A Online Appendix (not intended for publication)

A.1 A Simple Model of Editing Behavior

To support the empirical analysis, we set up a model of editing behavior in this section. The goal of the model is to outline a simple mechanism of user interaction that leads to consecutive edits of different users influencing each other. In particular, it reflects the type of dynamics we aim to capture in the empirical analysis, which is a positive effect of past contributions on current editing behavior. Note that it is not our aim to structurally estimate the model, therefore the model parameters do not directly map into specific coefficients in the regression framework. The primary aim of the model is to structure our thinking around possible threats to identification in a way that complements the (mostly verbal) exposition in the main text.

We consider the behavior of user i on article j in time period t. We assume the content in each article can be represented in a vertical quality space as $x_{jt} \in [0, \overline{j}]$, where $\overline{j}$ denotes the article-specific maximum attainable quality level. We also assume users are homogenous with respect to their preferences over content; that is content translates into a quality metric x_{jt} that does not vary across users (i.e. there is no i subscript on the quality level).

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

$$u_{it} = -\alpha_i(x^*_{it} - x_t),$$

where $\alpha_i \geq 0$ captures how strongly the user feels about the content of the specific article and x^*_{it} denotes the user's preferred quality level. All variables are article-specific, but for ease of exposition we suppress the the j subscript. We assume $x^*_{it} \geq x_t$. Either the consumer is able to improve the article by adding content and therefore his optimal quality level lies above the current one, or he has nothing to add and $x^*_{it} = x_t$. We think of x^*_{it} as reflecting both the user's preferences as well as his ability to actually make quality improving changes to the content. A low value of x^*_{it} could be due to two reasons: the consumer is either happy with the current quality level or he finds the quality insufficient but is not able to make any improvements based on his own knowledge of the topic. The utility expression u_{it} above therefore reflects the disutility incurred by the potential editor from being able to improve quality but not doing so. It does not necessarily reflect the utility from content consumption.

To edit the article, the consumer incurs an editing cost c_i. For simplicity, we assume the cost of editing to be independent of the length of the edit. Given this setup, a consumer will optimally decide to edit the article and re-position it to the optimal quality level x^*_{it} according to his preferences if

$$\alpha_i(x^*_{it} - x_t) > c_i$$

If the user decides to contribute, the quality level at the beginning of the next time period is altered: $x_{t+1} = x^*_{it}$. If the user does not edit the article, x_t will remain at its current level.

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

$$x_{it}^* = (1+\gamma_i)x_t + \xi_{it},$$

where $\xi_{it} \geq 0$ depends on the knowledge regarding relevant content that the user had before accessing the article and observing the already existing content. In case some of this knowledge is not yet incorporated into the article, this leads to $\xi_{it} > 0$. In other words, this component captures editing behavior that is triggered by the users inherent knowledge level on the respective topic and that is unaffected by the already existing content. $\gamma_i \geq 0$ instead captures that, due to heterogeneity in users' knowledge, the existing content will make areas for further contributions salient to the user visiting the article. We therefore think of the case in which $\gamma_i > 0$ not as creating new knowledge, but allowing the consumer to access existing knowledge more easily. γ_i is the key model component that captures the cumulative nature of the editing process, i.e. the extent to which existing content triggers further contributions to the same article.

We assume two types of consumers exist

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

Type 1 represents a user that draws inspiration from the current content and will augment it purely based on the knowledge already embedded in the current stock of content. We will refer to this type also as "inspired" users. Type 2 represents a user that brings new information to the article, but is not influenced by the existing content. Each time period carries a certain probability of a user of each type arriving. We denote the probability of arrival with λ_1 (λ_2) for users of type 1 (2).

Based on the equations above we can derive the edit probability in a given time period as

$$\begin{aligned} Pr(Edit_t = 1) &= \lambda_1 Pr(\overline{\gamma} x_t > \frac{c_i}{\alpha_i}) + \lambda_2 Pr(\overline{\xi} > \frac{c_i}{\alpha_i}) \\ &= \lambda_1 F(\overline{\gamma} x_t) + \lambda_2 F(\overline{\xi}), \end{aligned}$$

where $F(\cdot)$ denotes the CDF of $\frac{c}{\alpha}$, the editing cost relative to preferences for the topic, in the user-pool. The expression above decomposes the likelihood of an edit into a separate term for each type of user. For each type, the edit probability is equal to the arrival probability (λ_1 and λ_2) times the probability of editing conditional on arrival. The latter is equal to $F(\overline{\xi})$ for type-2 users and independent of current content. For type-1 users instead, the current content level x_t increases the edit probability. Our main focus in this paper is to estimate the magnitude of the effect of current content on editing behavior, here represented by $\lambda_1 \frac{\partial F}{\partial x_t}$.

We note that the model entirely ignores social interactions between users and assumes that there is random arrival from a pool of anonymous users. Past contributions influence current editing behavior only through providing inspiration to some subset of arriving users.

However, many papers on user interaction on Wikipedia document social interaction between users. For instance, some users explicitly collaborating on an article by coordinating their editing activity. In our model, such dynamics could be captured by an edit increasing the likelihood of a knowledgable user arriving in the next time period. I.e. in the presence of social interactions between users, individual edits can trigger edits by collaborating users in subsequent time periods. This would lead to very similar dynamics in editing behavior and also a positive effect of past contributions on current editing activity. For the sake of simplicity the version of the model outlined above does not feature such a channel.

Differences in Article Popularity

To relate the model to the empirical exercise in a simple way, we assume a uniform distribution for $F(\cdot)$, which simplifies that edit-probability expression to

$$Pr(Edit_{jt} = 1) \quad = \quad \lambda_1 \overline{\gamma} x_{jt} + \lambda_{2j} \overline{\xi},$$

Note that we now re-introduce the so far suppressed article subscript j. One way to think about a regression analog to this expression is a linear probability model based on the relationship above

$$Edit_{jt} \quad = \quad \lambda_1 \overline{\gamma} x_{jt} + \lambda_{2j} \overline{\xi} + \varepsilon_{jt},$$

where $Edit_{jt} \in \{0, 1\}$ is an indicator for whether an edit happened on article j in time period (week) t. ε_{jt} is the econometric error term, which captures the specific realization of $\frac{c_i}{\alpha_i}$ which leads to an edit occurring (or not) in any given time period. It reflects the random nature of what type of user (in term of his preferences α_i and editing costs c_i) arrives on the article.

In order to correctly identify the effect of x_{jt} on the edit probability we need to control for the second term which drives editing activity in the absence of the inspiration effect. More specifically, if the arrival probability λ_{2j} is higher on longer articles, then x_{jt} is correlated with the regression error term if we do not control for article fixed effects. One possible channel for such a correlation to occur is due to popularity differences across articles. If more popular articles see more edits and are longer, this implies a positive correlation between x_{jt} and λ_{2j}. In order to address this issue, we include article fixed effects which control for across article differences in λ_{2j}.

Platform-level Growth

Next, we consider the relationship described in the previous section, with the difference that we allow the arrival rate of knowledgable users to vary over time (but for the moment not across articles):

$$Edit_{jt} = \lambda_1 \overline{\gamma} x_{jt} + \lambda_{2t} \overline{\xi} + \varepsilon_{jt}.$$

A very similar logic highlights the threat to identification here: If article length x_{jt} is higher in time periods with a higher arrival rate λ_{2t}, this (in the absence of time period fixed effects) leads to a correlation of x_{jt} with the error term. Such a correlation is likely due to the general growth trend on the platform as a whole. Articles tend to be longer later in their life and activity is also higher in later years due to the increase in popularity of Wikipedia. For this reason we include a full set of week fixed effects.

Identification with Two-Way Fixed Effects

Whether we are able to recover the causal effect of article length on editing activity depends critically on our ability to control for differences in user arrival rates across both articles and time. Differences in article popularity as well as an increase in the general user-pool over time make it likely that arrival rates differ over time and across articles. More generally, we can decompose the article- and time-period-specific arrival rate (λ_{2jt}) as

$$\lambda_{2jt} = \widetilde{\lambda}_{2j} + \widetilde{\lambda}_{2t} + \widetilde{\lambda}_{2jt},$$

which we can plug into the edit regression above:

$$Edit_{jt} = \lambda_1 \overline{\gamma} x_{jt} + (\widetilde{\lambda}_{2j} + \widetilde{\lambda}_{2t} + \widetilde{\lambda}_{2jt}) \overline{\xi} + \varepsilon_{jt}.$$

The inclusion of article and week fixed effects, allows us to control for arrival rate differences across article ($\widetilde{\lambda}_{2j}$) and over time ($\widetilde{\lambda}_{2t}$). The key identifying assumption is that any factor that might affect both arrival rates and article length doesn't vary differentially over time across articles. In other words, article length is uncorrelated with article-specific changes in arrival rates over time: $Corr(x_{jt}, \widetilde{\lambda}_{2jt}) = 0$. Our main robustness checks are all centered around investigating the validity of this assumption.

A.2 The Evolution of Editing Behavior by User Type

The analysis in this section is an extension of section (5) of the main paper, which describes how edit length per user as well as other aspects of editing behavior change with article length. Table (A1) reports all the result described in this section. The top panel replicates Table (5) of the main paper for easier reference, the middle and bottom panel provide additional regression results.

The Effect of Article Length by Users Type

To dig deeper into the nature of edits being triggered by article length increases, we analyze the extent to which the triggered edits originate from new users versus users that previously edited the same article. In order to implement this analysis, we define an edit as belonging to a returning user if an edit was made previously on the same article using the same user-name. We group the remainder of edits into edits by new users with a registered user account and new users that are only identified by their IP address. The latter case occurs when users make an edit without registering for a user account.[1] We note that this categorization is not entirely without problems. First, IP address identification is not very precise. For instance, the same user could log on from different computers and therefore have different IP-addresses associated with his edits. Secondly, IP-addresses can change over time for the same device. We might therefore miss some returning users that appear under different IP-addresses in our data. Our sense is that frequent contributors do usually have a user account, so this issue might not be very severe. However, there is nothing in the data that allows us to directly back up this assertion. Second, even a user with a Wikipedia user account could in principle change his username and thus appear as a new user in our data although he previously made edits on the same article. Both issues would lead us to identify a smaller number of returning users than there actually are. We discuss below how this affects the interpretation of our results.

We assess the effect on edits by user type in two ways. First, we separately regress the weekly number of each user type on article length and the usual set of controls in columns (1) to (3) of the middle panel in Table (A1). Doing so we find an effect of 0.098 of article length on the number of users with an IP address as well as an effect of 0.058 and 0.048 on the number of new and returning users respectively. All effects are statistically significant. Relative to the effect on the total number of users of 0.203, this split implies that about 50% of the triggered edits originate from users with only an IP address. About 25% of edits originate from each of the other two groups. While this illuminates which types of users are making edits when article length increases, it does not describe whether the proportion of user types is systematically different on articles of different length. In order to shed light on this aspect, we regress the *fraction of users* of each type on article length (plus controls) separately in columns (4) to (6) of the middle panel. The results of those regressions show that the fraction of users with an IP address is unaffected by article length, but there is an almost one-to-one

[1] There are only very few repeat edits from the same IP address on the same article, we therefore do not treat this type of edit as a separate category. Instead we include those edits into the returning user category.

substitution between new and returning users with a larger share of edits by returning users on longer articles. 10,000 additional characters of article length increase the share of returning users by 1.6 percentage points. Relative to an average share of returning users of 25 percent and an increase in the share of about 10 percentage points between 2002 and 2009, this is a relatively modest effect.

Changes in the Type of Edits by User Type

In a next step, we explore the effect of article length on the type of edits being made separately for each user type. The results in the top panel of Table (A1) show that article length does not influence edit distance per user and has only a small effect on the extent of content addition/deletion and the share of reverted edits. As we show in this section, these small and insignificant effects for the average user mask some interesting compositional changes. As it turns out, different types of users change aspects of their editing behavior in opposite directions as article length increases. In order to analyze the effect of article length by user type, we use the fraction of users of each type (for each article/week) as computed in the previous section and interact the fractions for all three types with article length.[2] This allows us to trace out how different dimensions of editing behavior change with article length depending on what type of user is making the edit. The regressions, results from which are reported in the bottom panel, mirror the ones in the top panel and decompose the average effects in the top panel by user type.

In terms of edit distance per user, we find that this metric increases for returning users but decreases for the other two types. None of the three terms is significantly different from zero. However, the coefficients for IP and new users are both significantly different from the effect on returning users at the 1 and 5 percent level respectively. The results are slightly stronger when using the capped edit distance measure in column (2). The negative effects for IP-address and new users are significantly different from zero now and the difference relative to returning users is more pronounced. This pattern suggests some re-allocation of activity toward returning users on longer articles. Similarly, when decomposing the effect on the addition/deletion metric, there are some interesting differences across types. Here we find that the increase in content deletion relative to content addition originates from new and returning users, but behavior of IP-address users is unchanged. Finally, we find that the increase in reverted edits is entirely driven by a strong rise in reverted edits for IP-address users, whereas we actually see a slight decline for the other two types. The increase in edits being reverted for IP-address users is fairly substantial and constitutes about 25 percent of the standard deviation of the variable.

Overall, the results from these regressions paint a picture that the amount of editing activity as well as the impact on article content shifts toward returning users as articles grow. As we saw in the previous subsection, there is a slight increase in the fraction of returning users. Furthermore, returning users make longer edits and their edits are less likely

[2] When using all three interactions we cannot include article-length on its own as well. This is due to the fact that the fractions for all three types add up to one.

to be reverted as article length increases. Finally, returning users are more likely than other users to delete content on longer articles. Although the effects are not statistically and/or economically strong along all of those dimensions, the results do indicate a stronger influence on content creation by established users on longer articles, which is consistent with the notion of Wikipedia becoming increasingly hostile towards new users posited by **?**. Also note that due to the fact that assigning users to the different types is imperfect, we are likely to have included some returning users incorrectly in the other two categories. If this is indeed the case, we might be underestimating the shift in activity toward returning users.

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Edit-Distance Per User	Capped Edit-Dist. Per User	Addition/ Deletion Metric	Fraction of Reverted Edits		
Mean	413	317	0.460	0.083		
S.D.	2812	1213	0.629	0.248		
Article Length	-55.957	-73.383	-0.010**	0.011**		
	(134.237)	(66.667)	(0.005)	(0.005)		
Observations	33,953	33,953	33,953	33,953		
Dependent Variable	Number of IP Users	Number of New Users	Number of Returning Users	Fraction of Edits by IP Users	Fraction of Edits by New Users	Fraction of Edits by Returning Users
Article Length	0.098***	0.058***	0.048***	0.000	-0.017***	0.016***
	(0.034)	(0.012)	(0.010)	(0.004)	(0.004)	(0.005)
Observations	247,002	247,002	247,002	33,953	33,953	33,953
Dependent Variable	Edit-Distance Per User	Capped Edit-Dist. Per User	Addition/ Deletion Metric	Fraction of Reverted Edits		
Article Length * Fraction of IP Users	-122.356	-132.138**	0.004	0.061***		
	(112.240)	(61.667)	(0.005)	(0.009)		
Article Length * Fraction of New Users	-149.107	-162.254**	-0.010**	-0.008***		
	(196.008)	(79.427)	(0.004)	(0.003)		
Article Length * Fraction of Returning Users	92.181	64.381	-0.022***	-0.007***		
	(94.258)	(56.696)	(0.008)	(0.002)		
Observations	33,953	33,953	33,953	33,953		

Table A1: **Change in Editing Behavior as a Function of Article Length.** The unit of observation is a week/article pair. Standard errors are clustered at the article level. The dependent variable is defined only for article/week combinations with at least one edit in all regressions (except for the first three columns in the middle panel which are based on the full sample). The number of observations is accordingly smaller than in our baseline regression.