Authority Bias *

Marit Hinnosaar† Toomas Hinnosaar‡

August 2015

Abstract

People at positions with formal authority are often expected to make better decisions and fewer mistakes, and therefore their opinions and contributions are given higher weight. This can be an equilibrium effect: people may be selected to the positions with formal authority because of their knowledge or skills. But respect for authority could also be a behavioral bias. These two explanations have very different implications. Our goal is to measure the authority bias, which we define as the difference between perceived and true quality of contributions by people with formal authority. Identifying the authority bias is complicated by the fact that almost always the observable outcomes include both explanations. We propose a method of identifying the authority bias that allows us to separate it from the equilibrium effect. We estimate the authority bias using a novel dataset from Wikipedia. In Wikipedia, editors at high-rank positions are treated differently, and there is high regional variation. Our preliminary estimation results indicate that the authority bias does not exist in Western Europe, but is large in Eastern Europe. The authority bias more than doubles the time needed for the mistakes made by high-rank editors in Eastern Europe to be corrected.

Key words: media economics, Wikipedia, behavioral economics, authority, social learning.

1 Introduction

The likelihood of questioning authority is observed to differ across countries. This could be an equilibrium outcome that in some countries, those in the positions with authority are less likely to make mistakes, and therefore, it is seldom questioned whether they are being correct,

*We would like to thank Shane Greenstein for extensive comments and helpful discussions. The paper has benefitted from suggestions made by Alessandro Pavan, Asher Wolinsky, Chris Vickers, Ignacio Franceschelli, Javier Donna, Jeff Ely, Jose Miguel Abito, Ka Hei Tse, and Nicola Persico. We are also grateful to the audiences at EARIE (Rome), ESEM (Malaga), Internet Politics & Policy workshop (Oxford), and Wikipedia workshop (Mannheim).

†Collegio Carlo Alberto, marit.hinnosaar@gmail.com.

‡Collegio Carlo Alberto, toomas@hinnosaar.net.

while in other countries the opposite happens. Alternatively, the heterogeneity in questioning authority could be an exogenous cultural difference. Separating these two reasons of why authority is not questioned, has important policy implications.

We study questioning authority in the example of Wikipedia editing. Anyone can write and edit Wikipedia articles, but there exists hierarchy among Wikipedia editors. There are unregistered and registered users, and among the last group, there is a couple of thousand users, so called Wikipedia administrators, who are given substantial authority. We are interested in how the rest of the Wikipedia editors respect Wikipedia administrators' authority. Specifically, how likely they are to correct administrators' mistakes, and more importantly, whether this is an equilibrium outcome.

We build a model to describe the behavior of Wikipedia editors that allows for the possible authority bias, where the edits made by those at high rank positions are treated as more likely to be correct than they actually are. In the model, Editor's objective is to maximize the probability that when he leaves the page the page is correct. But verifying the information on the page is costly, and even when verifying, editors can make mistakes. Therefore, each editor chooses to verify the information only if he suspects that the page is incorrect, the cost of verification is small, and he is sufficiently certain that he won't make a mistake himself. Note that in the model, we allow the state of the world to change in time. This implies that even if the edit was initially correct, it will become outdated eventually, and therefore, there will always be new edits.

First, we look at the equilibrium outcome without any authority bias. We assume that there are two types of editors, who differ only by their probability of making mistakes. We call the ones who are less likely to make mistakes administrators, and the other ones regular users. This implies that, in equilibrium, administrators edit more often, because their expected payoff from editing is higher compared to the regular users, as they are less likely to make mistakes. Additionally, by assumption, administrators' edits are more often correct, and therefore, editors who arrive to the page after them, are less likely to edit the page. Hence, even without any bias in perception, administrators behave and are treated differently. This difference in treatment is what we call the equilibrium effect.

However, similar effects arise when other editors simply believe that the administrators make mistakes less often. The difference in the actual and perceived probability of making mistakes, is what we call the authority bias. Specifically, we assume that only regular users can potentially have authority bias towards administrators. In the model with authority bias, administrators are again more active than Wikipedia regular users. This is not because

administrators edit more than they should, but because regular users edit less than they should compared to the world without authority bias. Regular users edit less, because they believe administrators' edits are more likely to be correct, and therefore, when they arrive to a page edited by an administrator, they are less likely to edit.

The key observations that allow us to distinguish in the data between the equilibrium effect and authority bias, are the following. Likelihood of editing is affected by the perceived probability of making mistakes. Namely, regular user chooses whether to edit a page following an administrator, based on his belief about the probability of the administrator making mistakes. But whether or not the edit is correct depends only on the actual probability of making mistakes.

Our dataset consists of full editing histories of the English language Wikipedia pages of capitals and a couple of largest cities of each country in Europe, and some characteristics of Wikipedia users. The reasons we concentrate on this sample are the following. First, restricting attention to only the English language Wikipedia allows us to disregard differences in the rules of editing across Wikipedias in different languages. Our source of cross country variation comes from the city pages from different countries. We assume that users who edit a page of a city in some country, are more likely to be familiar with the cultural norms of that country (and we show that this assumption is reasonable in the dataset). Most of our empirical analysis concentrates on a specific type of fact, city population. The reasons we use this fact include: it is well defined, noncontroversial (editors do not have personal incentives to misrepresent the fact), and changes over time.

We first provide preliminary evidence that Wikipedia administrators are treated differently and part of the differential treatment comes from the authority bias. For this we use the institutional detail in Wikipedia, where some edits are marked as questionable. We proceed in two steps. In the first step, we look at the likelihood of edit begin marked as questionable. We argue that if administrators' edits are marked as questionable less often, then this captures the aggregate effect (could be either because of the equilibrium effect or the authority bias). We find that indeed there is evidence of differential treatment of administrators. In the second step, we look at the likelihood that edit is changed after being marked as questionable. We argue that if editors do not mark edits as questionable randomly, but instead have some unobserved information and choose the edits optimally, then without authority bias, the likelihood that edits will be changed after being marked as questionable should equal for both types. But with positive authority bias, conditional on begin marked as questionable the edits by administrators are more likely to be changed. Based on this, we find some evidence

of authority bias in our sample.

Then we go on to estimate the model using data on the edits from Wikipedia city pages, concentrating on changes in population measures. Our preliminary estimation results indicate that the authority bias does not exist in Western Europe, but is large in Eastern Europe. The authority bias more than doubles the time needed for the mistakes made by high-rank editors in Eastern Europe to be corrected.

Our findings have important policy implications. Not questioning authority can lead to dire consequences, unless this behavior is an equilibrium outcome. In our specific example, authority bias leads to a longer time for mistakes to be corrected in Wikipedia. But such finding also hints that perhaps the authority bias exists in other areas outside Wikipedia which would be important in management, investment, innovation, health care etc.

Literature: There is a growing literature on how culture and economic outcomes are related.[1]. Cross country differences in respecting and not questioning authority have been found in several studies both in management science and psychology. One of the first to document these was Hofstede (1980) based on a large scale survey in IBM international offices in more than 30 countries. The survey was used to construct a Power Distance Index that measures the extent to which the less powerful members of organizations and institutions accept and expect that power is distributed unequally. It is based on the survey questions related to how comfortable employees are in expressing critique of those higher in the decision making hierarchy, and respecting power of others simply based on hierarchical positions. Hofstede found that the power distance index had high scores indicating more respect for authority in Asia, Latin-America, Africa and Arab countries, and mostly low values indicating low respect for authority in Europe, and United States was somewhere in the middle. Among European countries, the index had lower values in the north, and higher values in Southern and Eastern Europe. Data on the Hofstede's Power Distance Index measures have been later collected in several independent surveys, extending the index database to almost 80 countries, and covering people from many different firms and professions with different education and decision making levels.[2]

[1]Most of the studies identify the impact of cultural differences either in laboratory settings doing experiments in different countries, or looking at immigrants and using the information about their country of origin. For an overview of the studies on the impact of culture on economic outcomes, see Guiso, Sapienza, and Zingales (2006) and Fernández (2011)

[2]For an overview of the studies related to Hofstede's index, see Kirkman, Lowe, and Gibson (2006). Based on the survey, Hofstede constructed several indexes, which have often been used as inputs in the economics literature. For example, recently Gorodnichenko and Roland (2011) found that for long-run growth, the most important among cultural dimensions, is the individualism-collectivism index.

In psychology and computer science literature, authority in Wikipedia has been studied by Pfeil, Zaphiris, and Ang (2006) and Hara, Shachaf, and Hew (2010), who both looked at the correlation between Hofstede's Power Distance Index, and Wikipedia editors behavior. Pfeil, Zaphiris, and Ang (2006) compare editorial history in case of one page (games) in four different language Wikipedias, and find that in Japan and France, which are countries with high power distance index values, compared to Germany and the Netherlands, there is less deleting actions (deleting a link or text) on the respective language Wikipedia pages. Hara, Shachaf, and Hew (2010) compare discussions in four different language Wikipedia talk pages and find that in Japanese and Malay Wikipedia, users tend to be more polite in discussions, than in English and Hebrew Wikipedia, which correlates with Hofstede power distance index values. They also found that conflict and disagreement in discussions was more frequent in English and Hebrew Wikipedia, but the differences were not statistically significant. Different from our study, these papers do not consider the equilibrium effect of why there is less deletion and less conflict in some countries.

Our paper adds to the few recent papers in the economics literature that use data from Wikipedia. Zhang and Zhu (2011) analyze the impact of group size on the incentives to contribute to public good, Greenstein and Zhu (2012) analyze the Democrat/Republican bias in Wikipedia, Piskorski and Gorbatai (2011) study norm violations, Ransbotham and Kane (2011) study how collaboration affects information quality.

2 Data and Institutional Background

2.1 Institutional Background: Wikipedia Administrators

Although, anyone can write and edit Wikipedia articles, there is hierarchy among Wikipedia users. At the lowest level there are anonymous (unregistered) users who have the most limited access. At the next level, there is about 15 million registered users, who can start new articles and edit almost all articles. At the top of the hierarchy, there is a small number (a couple thousand) users who can perform powerful functions. We concentrate on a group of those at the top of the hierarchy, Wikipedia administrators. This category of users can among other things block other editors from editing any Wikipedia article, and protect a specific article from being edited by other editors. Currently, the English language Wikipedia has about fifteen hundred administrators.

The powerful administrative functions given to the administrators, do not directly imply that the administrators should have more authority when they edit articles. But Wikipedia

itself states in the history of Wikipedia administrator's status that in the early stages, it was expected that "... administrators should be a part of the community like other editors, with no special powers or privileges when acting as editors. /.../ However, Wikipedia's worldwide cultural impact and visibility grew in the intervening years, and as the community grew with it, the role of administrators evolved. Standards for adminship have risen considerably and the community generally holds administrators to a higher standard of editorial and interpersonal conduct." More importantly, later we do find evidence in our data that administrators are treated differently by other users.

When one edits a Wikipedia article then he sees who has written the part that he wants to delete or change. The concern might be that unregistered users do not know how to distinguish between administrators and other registered users. Therefore, in the empirical analysis we mostly look at how registered users treat administrators, compared how they treat the other registered users.

2.2 Overview of Data

Our dataset consists of editing histories of Wikipedia pages. Specifically, our empirical analysis focuses on articles on cities in the English language Wikipedia. This is done for two reasons. First of all, restricting attention to only the English language Wikipedia allows us to disregard institutional differences, namely the differences in the rules of editing across the Wikipedia projects (Wikipedias in different languages). On the other hand, our source of cross country variation are the city pages from different countries. In the empirical analysis, we assume that users who edit a page of a city in some country, are more likely to be familiar with the cultural norms of that country, and we show in Appendix A why this assumption is reasonable in this dataset.

The second reason of using data from the pages of cities, is that all those pages include a specific type of information that we are analyzing in detail, namely the population of the city. In our analysis we are focusing on the population numbers for several reasons. First, it is well defined and included in the same way on all city pages. Second, it is a type of information that changes over time. Because there is no convergence to one particular true number, there is always exogenous reason to verify and update the numbers occasionally. This also means that it does not matter when the Wikipedia page for this particular city was created. Finally, it is a number (or a set of numbers) and therefore easier to collect and analyze than textual information.

Sample of Wikipedia Pages: Our dataset consists of full editing histories of the English language Wikipedia pages in our sample and some characteristics of the Wikipedia users that have edited these pages. Specifically, we look at the capitals and 2 largest cities of each country in Europe, and here the size of the city is defined by its population. That is, we take either 2 or 3 cities from each country, depending on whether the capital is among the 2 largest cities. [3]

Editors: Technically, there are 4 types of users in Wikipedia who edit articles: unregistered users, bots, administrators, and registered users who are neither bots nor administrators. We mostly look at only two categories of users: administrators and registered users who have edited the Wikipedia pages in our sample. In the estimation, registered users are the only ones we include among *Regular Users*.

Descriptive Statistics of Wikipedia Edits in the Preliminary Analysis: In the preliminary analysis we look at the sample of all edits on city pages. We take each section on a page as a separate observation. This is done, since we want to look at whether the edit is marked as questionable, and in practice section is more or less the smallest unit on a page that is marked as questionable.[4] Descriptive statistics of the sample are presented in Table 3. About 10% of the edits in the sample are created by Administrators. Almost 9% of the edits in the sample are marked as questionable.

Descriptive Statistics of Edits of City Population In the structural estimation we consider a specific type of information, namely city population in the information box. An observation in this dataset is an edit of a particular population measure (for example population of metro area). For each observation we observe the user who wrote it, time it was written, how long it took before someone updated the information box again, who updated it and whether she updated this population measure or not. Descriptive statistics are presented in Table 4.

[3] Our choice of cities is based on the following lists:
`http://en.wikipedia.org/wiki/List_of_national_capitals` and
`http://en.wikipedia.org/wiki/List_of_largest_cities_and_second_largest_cities_by_country`

[4] Sometimes a sentence is marked as questionable, but this is quite rare.

3 Preliminary Evidence

In this section, we present evidence on the differential treatment received by Administrators and authority bias, using the institutional detail in Wikipedia, where some edits are marked as questionable. We proceed in two steps, first testing the existence of aggregate effect (equilibrium effect plus authority bias) and second testing the existence of authority bias.

In the first step, we look at the likelihood of edit begin marked as questionable. If either the administrators write fewer questionable edits (equilibrium effect is positive) or the editors who choose which edits to mark as questionable simply believe that the administrators write less questionable edits (authority bias is positive), then we would observe that the proportion of edits marked questionable is lower for the administrators than for the regular users. Therefore the difference between the rates with which administrators' and regular users' edits are marked as questionable captures the aggregate effect.

In the second step, we look at the likelihood that edit is changed after being marked as questionable. Assuming that the editors who mark edits as questionable are doing it with some private information about the edits (otherwise there would be no need for marking), then the optimal way to mark edits as questionable is such that the posterior probability that edits actually need changing is the same for both types. If this was not the case, then either the quality of Wikipedia could be increased by marking more edits questionable or the editors who verify the marked edits would start to ignore marked items by one of the types. However, if the editors who mark the edits questionable believe that the administrators write less questionable edits than they actually do (authority bias), they will mark too few of administrators' edits questionable and therefore those edits that still get marked are more likely to be incorrect than in optimum. At the same time, if the editors do not have biased view towards regular users' edits, the probability that their edits get changed after being marked as questionable remains unchanged. This implies that with positive authority bias, conditional on begin marked as questionable the edits by administrators are more likely to be changed than the edits marked as questionable by the regular users.

The above argument holds even if the editors who verify the marked edits are also somewhat biased as long as their bias is smaller. The technical details are provided in Appendix B.

3.1 Aggregate Effect of Differential Treatment

First, let's look at the aggregate effect. So, the question is the following. Are the high-rank editors treated differently?[5] To test this, we run the following regressions using the sample of edits, where an edit is defined as a change in a paragraph on the page. The dependent variable is the indicator variable whether edit was marked as questionable. We regress it on an indicator variable *Administrator*, that takes value 1, when the edit was made by an Administrator, *Number of Revisions* is thousands of revisions of the whole page in 12 months after the edit is made, and *Editing experience* of editor on city pages that are in our sample so far.

The results from this regression are presented in column 1 in Table 5. The most important numbers in the table are in the first row, where we see that Administrators are questioned less often. When page has larger number of revisions, then the likelihood of an edit being questioned is smaller. Editing experience of the editor decreases the likelihood of his edit being questioned. Regression in Column 2, in addition to the above includes country fixed effects. Column 3 instead of the fixed effects includes country characteristics like *Internet Access* which is the percentage of population using internet and *Education* which is the average schooling received by males. The result that administrators' edits have lower probability of being questioned remains unchanged.

Columns 4 and 5 in Table 5 demonstrate that the differential treatment is larger in Eastern Europe. These regressions include an indicator variable for Eastern Europe, and an interaction term for Administrator and Eastern Europe. Edits in Eastern Europe are more likely to be questioned. But the interaction term is negative, which implies that in Eastern Europe, administrators' edits have relatively lower probability of being questioned.

Correlation of the aggregate effect and Hofstede Power Distance Index: From Table 6 we see that the differential treatment is related to other measures of authority, namely to the Hofstede's Power Distance Index. The regressions are the same as in table 5, except that we include Hofstede's Power Distance Index, and its interaction term with *Administrator* variable. The larger values of the index characterize countries with more respect for authority. The most interesting term for us 6 is the index and administrator interaction term, which has negative coefficient estimate. That is, we find that in countries that respect authority more according to the Hofstede's Power Distance Index, the differential treatment of administrators is larger. Note that the coefficient estimate of administrator variable itself becomes positive

[5]Appendix C presents additional evidence on the differential treatment received by Administrators.

but insignificant.

3.2 Authority Bias

In the second step, we test for the existence of authority bias by looking at the edits that are questioned. Remember that without authority bias, conditinal on being questioned, high-rank editors edits should not be different from the regular edits. But with authority bias administrators edits should be more likely to be changed.

Table 7 presents estimates from the regression where the dependent variable is the time until edit is changed conditional that it was marked as questionable. Regression in column 1, includes an indicator variable whether the edit is made by an Administrator, *Number of Revisions*, and editor's *Editing experience*. The estimates show that administrators edits survive a shoter time after being marked as questionable. Although, when country fixed effects and country characteristics of internet access and education are added to the regression in columns 2 and 3, then the effect becomes smaller and insignificant.

Regressions in columns 4 and 5, include an indicator variable for Eastern Europe, and an interaction term for Administrator and Eastern Europe. Estimates in column 4 demonstrate that in Eastern Europe, administrators' edits survive relatively shorter time. Although, when internet access and education are added to the regression in column 5, then the estimated administrator's effect becomes insignificant.

4 Model

4.1 Assumptions

Time t is continuous and the state of the world $S_t \in \mathbb{R}$ is unobservable. Initial state is S_0 and with with Poisson rate $\rho > 0$ it changes to a new value according to some unknown process, so that probability of changing to any particular number in $\mathbb{R}$ is non-atomic. Wikipedia page holds information $w_t \in \mathbb{R}$ about the state S_t. In particular, initial w_0 is correct with probability[6] $Pr(w_0 = S_0) = \mu_0 \in [0, 1]$. Information w_t changes only when it is changed by the Editors, otherwise it stays constant.

Editors arrive to a page with Poisson rate $\lambda > 0$. Since the probability of two editors arriving at the same moment is 0, we call the Editor who arrives at time t the "Editor t". The objective of each Editor is to make sure that when she leaves, the information on the

[6]This includes having pure noise if $\mu_0 = 0$ and having precise starting value $\mu_0 = 1$ as special cases.

page is correct. But (1) verifying the existing information is costly, (2) even after verification, editors sometimes make mistakes. In particular, payoff of the Editor is the following. Editor t gets value 1 when $w_t = S_t$ and 0 otherwise, minus the cost of verification. Different editors have different costs of verification and also probabilities of mistakes, which we will describe as qualities of their signals.

If Editor t chooses to do so, she verify the information on the page by receiving a private signal $s_t \in \mathbb{R}$ at a cost c_t. The cost is a private value with cumulative density function F such that its support includes $(0, \varepsilon)$ for some $\varepsilon > 0$. Conditional on true state S_t, the signal s_t is independent of everything else in the model and is such that $Pr(s_t = S_t) = q_t > 0$ and with $1 - q_t$ it takes non-atomic values in $\mathbb{R}$. We will call the probability q_t Editor t's signal quality and define its distribution later. For now it suffices to assume that it is fixed for each Editor and it becomes public every time an Editor verifies the information.

To summarize, each Editor t first chooses whether or not to receive the signal. After observing the value of the signal s_t (or not observing, if she chose not to verify), Editor t chooses whether to change the Wikipedia page or leave it unchanged. After that she exits the model.

In the analysis below we will concentrate on "Limited information equilibrium", each Editor t only observes (1) the time when the information was last verified, denoted by τ, (2) the identity of Editor τ who verified the information at τ, (3) the value w_τ at the Wikipedia page after Editor τ's verification.

In particular in this structure we make two simplifying assumptions, both limit the amount of learning in the model. First, we assume that the editors observe all verifications, including those in which it turned out that the information on the page was correct. This may not be true in practice, since when an editor finds that the whole page is perfectly correct, she may choose not to update the page at all. In our empirical analysis we assume that editors verify the information in the part of the Wikipedia page that we are looking (city information box) if and only if they change something in this part. Second, the Limited information equilibrium means in particular that the editors do not know whether previous editors' signals confirmed or did not confirm the information on the page. We discuss how the process would change without these assumptions in the next subsections.

4.2 Equilibrium behavior

Suppose that the page was last verified was at time τ, either by an Editor τ or it was the initial value $\tau = 0$, and the probability that the information was correct after the last edit

was μ_τ. Suppose now that at $t > \tau$, next Editor t arrives.

According to the stochastic process, the probability that the information verified at τ is still correct at t is $Pr(w_\tau = S_t) = \mu_\tau e^{-\rho(t-\tau)}$. If Editor t chooses not to receive the signal, then she would never change the page, since probability that any particular change is correct is negligible and therefore strictly lower than the probability that the page is still correct. Therefore, if Editor t chooses not to verify the information, she gets payoff $\mu_\tau e^{-\rho(t-\tau)}$.

Suppose now that Editor t chooses to receive signal with quality q_t for a cost c_t. Now there are two possibilities: her signal s_t either confirms the information in the page or differs from it. Suppose first that the signal confirms that page is correct. That is, $s_t = w_\tau$. Since we assume that all incorrect values have non-atomic probabilities, this means that the posterior probability that $w_\tau = s_t = S_t$ is equal to $Pr(s_t = S_t | s_t = w_\tau) = 1$. Clearly, it is optimal to not to change the value of the page, so $w_t = w_\tau$. Probability that signal confirms the value at the page is

$$Pr(s_t = w_\tau) = Pr(s_t = S_t, w_\tau = S_t) = q_t \mu_\tau e^{-\rho(t-\tau)}.$$

Now consider the case when the signal differs from the information in the page. That is $s_t \neq w_\tau$. This is because either signal or page or both are incorrect. The probability that this happens is

$$Pr(s_t \neq w_\tau) = 1 - Pr(s_t = w_\tau) = 1 - q_t \mu_\tau e^{-\rho(t-\tau)}.$$

In this case the signal can be correct only if the page is incorrect, which gives us that the posterior probability that signal is correct is[7]

$$Pr(s_t = S_t | s_t \neq w_\tau) = \frac{Pr(s_t = S_t, w_\tau \neq S_t)}{Pr(s_t \neq w_\tau)} = \frac{q_t[1 - \mu_\tau e^{-\rho(t-\tau)}]}{1 - q_t \mu_\tau e^{-\rho(t-\tau)}},$$

Depending on the relative sizes[8] of q_t and $\mu_\tau e^{-\rho(t-\tau)}$ the Editor t whose signal does not confirm the information on the page may either update it or not, but if q_t is so small that it is not optimal to use the signal, then it is clearly not optimal to receive the signal in the first place. This implies that Editor t only chooses to receive the signal if she will update the

[7] This posterior is strictly positive for all $q_t > 0$ and strictly less than q_t whenever $\mu_\tau > 0$.

[8] We can similarly compute the posterior probability that the page is correct and get

$$Pr(s_t = S_t | s_t \neq w_\tau) > Pr(w_\tau = S_t | s_t \neq w_\tau) = \frac{(1 - q_t)\mu_\tau e^{-\rho(t-\tau)}}{1 - q_t \mu_\tau e^{-\rho(t-\tau)}} \iff q_t > \mu_\tau e^{-\rho(t-\tau)}.$$

There is also strictly positive probability that both s_t and w_τ are wrong, but this does not matter in the updating decision, since changing the value of the page to anything other than s_t and w_τ is correct with zero probability.

page according to the signal. Before knowing the value of signal, the probability that it will be correct is[9] q_t, therefore if Editor t always sets $w_t = s_t$ the ex-ante probability that w_t will be correct is q_t.

Therefore the Editor t chooses the verify the page and set $w_t = s_t$ if and only if $q_t - c_t > \mu_\tau e^{-\rho(t-\tau)}$. That is, she chooses to verify the information when her signal is more precise than the current information and the difference is large enough to cover c, the cost of verification. The probability that she verifies the page is $Pr(q_t - c_t > q_t - \mu_\tau e^{-\rho(t-\tau)}) = F(q_t - \mu_\tau e^{-\rho(t-\tau)})$.

Finally we need to discuss how the posterior belief μ_t is formed. With full information, next Editors would observe whether the Editor t verified the information and whether updated the page or not and therefore their posterior would be the same as Editor t's computed above,

$$\mu_t^f = \begin{cases} 1 & w_t = w_\tau, \text{and verified,} \\ \frac{q_t[1-\mu_\tau e^{-\rho(t-\tau)}]}{1-q_t\mu_\tau e^{-\rho(t-\tau)}} & w_t \neq w_\tau, \\ \mu_\tau e^{-\rho(t-\tau)}, & \text{did not verify.} \end{cases} \tag{4.1}$$

In Limited information equilibrium, we assume that the next Editors will only observe either (1) that Editor t verified the page at time t and had signal quality q_t, (2) the value that she wrote on the page was w_t. In particular, this means that they do not know whether $w_t = w_\tau$ or not. This means that their posterior probability cannot be conditioned on the updating, which gives us posterior for the limited information case after verification is $Pr(s_t = S_t) = q_t$. The reason is simple—since Editor t ignores the value w_τ, its precision does not add any new information to the model. Therefore we get that in the Limited information equilibrium,

$$\mu_t = \begin{cases} q_t & \text{If Editor } t \text{ verified,} \\ \mu_\tau e^{-\rho(t-\tau)}, & \text{did not verify.} \end{cases} \tag{4.2}$$

From (4.2)[10] we see that when the editor did not verify the information, then it does not matter whether she arrived or not, the belief at time t would be the same $\mu_t = \mu_\tau e^{-\rho(t-\tau)}$ in both cases.

Note that we are also using the assumption that next Editors observe verifications rather

[9]Since Editor t will ignore w_τ, its precision cannot affect the probability. Alternatively, before knowing whether $s_t = w_\tau$ or not, her expectation is

$$Pr(s_t = S_t|w_\tau) = Pr(s_t = S_t|s_t = w_\tau)Pr(s_t = w_\tau) + Pr(s_t = S_t|s_t \neq w_\tau)Pr(s_t \neq w_\tau) = q_t\mu_\tau e^{-\rho(t-\tau)} + q_t[1-\mu_\tau e^{-\rho(t-\tau)}] = q_t$$

[10]The same is true for (4.1).

than simply updates here. If we would alternatively assume that the next editors only observe the fact that Editor t verified the page (and her signal quality q_t) when $w_t \neq w_\tau$, there would always be a possibility that there were several agent who verified the information in the meantime and all their signal confirmed w_τ.

In the following analysis we therefore know that the beliefs are updated according to (4.2) and each new agent verifies the information if and only if $c_t < q_t - q_\tau e^{-\rho(t-\tau)}$. This gives us two observations:

(i) Editors sometimes verify the information. Since we assume that $\rho > 0$as time passes, eventually $\mu_\tau e^{-\rho(t-\tau)} \to 0$. Editors arrive with positive probability and some of them have low costs, so the probability that at there will never be Editor t with cost smaller than $q_t - \mu_\tau e^{-\rho(t-\tau)} \to q_t$ is equal to 0.

(ii) Editors do not verify the information very often. First of course, they arrive according to Poisson process, so the probability of having several editors arriving in a short period is small. Moreover, even if it happens, it cannot be that all of them verify the information. Each time when an Editor τ verifies the information, the posterior jumps to q_τ and if another Editor arrives soon after, the information is still close to q_τ, so the editor verifies the information only if her cost is very small (which happens with low probability) or her signal has much higher quality, $q_t > q_\tau + c$.

The dynamic of the is therefore model simple. Each time when an editor verifies the information, the posterior belief that the information is correct, is relatively high and even if another editor arrives soon after, she will not find it optimal to verify the information. As time passes, it becomes less and less likely that the information is still correct and therefore more likely that next Editor finds it optimal to verify the page.

4.3 Two types of editors

So far we have not specified how the signal qualities q_t are generated. In Wikipedia, there are two types of Editors—Administrators and Regular Users. Usually the Administrators are Regular Users who are promoted after some time. Therefore it is likely that they can verify the information with somewhat higher precision.

In particular we assume that q_t can only have one of two values. Either Editor t is an Administrator and has $q_t = q_A$ or a Regular User and has $q_t = q_R$, where $q_A \geq q_R$ are model

parameters. We assume that the arrival rate of Administrators[11] is λ_A and the arrival rate of Regular Users is λ_R, so that $\lambda = \lambda_A + \lambda_R$.

Note that at time t, both the optimal behavior and the updating rule (4.2) only depend on previous Editors through two variables: q_τ and $t - \tau$. Therefore we simplify the notation by defining everything as a function of t and u, where t now describes the time since the last verification ($t - \tau$ in previous notation) and $u \in \{A, R\}$ is the type of the last Editor who verified the information.

In this notation new Editor of type θ updates the page if and only if her cost c is strictly lower than $q_\theta - q_u e^{-\rho t}$ and the updated posterior belief is q_θ if she verifies the information and $q_u e^{-\rho t}$ if she did not.

It means at (t, u) the type $\theta \in \{A, R\}$ who will verify the information arrive at rate $\lambda_\theta F(q_\theta - q_u e^{-\rho t})$ and whenever this happens, the page is verified and belief changed to q_θ. If this happens, the probability the new user changes the value of the page ($w_t \neq w_\tau$ in previous notation) is

$$Pr(\text{Change}|\theta, \text{Verified}, t, u) = 1 - q_\theta q_u e^{-\rho t}. \tag{4.3}$$

We can now make three observations about the differences between the Administrators and Regular Users.

(i) Administrators are more likely to verify the information, because for any (t, u) we have that if $c < q_R - q_u e^{-\rho t} \leq q_A - q_u e^{-\rho t} >$, so whenever Regular User would verify, Administrator would also verify (but not necessarily the opposite).

(ii) Administrators are less likely to make mistakes. The posterior after Administrator's verification, q_A, is weakly higher than the posterior after Regular User's verification, q_R.

(iii) Administrator's verifications survive longer (are less often verified). If Administrator verified the information t time units ago and new user with type θ and cost c arrives, she verifies the page only if $c < q_\theta - q_A e^{-\rho t} < q_\theta - q_R e^{-\rho t}$. Therefore she would also verify Regular User's edit (but not vice versa).

4.4 Authority

We define "authority bias" as a cognitive bias of having incorrectly high belief that the information verified by a person with formal authority is correct. Of course, Administrators

[11] Equivalently, any particular Editor is an Administrator with probability $p_A = \frac{\lambda_A}{\lambda}$ and Regular User with probability $p_R = 1 - p_A = \frac{\lambda_R}{\lambda}$.

in Wikipedia are selected to be Administrators for a reason and on average they can verify the information with objectively higher precision. As we see from the observations in previous section, this means that Administrators would behave and would be treated differently in equilibrium even in absence of any biases. To separate the differences in behavior and treatment that come from authority bias from the equilibrium effect, we add this bias to the model and show that empirical analysis can distinguish authority bias from equilibrium effects.

We assume that each agent knows her own signal quality $q_\theta \in \{q_A, q_R\}$, but the beliefs about previous verifier's signal quality are given by Table 1. That is, all beliefs are correct in all cases except when the new Editor is Regular User and previous Editor was Administrator. In this case, instead of correct belief $q_u = q_A$, the current Editor places additional probability a that the information was correct. We call a the authority bias term. If $a = 0$, the users are not biased, whereas if[12] $a > 0$ they are.

	$u = A$	$u = R$
$\theta = A$	$q_u = q_A$	$q_u = q_R$
$\theta = R$	$q_u = q_A + a$	$q_u = q_R$

Table 1: Perceived signal qualities of previous verifiers.

The authority bias affects the Editors' decisions to verify the page. In particular, at (t, A), because the previous verification was made by an Administrator, Regular Users perceive the information to be more precise than it actually is and verify the information if and only if $c < q_R - (q_A + a)e^{-\rho t}$. This means that with strong authority bias the probability to observe verifications by Regular Users soon after verification by Administrator would be relatively low, but this difference disappears as the time passes. Poisson verification rates by two types of users for any (t, u) are given by Table 2.

	$u = A$	$u = R$
Verification rate by $\theta = A$	$\lambda_A F(q_A - q_A e^{-\rho t})$	$\lambda_A F(q_A - q_R e^{-\rho t})$
Verification rate by $\theta = R$	$\lambda_R F(q_R - (q_A + a)e^{-\rho t})$	$\lambda_R F(q_R - q_R e^{-\rho t})$

Table 2: Verification rates

On the other hand, since we assume that the authority bias a is only a cognitive bias and true signal qualities are still q_A and q_R, if an agent chooses to verify the information, the probability of getting a signal that differs from the page is still the same as before, given by

[12]It is also possible to have $a < 0$, which can be interpreted as strong suspicion against authority.

(4.3), which we rewrite here for convenience as (4.4). That is, if a new user decides to verify the information, then the probability of finding a that information was incorrect is comes from the true probability rather than perceived probability.

$$Pr(\text{Change}|\theta, \text{Verified}, t, u) = 1 - q_\theta q_u e^{-\rho t}. \tag{4.4}$$

4.5 Implications

The first implication of the model is that we can now define differential treatment. Let us focus on the decision of a Regular User whether to verify the information that was verified by a previous user t time units ago. If the previous editor was a Regular User, the current user verifies the information if the cost of verification $c_t < q_R - q_R e^{-\rho t}$, whereas if the previous editor was Administrator, the condition is $c_t < q_R - (q_A + a)e^{-\rho t}$, so the difference in threshold level for the cost is $(q_A + a - q_R)e^{-\rho t}$. If this difference is strictly positive, the Administrator's edits are verified less often. Part of this difference comes from the differences in real accuracy of verifications, $q_A - q_R$, but part from the authority bias. In particular, the authority bias accounts for $\frac{a}{q_A+a-q_R}$ share of differential treatment.

The second implication is that we can now quantify the effect of authority bias to the outcomes. In particular, the interesting characteristics are how the authority bias affects the expected time until the information is verified and how it affects the accuracy of Wikipedia.

The probability that information written by an Administrator is verified exactly t time units later is $S_A(t, A)S_R(t, a)[h_A(t, A) + h_R(t, R)]$, where $h_\theta(t, u)$ is the probability that type θ user arrives and verifies information written by type u user t time units ago (hazard rate) and $S_\theta(t, u)$ is the probability that in t time units after type u user verified the information, no type θ user has arrived and verified the information (survival rate). Since $S_\theta(t, u) = e^{-\int_0^t h_\theta(\tau,u)d\tau}$, we get that

$$\begin{aligned} E[T|A] &= \int_0^\infty tS_A(t, A)S_R(t, A)[h_A(t, A) + h_R(t, A)]dt \\ &= \int_0^\infty S_A(t, A)S_R(t, A)dt = \int_0^\infty e^{-\int_0^\infty [h_A(\tau,A)+h_R(\tau,A)]d\tau}dt. \end{aligned} \tag{4.5}$$

By (4.5), the expected time until verification is decreasing in both hazard rates h_A and h_R. Verification rate by Administrators, h_A, is indepedent of authority bias a and verification rate by Regular Users $h_R(t, A) = \lambda_R F(q_R - (q_A + a)e^{-\rho t})$ is strictly decreasingin a. Therefore the expected verification time is strictly increasing function of authority bias.

Expected probability of finding correct information.

Work in progress.

4.6 Extension: unobservable verifications

In this section we discuss the equilibrium in the case when the verifications by editors are unobservable. That is, the editors only observe the edits by previous editors when they found a mistake in the page and changed the page. This complicates the analysis, since now as the time passes the information gets less precise for exogenous reasons, but this is partly balanced by the fact that it is more likely that some editors verified the information in the meantime and simply found that it is still accurate. For example when the information is constant ($\rho \to 0$) then the fact that there has not been any edits over a long period means that the information in the page has probably converged to the true value.

Work in progress.

5 Empirical Strategy

We estimate the model using data on all the revisions that changed the information box on the Wikipedia city pages. We are specifically interested in revisions that change population measures.

The key assumption we make in the empirical analysis, is that, verifying information is observable (both to Editors and the econometrician). Specifically, we assume that an Editor verifies the information about the population measures if and only if he creates a revision that changes the information box.[13]

The data is in the following form. For each revision, we observe the Editor who created the revision and the time it was created, and the Editor who verified the information in the revision and the time it was done. We also observe other characteristics regarding the page and editor, including the region. Descriptive statistics of this sample of edits was presented in section 2.2.

5.1 Parametric assumptions

We make the following assumption about the distribution of costs.

Assumption 5.1. *Editing cost c is independently and identically distributed across editors in the following way: (1) with probability γ_0 editor always edits (that is, his editing cost*

[13]The change could be a new link to the reference or a new date when it is checked, among other things.

equals minus infinity), (2) with probability $(1-\gamma_0)$ *his editing cost is distributed according to exponential distribution with parameter* $\gamma_1 > 0$.

The assumption simplifies the estimation of the model. Let's denote the verification rate, the rate at which a page is verified by Editor of type θ given state (t, u), by $h_\theta(t, u)$. These rates for different states and Editors were presented in Table 2. With the assumption above, verification rate by Administrator becomes:

$$h_A(t,u) = \lambda_A\gamma_0 + \lambda_A(1-\gamma_0)\cdot[1-\exp\left(-\gamma_1\cdot(q_A - q_u\cdot\exp(-\rho t))\right)] \tag{5.1}$$

The verification rate by Regular User in case of edits made by Regular users, takes a similar form, but in case of edits made by Administrators, there is the authority bias parameter a:

$$h_R(t,A) = \lambda_R\gamma_0 + \lambda_R(1-\gamma_0)\cdot[1-\exp\left(-\gamma_1\cdot(q_R - (q_A+a)\cdot\exp(-\rho t))\right)] \tag{5.2}$$

From verification rate we can construct the survival function, which is the probability that the edit made by editor of type u has not been verified by editor of type θ by time τ, $S_\theta(\tau, u) = \exp\left(-\int_0^\tau h_\theta(t,u)dt\right)$ For the ease of exposition, let's look at the minus logarithm of the Survival function. The survival function takes the following form (except for the case with authority bias):

$$\begin{aligned}
-\log S_\theta(\tau,u) &= \int_0^\tau h_\theta(t,u)dt \\
&= \lambda_\theta\tau - \lambda_\theta(1-\gamma_0)\cdot e^{-\gamma_1 q_\theta}\int_0^\tau e^{\gamma_1 q_u\cdot\exp(-\rho t)}dt \\
&= \lambda_\theta\tau + \frac{\lambda_\theta(1-\gamma_0)}{\rho}\cdot e^{-\gamma_1 q_\theta}\int_{-\gamma_1 q_u}^{-\gamma_1 q_u\cdot\exp(-\rho\tau)}\frac{e^{-u}}{u}du \\
&= \lambda_\theta\tau + \frac{\lambda_\theta(1-\gamma_0)}{\rho}\cdot e^{-\gamma_1 q_\theta}\left[E_1(-\gamma_1 q_u) - E_1\left(-\gamma_1 q_u e^{-\rho\tau}\right)\right]
\end{aligned} \tag{5.3}$$

where the third inequality follows by the change of variable in the integral, such that $u = -\gamma_1 q_u\cdot\exp(-\rho t)$, and E_1 denotes the Exponential Integral[14] For the case with authority bias, that is, for verification rate by Regular User in case of edits made by Administrators,

[14]Exponential integral is defined as $E_1(x) = \int_1^\infty \frac{e^{-xt}dt}{t}$. Alternatively, $E_1(x) = \int_x^\infty \frac{e^{-u}du}{u}$.

the minus logarithm of survival function is:

$$-\log S_R(\tau, A) = \lambda_R \tau + \frac{\lambda_R(1-\gamma_0)}{\rho} \cdot e^{-\gamma_1 q_R} \left[E_1(-\gamma_1(q_A + a)) - E_1\left(-\gamma_1(q_A + a) \cdot e^{-\rho\tau}\right)\right] \quad (5.4)$$

5.2 Identification

The parameters that we want to estimate are the authority bias a, updates arrival rate ρ, and for both types of users, Administrators and Regular Users, the quality of their signals q_A, q_R, and arrival rates, λ_A, λ_R, and, finally, the cost distribution parameters γ_0 and γ_1.

Recall from the previous section that the model predicts that the probability that the value on the page changes when an Editor of type θ verifies the information on the page, given state (t, u), is given by Equation 4.4. Our data allows us to construct the empirical counterparts to the probability, the left hand side of the equation. Data allows us to identify all three parameters (q^A, q^R, ρ) in the equation.[15] First, the variation in time t when Editor verifies the information given the same u and θ, identifies ρ. Then variation in the type of Editor and u for given t, identifies q_A and q_R.

The model predicts the verification rates by Editor of type θ in state (t, u), as given by equations 5.1 and 5.2. These rates are observable in our data. Let's first consider the verification rates by Administrators. Note that for given u, the verification rate increase in time t, since the editing cost threshold, up to which editor verifies, increases in time. Variation in t identifies the cost function parameters γ_0 and γ_1. First for γ_1, let's take two durations $t_1 < t_2$, which then using the parameters we know from above, give us two cost thresholds, which we denote by $\bar{c}_1 < \bar{c}_2 = k \cdot \bar{c}_1$, where $k > 1$. Then the following ratio of verification rates:

$$\frac{h_A(t_2, u) - h_A(0, u)}{h_A(t_1, u) - h_A(0, u)} = \frac{1 - e^{-\gamma_1 k \bar{c}_1}}{1 - e^{-\gamma_1 \bar{c}_1}}$$

identifies γ_1, since the left hand side is observable from data and the right hand side is strictly

[15] Moreover, since there are only three parameters (q^A, q^R, ρ), but the equation above provides infinitely many moments, for each time period $t \in [0, \infty)$ and both Editor types (A, R), the model is overidentified.

decreasing in γ_1.[16] Knowing γ_1, the ratio of verification rates at time t_1 and zero

$$\frac{h_A(t_1,u)}{h_A(0,u)} = 1 + \frac{1-\gamma_0}{\gamma_0}\left[1 - e^{-\gamma_1 k\bar{c}_1}\right]$$

identifies γ_0, since $\frac{1-\gamma_0}{\gamma_0}$ is strictly decreasing in γ_0. Knowing the cost function parameters, allows us to identify λ_A from the verification rate by Administrators. Comparing verification rates by Administrators and Regular Users in case of edits made by Regular Users, identifies λ_R. Finally, authority bias a is identified from the verification rate by Regular Users of edits made by Administrators.

5.3 Estimation

We estimate the parameters using Maximum Likelihood. We break up the likelihood function into three components, and estimate the first component separately from others. First, using data on whether user makes a change conditional that he verified, we estimate user signal qualities, and information persistence, (q^A, q^R, ρ). Then in the second stage, using data on the survival of each edit, we estimate the rest of the parameters $(\lambda_A, \lambda_R, a, \gamma_0, \gamma_1)$.

The 2nd stage likelihood has two components, one for each editor type. The likelihood is constructed using the verification rates given by equations 5.1 and 5.2 and the respective survival functions given by equations 5.3 and 5.4. The log-likelihood function for Administrator takes the following form:

$$\log L_A = \sum_i 1[AdministratorVerified_i = 1] \cdot \log h_A(t_i, u_i) + \sum_i \log S_A(t_i, u_i)$$

where $h_A(t_i, u_i)$ is the verification rate by Administrator, and $S_A(t_i, u_i)$ is the respective survival function. We construct a similar log-likelihood function for Regular User $\log L_R$ and maximize the sum of these log-likelihood functions.

[16] Right hand side of the hazard ratio is decreasing

$$\frac{\partial RHS(\gamma_1)}{\partial \gamma_1} = \frac{k\bar{c}_1 e^{-\gamma_1 k\bar{c}_1}(1 - e^{-\gamma_1\bar{c}_1}) - \bar{c}_1 e^{-\gamma_1\bar{c}_1}(1 - e^{-\gamma_1 k\bar{c}_1})}{(1 - e^{-\gamma_1\bar{c}_1})^2} < 0$$

since $ke^{\gamma\bar{c}_1} - e^{-\gamma k\bar{c}_1} - k + 1 < 0$ for $\gamma\bar{c}_1 > 0$. Note the expression above takes value 0 at $\gamma\bar{c}_1 = 0$, and is strictly decreasing in $(\gamma\bar{c}_1)$ from thereafter.

6 Quantifying the Equilibrium Effect and the Authority Bias

6.1 Estimation Results

Table 8 presents our preliminary estimation results. In Western Europe, Administrators verifications are objectively more precise ($q_A > q_R$), and the authority bias is small and negative ($a < 0$). In contrast, the estimates from Eastern Europe show that edits by Administrators are less often correct than edits by Regular Users, but the authority bias is large and positive, which means that Administrators' edits are treated as very accurate.

There is about 23–28% of the editors who verify the information regardless of the situation (γ_0). In Western Europe, 5% of the Editors are Administrators ($\frac{\lambda_A}{\lambda_A+\lambda_R}$), in Eastern Europe, the percentage is 20%.

6.2 Discussion

From estimation results in Table 8 we can compute the share of differential treatment due to authority bias, $\frac{a}{q_A+a-q_R}$. In Western Europe, the authority bias is negative and very small, whereas in Eastern Europe, it is almost 180%. That is, not only it accounts for the whole difference in treatment, it even reverses the sign of differential treatment. Namely, in Eastern Europe, verifications by Administrators are less precise than verifications by Regular Users, but because of the authority bias are treated as more precise.

The authority bias has significant implications for the information quality in Wikipedia. In particular, if we compare the average time between edits (as characterized by (4.5)) with and without authority bias, we get that the authority bias increases the time between verifications by 141% in Eastern Europe. Whereas, in Western Europe, the verification time is decreased by 12.4%.

References

Fernández, R. (2011): "Does Culture Matter?," in *Handbook of Social Economics*, ed. by J. Benhabib, M. O. Jackson, and A. Bisin, vol. 1, pp. 481–510. North-Holland.

Gorodnichenko, Y., and G. Roland (2011): "Which Dimensions of Culture Matter for Long-Run Growth?," *American Economic Review*, 101(3), 492–498.

Greenstein, S., and F. Zhu (2012): "Is Wikipedia Biased?," *American Economic Review (Papers and Proceedings)*.

Guiso, L., P. Sapienza, and L. Zingales (2006): "Does Culture Affect Economic Outcomes?," *Journal of Economic Perspectives*, 20(2), 23–48.

Hara, N., P. Shachaf, and K. F. Hew (2010): "Cross-cultural analysis of the Wikipedia community," *Journal of the American Society for Information Science and Technology*, 61(10), 2097–2108.

Hofstede, G. H. (1980): *Culture's consequences, international differences in work-related values*. Sage Publications.

Kirkman, B. L., K. B. Lowe, and C. B. Gibson (2006): "A Quarter Century of "Culture's Consequences": A Review of Empirical Research Incorporating Hofstede's Cultural Values Framework," *Journal of International Business Studies*, 37(3), 285–320.

Pfeil, U., P. Zaphiris, and C. S. Ang (2006): "Cultural Differences in Collaborative Authoring of Wikipedia," *Journal of Computer-Mediated Communication*, 12(1), 88–113.

Piskorski, M. J., and A. D. Gorbatai (2011): "Testing Coleman's Social-Norm Enforcement Mechanism: Evidence from Wikipedia," *Harvard Business School Strategy Unit Working Paper*, 11(055).

Ransbotham, S., and K. Kane (2011): "Membership Turnover and Collaboration Success in Online Communities: Explaining Rises," *Management Information Systems Quarterly*, 35(3), 613–627.

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

A Editors of a City Page Are More Likely to Come from That Country

In our empirical analysis we assume that the editors of a city page in some country are more likely to be affected by the cultural norms of that country. But unfortunately, for many of

the editors we cannot observe their location. Therefore, the question arises, how realistic is that assumption. To check this, we do the following.

First of all, we might be worried whether the editors who edit city pages do that in many countries. So, we first ask what percentage of editors of the city pages in some country, edit articles of the city pages in other countries. Of the registered users in our sample who are not administrators, 81% edit only the city pages of one country, 91% edit city pages in no more than two countries, and 99% in no more than nine countries.

Then we look at the unregistered editors physical location as determined by their IP address. We compute the percentage of editors of the city pages in some country, that are located in that country as determined by their IP address. Unfortunately, this can be done only in case of Wikipedia unregistered users, since Wikipedia does not collect that information regarding registered users. We find that 33% of all the edits by unregistered users are made from local (same country) IP addresses. The percentage of local users across countries is presented on Figure C.1.

Third, we look at what percentage of editors of the city pages of some country, edit Wikipedia in the official language of that country. Fortunately, this can be done for the editors in our sample: administrators and registered users.

B Marking Edits as Questionable

In this appendix, we formalize the argument in Section 3. Suppose there are two types of users whose goal is to improve the quality of Wikipedia articles sequentially. We will first introduce the model with unbiased agents and in the next subsection add authority bias. Let's call a representative agent from the first type a *Flagger* and the second type a *Verifier*.

In the first stage, Flagger enters and chooses whether to *flag* the statement for verification or not. We model this decision as a *flagging rule* $\{\mathcal{F}_\theta \subset [0,1]\}_{\theta\in\Theta}$, where $\theta \in \Theta$ denotes publicly observable information about the statement (which we will refer as *type*) and $s \in [0,1]$ the information that is only observable to Flagger (*signal*). Moreover, the posteriors $Pr(\text{Correct}|s,\theta)$ are increasing in s and the signals are distributed with continuous differentiable cumulative density function F independently of type θ.[17]. We assume that acquiring the information and flagging is costless and Flagger's goal is to maximize the probability that the statement is correct after the verification process.

In the second stage, Verifier enters and chooses whether to *verify* the statement or not,

[17]Of course $Pr(\text{Correct}|s \in \mathcal{F},\theta) = \int_{s\in\mathcal{F}} Pr(\text{Correct}|s,\theta)dF(s)$ depends on θ.

after observing whether the statement is flagged and sees the public information θ. We assume that the cost of verifying (and potentially editing) the statement has a cost c and Verifier maximizes the difference between probability that the statement is correct and the cost. Therefore, for a given flagging rule, Verifier chooses to verify the statement if and only if the improvement is higher than the cost, or formally $1 - c \geq Pr(\text{Correct}|s \in \mathcal{F}_\theta, \theta)$.

Note that we can without loss of generality focus on the case where the statement is verified if and only if it is flagged, since there always exists equivalent flagging policy[18] Let us assume additionally that the distribution of s has no mass points and for all $\theta \in \Theta$ we have $Pr(\text{Correct}|\theta) > 1 - c > Pr(\text{Correct}|s, \theta)$ for some $s \in [0, 1]$. This guarantees interior solutions where Flagger flags some, but not all statements and we do not have to worry about non-discrete jumps in probabilities.

Flagger has a constrained maximization problem

$$\max_{\mathcal{F}_\theta : Pr(\text{Correct}|s \in \mathcal{F}_\theta, \theta) \leq 1-c} \sum_\theta Pr(\theta) \left[Pr(s \in \mathcal{F}_\theta) + Pr(\text{Correct}|s \notin \mathcal{F}_\theta, \theta) Pr(s \notin \mathcal{F}_\theta) \right].$$

The problem is solvable separately for each $\mathcal{F}_\theta$. The optimal flagging rules must be such that $\mathcal{F}_\theta = [0, \phi_\theta]$ for some $\phi_\theta \in (0, 1)$ and $Pr(\text{Correct}|s \in \mathcal{F}_\theta, \theta) = 1 - c$.

Finally note that with correct beliefs, the probability that the statement is actually changed must be the same as the probability that it is incorrect, so $Pr(\text{Changed}|s \in \mathcal{F}_\theta, \theta) = 1 - Pr(\text{Correct}|s \in \mathcal{F}_\theta, \theta) = c$. That is, the higher is the cost of verification, the higher must be the probability that changes are needed.

B.1 Equilibrium effect

Suppose that the only publicly available information about the statement is the type of the author of the statement, who can be either Administrator (A) or Regular user (R). So, $\Theta = \{A, R\}$.

We define the *equilibrium effect* so that for all $s \in [0, 1]$, $Pr(\text{Correct}|s, A) > Pr(\text{Correct}|s, R)$.

[18] For example, if $Pr(\text{Correct}|s \in \mathcal{F}_\theta, \theta) > 1 - c$ and $Pr(\text{Correct}|s \notin \mathcal{F}_\theta, \theta) \leq 1 - c$, so that Verifier chooses not to verify flagged items and verify items that are not flagged for type θ, then the new flagging policy would be such that flags are switched for type θ.

Then we have that $\mathcal{F}_A \subsetneq \mathcal{F}_R$, because otherwise we would have

$$\begin{aligned} 1-c &= \int_0^{\phi_R} Pr(\text{Correct}|s,R)dF(s) \leq \int_0^{\phi_A} Pr(\text{Correct}|s,R)dF(s) \\ &< \int_0^{\phi_A} Pr(\text{Correct}|s,A)dF(s) = 1-c. \end{aligned}$$

To summarize, under the equilibrium effect assumption, we get two testable implications.

1. Statements by type A are flagged more often than those by type R.
2. Flagged statements of either type A and R, are changed with equal probability.

B.2 Authority bias

Suppose now, that Flagger and Verifier have incorrect beliefs about type A. In particular, let's denote the perceived probabilities[19] by $Pr^f(\cdot)$ and $Pr^v(\cdot)$ respectively and assume that for all $s \in [0,1]$

$$Pr^f(\text{Correct}|s,A) > Pr^v(\text{Correct}|s,A) > Pr(\text{Correct}|s,A),$$

$$Pr^f(\text{Correct}|s,R) = Pr^v(\text{Correct}|s,R) = Pr(\text{Correct}|s,R).$$

Then Flagger uses biased flagging rule $\mathcal{F}_A^f = [0,\phi_A^f]$ such that

$$1-c = Pr^f(\text{Correct}|s \in \mathcal{F}_A^f, A) > Pr(\text{Correct}|s \in \mathcal{F}_A^f, A)$$

and therefore $\mathcal{F}_A^f \subsetneq \mathcal{F}_A$. This gives

$$Pr(s \in \mathcal{F}_A^f) < Pr(s \in \mathcal{F}_A) < Pr(s \in \mathcal{F}_R).$$

Interpretation: the observed flagging frequency for A could be lower than flagging frequency for R for two reasons: authority bias or equilibrium effect.

The verifier's bias does not affect the flagging behavior, but has an effect to observed

[19] Where each agent believes that his or her perceived probability is the true one and is not aware of the possible bias of other agents.

frequency of changing. Namely,

$$\begin{aligned} Pr^v(\text{Changed}|s \in \mathcal{F}_A^f) &= E_s\left[1 - Pr^v(\text{Correct}|s, A)|s \in \mathcal{F}_A^s\right] \\ &> E_s\left[1 - Pr^f(\text{Correct}|s, A)|s \in \mathcal{F}_A^s\right] = Pr^f(\text{Changed}|s \in \mathcal{F}_A^f) \\ &= c = Pr(\text{Changed}|s \in \mathcal{F}_R, R). \end{aligned}$$

That is, as long as the verifier is less biased than the flagger, the probability with which we observe changes after flags to edits by type A is strictly higher than the probability with which there are changes to type R's edits. The intuitive reason is that for any flagging rule Flagger, who is more biased, believes that the change rate is lower than Verifier believes (which is the rate we observe). Since the optimal flagging rule sets perceived change rates for two groups equal from Flagger's viewpoint, the actual change rates will be different.

Summary of observable implications with both equilibrium effect and authority bias.

1. Statements by type A are flagged more often than those by type R (for both reasons).
2. Flagged statements by type A are changed more often than flagged statements by type R.

C Evidence of Aggregate Effect from Survival of Edits

Additional evidence for the aggregate effect of differential treatment of Administrators comes from the survival time of edits.

To test whether there is differential treatment, we estimate the following regressions. For each edit we regress *Survival Time* (time until deleted, measured in months) on the indicator variable of whether the edit is made by the Administrator and other observable characteristics of the edit and the city page X. The estimation results are presented in Table 9.

The first set of results from these preliminary regressions are from a linear regression of this form:

$$SurvivalTime_i = \alpha \cdot Admininistrator_i + \beta \cdot X_i + \varepsilon_i$$

The observable characteristics X_i of each edit i include the following. *Administrator* is the indicator variable which takes value 1, when the edit was made by an Administrator, *Eastern Eur* is the indicator variable for edits made on Eastern European city pages, and *Admin Eastern Eur* is the indicator variable which takes value 1, when the edit was made on an Eastern European city page by an Administrator. In addition to the above, Column 1,

presents results from a regression that includes the number of revisions the city page has. In Column 2, we add characteristics of the edit (whether the edit reverts someone else's edit, whether it is described as a minor edit, the length of the edit), and editor's experience (no of edits made on city page so far).

Columns 3-6 present estimation results, where the same regression is estimated as a survival model. First, the parametric log-logistic survival model:

$$\log SurvivalTime_i = \alpha \cdot Admininistrator_i + \beta \cdot X_i + \sigma\varepsilon_i$$

where the error term ε is assumed to have standard logistic distribution, and σ is the scale parameter. The final two columns are from Cox semi-parametric proportional hazard survival model. Note that the dependent variable in columns 1-4 is *Survival Time*, and in columns 5-6, it is *Hazard Rate of Deletion*, and therefore, we would expect all the signs of coefficient estimates in last two columns to be reversed.

All the regressions are estimated using data on the survival of edits up to 12 months. The sample includes only these Administrators who have made edits in both Eastern and Western Europe.

In all the specifications, if the edit was made by an Administrator in Eastern Europe, it survives significantly longer. Pages that get more revisions usually have shorter survival of edits. Minor edits survive longer, and longer edits survive shorter period of time. Revert edits which are probably more controversial survive shorter period of time. Edits by editor with more experience survive longer.

Tables and Figures

Table 3: Descriptive Statistics of Edits

	Obs	Min	Max	Mean	Sd	P25	Median	P75
Marked as questionable	208195	0.000	1.000	0.086				
Administrator	208195	0.000	1.000	0.101				
Number of Revisions / 1000	208195	0.001	4.040	0.935	0.744	0.426	0.711	1.207
Editing Experience	208195	0.001	30.446	1.367	4.455	0.005	0.044	0.257
Internet Access	188826	0.000	0.096	0.056	0.021	0.040	0.058	0.072
Education	181452	0.007	0.014	0.012	0.001	0.010	0.012	0.013
Hofstede Power Distance Index	157219	0.110	1.040	0.548	0.227	0.350	0.570	0.680
Survival in days since questioning	17968	0	365	127	121	27	82	204

Table 4: Descriptive Statistics of Edits of Population Measures by Region

	AA	AR	RA	RR
	Median Duration in Days			
Western Eur	2.0	4.0	1.0	3.0
Eastern Eur	0.0	7.0	1.0	4.0
	Share Not Changed			
Western Eur	0.920	0.866	0.936	0.861
Eastern Eur	0.690	0.890	0.858	0.870
	No of Observations			
Western Eur	112	1522	642	12260
Eastern Eur	29	337	148	3025

Table 5: Dependent variable: indicator whether an edit was marked as questionable

	(1)	(2)	(3)	(4)	(5)
Administrator	-0.053***	-0.036***	-0.057***	-0.030***	-0.025***
	(0.019)	(0.007)	(0.020)	(0.006)	(0.005)
Number of Revisions	-0.021**	0.008	-0.023**	-0.007	-0.002
	(0.009)	(0.018)	(0.009)	(0.013)	(0.015)
Editing Experience	-0.004***	-0.003***	-0.004***	-0.003***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Internet Access			-0.027		2.292**
			(0.510)		(0.991)
Education			24.817**		-11.943
			(11.893)		(10.836)
Eastern Europe				0.126**	0.189***
				(0.058)	(0.069)
Eastern Europe*Administrator				-0.099**	-0.084**
				(0.043)	(0.035)
Country Fixed Effects		Yes			
Adj R-squared	0.010	0.103	0.024	0.042	0.064
No of obs	208195	208195	181452	208195	181452

Notes: Each column reports estimates from a separate linear regression. *Administrator* is the indicator variable which takes value 1, when the edit was made by an Administrator. *Number of Revisions* is thousands of revisions of the whole page in 12 months after the edit is made. *Editing experience* of editor on city pages that are in our sample so far. *Internet Access* is the percentage of population using internet and *Education* is the average schooling received by males, and both variables describe the country of the city page. *Eastern Europe* takes value 1, when the city is in Eastern Europe. *Eastern Europe*Administrator* is the indicator variable which takes value 1, when the edit was made by an Administrator to a city page in Eastern Europe. Robust standard errors, clustered on city level, are in parenthesis. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 6: Dependent variable: indicator whether an edit was marked as questionable

	(1)	(2)
Administrator	0.059	0.024
	(0.051)	(0.026)
Hofstede Power Distance Index	0.236	0.358***
	(0.142)	(0.133)
Index*Administrator	-0.213*	-0.131**
	(0.120)	(0.064)
Number of Revisions	-0.016	-0.012
	(0.015)	(0.014)
Editing Experience	-0.004***	-0.004***
	(0.001)	(0.001)
Internet Access		2.278*
		(1.153)
Education		18.744*
		(10.434)
Adj R-squared	0.045	0.077
No of obs	157219	156622

Notes: Each column reports estimates from a separate linear regression. *Administrator* is the indicator variable which takes value 1, when the edit was made by an Administrator. *Index*Administrator* takes the value of Hofstede Power Distance Index when the edit is made by an Administrator. *Number of Revisions* is thousands of revisions of the whole page in 12 months after the edit is made. *Editing experience* of editor on city pages that are in our sample so far. *Internet Access* is the percentage of population using internet and *Education* is the average schooling received by males, and both variables describe the country of the city page. Robust standard errors, clustered on city level, are in parenthesis. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 7: Dependent variable: time until change

	(1)	(2)	(3)	(4)	(5)
Administrator	-18.552*	-14.984	-16.568	0.983	-0.293
	(9.621)	(9.379)	(10.035)	(7.200)	(7.741)
Number of Revisions	-94.859***	-89.898***	-95.277***	-94.691***	-93.920***
	(14.064)	(22.570)	(15.631)	(17.475)	(17.751)
Editing Experience	0.268	0.707	0.717	0.707	0.892
	(1.139)	(0.864)	(1.149)	(1.004)	(1.130)
Internet Access			-592.039		-131.254
			(494.481)		(575.828)
Education			8459.692		192.760
			(5854.068)		(8396.414)
Eastern Europe				30.499**	27.660
				(13.774)	(19.858)
Eastern Europe*Administrator				-43.794*	-37.289
				(22.567)	(23.558)
Country Fixed Effects		Yes			
Adj R-squared	0.192	0.292	0.210	0.207	0.217
No of obs	17968	17968	15836	17968	15836

Notes: Each column reports estimates from a separate linear survival time regression. *Administrator* is the indicator variable which takes value 1, when the edit was made by an Administrator. *Number of Revisions* is thousands of revisions of the whole page in 12 months after the edit is made. *Editing experience* of editor on city pages that are in our sample so far. *Internet Access* is the percentage of population using internet and *Education* is the average schooling received by males, and both variables describe the country of the city page. *Eastern Europe* takes value 1, when the city is in Eastern Europe. *Eastern Europe*Administrator* is the indicator variable which takes value 1, when the edit was made by an Administrator to a city page in Eastern Europe. Robust standard errors, clustered on city level, are in parenthesis. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Table 8: Estimation Results

Parameter	Western Europe	Eastern Europe
q_A	0.9657	0.9155
q_R	0.9351	0.9444
ρ	0.0019	0.0027
λ_A	0.0337	0.1050
λ_R	0.6874	0.4315
a	-0.0057	0.0654
γ_0	0.2318	0.2874
γ_1	0.0442	0.0000
Observations	14633	3566
$\log L$	-39314.5	-9146.1

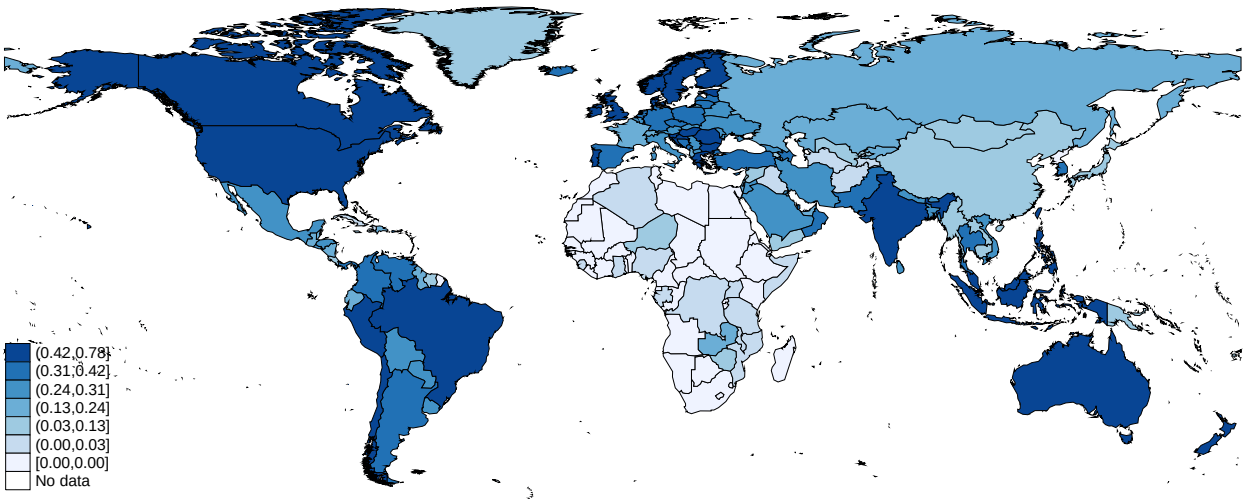

Figure C.1: Local users as a percentage among unregistered users, defined by IP address, 2009 – 2010

Table 9: Survival of Wikipedia Edits in Eastern and Western Europe.

	Linear		Log-logistic		Cox PH	
	(1)	(2)	(3)	(4)	(5)	(6)
Administrator	-0.055	0.004	-0.033	-0.006	0.022	0.010
	(0.157)	(0.154)	(0.080)	(0.079)	(0.055)	(0.055)
Admin Eastern Eur	0.881***	0.893***	0.488***	0.497***	-0.340***	-0.347***
	(0.294)	(0.297)	(0.150)	(0.152)	(0.101)	(0.102)
Eastern Eur	-0.228	-0.213	-0.104	-0.097	0.065	0.063
	(0.182)	(0.183)	(0.089)	(0.090)	(0.060)	(0.061)
Revisions (thousands)	-1.235***	-1.208***	-0.553***	-0.538***	0.352***	0.342***
	(0.220)	(0.218)	(0.104)	(0.103)	(0.057)	(0.056)
Revert Edit		-0.826*		-0.384*		0.252**
		(0.458)		(0.201)		(0.127)
Minor Edit		0.527***		0.213***		-0.120***
		(0.135)		(0.062)		(0.038)
Length of Edit		-0.271***		-0.132***		0.035***
		(0.071)		(0.043)		(0.010)
Editing Experience		0.488***		0.255***		-0.171***
		(0.056)		(0.036)		(0.027)
Constant	9.287***	9.284***	3.299***	3.289***		
	(0.145)	(0.150)	(0.082)	(0.084)		
ln(sigma)			0.094***	0.090***		
			(0.013)	(0.013)		
Adj R-squared	0.019	0.025				
Log-likelihood	-118854	-118741	-47767	-47658	-166303	-166243
No of obs	39848	39848	39848	39848	39848	39848

Notes: Each column reports estimates from a separate regression. Columns 1-2 are from linear regression, columns 3-4 are from parametric survival model with Log-logistic distribution, columns 5-6 are from Cox semi-parametric proportional hazard survival model. The dependent variable in columns 1-4 is *Survival Time*, and columns 5-6 *Hazard Rate of Deletion*. *Administrator* is the indicator variable which takes value 1, when the edit was made by an Administrator. *Admin Eastern Eur* is the indicator variable which takes value 1, when the edit was made in Eastern Europe by an Administrator. Robust standard errors are in parenthesis. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.