Cumulative Knowledge and Open Source Content Growth: The Case of Wikipedia

Aleksi Aaltonen*, Stephan Seiler†

December 22, 2013

Abstract

We analyze content growth on one of the largest open source platforms: Wikipedia. Using edit-level data over 8 years across a large number of Wikipedia pages, we find that content is still growing substantially even in later years. Less new pages are created over time, but at the page-level we see very little slow-down in activity. One key driver of growth is a positive spill-over effect of past edits on current activity: we find that longer pages experience significantly more editing activity while controlling for a host of confounding factors such as popularity of the topic and platform-level growth trends. The magnitude of the externality is economically important and growth in editing activity on the average page would have been at least 50 percent lower in its absence.

1 Introduction

There has been a substantial growth in content on the internet that arises outside of traditional firms. Known as user-generated content, Web 2.0, social media and crowdsourcing, internet-based platforms represent new ways to organize production of online content. Some platforms are primarily used to share individually produced content, such as blogs or social networks like Facebook or Twitter. In other cases, there is a more direct interaction between users in the production of content and the end product is the result of a collaborative process with different people contributing pieces of the overall product. A leading example for this type of joint-production is Wikipedia, an online encyclopedia that now contains almost 4.4 million individual articles (on the English language version of the webpage) and has been edited by over 20 million users since its inception in 2001. Besides being one of the most frequently visited websites, the impact of Wikipedia has extended beyond the platform itself.

*London School of Economics, Department of Management.

†Stanford University and Centre for Economic Performance.

Many firms are using private, "wiki"-style platforms in order to store and share knowledge within the company. Furthermore, there are also other public open source projects such as an online dictionary and a collection of open source teaching material that use the same interface as Wikipedia.

In this paper we study the production process of content on Wikipedia. The analysis sheds light on the dynamics of editing behavior on the world's most popular reference tool and also allows us to learn more broadly about the drivers of content growth on open source platforms. English Wikipedia is currently made of over 31 million pages. 4.4 million pages represent the encyclopedia articles, while the rest describe user profiles, policies and guidelines, discussions etc. We focus on article pages that are of an encyclopedic nature and for which the existing stock of human knowledge changes little over time. More specifically, we are interested in the process by which a given level of information knowledge on a topic that exists outside of Wikipedia, is converted into online content. For many pages on Wikipedia that mirror efforts of more traditional encyclopedias, the incorporation of a given knowledge stock is likely to be the main driver of content growth.[1] Specifically, we focus our attention on pages of the "Roman Empire" category.

We start by documenting the growth process of Wikipedia at a granular level for pages within this category. A few stylized facts emerge from this: (1) There is an enormous increase in editing activity from around 5,000 sentences of edits contributed by 200 users in 2002 to a peak of 130,000 contributed sentences from over 6,000 users in 2006 across all pages in our sample. This is followed by a modest slow-down to an editing activity of 100,000 sentences in 2009. Although less new pages are created over time, editing activity within existing pages remains at a high level even in later years. One might have suspected to see editing level-off as most of the existing knowledge stock is incorporated into the page and it becomes increasingly difficult to make further contributions. However, at the very least the process of saturation is very slow and even for a category with fairly stable knowledge such as the "Roman Empire" even pages created in 2002 are still edited heavily in 2009. (2) Pages created in earlier years receive more edits than later pages. This type of selection is most likely due to pages on more popular and broader topics being created first and not in itself very surprising. The effect is however quite pronounced and very long-lasting. Pages of a 2002 vintage are edited by more than 3-times as many users in *any* given year, i.e. even 7 years later in 2009, relative to pages created in later years. (3) We find that the growth process is largely driven by an increase in the number of contributors. The number of edits per contributor as well as the length of the average edit are fairly stable over time.[2]

We then focus on a key driver of growth that is specific to an open source environment like Wikipedia

[1] We do not concern ourselves with the interesting question how Wikipedia incorporates new information. For pages on current political event for instance new information plays a key role and the knowledge stock is constantly changing.

[2] We will use the terms "user" and "contributor" interchangeably going forward. We use both to denote a person that is editing rather than merely reading a Wikipedia article.

and constitutes an important advantage over more traditional production processes: having a large pool of potential editors allows individual contributors to add small pieces of information to a page and rely on subsequent users to develop the content further. In contrast to more traditional editorial processes, a user does not need to provide the entire content on a particular topic. Neither is it necessary to explicitly organize and coordinate the editing activity. Instead, a large set of anonymous users interact in the creation of content. A change in page content might therefore inspire other users to build on past edits. The mechanism is very similar to the process of knowledge accumulation analyzed in the R&D literature (Scotchmer (1991)). Innovators will make use of prior knowledge allowing them to "stand on the shoulder of giants". Weitzman for instance proposes a theoretical model of innovation production where "new ideas arise out of existing ideas in some kind of cumulative interactive process" (Weitzman (1998)). Similarly here, users will draw on the current knowledge when contributing themselves. Current content might influence them by providing new information or by making missing pieces of information salient to them.[3] This type of positive externalities is a source for growth in the production process that is particularly relevant due to the open source nature of Wikipedia, allowing any user to add content to existing pages.

In order to identify the existence of positive externalities, we regress measures of weekly editing activity on current page-length, while controlling for a host of confounding factors. Specifically, we control for inherent popularity differences across topics by including a set of page fixed effects. Furthermore, we allow for an aggregate growth-trend for the Roman Empire category as a whole. We do this in a very flexible way by including a separate dummy for every week in our eight year sample period (a week/page combination is our unit of observation). In sensitivity checks we also show robustness to including relatively flexible page-specific time-trends and run a specification which uses only changes in editing behavior following drastic changes in page-length. Finally, we use an IV-strategy in order to control for the presence of information shocks which are correlated over time.

We find evidence for the existence of positive editing externalities and the magnitude of the estimated effect of page-length on editing activity is economically important. In the absence of the spill-over effect, growth in editing activity between 2002 and 2010 would have been halved. Moreover, we find the externality to be particulary strong for pages created at the beginning of Wikipedia's existence, which are mostly pages that cover topics of broader interest. Furthermore, page-length leads to more editing activity by increasing the number of users editing a particular page. However, we find no evidence that the amount of editing *per user* changes as pages grow. Finally, we find that edits on longer pages are more likely to involve deletion of content and they are more likely to be

[3]Olivera, Goodman, and Tan (2008) propose a theory of contribution behavior which involves searching and matching of potential contributors to contribution opportunities. In our context edits by other users would constitute the creation of such an opportunity that some of the potential users might match with.

reverted by subsequent edits. However, both effects are very small in magnitude. These findings can inform the design of other open-source platforms such as within-firm Wikis or other large Wiki-style projects. The presence of the editing externality suggests that it might be beneficial to incentivize users to contribute content in order to trigger further contributions. We also find suggestive evidence that the spill-over effect varies with the total number of users active on the platform. This suggests that achieving a critical mass of potential contributors is important in order to increase the spill-over effect triggered by the editing externality.

One important caveat of our analysis is the fact that we can only measure the amount of activity but are not able to assess the evolution of page quality directly. Assessing quality is generally difficult and no metric is readily available to measure it consistently across pages and time.[4] One might suspect that a larger amount of editing will increase the final quality of Wikipedia articles which tends to be quite high (Giles (2005)). Furthermore, several studies across a wide range of topic areas find that Wikipedia contains very few outright mistakes, but articles often contain significant omissions (Bragues (2007), Devgan, Powe, Blakey, and Makary (2007) and Brown (2011)). This would suggest that editing activity which is likely to fill in some of the omissions will tend to improve quality. While this is encouraging, we have no way to directly assess how editing activity maps into quality improvement.

The paper relates to the literature documenting the growth process on Wikipedia such as Almeida, Mozafar, and Cho (2007), Suh, Convertino, Chi, and Pirolli (2009) or Voss (2005). However, contrary to those paper we describe the growth process at a more granular level. In particular, we document that there are subtle difference between page-level growth and category-level growth which combines within-page growth and the addition of new pages. Furthermore, we isolate a specific driver of growth due to spill-over effects in editing activity. One paper that look at a similar but more narrow issue is Gorbatai (2011) who analyzes whether expert editors become more active when observing prior edits by novice users. Second, the paper contributes to the emerging literature on Wikipedia more broadly such as Greenstein and Zhu (2012b) and Greenstein and Zhu (2012a) who document the extent of political slant on Wikipedia, Zhang and Zhu (2011) and Ransbotham and Kane (2011) who analyze the effect of the social network structure within Wikipedia or Nagaraj (2013) who uses Wikipedia data to assess the effect of copyright on creative reuse. Our study is also related to the concepts of knowledge accumulation and knowledge spill-overs which are central to the endogenous growth literature (Romer (1990), Jones (1995), Furman and Stern (2011)). At the micro-level Jaffe (1986) takes the fact that competing firms' R&D effect a firm's own activity as evidence for spill-over effects. Using data from patent citations, several papers explore the specific nature of the spill-over effect and how its magnitude varies with distance (Henderson, Jaffe, and Trajtenberg (1993)), within and

[4] Arazy, Nov, Patterson, and Yeo (2011) measure quality for a small number of pages at one point in time by having 3 librarians assess quality for each page.

across firms (Belenzon (2012)) as well and between countries (Jaffe and Trajtenberg (1999)). In this paper we quantify the effect of a spill-over effect within Wikipedia of accumulated past knowledge, as embodied by the page-length, on new knowledge creation, which we capture by measures of current editing activity.

The structure of the paper is as follows. In the next section we provide a description of the data followed by descriptive statistics in Section(3). In Section (4) we illustrate some important features of the growth process with a simple theoretical model. Section (5) presents page-level growth patterns. Sections (6) and (7) present the main empirical results as well as robustness checks and extensions. In Sections (8) to (10) we explore heterogeneity in the effect, changes in the type of edits being made and put the magnitude of the estimated effect into the broader context. Finally, some concluding remarks are provided.

2 Data

We use the English language Wikipedia database exported as "XML dump" on 30 January 2010 and made freely available by the Wikimedia Foundation.[5]. The massive data-set contains the full text of every revision of every surviving page the English version of Wikipedia from the beginning of the website on 16 January 2001 to January 2010. The raw data allows us to measure the exact content for every version of each page and attributes edits to individual contributors. Every time a page is edited and saved this creates a new XML record.[6] The original XML records were preprocessed using Python scripts into a tabular data-set representing 19,376,577 pages and 306,829,058 revisions.

We transform the raw XML records into a numerical format and focus on the length of the article at each revision as well as the amount of change in content, measured by the number of characters that were changed by a particular edit of the page. More precisely, for two consecutive versions of the same page, we compute the number of characters that needs to be added, deleted or changed (each one of these actions is counted equally) in order to convert one version of the page into the next. For ease of exposition we will refer to this metric simply as "edit-distance" in the remainder of the paper. The procedure to calculate edit-distance is computationally quite burdensome and in order to implement it across a large set of edits we employ the Levenshtein algorithm (see Levenshtein (1966)). More details on this procedure are provided in the appendix. The number of characters changed is arguably the most direct measure of the extent of an individual edit. Consider for instance the case of an edit which *replaces* large parts of a page with new content and might entail little change in page-length despite

[5] enwiki-20100130-pages-meta-history.xml.7z (34,248,021,709 bytes)

[6] Often a page is saved multiple times during one continuous editing process. We therefore consider any changes by the same user within an 8 hours windows (without any other user editing the page within the same time-window) as a single edit.

substantial content changes. Our edit-distance measure is able to capture such changes, which one would miss if looking only at changes in page-length. Finally, we are also able to track users across multiple edits.

Our analysis focuses on pages that belong to one particular category: the "Roman Empire". We choose this category, which comprises 1403 unique pages, due to the fact that knowledge on the topic is presumably undergoing relatively little change during our sample period. This helps us in terms of our identification strategy and also removes an additional layer of complexity which is the incorporation of new information into Wikipedia. For most of our analysis we assume that the knowledge stock with regards to topics in the "Roman Empire" category is stable. In order to define pages which belong into the category, we first select all 1571 pages which are linked to from the Roman Empire category page. However, Wikipedia does not assign each page to a unique category in a hierarchical fashion. Instead, many pages are categorized under multiple overlapping categories. We therefore manually reviewed the titles of those 1571 pages and eliminated the ones which only tangentially pertain to the Roman Empire. Note that it is not of major importance to our analysis to define a set of related pages, we simply need a set of pages for which the stock of human knowledge can be assumed relatively stable. We therefore exclude, for instance, several pages on video games and movies from the analysis as those do not belong to the Roman Empire category in a strictly historical sense.[7] Furthermore we drop a set of pages about geographical locations that still exist under the same name today as they might be edited due to more recent events taking place at those locations. In the appendix we provide more detail on the set of pages that we remove. The exclusion of the pages mentioned above narrows our sample down from 1571 to 1403 pages.

Finally, we have to deal with the fact that there is a certain amount of activity on Wikipedia coming from automatic "bots" rather than human contributors. These are user accounts controlled by software programs which are primarily used to fulfill relatively mechanical tasks such as correcting spelling and punctuation mistakes. A second purpose of bots is to detect vandalism of pages and to revert the vandalized page to its pre-vandalism state. Bot activity needs to be declared and the Wikipedia community might block users which use their account for undeclared bot activity. Bot activity can therefore be usually identified from user-accounts. We use both the bot user group which contains a list of bot user-account ids and investigate manually contributors with very large amounts of edits to check whether their user-page declares them as a bot. Although there might be some undeclared bot activity that we might be missing, we do believe that we are able to capture the majority of bot activity in our data.[8] For the empirical analysis we do not consider contributions by

[7]For example the movie "Monty Python's Life of Brian" appear in the Roman Empire category and receive a substantial amount of edits.

[8]As one would expect, we find that the average edit-distance of a bot edit is only about 10 percent of the average edit-distance for human contributors. This excludes cases where a bot is reverting a vandalized page to a previous

bots as part of editing activity. However, we do keep track of the aggregate page-length at every point in time regardless of whether the page has been edited by bots or human users. In other words we are only looking at human user contributions to the individual articles and will ignore bot contributions when computing our dependent variable, editing activity. The current knowledge stock captured by the page-length instead will reflect cumulative edits by both humans and bots.

3 Descriptive Statistics

We start by providing some descriptive statistics on the observed editing behavior. Our sample contains a total of 77,671 (non-bot) edits across all 1403 Roman Empire pages. The first two lines of Table (1) report the extent of individual edits measured by the change in page-length in units of characters. For ease of exposition we split the sample into positive and negative length changes, the former representing about two-thirds of all edits. Taken together the two rows display a large degree of heterogeneity in the length of edits. While the median length change for positive and negative changes is 36 and 29 characters respectively, the length of edits increases exponentially with length changes of over 10,000 characters at the 99th percentile. A very similar picture emerges when we use edit-distance as a measure of editing activity. Note that this metric is based on the number of characters that changed between two versions and is therefore by construction always positive. Again, we observe edits in the tail of the distribution that are orders of magnitudes larger than the median edit. The median edit-distance of 40 characters corresponds to about half a sentence (a typical English sentence has 73 characters) and constitutes a fairly small change in page content. In other words, although the growth process of pages is smooth and incremental for the most part, occasional large edits can change page content dramatically in a short amount of time. Similar patterns are also documented in. These type of discrete jumps in content is something we explicitly exploit in one of our empirical tests later.

Note also that while the majority of edits are net additions of content, there is a large fraction of edits which decrease page-length. This does not necessarily imply that these are "destructive" edits. The net effect of re-writing a paragraph for instance might be to lower total page length. In order to dig deeper into the heterogeneity between edits we compute a direct measure of the extent of addition and/or deletion of content. We use a very simple metric in order to capture the nature of edits in this respect by combing 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 $-\Delta Length = EditDistance$, i.e. the number of characters that were changed is equal to the reduction in length. The length change (in either direction) cannot exceed the number of characters changed. Based on this we compute

version. Edit-distance in those cases can be very large.

$\Delta Length/EditDistance \in [-1, 1]$. We find that about 37 percent of edits are pure additions of content (i.e. $\Delta Length/EditDistance = 1$), whereas 15 percent are pure deletions. The remaining edits are intermediate cases in which some existing content was deleted, but new one was also added. Edits within the intermediate range are roughly uniformly distributed over the range of our metric.

Finally, we report the number of edits which are involved in the reversion of a past edit. This can happen if a user decides to "undo" a previous edit, effectively returning the page to its state before the edit was made. This happens quite frequently and has even lead to research that focuses entirely on the dynamic of reverting edits such as Halfaker, Kittur, Kraut, and Riedl (2009) and Piskorski and Gorbatai (2013). In our data we find that on pages within the Roman Empire category about 29 percent of edits are involved in a reversion. Out of those, 14 percent are edits that have been reverted and 13 percent are reverting edits that restore a previous version of the page. About 2 percent of edits are reverting edits that are themselves subsequently reverted. These are mostly part of longer spells of "edit wars" where users go and back forth between reverting each other's edits repeatedly. There are two main sources of reversions, disagreement over newly added content which then gets removed as part of a reversion and vandalism. The latter usually involves the deletion of a large amount of content which is subsequently restored by a reverting edit. How to deal with *reverted* edits as well as the *reverting* edits is important for our empirical analysis. Consider for instance the unsuccessful attempt to add 1,000 characters worth of content. In the data this will be recorded as two edits (the attempt of adding content and the reverting edit) with an edit-distance of 1,000 each. This would lead to seemingly large amount of editing activity, but actually left the page unchanged. Similarly, "edit wars" can contain a large amount of edits that add and remove the same piece of content multiple times. At the bottom of Table (1) we report edit-distance separately for edits which are not involved in a reversion and reverted/reverting edits. We find edits involved in reversions to be substantially larger presumably due to vandalism involving big changes in content. For most of our analysis we drop all edits that are overturning a prior edit by restoring the page to an earlier version, i.e. all *reverting* edits. However we do keep most *reverted* edits in our sample because they constitute legitimate editing activity despite the fact that they do not have a lasting impact on the page. Indeed, many edits even if they are not deleted immediately are removed at least partially by later edits. The only exception to the above rule are edits which we consider to be acts of vandalism. We define vandalism as an edit that is a pure deletion of content that was later reverted. Going forward all descriptive statistics and other empirical analysis will be based on the subsample of (non-bot) edits which are neither reverting nor vandalizing edits, unless mentioned otherwise. The last row of the descriptive statistics table shows the distribution of edit-distance, our main measure of editing activity, for the final sample of edits.

For most of our empirical analysis later, we aggregate editing activity at the page/week-level. This

allows us to measure the number of users editing the page in any particular week as well as other measure of editing activity. Importantly, there are often long spells of inactivity on individual pages, something that the summary statistics at the edit-level in the previous table do not capture. We document the distribution of two key variables that measure editing activity in the lower panel of Table (1): the number of users[9] and cumulative edit-distance per week (added up across individual edits if there are multiple ones within a week). The unit of observation is a page/week combination of which we have a total of 265,707 across the 1403 pages and up to 434 weeks. In about 86 percent of page-weeks we observe no editing activity. The average number of users is equal to 0.215 and there is rarely more than one user editing a page in any given week. We also explore to what extent the number of users is explained by differences across pages as well as the overall growth trend of Wikipedia. We find that about 29 percent of the variation in number of users is explained by across page variation, but the aggregate time trend explains relatively little. In terms of the weekly cumulative edit-distance we find a skewed distribution with large edits in the right tail of distribution consistent with Table (1). Interestingly, edit-distance is explained by across page variation to a much smaller extent than the number of users. This suggests that the very large edits do not systematically occur repeatedly on a particular set of pages.

We postpone the discussion of page growth patterns until after introducing a simple model to guide our thinking about the content production process.

4 A Simple Model of Editing Behavior

We consider the behavior of user i on page j in time period t. A user in our terminology denotes a potential editor of the webpage, we do not model the consumption of content. We assume that the content on each page can be represented in a vertical quality space as $x_{j,t} \in [0, \infty)$. For simplicity we also assume that users are homogenous with respect to their preferences over content, i.e. the same content will translate into a quality metric $x_{j,t}$ that does not vary across consumers. We assume that decisions on different pages are taken independently and therefore drop the j subscript for expositional purposes from now on.

When a user visits a particular page j in time period t, he receives the following utility

$$u_{i,t} = -\alpha(x^*_{i,t} - x_t)$$

[9]Users rarely make multiple edits per week and number of users and number of edits are therefore highly correlated (correlation coefficient of 0.9785). Furthermore the number of edits is hard to define because in the raw data an edit is an instance of saving a new version of the page. Sometimes users save a page multiple times in a short time interval and it might be reasonable to consider all consecutive saved versions by the same user as a single edit. Any type of aggregation is always somewhat arbitrary however. Due to the high correlation with the number of users per week (which is not affected by multiple saved versions) we therefore focus on this measure instead.

Where $\alpha \geq 0$ captures how strongly the user feels about the content on the page. $x^*_{i,t}$ denotes the user's preferred quality level of content on the page. We assume that $x^*_{i,t} \geq x_t$. Either the consumer has knowledge that would improve the page and therefore his optimal quality level lies above the current one or he has nothing to add and $x^*_{i,t} = x_t$. As there is no cost of editing the page, the user will always re-position it to the optimal position according to his preferences. This will lead to a different quality value at the beginning of the next time period: $x_{t+1} = x^*_{i,t}$. If the user does not edit the page, x_t will remain at its current position.

We assume that a user's optimal quality level is determined by the following relationship

$$x^*_{i,t} = (1 + \gamma_i)x_t + \xi_{it}$$

Where $\gamma_i \geq 0$ captures the extent to which content on the page triggers any further contributions by the user. $\xi_{it} \geq 0$ represents any information that affect the optimal quality level that is derived from sources outside of Wikipedia. Put differently, γ_i and ξ_{it} represent internal and external information provision respectively. External information might not be incorporated into the page yet which would lead to $\xi_{it} > 0$. In the case of internal information, this information is by definition already incorporated in the page. However, due to heterogeneity in users' knowledge the existing content might help the user to remember additional knowledge he has regarding the topic. We therefore think of the case where $\gamma_i > 0$ not as creating new knowledge, but allowing the consumer to access existing knowledge more easily.

We assume that there are two types of consumers

$$\begin{aligned} &\text{Type 1:} \qquad \gamma_i = \overline{\gamma} > 0 \ , \ \xi_i = 0 \\ &\text{Type 2:} \qquad \gamma_i = 0 \ , \ \xi_i = \overline{\xi} > 0 \end{aligned}$$

In each time period there is a chance of λ_1 (λ_2) that a user of type 1 (2) arrives on the page. We further assume that $(\lambda_1 + \lambda_2) < 1$, in other words there is a strictly positive probability that no user arrives in any given time period.[10] Type 1 denotes a user that is able to draw inspiration from the current content level and will augment the content purely based on the knowledge already embedded in the current content. We will refer to this type also as "inspired" users. Type 2 represents a user

[10]More generally one can think of $1 - (\lambda_1 + \lambda_2)$ as the probability of either no users visiting the page or a user visiting who does not have anything to contribute to the page. For instance, one could easily extend the model to a more general case where consumers have to incur a cost (drawn from some distribution) to edit the page. In this case a consumer with $\gamma_i > 0$ and/or $\xi_i > 0$ might still decide not to edit if his edit-costs are sufficiently high. For the sake of simplicity we capture all time periods without an edit (for whatever reason) as no user visiting the page.

that brings new external information to the page.

It is easy to see that when a type 2 user visits the page in time period t the growth in content is equal $\Delta x_t = x_{t+1} - x_t = \overline{\xi}$, which is the new information that the user incorporates into the page. However, in our model the difference in content change does not only affect the current time period t put has a knock-on effect on future time periods. For instance one time period ahead the expected level of content growth (relative to no user visiting in time period t) will be

$$E(\Delta x_{t+1}|Type_t = 2) - E(\Delta x_{t+1}|Type_t = \varnothing) = \lambda_1 \overline{\gamma}\overline{\xi}$$

In total there will therefore be a relative increase of $(1+\lambda_1\overline{\gamma})\overline{\xi}$ comprised of the initial incorporation of $\overline{\xi}$ and the positive externality on editing in the next period $\lambda_1\overline{\gamma}\overline{\xi}$. The magnitude of the externality is very intuitively determined by the probability that an "inspired" user arrives on the webpage λ_1 and the magnitude of the inspiration effect $\overline{\gamma}$. Although our estimation does not map onto the model in a structural sense, our main focus will be to estimate the magnitude of this page-specific positive externality.

4.1 Page-level and Aggregate Growth

In the case of a platform experiencing such a rapid growth process as Wikipedia it is important to consider factors driving growth at the page as well as at a more aggregate level. This distinction will play an important role for our empirical identification strategy. For each page individually the existence of some type 1 users (i.e. with $\gamma_i > 0$) will lead to higher activity on pages with a higher content quality level x_t. However, it is likely that the pool of potential users grows over time as Wikipedia's aggregate content and the level of visibility of the platform grows. With the likelihood of a page visit being higher as the platform grows this will have a feedback effect on content provision.

In our model we can think of this mechanism as the aggregate content shifting the probability of a "knowledgable" user visiting the webpage. Formally, we assume that probability of page j being visited λ_{2jt} is a function of aggregate content across all pages $X_t = \sum_j x_{jt}$. More specifically, we model the effect of an increase in the user-pool with the assumption that $\frac{\partial \lambda_{2jt}}{\partial X_t} > 0$. In other words, the visit probability increases with X_t for type 2 users that are able to contribute external knowledge to the page.[11]

To see how this affects the analysis consider the following expressions for ex-ante expected growth rates in consecutive periods

[11]Note that the visit probability λ_{2jt} is now specific to the time-period, which was not the case previously.

$$
\begin{aligned}
E\Delta x_{jt} &= \lambda_{1j}\overline{\gamma_j}x_{jt} + \lambda_{2jt}\overline{\xi_j} \\
E\Delta x_{jt+1} &= \lambda_{1j}\overline{\gamma_j}x_{jt+1} + \lambda_{2jt+1}\overline{\xi_j}
\end{aligned}
$$

Note that in the absence of an effect of platform-level growth on the visit probability ($\lambda_{2jt} = \lambda_{2jt+1}$) we will see an increase in activity over time ($E\Delta x_{jt+1} > E\Delta x_{jt}$) only if the content stock increased ($x_{jt+1} > x_{jt}$) and some positive externality exists ($\overline{\gamma_j} > 0$). Instead, in the case of an increase the visit probability caused by an increase in aggregate content ($X_{t+1} > X_t$ and therefore $\lambda_{2jt+1} > \lambda_{2jt}$) we could see an increase in activity even in the absence of an externality from editing ($\overline{\gamma_j} = 0$). In this case we would observe an increase in editing activity over time as well as an increase in the content stock. This correlation is due to the fact that content on other pages grows which will increase λ_{2j} and at the same time x_j increases as new external information $\overline{\xi_j}$ is used to update the page. This relationship is not due to a causal effect. Instead, later in the platform's live pages tend to be longer and at the same time the user-pool is larger due to platform, but not necessarily page-level growth. In order to avoid picking up this purely correlational effect we control carefully for the time-trend in aggregate growth.

4.2 Crowding-out Effect in Editing

For simplicity we have so far assumed that a type 2 user always contributes the same amount $\overline{\xi}$ regardless of the current content level x_t. More realistically there will be some correlation in the knowledge stock among different users regarding a particular topic. We would therefore expect that as page-length increases there will be less additional content that an individual user can contribute to the page. Put differently there is likely to be some extent of crowding-out between edits as an edit will prevent somebody else from contributing the same piece of information later on. In our model we can capture this by assuming that $\frac{\partial \xi}{\partial x_t} \leq 0$ for type 2 users. Our baseline model represents the extreme case of $\frac{\partial \xi}{\partial x_t} = 0$ where knowledge is mutually exclusive between users and a user always contributes the same amount if he visits the page regardless of any prior editing activity on that page. For the case of $\frac{\partial \xi}{\partial x_t} < 0$ instead, it will be the case that longer pages receive less editing activity due to some of the potential contributions having been already incorporated into the page. In the empirical application we will not be able to separate this effect from the positive editing externality. Our estimate of the effect of page-length on editing activity will therefore capture the net effect of both mechanisms. However, as most of our data comes from a period of strong growth, the crowding out channel is likely to be less important. We also present some evidence that both mechanisms might be at work when investigating

heterogeneity in the externality across page vintages. We find that pages which were created in later years tend to pertain to more narrow topics. They are characterized by a lower net effect of page-length on editing, possibly due to the fact that the crowding-out effect is relatively more important for those pages relative to ones on broader and more popular topics.

4.3 External Information Shocks

A final aspect of page-growth that will inform our empirical analysis is the presence of correlated information shocks. It is generally quite likely that as new knowledge regarding a particular topic is discovered, multiple users will try to incorporate this new information. Possibly each one will contribute part of the increase in the external knowledge stock which can lead to temporary bursts in editing activity, which are unrelated to any editing externality within the page. Correlated information shocks are an issue to the extent that we might falsely interpret later edits in the activity burst to be reacting to previous edits whereas in reality all editing activity within a certain time-window is driven by the same external information shock.

In order to illustrate the pattern with a simple example, consider the case of a temporary information shock which increases external information provision of type 2 users (if one such user visits the page in a particular time period) to $\overline{\xi} + \theta$ for several periods starting in $t + 1$. In the absence of any editing externality ($\overline{\gamma} = 0$) expected content growth[12] in t and subsequent periods is equal to

$$\begin{aligned} E\Delta x_t &= \lambda_2(\overline{\xi}) \\ E\Delta x_{t+\tau} &= \lambda_2(\overline{\xi} + \theta) \end{aligned}$$

Where $\tau \in [1, T]$ denotes the set of time periods which are affected by the information shock. In $t+1$ page-length is higher ($Ex_{t+1} = x_t(1 + \lambda_1\overline{\gamma}) + \lambda_2\overline{\xi}$) and at the same time expected content contribution is higher by $\lambda_2\theta$. The same logic applies when comparing any of the later periods with higher editing activity due to the external shocks with time period t. This leads to a positive correlation between page-length and new content contribution when considering the time periods affected by the shock with the ones before. Note however, that after the information shock is fully incorporated into the page in $t + T$, the extent of contributions will go back to its original lower level. A comparison of any post-information shock period with periods with a higher contribution level of $(\overline{\xi} + \theta)$ will be characterized by a negative correlation of page-length and contribution level. It is therefore not clear in which direction correlated information shocks would bias our estimate. Nevertheless, any kind of

[12] Expectations are taken from the perspective of the beginning of the respective time-period. I.e. the knowledge stock at the beginning of the time-period is know, but page-visits have not realized yet.

correlation which is caused by something other than the editing externality is in principle problematic. To a large extent, the selection of pages from the Roman Empire category helps us mitigate this issue as it seems unlikely that knowledge about the topics within the category is subject to major shifts. Correlated information shocks are therefore unlikely to be of great concern in our specific setting. We do however also run a set of robustness checks to deal with this issue specifically.

5 Patterns of Content Growth

5.1 Content Growth at the Category-Level

Before we analyze the drivers of content growth, we first start by reporting the evolution of content for the Roman Empire category as a whole. Table (2) reports the number of pages created each year as well as the amount of editing activity on those pages. We find that the number of new pages created increases almost monotonically until 2005 and decreases afterwards. The second and third column report the total number of users active each year and the number of edits on any page within the category. For both measures we see a very substantial increase in activity peaking in 2007. Finally, we look at the amount of editing captured by the cumulative annual edit-distance across all pages. The pattern for this variable is quite similar to the other measures of editing activity: we see a strong increase early on with and a slight decrease in the later years. In the case of all three metrics the eventual slow-down and decrease is substantially smaller than the initial "ramp-up", especially in the very early years. For example, the number of edits increased from 556 edits in 2002 to 13,874 in 2007 and then decreased very slightly over the next 2 years to 13,122 edits in 2009. In other words, we seem to be seeing some level of maturity and possibly saturation in terms of content. But, despite the long time-horizon the level of activity is still quite high in the Roman Empire category. The growth patterns are consistent with finding elsewhere such as Suh, Convertino, Chi, and Pirolli (2009) who document exponential growth patterns up to 2007 and a slow-down afterwards.

Out of our three activity measures, edit-distance is presumably the most direct one as it captures both how many users engage in editing as well as how much each one contributes. However, similar to Almeida, Mozafar, and Cho (2007) we find that the ratio of edits per user as well as the edit-distance per edit is very stable over time. Therefore, most of the growth process on Wikipedia seems to be driven by an increase in the user-pool rather than changes in editing behavior of existing users. We report a larger set of edit activity measures in Table (B1) in the appendix. Note that the number of edit per users on a yearly basis is very stable with roughly 2 edits per user, with the possible exception of the first year for which there are 3 edits per user on average. Edit-distance per edit does fluctuate more over the years, but it also does not show any clear time-trend. Because cumulative edit-distance

on a yearly basis can be strongly affected by a few very "heavy" edits, we also report a version of the edit-distance which caps individual edits at 10,000 characters (roughly the 98th percentile of the edit-distance distribution). The capped metric exhibits a similar growth pattern over the years as the other measures of editing activity.

5.2 Content Growth at the Page-Level

One issue with the aggregate analysis is the fact that the composition of pages changes over time. The time trends reported in the previous section therefore combine the effects of changes in activity on existing pages as well as the effect of adding more pages over time. In order to provide a more detailed analysis, we split the the pages into different groups depending on the year in which the first edit was made. Each category therefore consists of pages with a similar age at any point in time. For each set of pages we report the average cumulative edit-distance per page in each year of the pages' existence. Tracing out the evolution of these different page "vintages" gives a clearer picture of the editing dynamics over time at the page-level.

Table (3) reports the average page-level number of users as well as the cumulative edit-distance for pages of the same vintage within a given year. The first thing to note is that the activity on pages started in 2002 (the first year of activity)[13] dwarves the activity on pages of any later vintage. Editing activity generally decreases across vintages for most years and the differences in editing activity are extremely long-lived. Even in 2009, 7 years after the earliest pages were started, the 2002 vintage pages still receive over 3-times more editing activity than pages of any later vintage. We also find that later vintages peak earlier in their lifetime and at a lower level. The patterns look very similar for both measures of editing activity, but differences between vintages are slightly less pronounced for the number of users. As an additional metric we also compute the same table using edit-distance capped at 10,000 characters and find qualitatively similar results which are reported in Table (B2) in the appendix. The decrease in activity across vintages is most likely due to the fact that pages on the most interesting / broad / relevant topics were started early on and these pages are therefore edited by a larger number of users. To illustrate this pattern we report the five pages with the largest number of edits for each vintage. The top five pages created in 2002 all concern very broad topics such as "Holy Roman Empire" or "Saint Peter". In contrast, among the five most edited pages of the 2009 vintage are more narrow topics such as "Principality of Stavelot-Malmedy" and "Siege of Godesberg (1583)".

The descriptive statistics in Table (3) also show that some interesting editing dynamics were masked at the category-level by the aggregation over different page vintages. We find that at the individual

[13] Wikipedia was started in early 2001, however for the Roman Empire category we observe only a very small level of activity at the end of 2001. The 2001 pages (there are 3) are included in the 2002 vintage.

vintage-level, i.e. holding fixed the number of pages over the years,[14] a somewhat stronger slow-down is taking place (relative to the category-level). For instance, pages created in 2002 did experience a decrease from a peak of 57 users to 39 in 2009. The edit distance decreased from 53,000 to 31,000 characters and the capped edit distance decreased from 23,000 to 17,000 characters over the same time horizon. Other vintages experienced a similar or more modest slow-down in activity. Nevertheless, similar to the category-level growth patterns, the initial increase is still substantially larger than the subsequent decline for the earlier vintages. This asymmetry in the growth-pattern is less pronounced or absent for later vintages.

Finally, we also report the evolution of average page-length, which is the stock-variable that the editing activity contributes towards. We find that average page-length for the earliest pages has increased roughly 10-fold and by about 200 to 400 percent for most other vintages. The difference between vintages in terms of page-length is not as pronounced as the differences in editing activity. This is to some extent due to the fact that on the earlier pages a larger fraction of edits did not add new content, but rather deleted or removed prior content. Secondly, the number of reverted edits is also larger for pages created earlier. We investigate both channels directly in Table (B2) in the appendix and find that edits involved more deletion of content over time in particular for the earlier vintages. More specifically, our measure of content addition/deletion ($\Delta Length/EditDistance \in [-1, 1]$) drops from an average of 0.53 to 0.36 for 2002 vintage pages over time. The relative proportion of addition of content versus deletion decreases over time for all vintages. Also, in any given calendar year older vintages tend to have more deletions. Similarly, the amount of reverted edits increases over time and more strongly affects earlier vintages. In 2009, 28 percent of edits on 2002 vintage pages were reverted compared to 13 percent for the 2003 vintage and even less for later vintages. An alternative way to look at this is to compare total cumulative edit-distance over a page's life-time with its length. We compute such a measure for each page in the last week of our sample in January 2010 and report the results grouped by page vintage in Table (B4) in the appendix. In line with the findings above, we find that the ratio of length to cumulative edit-distance is as low as 25 percent for 2002 vintage pages and increase for younger vintages with a ratio of 67 percent for 2005 pages and 91 percent for the youngest pages created in 2009. In other words, out of all contributions made on pages created in 2002 only one quarter are still part of the page's content in 2010.[15]

[14] Across the rows of any given column of Table (3), the number of pages does not change. This is different from Table (2), where across columns both the number of pages and the activity on each page changes.

[15] We also investigate the role of edits executed by bots on page-growth patterns. The fraction of edits done by bots at the vintage/year-level are reported in the bottom panel of Table (B2). We find that the share of bots increased over time and is larger on pages that were created later. In other words the difference in editing activity between vintages would be even more pronounced if we considered only non-bot edits.

6 Content Growth and Externalities from Editing

The previous sections have shown that Wikipedia experienced substantial growth both in the number of pages as well as the level of activity on each individual page. Our analysis documents a slow-down in editing activity, perhaps indicative of a certain level of maturity in content within the Roman Empire category. However, with activity peaking around 2006 / 2007, most of our data (2002 to January 2010) stems from a period of rapid growth. In the remaineder of the paper we set out to identify a key driver of content growth: page-specific externalities of editing that result from users "building-up" on previous edits by other users.

More specifically, we analyze whether current page-length has an effect on editing activity. We would expect to see such an effect if users read the current content of the page, which is the cumulative of surviving past edits, and draw inspiration from it. This mechanism is indeed often mentioned by Wikipedia users: the open source nature of the platform allows people to add small pieces of knowledge to the existing body when they see an omission. Many users might not have created a particular page from scratch, but the fact that some version of it existed triggered their contribution. In our model this effect is captured by the presence of a positive mass of users with $\gamma_i > 0$ that draw inspiration from current content and update the page based on this inspiration.

In our main specification we regress the number of weekly users on the current length of the page (in units of 10,000 characters). In order to implement this, the edit-level data is aggregated at the weekly level. In other words an observation is now a page/week combination. Leaving out pages that were started in 2009 or later due to a short time-series, we have 1267 pages and up to 434 weeks of data for the earliest page. This yields a total of 265,706 observations.[16] We include a set of page fixed effects into the model in order to control for the general appeal and popularity of the particular topic of the page. We also control very flexibly for a general time trend in editing behavior within Wikipedia as a whole. This is important in our context as the predictions from the theoretical model in Section (4.1) illustrate. Pages will tend to be longer later on in their lifetime and at the same time in later years more users were active on Wikipedia. Table (3) highlights this feature of the data: both the length-stock and editing activity have a positive time trend. We want to avoid picking up this general platform-level growth effect and instead isolate the effect of page-level variables on editing activity. To this end, we include a set of weekly dummies for the roughly 8 year long (434 weeks) sample period in each regression, which is the most flexible way to control for general time-effects.[17] We cluster standard errors at the page-level in order to allow for an arbitrary within-page error correlation. Formally, we

[16] We drop the first week for each page because by construction the founding week contains at least one edit (and has zero length).

[17] We do not include dummies for the first 20 weeks of our sample as only a few pages exist during that time period and weekly dummies are therefore hard to identify together with page fixed effects.

run the regression

$$UserNum_{jt} = \beta PageLength_{jt} + \theta_j + \psi_t + \varepsilon_{jt} \quad (1)$$

Where j denotes a specific page and t denotes a week. θ_j and ψ_t are a set of page-/ week-fixed effects respectively. ε_{jt} denotes the error term.

The first column of Table (4) reports the coefficient on page-length which is equal to 0.204 and highly significant. In other words about 50,000 additional characters (about 700 sentences) of page length are associated with one more active user per week. To get a sense of the magnitude of the effect, note that the average page in 2009 is about 7,500 characters long. The page will therefore be edited by about 0.15 additional users *per week* compared to when it started. The average page that was created in 2002, the first year in our data, was about 17,000 characters longer in 2009. This length change will lead to an additional 0.35 users each week. Given an average of 0.207 weekly users (the median is zero) and a standard deviation of 0.813 in 2009, this is a fairly substantial effect.

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 277.9, which can be interpreted as 10,000 characters of page-length (about 140 sentences) leading to almost 300 characters or 4 sentences of additional weekly editing activity. For the average 2009 page this would entail an additional 500 characters / 6.5 sentences being contributed each week. This is a large effect relative to an average weekly edit-distance of 400 characters. However, the effect might seem small relative to the large standard deviation of the edit-distance variable which is equal to 19,000 characters.

Because the distribution of weekly total edit-distance is extremely skewed, it is not clear whether its standard deviation is the best benchmark for the effect size. For this reason and to test whether our results are driven by large outlier values, we re-run the regression using the capped edit-distance defined before as our dependent variable. When we switch the dependent variable we obtain a positive and significant coefficient, but of smaller magnitude than for our baseline case. 10,000 characters of additional page-length lead to 117 characters of additional edits rather than 278. Note however that in terms of standard deviations of the underlying variable (reported in the first row of Table (4)) the effect is actually substantially 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 as those edits do have a very strong impact on the respective page. The capped measure simply provides evidence that it is not only the very heavy edits that are driving 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 a page, however it is quite noisy 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 which is much less affected by outliers. Furthermore we find that average edit-distance per user is fairly stable over time and most of the growth in activity is driven by an increase in the number of users. This is true for the general category as well as page-level time-trends reported in Tables (2) and (3). Later we also test explicitly whether increases in page-length lead to relatively longer or shorter edits and do not find this to be the case. We are therefore able to focus on the number of users as our main measure of editing activity without missing any important growth dynamics.

7 Robustness Checks

7.1 Page Age and Visibility

One alternative explanation for the patterns we see in the data might be that the pages created earlier are more visible simply due to the fact that they have existed for longer and more users had a chance to come across them. A simple way to test this hypothesis would be to include a control for page-age. However, note that the two-way fixed effects for pages as well as time-trends already control for this implicitly. To see this consider the following specification with only a linear time-trend and page-age as control

$$\begin{aligned} UserNum_{jt} &= \beta * PageLength_{jt} + \theta_j + \gamma * t + \delta * PageAge_{jt} + e_{jt} \\ &= \beta * PageLength_{jt} + \theta_j + \gamma * t + \delta * (t - PageBirthYear_j) + e_{jt} \end{aligned}$$

Note that page-age can be decomposed into a page-specific component (year of creation of the page) and a linear time-component. It is easy to see that δ in the above specification cannot be separately identified from θ_j and γ due to co-linearity of the variables. The same intuition extends to the case of our baseline specification which includes more flexible time controls. In other words, the estimated β is not affected by increased editing activity which originated from a page's length of existence on Wikipedia.

7.2 Page-specific Time-trends

Another issue one might worry about is the presence of page-specific time-trends. If pages have different inherent levels of activity growth this would lead to some pages growing faster than predicted by the average time trend (captured by the ψ_t-terms). These pages would be longer than average and have a larger number of active users. The inverse would be true for pages with a below-average growth-trend.

In terms of our model we can think of the visit probability of users λ_2 increasing over time due to the larger visibility of Wikipedia and a larger pool of potential users as outlined in Section (4.1). If the visit probability increases disproportionately for some pages relative to others this could introduce spurious correlation between editing activity and page-length. We tackle this issue in two ways.

First, we re-estimate our baseline model including a page-specific cubic time-trend. In other words we estimate

$$UserNum_{jt} = \beta PageLength_{jt} + \theta_j + \gamma_j * t + \delta_j * t^2 + \zeta_j * t^3 + \nu_{jt}$$

where t, as before, denotes the time subscript.[18] Note that this specification includes 1403 page fixed effects as well as 3∗1403 coefficients (!!!) to capture page-specific trends. Results from a regression with only a linear time-trend as well as higher order controls are reported in columns (2) to (4) of Table (5). For easier comparison, the baseline coefficient is reported in the first column of the table. The coefficient on page-length is very similar across specification with slightly lower but always statistically significant coefficients when adding page-specific time-trends. With the exception of the inclusion of cubic time-trends, the coefficient on page-length is never significantly different from the baseline. Given the shape of the aggregate and page-level growth patterns which are characterized by an initial steep increase and later slowdown, we believe that the cubic page-specific time-trends do a good job of controlling for page-specific growth dynamics.

The test also highlights a key source of variation in the data which allows us to get precise estimates even after including very rigorous time-trend controls: discrete and large jumps in page-length due to individual "heavy" edits. More specifically, even if pages had their own time-trends due to difference in popularity between topics, it is arguably reasonable to treat the specific timing of very large edits as exogenous. Table (1) provides some evidence in this respect: while page fixed effects have predictive power for the number of users, this is not the case for the edit-distance which is more strongly influenced by a few large edits. The test above shows that after controlling for the "smooth part" of a page's growth-process, editing externalities can be identified from jumps in page-length.

In order 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 page-length of more than 1,000 characters. This limits us to only 2,227 observations, which constitutes a little under one percent of the full sample. For each instance of a large change in length we compute the number of users in the week preceding the

[18] Note that we do not include a week-specific fixed effect ψ_t as we did previously. While week FEs are in principle identified they turn out to be highly co-linear with the page-specific growth-trends and we therefore prefer to leave them out. We re-ran the regression with page-specific time-trend and year (instead of week) fixed effect and find very similar results.

change as well as the week following the length increase. We then regress the change in the number of users on the change in page-length. Formally this is similar to a differenced version our original regression (1):

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

Note that we omit the week that contains the large edit itself in order to compare time-periods that are strictly before / after the jump in page-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. As we are using a very narrow time-window this omission should have a negligible impact on the regression. Similarly, any page-specific growth-trend would also presumably have a minimal effect. Similar to a regression discontinuity type of approach, we are relying on the fact that other than the page-length increase, nothing else changed that could have an effect on editing activity. While we consider this a fairly strong test, it suffers from the fact that we are only able to work with a small sub-sample of the full data-set. We find a positive and highly significant coefficient which is reported in the final column of Table (5). In terms of magnitude the estimated of coefficient of 0.186 is very similar to the baseline coefficient of 0.204.

Finally, we run a placebo test to further probe whether the coefficient on page-length is picking up general page-level growth trends. If it was the case that some pages get more editing activity and grow faster due to their inherent popularity we should see a high correlation between current and *all* past editing activity. Instead, if we are correctly identifying the editing externality as the mechanism, then current editing activity should only be responsive to the past editing activity that is still embodied in the current page-content. Put differently, there is no reason why content that once existed on the page but was later deleted should in any way inspire current users to contribute. The externality should therefore only lead to a response of editing behavior to *surviving* edits rather than all past editing activity. As we show in Table (B4), pages' cumulative edit-distance is often substantially larger than their page-length later in their life because content is often replace or deleted. This allows us to run a regression where we include both current page-length as well as cumulative past editing activity in the regression.[19] Coefficient estimates from this regression are reported in the final column of Table (5). We find that after controlling for page length the cumulative past 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 page-length and edit-distance). Furthermore, the

[19]Note that if there is no deletion or replacement of content on a page the two measures would be identical. For most pages the metrics diverge at some point in their lifetime.

coefficient estimate on page-length remains almost unchanged relative to the baseline specification. This shows 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 spill-over effect.

7.3 Correlated Information Shocks

A further threat to a causal interpretation lies in the presence of information shocks that are persistent over time. For instance, new information could become available to users outside of Wikipedia at a particular point in time, but users might not all respond to the news at the same time. Instead, it is entirely possible that over an extended period of time different users will slowly incorporate the new information into the Wikipedia article. Section (4.3) of the theoretical model outlines the consequences of such an external information shock. In short this kind of shock will lead to an increase in both page length as well as current editing behavior. We therefore explicitly chose a set of pages that was presumably not particularly affected by new information. Most likely, the stock of knowledge regarding historic topics such as the Roman Empire among the user pool does not change very much over time. It is however not inconceivable that information shocks through for instance media consumption (such as a TV documentary) could create the type of endogeneity problem just described.

Although we think that an endogeneity problem is unlikely to be present for the set of pages considered, we test whether our estimates are robust to an IV-strategy where we instrument the current length of the page with lagged page-length. The idea is to use page-length from a time period far enough away that the effect of any information shock that affected lagged page-length will have no effect on current editing anymore. However, page-length is highly persistent over time, which leads to a high correlation of current with lagged length. We experiment with various lags and find similar results. Columns (2) to (4) of Table (6) report results using page-length 3 months prior. The first stage is highly significant and the second stage coefficient on page-length is not significantly different from the OLS coefficient in column (1). This is not just due to a large standard error, but the fact that the two point estimates are extremely similar.[20] We also replicate the IV using an even larger lag of 6 months in Columns (5) to (7), which yields again very similar results.

[20]Column (2) replicates the OLS specification of the baseline case with the reduced number of observations that is available for the IV. Due to the usage of a lagged instrument the first few observations for each page cannot be used in the regression (the instrument is not defined for those observations).

8 Effect Heterogeneity

As we saw in Section (5.2), pages created in earlier years are edited more heavily, presumably because pages on the most interesting and appealing topics were created first. If it is indeed broader appeal that led to more edits on those pages, it is quite likely that editing externalities also differ across page vintages. Specifically, we would expect to see larger externalities for the earlier pages as there will be a larger pool of potential users that might be inspired by previous users' contributions. We investigate exactly this by including a set of interaction effects of page-vintage (i.e. the year of creation) and current page-length. The results are reported in column (2) of Table (7). For easier reference our baseline regression is replicated in the first column. We find that there is a large amount of heterogeneity across page-vintages with a significant coefficient of 0.495 for the earliest pages created in 2002, which is substantially larger than our baseline effect of 0.204. As expected the magnitude of the externality decreases almost monotonically across vintages with an insignificant effect for pages created in 2007 and 2008. As outlined in the model in Section(4.2), there might be two forces at work in terms of externalities: on the one hand increases in page-length might lead to more edits due to other users building up on past contributions. On the other hand there can also be a crowding-out effect of current edits, in particular if the page content is already near completion in terms of the current knowledge on the topic. The latter might be causing a negative net effect for later vintages which contain pages on more narrow topics (see Table (B3)) which might be exhausted after a smaller number of edits.

Note that when including vintage interactions we are mixing up the effect of inherent differences in the popularity of pages and a general time-trend in editing activity and spill-overs. Possibly there are changes in the extent of spill-over effect over time that might lead to lower editing externalities in later years for all pages. However, because later vintages by definition only exist in those later years they might have lower spill-overs not due to inherent page differences, but due to the different time horizon of their existence. In order to disentangle those two forces we interact page-length with a set of vintage as well as calendar year dummies. As both set of dummies add up to one, we have to exclude one interaction term. We therefore omit the interaction of page-length with the calendar year 2006, but keep the full set of vintage interactions. When doing so we find even stronger spill-over effects for some vintages, which can be interpreted as the vintage specific effect in 2006 (the peak year in terms of the magnitude of the spill-over effect). The decline across vintages is more modest for this specification and we find positive and significant effects for all vintages. In terms of the calendar year interactions we find an inverse U-shape with a peak in 2006 and substantially lower spill-overs in other years. A possible explanation for this pattern is variation in the total user pool which also varies in an inverted-U shape over time. This relationship is quite intuitive and predicted by our model: holding page popularity fixed, years in which there are more active users on Wikipedia should be characterized

by bigger spill-overs. In our model this is captured by changes in the arrival probabilities of users over time. In Figure (1) we investigate the issue by plotting the evolution of the total number of users across all Roman Empire pages against the magnitude of the estimated year-specific spill-over effect.[21] The figure shows that both vary in a very similar way over time lending support to the notion that spill-over effects are magnified by the size of the user pool. This mechanism is also consistent with the finding in Zhang and Zhu (2011) that an exogenous reduction in the user pool lowers contributions from users that are still active. Alternatively, the crowding-out effect described above might be another factor leading to a decline in the effect size in later years when page content is presumably relatively closer to completion.

We also explore non-linearities in the effect of page-length on editing activity, but find little evidence for any such effects. When adding higher order terms, we find the coefficients on both a square and a cubic term on page-length to be insignificant.

9 Changes in the Type of Edits

Having established how the number of weekly users changes as a function page-length, we now try to unpack whether and how the type of edits changed that users are making. Results using various measures to characterize different dimensions of editing behavior are reported in Table (8). We first test whether longer pages are characterized by edits that contain relatively more or less addition / deletion of content. For this purpose we use the measure for the importance of content addition versus deletion introduced earlier: $\Delta Length/EditDistance \in [-1, 1]$. We use the same setup as our baseline regression with the only difference that we are only able to use page-week pairs that contain at least one edit. For these weeks we compute the average addition / deletion metric across all edits and regress it on page-length. When running this regression we find a negative and significant coefficient of -0.025, which implies that edits on longer pages 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 0.413 (0.621). As a further point of reference, note that the metric falls by about 0.2 for 2002 vintage pages between 2002 and 2009 as shown in Table (B2). This is an order of magnitude larger than the 0.025 change induced by an increase of 10,000 characters in page-length. This constitutes a large difference in particular because the average page in 2009 is only 7,500 characters long.

A similar pattern emerges when using the fraction of reverted edits as the dependent variable. We find a negative and significant effect which shows that edits on longer pages are more likely to

[21] Note that because the spill-over effect is estimated at the individual page level it is not by construction correlated with the total user pool (defined as the number of users that edited at least one of the 1403 pages within the category).

overturned by subsequent edits of other users. However, the magnitude is again quite small compared to the variable's mean and standard deviation as well as the increase in the metric over time reported in Table (B2)

Furthermore, we also analyze whether the length of edits changed as a function of page-length. This is particularly important for our purpose as we focused on the number of users as our main measure of editing activity. While we found that longer pages are edited by more users, it could be the case that this effect is counteracted by users making shorter edits which would weaken the spill-over effect. We do however find no evidence that the weekly edit-distance per user changed as a function of page-length. The coefficient in column (3) of Table (8) is insignificant and small in magnitude. In order to be sure that the noisiness induced by outlier values is not the only reason for not finding an effect, we also compute the capped edit-distance per user and use it as the dependent variable in column (4). Again, we find no significant effect.

10 Editing Externalities and Platform-Level Growth

In order to assess the overall relevance of the spill-over we compare its magnitude with the general category-level growth trend. Doing so will allow us to quantify to what extent the growth-process would have been slowed down in the absence of the externality. Even in the absence of the spill-over effect, editing activity would still have grown over time due to the overall growth trend which is captured by the set of weekly dummies in our estimation. We start by plotting out the dummies over time in Figure (2). We find an initial increase and a modest slow-down in later years which is a very similar pattern as the category-level growth rates in activity reported earlier in Table (2).[22] Between 2002 and 2010 the number of weekly users increased by about 0.4. Pages created in 2002 are the only pages to have been affected by the category-level growth trend over the whole sample-period. The average length change for those pages between 2002 and 2010 is equal to 19,000 characters which translates into an increase of $0.204 * 1.9 = 0.39$ in terms of weekly users. In other words the category-level growth-trend and the spill-over effect contributed roughly equally to the increase in editing activity over time. For pages at higher percentiles of the length distribution the spill-over effect can even dominate. For instance the page-length of 2002 pages at the 75-th percentile is equal to 23,000 characters making the effect of the spill-over larger than the general growth-trend.

This quantification suggests a substantial impact of the spill-over on editing behavior on Wikipedia. Without it the growth in editing would have been lowered by about 50 percent. Furthermore, the content creation caused by the spill-over effect is likely to have improved the quality of content on

[22]Note that there are negative values at the beginning of the time series which constitute a decrease relative to the omitted category which are the first 20 weeks of the sample for which on weekly fixed effect is included.

Wikipedia which in turn will have increased site-traffic thereby increasing the pool of potential editors. For instance Antin and Cheshire (2010) document the fact that readership of Wikipedia makes becoming a future editor more likely. If readership is correlated with article and platform quality this will lead to an effect of content provision on the number of editors. In other words it is possible that part of the aggregate growth-trend is itself partially caused by past spill-over induced editing activity. Our model of editing activity captures exactly this channel by making the page-visit probability λ_{2jt} a function of aggregate category-level content X_t in Section (4.1). Therefore, the cumulative of all contributions on individual pages x_t triggered by the spill-over will lead to a feedback effect on content growth via increasing the user-pool. In this case our estimate represents a lower bound on the importance of the editing externality.

11 Conclusion

In this paper we documented the growth process of open source content using a large set of pages on Wikipedia over an 8 year time-horizon. Using very detailed edit-level data we find substantial growth in editing activity in earlier years with a modest slow-down towards the end of our sample period. Pages that were created earlier receive significantly more edits than later vintages at any point during their lifetime, most likely due to the fact that they concern broader topics. We identify one key driver of content growth which is a spill-over effect of past edits on current editing activity. More specifically, we find that page-length has a positive effect on the number of weekly users as well as total weekly contributions as measured by the cumulative edit-distance while controlling for page popularity as well as a flexible category-level growth trend. The result is robust to a whole battery of robustness checks suggesting that we are able to identify a causal effect of the content stock on editing activity. In terms of magnitude the effect is economically important with about half of the growth in editing activity over time being caused by the externality. Furthermore, we find that most of the growth process is driven by an increase in the number of users whereas the amount of editing activity per user is fairly stable. We also find evidence that edits involve relatively more deletions and are more likely to be subsequently reverted as page-length increases. Both effects are, albeit statistically significant, of very small magnitude.

Our findings imply that when designing an open source platform one might want to incentivize users to contribute in order to trigger further edits via the spill-over effect. Our results also suggest that a large pool of potential editors is important in order to benefit from the externality. This assertion is supported by the fact that years with a larger category-level user pool are characterized by bigger spill-over effects as are pages on topics with a broader appeal. The extent to which the spill-over can be

harnessed might therefore to a large extent depend on the focus and scope of the project in question.

References

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

Antin, J., and C. Cheshire (2010): "Readers are Not Free-Riders: Reading as a Form of Participation on Wikipedia," in *Proceedings of the CSCW*, Savannah, Georgia.

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.

Belenzon, S. (2012): "Cumulative Innovation and Market Value: Evidence from Patent Citations," *Economic Journal*, 122(559), 265–285.

Bragues, G. (2007): "Wiki-Philosophizing in a Marketplace of Ideas: Evaluating Wikipedia's Entries on Seven Great Minds," *MediaTropes eJournal*, 2(1), 117–158.

Brown, A. R. (2011): "Wikipedia As a Data Source for Political Scientists: Accuracy and Completeness of Coverage," *Political Science & Politics*, 44, 339–343.

Devgan, L., N. Powe, B. Blakey, and M. Makary (2007): "Wiki-surgery? Internal validity of Wikipedia as a medical and surgical reference," *Journal of the American College of Surgeons*, 205, S76–S77.

Furman, J., and S. Stern (2011): "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Knowledge Production," *American Economic Review*, 101(5), 1933–63.

Giles, J. (2005): "Internet encyclopaedias go head to head," *Nature*, 438, 900–901.

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., 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.

Henderson, R., A. Jaffe, and M. Trajtenberg (1993): "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *Quarterly Journal of Economics*, 119(434), 578–598.

Jaffe, A. (1986): "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Value," *American Economic Review*, 76, 984–1001.

Jaffe, A., and M. Trajtenberg (1999): "International Knowledge Flows: Evidence from Patent Citations," *Economics of Innovation and New Technology*, 8, 105–136.

Jones, C. I. (1995): "R&D-Based Models of Economic Growth," *Journal of Political Economy*, 103(4), 759–784.

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

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

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.

Romer, P. M. (1990): "Endogenous Technological Change," *Journal of Political Economy*, 98(5), S71–102.

Scotchmer, S. (1991): "Standing On the Shoulders of Giants: Cumulative Research and the Patent Law," *Journal of Economic Perspectives*, 5(1), 29–41.

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.

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

Weitzman, M. L. (1998): "Recombinant growth," *Quarterly Journal of Economics*, 113(2), 331–360.

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
Length-Change	(if $\Delta > 0$)	63.17	1023	15976	36	147	879	2350	18082
Absolute Length Change	(if $\Delta < 0$)	36.83	2045	22924	29	133	1120	4692	52791
Edit-Distance	Full Sample		1363	17888	40	181	1175	3240	30796
Adding / Deletion Measure	Addition	37.08							
	Deletion	14.75							
	Mix	48.17	0.17	0.60	0.17	0.73	0.94	0.98	0.997
Reverted Edits	All	29.28							
	Reverted	14.38							
	Reverting	13.02							
	Both	1.88							
Edit-Distance	Non-Reverted Edits	70.72	476	2433	37	153	795	1995	9231
	Reverted Edits	29.28	3503	32737	49	310	3098	12797	73739
	Final Sample	82.56	608	13208	37	152	800	2057	10001

WEEK LEVEL	Weeks with no Edit	Mean	S.D.	75th	90th	95th	99th	R-square Page FEs	R-square Week FEs
Number of Users	86.28	0.215	0.822	0	1	1	3	0.2940	0.0093
Edit-Distance	86.28	146	8216	0	22	113	2205	0.0078	0.0015

Table 1: **Descriptive Statistics**

Year	Number of Pages Created	Number of Users	Number of Edits	Cumulative Edit Distance (Unit: Characters)	Cumulative Edit Distance (Unit: Sentences)
2002	85	182	556	394,967	5,411
2003	72	414	973	527,520	7,226
2004	121	1,252	2,714	1,100,098	15,070
2005	**337**	3,215	7,390	4,412,004	60,438
2006	216	6,138	12,622	**9,361,682**	**128,242**
2007	239	**7,138**	**13,874**	8,005,666	109,667
2008	197	6,213	12,874	7,621,270	104,401
2009	136	5,768	13,122	7,539,501	103,281

Table 2: **Content evolution at the category level.**

		Year in which the page was started							
		2002	2003	2004	2005	2006	2007	2008	2009
NUMBER OF USERS	2002	5.27							
	2003	6.46	3.71						
	2004	15.71	7.06	3.80					
	2005	31.76	12.96	8.22	3.84				
	2006	52.53	**15.67**	**12.07**	**5.47**	4.84			
Calendar Year	2007	**57.41**	15.52	11.51	5.35	**5.86**	3.26		
	2008	44.22	13.01	10.07	4.82	5.38	**3.89**	4.14	
	2009	39.48	13.14	8.59	5.21	4.80	4.56	**5.21**	**4.01**
EDIT-DISTANCE	2002	4647							
	2003	3123	3726						
	2004	7825	1966	2453					
	2005	34524	5882	3094	2132				
	2006	31723	**13746**	**36700**	1841	**3856**			
Calendar Year	2007	**52793**	6394	5062	2584	3803	3987		
	2008	17600	5941	4842	**2760**	2209	**4731**	**14812**	
	2009	31728	5608	5826	2166	1946	2682	4388	**10142**
PAGE LENGTH	2002	2406							
	2003	4012	3168						
	2004	6335	4511	2079					
	2005	9881	6904	3461	1699				
	2006	13084	7930	4755	2746	3024			
Calendar Year	2007	17376	9845	6087	3915	4674	3214		
	2008	18738	11293	7720	5010	5716	4785	7352	
	2009	21414	12921	11059	5791	6319	6426	8573	7838

Table 3: **Content Evolution at the Page-Level.** The table documents the evolution of the average page-level number of users, edit-distance and page-length by page vintage and calendar year.

Dependent Variable	Number of Users	Edit-Distance	Capped Edit-Dist.
S.D. of the DV	0.8138	8265	1025
Page Length	0.204***	277.9***	116.5***
	(0.054)	(105.5)	(43.4)
Page FEs	Yes	Yes	Yes
Week FEs	Yes	Yes	Yes
Observations	265706	265706	265706
Pages	1267	1267	1267
Weeks	433	433	433

Table 4: **The Effect of Page-Length on Editing Activity.**

Sample	Full Sample	Full Sample	Full Sample	Full Sample	Large Edits Only	Full Sample
Dependent Variable	Number of Users	Number of Users	Number of Users	Number of Users	Δ Number of Users	Number of Users
Page Length (Unit: 10,000 characters)	0.204***	0.137***	0.147***	0.086***		0.186***
	(0.054)	(0.040)	(0.050)	(0.024)		(0.048)
Page Length (Unit: 10,000 characters)					0.187***	
					(0.068)	
Cumulative Edit-Distance (Unit: 100,000 characters)						0.007
						(0.005)
Page-Specific Time-Trend:						
Linear	No	Yes	Yes	Yes	No	No
Square	No	No	Yes	Yes	No	No
Cubic	No	No	No	Yes	No	No
Page FEs	Yes	Yes	Yes	Yes	No	Yes
Weeks FEs	No	No	No	No	No	Yes
Observations	265706	265706	265706	265706	2227	265706
Pages	1267	1267	1267	1267	690	1267
Weeks	433	433	433	433	335	433

Table 5: **Robustness Check: Page-specific Time-trends.**

	Baseline		3 month lag			6 month lag	
Estimation Method	OLS	OLS	IV 1st Stage	IV 2nd Stage	OLS	IV 1st Stage	IV 2nd Stage
Dependent Variable	# Users	# Users	Page Length	# Users	# Users	Page Length	# Users
Page Length	0.204***	0.199***		0.195***	0.189***		0.203***
	(0.054)	(0.054)		(0.059)	(0.052)		(0.066)
Lagged Page Length (3 Months)			0.846***				
			(0.053)				
Lagged Page Length (6 Months)						0.697***	
						(0.109)	
Page FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	265706	242872	242872	242872	226401	226401	226401
Pages	1267	1267	1267	1267	1267	1267	1267
Weeks	433	420	420	420	407	407	407

Table 6: **Robustness Check: Correlated Information Shocks** Lagged instruments are used in all IV-specifications. This alters the sample because lagged values are not defined for a set observations in the beginning of each page's time series. The OLS is replicated each time for the reduced sample for which the instrument is available.

Dependent Variable	Number of Users	Number of Users	Number of Users
Page-Length	0.204***		
	(0.054)		
Page-Length		0.495***	0.696***
* 1(Vintage==2002)		(0.092)	(0.148)
Page-Length		0.080***	0.167*
* 1(Vintage==2003)		(0.012)	(0.094)
Page-Length		0.099***	0.406***
* 1(Vintage==2004)		(0.018)	(0.114)
Page-Length		0.113***	0.394***
* 1(Vintage==2005)		(0.011)	(0.110)
Page-Length		0.100***	0.360***
* 1(Vintage==2006)		(0.035)	(0.112)
Page-Length		-0.024	0.290**
* 1(Vintage==2007)		(0.052)	(0.122)
Page-Length		0.010	0.318***
* 1(Vintage==2008)		(0.029)	(0.120)
Page-Length			-1.013*
* 1(Year==2002)			(0.552)
Page-Length			-0.638***
* 1(Year==2003)			(0.243)
Page-Length			-0.337***
* 1(Year==2004)			(0.121)
Page-Length			-0.126
* 1(Year==2005)			(0.082)
Page-Length	Omitted		
* 1(Year==2006)	Category		
Page-Length			-0.009
* 1(Year==2007)			(0.048)
Page-Length			-0.240**
* 1(Year==2008)			(0.105)
Page-Length			-0.331***
* 1(Year==2009)			(0.118)
Page FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
Observations	265706	265706	265706
Pages	1267	1267	1267
Weeks	433	433	433

Table 7: **Heterogeneity Across Page-Vintage and Calendar Year.** We cannot allow for a full set of both vintage and year interaction effect. We therefore omit one year interaction for the peak-year 2006, but keep all vintage interaction terms. Standard errors are clustered at the page-level.

	(1)	(2)	(3)	(4)
Dependent Variable	Addition/ Deletion Metric	Fraction of Reverted Edits	Edit-Distance Per User	Capped Edit-Distance Per User
Mean	0.413	0.083	394	343
S.D.	0.621	0.246	2513	1342
Page Length	-0.025***	0.008*	-28.243	-76.724
	(0.007)	(0.005)	(133.562)	(74.562)
Page FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Observations	34305	34305	34305	34305
Pages	1267	1267	1267	1267
Weeks	415	415	415	415

Table 8: **Change in Editing Behavior as a Function of Page-Length.** Note that we can only use page/week combinations with at least one edit in this regression. The number of observations is accordingly smaller than in our baseline regression.

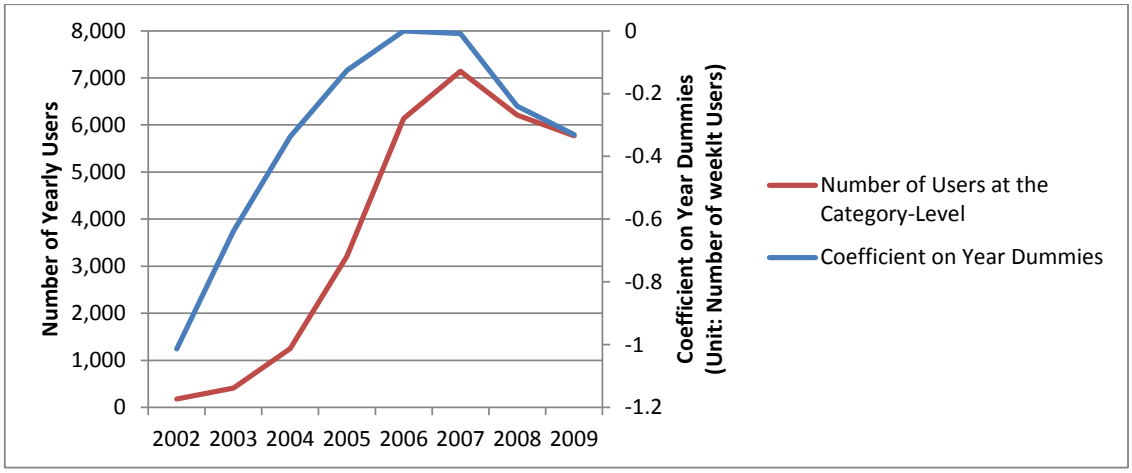


Figure 1: **Time-Series of the Spill-Over Effect and the Total User Pool**

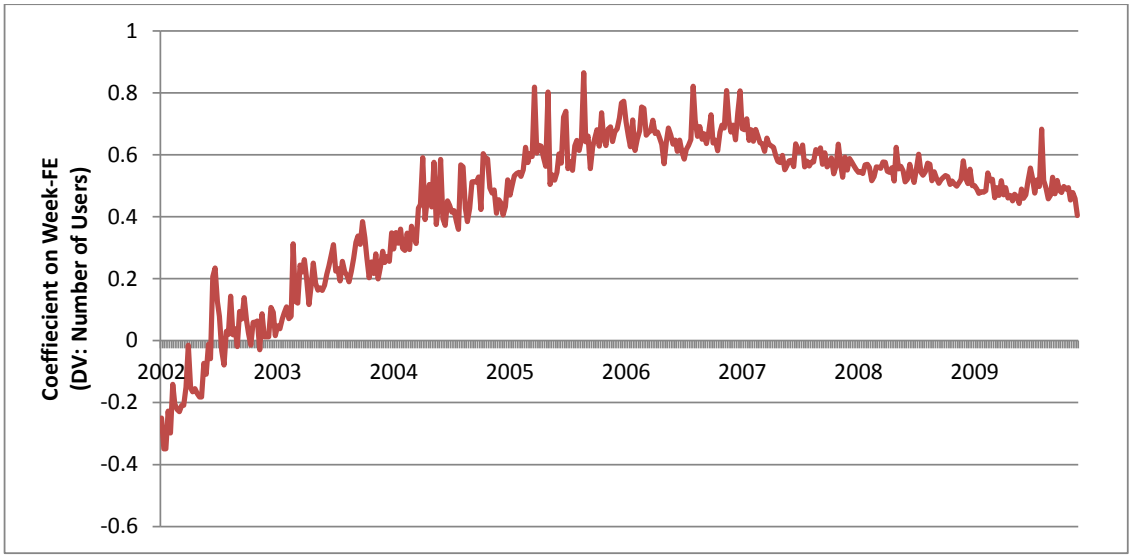


Figure 2: **Flexible Estimates of the Effect of Time on Edit Distance and the Number of Users**

A Appendix: Data Construction

A.1 Edit-Distance Calculation

Our edit-distance metric is a measure of dissimilarity between two character strings. We use the Levenshtein edit distance that is one of the most common algorithms for calculating string dissimilarities. The procedure is relatively complex to implement and computationally heavy. Therefore, we used a Python code from the google-diff-match-patch software package which is a relatively mature and well-tested implementation of Levenshtein algorithm.

A.2 Page Selection

In order to only use pages belonging to the Roman Empire we manually investigated all 1571 pages linked to from the Roman Empire category page. Through this process we identified 168 pages that were incorrectly categorized. The main goal of our selection was to eliminate pages which involve more recent events which our about the Roman Empire in a more narrow sense. The reason for this was to end up with a set of pages that contained purely historic content and therefore would not be subject to major changes in the knowledge regarding the topics covered. We therefore maintain pages on historical figures for instance which one might primarily assign to a different category. This includes for instance religious figures such as Saint Peter. Also, we keep pages both on Antique Rome as well as the Holy Roman Empire. We eliminate all pages on video games, movie and books. Furthermore our original list contains many geographic locations (cities, counties etc.). We maintain all denominations which have ceased to exist, but drop all location whose name is still in use currently. For example, we drop the page Bremen (the city in Germany), but keep Archbishopric of Bremen (a region that did exist during the Holy Roman Empire). Through this process we eliminate 168 pages and are left with a final set of 1403 unique pages.

B Appendix: Tables

Year	Number of Pages Created	Number of Users	Number of Edits	Cumulative Edit Distance (Unit: Characters)	CAPPED Edit Distance	Edits Per User	Edit Distance Per User	CAPPED E-Dist. Per User
2002	85	182	556	394,967	338,981	3.05	710	610
2003	72	414	973	527,520	486,815	2.35	542	500
2004	121	1,252	2,714	1,100,098	949,354	2.17	405	350
2005	337	3,215	7,390	4,412,004	3,000,223	2.30	597	406
2006	216	6,138	12,622	9,361,682	4,436,502	2.06	742	351
2007	239	7,138	13,874	8,005,666	5,166,570	1.94	577	372
2008	197	6,213	12,874	7,621,270	6,445,102	2.07	592	501
2009	136	5,768	13,122	7,539,501	5,727,252	2.27	575	436

Table B1: **Content evolution at the category level: More descriptive statistics.** The cap for the "capped edit-distance" variable is implemented at 10,000 characters. 97.5 percent of edits are below this threshold.

		Year in which the page was started							
		2002	2003	2004	2005	2006	2007	2008	2009
CAPPED EDIT-DISTANCE	2002	3988							
	2003	2752	3589						
	2004	6233	1932	2346					
	2005	18990	5434	3043	1972				
	2006	22360	**8741**	4642	1822	**3651**			
Calendar Year	2007	**22851**	5919	**4680**	2254	3566	3689		
	2008	16574	5743	4118	**2723**	2035	**4417**	**10343**	
	2009	16783	5107	5390	2064	1895	2376	3748	**8405**
ADDITION / DELETION METRIC	2002	0.53							
	2003	0.57	0.64						
	2004	0.49	0.54	0.63					
	2005	0.41	0.55	0.49	0.65				
	2006	0.40	0.50	0.40	0.48	0.57			
Calendar Year	2007	0.39	0.42	0.41	0.50	0.46	0.58		
	2008	0.37	0.36	0.42	0.46	0.41	0.37	0.47	
	2009	0.36	0.35	0.44	0.44	0.46	0.43	0.43	0.56
SHARE OF REVERTED EDITS	2002	0.01							
	2003	0.02	0.01						
	2004	0.06	0.02	0.01					
	2005	0.09	0.03	0.05	0.01				
	2006	0.23	0.06	0.16	0.04	0.03			
Calendar Year	2007	0.30	0.11	0.22	0.10	0.09	0.03		
	2008	0.31	0.13	0.25	0.07	0.07	0.04	0.04	
	2009	0.28	0.13	0.13	0.06	0.07	0.05	0.06	0.02

Table B2: **Content Evolution: Page-Length and Type of Edit.**

Year of Page Creation	Page Title	Number of Life-time Edits
2002	Roman Empire	4380
	Paul of Tarsus	3952
	Saint Peter	3323
	Pompeii	2811
	Holy Roman Empire	2059
2003	Praetorian Guard	485
	Great Fire of Rome	451
	List of states in the Holy Roman Empire	391
	Nine Years' War	386
	Peace of Augsburg	259
2004	Decline of the Roman Empire	1780
	Western Roman Empire	763
	Roman art	761
	War of the League of Cambrai	293
	Kingdom of Armenia	238
2005	Battle of Ceresole	369
	Ostsiedlung	279
	Siege of Jerusalem (70)	261
	Italian War of 1521-1526	242
	Diocletianic Persecution	236
2006	Census of Quirinius	400
	Italian War of 1542-1546	225
	Prince or Princess Belmonte	210
	Ulpiana	140
	Roman conquest of Hispania	95
2007	Late Roman army	440
	Persecution of Christians in the Roman Empire	216
	Ottoman - Habsburg wars	124
	Armorial of the Holy Roman Empire	72
	Conquest of Tunis (1535)	70
2008	Comparison between Roman and Han Empires	289
	Roman Senate	247
	Pederastic relationships in classical antiquity	135
	Alpine regiments of the Roman army	101
	Vulgar Latin vocabulary	101
2009	Philip the Arab and Christianity	68
	Legacy of the Roman Empire	49
	Principality of Stavelot-Malmedy	39
	Siege of Godesberg (1583)	37
	History of Rijeka	33

Table B3: **Title of "top 5" pages (measured by life-time edits) by year of creation.**

	Fraction	S.D.	Number of Pages
2002	0.26	0.17	85
2003	0.40	0.22	72
2004	0.54	0.22	121
2005	0.67	0.22	337
2006	0.68	0.22	216
2007	0.75	0.19	239
2008	0.78	0.21	197
2009	0.91	0.15	136

Table B4: **Page-Length Relative to Cumulative Edit-Distance at the End of the Sample Period.**