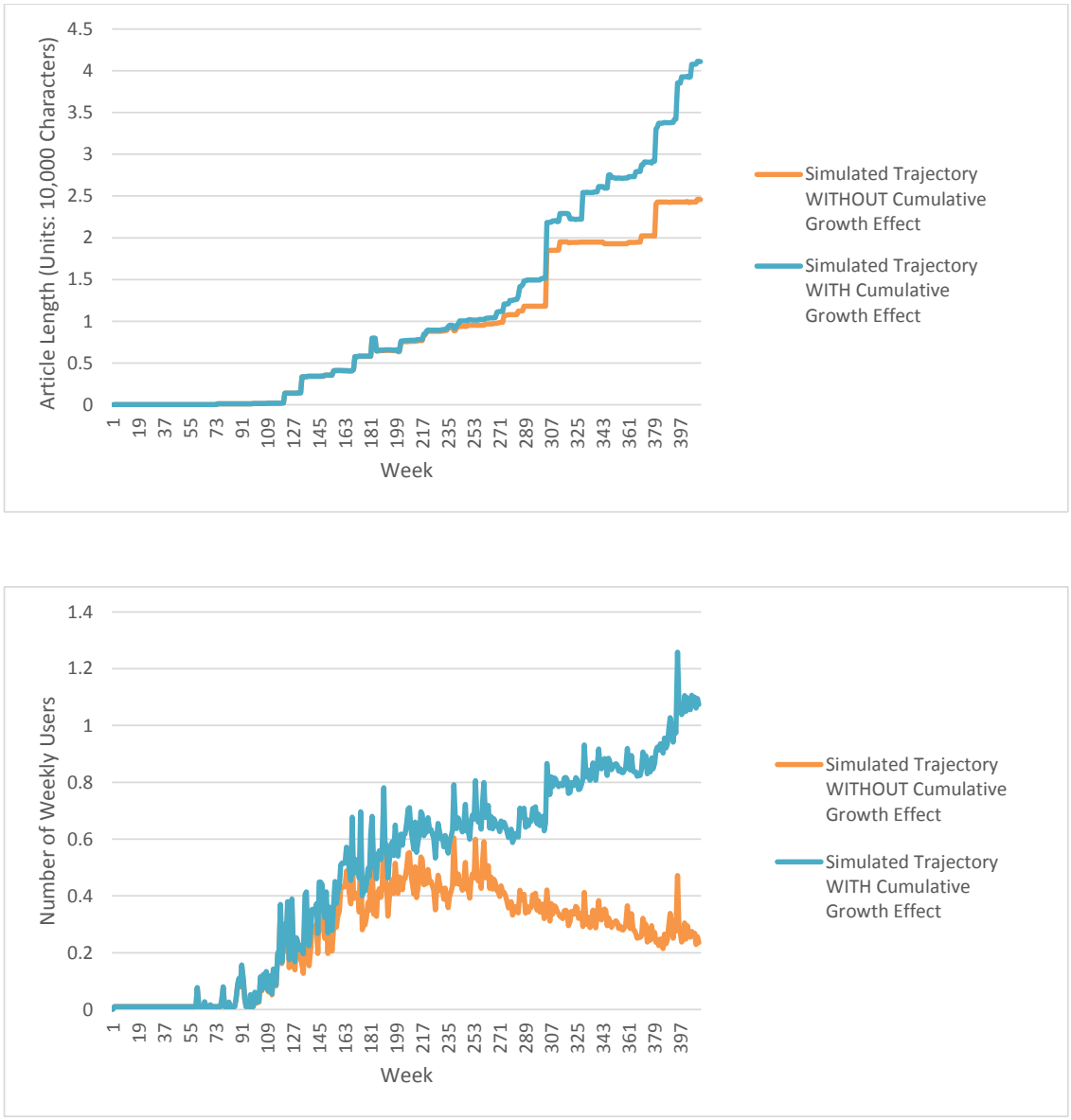


Figure 2: **Simulated Trajectories of Article Length and the Number of Weekly Users.** The top graph plots article length for one simluated article in the absence of and including the cumulative growth effect. The bottom panel reports the expected number of users in each week predicted by the regression estimates.

A Appendix: Data Construction

A.1 Article Selection

To define articles that belong into the Roman Empire category, we use the "enwiki-20100130-categorylinks.sql" file. This file represents the category structure on Wikipedia as SQL statements at the end of the sample period in January 2010. More specifically, the SQL dump contains a separate record for every link pointing to a category page within Wikipedia, from which we filtered records pointing to "Roman Empire." This selection yields a set of 1,571 articles, which we then manually reviewed to eliminate the ones that only tangentially pertain to the historical Roman Empire. Note that identifying a set of related articles is not of major importance to our analysis; we simply need a set of articles for which we can assume the stock of human knowledge to be relatively stable. Through this process, we identified 168 articles that were incorrectly categorized. The main goal of our selection was to eliminate articles that involve more recent events that do not pertain to the Roman Empire in a more narrow sense. The reason for such elimination was to end up with a set of articles that contained purely historical content and therefore would not be subject to major changes in the knowledge regarding the topics covered. We therefore maintain articles on historical figures, for instance, that one might primarily assign to a different category, for example, religious figures such as Saint Peter. Also, we keep articles both on Antique Rome as well as the Holy Roman Empire. We eliminate all articles on video games, movies, and books (e.g., the movie "Monty Python's Life of Brian" appears in the Roman Empire category and receives a substantial amount of edits). Furthermore, our original list contains many geographic locations (cities, counties, etc.). We maintain all denominations that have ceased to exist, but drop all locations whose name is still in use. For example, we drop the article on Bremen (the city in Germany) but keep Archbishopric of Bremen (a region that existed during the Holy Roman Empire). Finally, we also drop pages that are re-directs or disambiguation pages. These types of Wikipedia pages contain little content and their primary purpose is to provide a link to another (related) article. We remove pages that are re-directs / disambiguations for their entire lifetime and articles that turn into (and stay) re-directs / disambiguations at some point during our sample period. We maintain articles that are temporarily turned into re-directs / disambiguations (usually for a short period of time). This process leads us to eliminate a further 93 articles and leave us with a final set of 1,310 articles.

A.2 Edit Distance Calculation

We measure the difference between two consecutive versions of article content using an edit-distance metric. Measuring edit distance is a general approach for string-matching problems, which has applications in fields such as computational biology, signal processing, and information retrieval (Myers (1986), Navarro (2001), Spiliopoulos and Sofianopoulou (2007)). For instance, in computing, edit-distance calculations are used to correct spelling mistakes, patch (update) files, and cleanse and de-duplicate database entries. The edit-distance metric can be understood as the cost of transforming a string to another string or a measure of dissimilarity

between strings.

A number of edit-distance algorithms are optimized for different data and conditions. We use a simple edit-distance calculation that is defined as "the minimal number of insertions, deletions and substitutions to make two strings equal" with the cost of each operation being equal to 1 (see Navarro (2001) and Levenshtein (1966)). This calculation is also known as the Levenshtein distance. The value of this metric is zero if and only if the compared strings are equal and otherwise strictly positive. At the maximum, the edit distance is equal to the number of characters in the longer string.

We implement the calculation of edit distances using Python code from the google-diff-match-patch (see https://code.google.com/p/google-diff-match-patch/) software package that provides a set of mature and well-tested tools. The package is based on an algorithm presented in Myers (1986). The initial transformation of the raw XML records to a tabulated data set includes 87,346 edit-distance calculations (for all edits on Roman Empire articles including edits made by bots), which took about 15 hours to complete using a relatively modest multiprocessor environment.

A.3 Bot Activity

We have to deal with the fact that a certain amount of activity on Wikipedia comes from automatic "bots" rather than human contributors. These bots are user accounts controlled by software programs that are primarily used to execute tasks that can be automatized, such as correcting spelling and punctuation mistakes. Bots are also used to detect vandalism (attempts to intentionally destroy content) and to revert the vandalized article to its earlier state. Bot activity needs to be declared and the Wikipedia community might block users that use their account for undeclared bot activity. Bot activity can therefore usually be identified from user accounts. We use both the Wikipedia bot user group, which contains a list of bot user account IDs, and manually investigate contributors with very large amounts of edits to check whether their user account declares them as a bot. Although we might be missing some undeclared bot activity, we do believe we are able to capture the majority of bot activity in our data. Quantitatively, we find that 11% of edits on articles of the "Roman Empire" are done by bots.

A.4 Article Quality

We parse variables on article quality and references from Wikipedia article revisions using the Python regular expression module to identify special codes in article content. For the quality variable, we use Wikipedia's own quality grading as our metric. We look for quality tags such as "featured article" or "class = FA". The search pattern includes valid alternative spellings. Because a quality tag may appear both in the actual article page and its talk page, we combine the data from both sources. For the reference-count variable, we identify references by looking for "<REF> ... </REF>" entities in the article's markup.

B Appendix: Tables

	(1)	(2)	(3)	(4)
Estimation Method	IV 1st Stage	IV 2nd Stage	IV 1st Stage	IV 2nd Stage
Dependent Variable	Article Length	# Users	Article Length	# Users
Article Length		0.200***		0.209***
		(0.060)		(0.067)
Lagged Article Length (3 Months Lag)	0.836***			
	(0.052)			
Lagged Article Length (6 Months Lag)			0.684***	
			(0.105)	
Excluded Instruments F-stat	262.89		41.98	
Article FEs	Yes	Yes	Yes	Yes
Week FEs	Yes	Yes	Yes	Yes
Observations	231,519	231,519	216,036	216,036
Articles	1,191	1,191	1,191	1,191
Weeks	419	419	406	406

Table B1: **Robustness Check: Correlated Information Shocks** The unit of observation is an article-week pair. Standard errors are clustered at the article level. Lagged instruments are used in all IV specifications. The sample size is reduced relative to our baseline regression, because lagged values are not defined for a set observations in the beginning of each article's time series.