

# Smart but Fun: A Data-Driven Portrait of Wikipedia Editors\*

Robert West  
Stanford University  
Stanford, California, USA  
west@cs.stanford.edu

Ingmar Weber  
Yahoo! Research  
Barcelona, Spain  
ingmar@yahoo-inc.com

Carlos Castillo  
Yahoo! Research  
Barcelona, Spain  
chato@yahoo-inc.com

## ABSTRACT

While there has been a substantial amount of research into the editorial and organizational processes within Wikipedia, little is known about how Wikipedia editors (*Wikipedians*) relate to the on-line world in general. We attempt to shed light on this issue by using aggregated log data from Yahoo!'s browser toolbar in order to analyze Wikipedians' editing behavior in the context of their on-line lives beyond Wikipedia.

We broadly characterize editors by investigating how their on-line behavior differs from that of other users; e.g., we find that Wikipedia editors search more, read more news, play more games, and, perhaps surprisingly, are more immersed in pop culture. Then we inspect how editors' general interests relate to the articles to which they contribute; e.g., we confirm the intuition that editors show more expertise in their active domains than average users.

Finally, we analyze the data from a temporal perspective; e.g., we demonstrate that a user's interest in the edited topic peaks immediately before the edit and characterize typical chains of events leading to an edit. Our results are relevant as they illuminate novel aspects of what has become many Web users' prevalent source of information.

## 1. INTRODUCTION

Wikipedia is arguably one of the technological wonders of our era, an ambitious project that, as its supporters usually say, 'can only work in practice, but will never work in theory'. It is a prime example of a peer-production community [4] with a broad user base including a group of around 300K editors who edit Wikipedia every month, containing a core group of around 5K editors who do more than 100 edits every month.<sup>1</sup>

What we know about Wikipedia editors, often referred to as *Wikipedians*, we mostly know either through user surveys, or by looking at their activity in Wikipedia, including edits and discussions with other editors. In this paper, we introduce a new source of information: traces from browsing behavior. We use browsing

\*Under review—do not cite.

<sup>1</sup><http://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedians&oldid=457802123>

data obtained by the Yahoo! Toolbar and look for specific URLs corresponding to Wikipedia edits. This way we can obtain insights into the browsing behavior of Wikipedia editors both in general and during the time period immediately preceding an edit event. These are the main findings among the observations we present in the following sections:

- We find that, on broad average, Wikipedia editors seem, on the one hand, more sophisticated than usual Web users, reading more news, doing more Web search, and looking up more things in dictionaries and other reference works; on the other hand, they are also deeply immersed in pop culture, spending much online time on music- and movie-related websites.
- We then show that one of the main lines of distinction within the group of editors is their use of social networking sites. While those editors that spend much time on such sites tend to contribute more to entertainment-related articles, they are less involved in the Wikipedia community, with shorter and fewer edits per user.
- Next we introduce a novel notion of expertise based on user's search query histories and show that across all topical domains Wikipedia editors show significant expertise. We also nuance the first impression of all editors' being entertainment lovers, by showing that the latter form only a highly specialized subgroup that contributes many edits. We also show that more substantial edits tend to come from experts, and that editors with a Wikipedia account expose more expertise than other editors.
- Finally, we investigate how editors arrive at the pages they edit, by analyzing click referral chains. About half of the click chains culminating in an edit start with a Web search, with the other half originating on Wikipedia's main page.

Apart from being interesting in its own right, characterizing Wikipedia editors may be useful from a practical perspective. In order to target the promising readers for converting them into editors, it can help a lot to know what a typical editor is like. In this respect, organization like the Wikimedia Foundation can directly profit from the results of our research.

The rest of the papers is organized as follows. In Section 2 we discuss related work and summarize what is known about Wikipedians. Section 3 describes details of our data set, Yahoo! Toolbar data, and preprocessing steps. The question of who Wikipedia editors are and how their online behavior differs from mere Wikipedia readers or non-readers is analyzed in Section 4. Section 5 then looks at the expertise of editors and how it correlates with the size of edits or the topic of the edited article. Section 6 describes how

editors navigate to the article edit. Finally, Section 7 discusses future work and concludes this paper.

## 2. RELATED WORK

A significant amount of research has been done on Wikipedia both because of its significance (it is the 7th most visited site on the Web<sup>2</sup>) as well as because of its availability, with most of the data being released under a free content license. The literature on Wikipedia is vast and there are many studies characterizing its contents; for a recent report on this subject, see Ortega [13].

**Wikipedians.** A key source of information on Wikipedia editors (or *Wikipedians*) are semi-annual surveys conducted by the Wikimedia Foundation. According to the 2011 yearly survey [19] answered by more than 5K editors, they are well educated, with 61% having a college degree and 72% of them reading Wikipedia in more than one language. The median age is 28 years. The most cited ideological reasons for contributing to Wikipedia at all are volunteering and a belief that ‘information should be free’ [19, 12]. Reasons to edit a particular Wikipedia article are varied. According to one hypothesis [6, 11] editors contribute to solve cognitive dissonances between the current state of a Wikipedia article and their own knowledge. This supports the finding that looking for mistakes, bias, and incomplete articles is cited as a reason to contribute to Wikipedia by over 50% of surveyed editors [19]. Concerning the personality of Wikipedians, Hamburger *et al.* [8] found that Wikipedia editors tend to locate their real ‘me’ more often on the Internet than non-editors and that they have lower levels of agreeableness, openness, and conscientiousness.

**Usage analysis and edit patterns.** The usage of toolbar data is a well-established paradigm for studying user behavior on the Web; for a recent study including over 50 million page-views see Kumar and Tomkins [10].

Usage analysis has been applied to access logs of Wikipedia itself [14] to establish, among other findings, that less than 7% of the page views that Wikipedia serves are related to editing actions; and very often users click on the ‘edit’ button but do not make any changes to pages.

The authors of [15] present a color-coding technique called *Chromograms* for visualizing Wikipedia edit patterns. Different Wikipedia editor social roles are discussed in [17] where the following four key roles are identified: substantive experts, technical editors, vandal fighters, and social networkers.

**Expertise.** In small-scale analyses, experts can be identified by using surveys or looking for academic or technical qualifications. In a large-scale analysis, proxies for expertise need to be used. In this paper, we identify expertise with being *familiar* with a topic, more than being *proficient* at topic-related tasks. This is the preferred method used by expert-finding methods that rely on traffic analysis, e.g. White *et al.* [18] identify as experts in the medical domain users who visit the Medline website (a portal for medical literature search).

## 3. DATA SET DESCRIPTION

Since early 2008 users of Yahoo! Toolbar<sup>3</sup> have the option to allow Yahoo! to collect information about the websites they visit. The basic unit of the recorded toolbar data is a *pageview*, of which the following properties are relevant to us: the unique toolbar id, the timestamp, the URL of the page visited (in case the HTTPS protocol was used, only the domain part is available), the referrer

<sup>2</sup><http://www.alexa.com/siteinfo/wikipedia.org>

<sup>3</sup><http://toolbar.yahoo.com/>

URL from which the page was reached, a redirect flag, and locale information.

In the following we use toolbar ids as user ids. Though it is possible that Wikipedia editors use several distinct computers, with or without a Yahoo! Toolbar installed, to make edits this will not substantially affect our analysis unless their behavior differs hugely on each machine. Similarly, we assume that a single computer/toolbar is not used by several users but if this is the case then the true differences between editors and non-editors (see Section 4) and the observations concerning expertise (Section 5) would only be more pronounced and the trends we identify are expected to hold true.

### 3.1 Editors, readers-only, and non-readers

We consider toolbar data for the 10-month period from September 2010 to June 2011. To avoid undue sampling biases, we exclude all users with less than 1K or more than 1.2M pageviews. The set of all users is divided into three groups: editors of the English Wikipedia (.089% of all users), readers-only of the English Wikipedia (58%), and those that do not read any language version of Wikipedia (41%).

We make the assumption that editors of the English Wikipedia also speak English (although not necessarily as a first language) and attempt to control for cultural bias in the two non-editor groups by sampling representative subgroups of such users from primarily English-speaking locales. Our data contains 1.9K editors, and we subsample 5K readers-only and 10K non-readers, in order to have roughly equal numbers of pageviews.

### 3.2 Reliably determining edits

When referring to editors, we mean all users with at least one Wikipedia edit in the toolbar logs. We identify edits in the data by searching for the URL pattern `http://en.wikipedia.org/w/index.php?title=* & action=submit*` with the redirect flag set to true, both of which are necessary conditions for an edit. In order to eliminate false positives, and to collect additional information about the edit (such as its size and the user’s Wikipedia name), we use the timestamps to look up all candidates in the Wikipedia edit logs<sup>4</sup> and keep only those for which we find a match. We ignore 113 mere revert edits by checking the edit summary for specific substrings such as ‘revert’ or ‘rv’.<sup>5</sup> The rationale is to discard the cases where users click on the “edit” link but make no changes [14]. For the same reason, we will sometimes (where noted) also restrict ourselves to edits of a minimum size. Some edits are committed by automatic agents that do maintenance task on Wikipedia (bots); we exclude them using the official registry of Wikipedia bots.<sup>6</sup>

### 3.3 Editor–article pairs (EAPs)

Wikipedians often make several small edits to the same article in a row, possibly in an effort to avoid losing their work by saving often, and also to prevent versioning conflicts with other editors. In order not to give undue importance to these series of micro-edits, we use as our fundamental unit of analysis that of an *editor–article pair* (EAP), which collapses all edits a given user made to a given article.

<sup>4</sup>using the Wikipedia API at <http://www.mediawiki.org/wiki/API:Properties>

<sup>5</sup>The most important patterns are listed at [http://en.wikipedia.org/w/index.php?title=Wikipedia:Edit\\_summary\\_legend&oldid=458695721](http://en.wikipedia.org/w/index.php?title=Wikipedia:Edit_summary_legend&oldid=458695721).

<sup>6</sup>[http://en.wikipedia.org/w/index.php?title=Wikipedia:List\\_of\\_bots\\_by\\_number\\_of\\_edits&oldid=447262313](http://en.wikipedia.org/w/index.php?title=Wikipedia:List_of_bots_by_number_of_edits&oldid=447262313)

704	ENTERTAINMENT / TELEVISION_SHOWS	108	ENTERTAINMENT
656	ENTERTAINMENT / MUSIC	100	GOVERNMENT / MILITARY
492	ARTS / HUMANITIES/HISTORY	99	SOCIAL_SCIENCE
385	ENTERTAINMENT / MOVIES+FILM	94	RECREATION / TRAVEL
208	SOCIETY+CULTURE / RELIGION+SPIRIT.	91	SCIENCE / ECOLOGY
190	RECREATION / GAMES	60	RECREATION / SPORTS / SOCCER
179	ENTERTAINMENT / COMICS+ANIMATION	56	BUSINESS+ECONOMY / FINANCE+INVESTMENT
171	ARTS / HUMANITIES / LITERATURE	55	RECREATION / SPORTS
152	NEWS+MEDIA	51	RECREATION / SPORTS / BASEBALL
144	SOCIAL_SCIENCE / POLITICAL_SCIENCE	50	GOVERNMENT / LAW
108	EDUCATION	50	SOCIETY+CULTURE / FOOD+DRINK

**Table 1: A list of the 20 most frequent categories for the 5.3K editor–article pairs in our data set.**

We define the *edit size* of an EAP as the maximum edit size<sup>7</sup> over all its constituent edits, measured as the number of bytes in the article after the edit, minus before the edit. Note that this notion of edit size is really only a lower bound on the size of the change; e.g., if the editor deleted 100 bytes and added 101, we will count an edit size of 1. The distribution of edit sizes is highly skewed and roughly follows a power law of negative exponent for both positive as well as negative edit sizes. As observed by previous works [2], most edits are small.

In summary, we have around 13K atomic edit events on 5.1K unique articles, stemming from 1.9K editors and grouped into 5.3K EAPs, 77% of which have a positive edit size, 17% a negative one, and 6.5% one of 0.

### 3.4 Edit topic distribution

We map Wikipedia articles to categories such as ENTERTAINMENT / MUSIC or SCIENCE / ZOOLOGY. We do so by using the article name as a query to the Yahoo! search engine and inspecting the top 10 results, each of which is labeled with one Yahoo! Directory<sup>8</sup> category. Then we aggregate by Borda count, attributing  $11 - i$  votes to the  $i$ -th result and performing weighted majority voting to obtain a category for the article [16]. For instance, ANTHOLOGY gets the category ARTS / HUMANITIES / LITERATURE, and CARNIVOROUS PLANT is classified as SCIENCE / BIOLOGY / BOTANY.

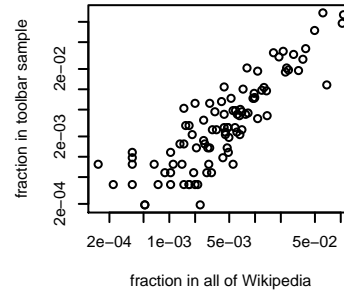
By looking only at data from a specific source, such as toolbar logs, one might obtain a biased user sample whose behavior and interests differ from other Wikipedia editors. We investigate this bias by comparing two distributions over categories of edited articles, one computed from our toolbar sample, the other computed from a representative sample of recent Wikipedia edits, collected on the live Wikipedia over a period of two days, by fetching 500 recent edits every hour.<sup>9</sup> In both cases articles were mapped to Yahoo! Directory categories as described above. As depicted in Figure 1, the topic distribution of our sample of toolbar users is similar to that of editors in general (Pearson’s correlation coefficient  $r = 0.88$ ). While this does not foreclose all potential types of bias, we argue that, if the types of edits are virtually identical to full Wikipedia, the people making those edits may be assumed to be reasonably similar as well.

It is quite revealing to take a look at the edit category distribution, the head of which is listed in Table 1. Note how strongly entertainment-related edits on topics such as music, TV, or games are featured. This is in line with previous works, e.g. in [9] it is shown that 7 of the top 10 larger categories by number of articles in Wikipedia are related to music, films, or television. This enter-

<sup>7</sup>In practice, the largest edit is typically the first one in a series of micro-edits, followed by small corrections.

<sup>8</sup><http://dir.yahoo.com>

<sup>9</sup><http://en.wikipedia.org/wiki/Special:RecentChanges>



**Figure 1: Log-log scatter plot of the distributions of categories of edited articles. There is high correlation between all of Wikipedia and the toolbar sample used in this paper. The outlier to the right is RECREATION/SPORTS/SOCCER, which is due to a tournament that was ongoing at sampling time.**

tainment bias will be a recurring theme in many places throughout our analysis.

## 4. WHO ARE THE WIKIPEDIANS?

What is currently known about Wikipedians comes mostly from examining their contributions and from surveys (see Section 2). Instead, we examine how three user groups—Wikipedia editors, readers-only and non-readers—differ in terms of their online behavior. One first striking observation is that editors spend more time online: In our data, editors have on average 3 times as many pageviews as readers-only, and 9 times as many as non-readers. A natural next question is how they spend their online time.

### 4.1 How do editors spend their online time?

To answer this question, we look at how the three groups differ in terms of the Web domains they frequent. We represent each user by a *relative domain frequency vector*, which counts for each candidate domain what fraction of all their pageviews they spent on it.<sup>10</sup> We consider relative rather than absolute domain frequencies, since, as mentioned above, the absolute numbers of pageviews vary a lot between the three user groups. Our set of candidate domains consists of the 10K most visited domains as of September 2010, according to Alexa. To have interpretable results, these domains were then grouped into categories. In some cases, this grouping was done by simply taking the top-level domain (e.g., .edu) or by searching URLs for a particular pattern. But in most cases we used all domains listed in the Yahoo! Directory for the respective category (e.g., ENTERTAINMENT / GAMES). Details are as follows.

- Reference: from Yahoo! directory; this class contains dictionaries, Q&A sites, and encyclopedias; phonebooks, Web directories, etc., were dropped; also Wikimedia projects such as Wikipedia, Wiktionary, Wikiquote, etc., were removed.
- Search: for search, we use a broad definition, not only Y! and Google, but also site search, product search, etc. To this

<sup>10</sup>For all analyses of Section 4, we consider only editors from primarily English-speaking locales, in order to reduce the language influence on the choice of domains visited, which leaves us with 1.8K of the original 1.9K editors. (Recall that readers-only and non-readers were sampled only from such locales in the first place.)

end, we assembled a list of URL patterns that contain elements such as `q=`, `p=`, `search=`, and so on. We constructed this list via a bootstrapping mechanism similar to [5], starting from a list of such fields for the major web search engines.<sup>11</sup>

- News: all domains listed on <http://listorious.com/GibertPascal/digital-newspapers>, filtered manually.
- .edu: all domains under the .edu top level domain (educational).
- .mil: all domains under the .mil top level domain (military).
- Games: all domains that, using a Yahoo!-internal machine-learned classifier, were classified into ENTERTAINMENT / GAMES in the Yahoo! Directory hierarchy. The same classifier was used for several of the following categories.
- Programming: all domains classified as COMPUTERS&INTERNET / PROGRAMMING&DEVELOPMENT.
- Sports: all domains classified as RECREATION&SPORTS / SPORTS.
- Torrents: all domains containing the substring `torrent`.
- Music: all domains classified as ENTERTAINMENT / MUSIC.
- Movies & TV: all domains classified as ENTERTAINMENT / MOVIES or ENTERTAINMENT / TELEVISION SHOWS.
- YouTube: all pageviews on [youtube.com](http://youtube.com).
- .org (non-Wiki): All domains under the .org top level domain (non-profit organizations), except [wikipedia.org](http://wikipedia.org) itself.
- Adult: All domains listed on <http://www.tblop.com>, combined with all domains classified as SOCIETY&CULTURE / SEXUALITY.
- Social network: all page views from [facebook.com](http://facebook.com), [orkut.com](http://orkut.com), [myspace.com](http://myspace.com), [friendster.com](http://friendster.com), or [hi5.com](http://hi5.com).

Fig. 2 contains a summary of the differences between the 3 groups with respect to the most interesting domain classes. In all figures, error bars correspond to 95% confidence intervals estimated by bootstrap resampling.

In most cases, the share of visits to a type of Web site for readers-only is in a middle ground between editors and non-reader. It is expected that readers-only are close to the average, since they represent the largest group (58% of users). It is, however, telling that editors and non-readers are typically on opposite sides of the spectrum.

Observations from Figure 2 include the following. Wikipedia editors are smart (more news, more educational domains, more reference, and more searches) but fun (more YouTube, more music, more games, and more TV). They also have a lower fraction of pageviews on adult content and social networking sites. Interestingly, Wikimedia’s most recent editor survey claims that ‘a typical Wikipedia editor [...] does not actually spend much time playing games’ [19, p. 3]; however, we find that editors have a significantly

higher fraction of pageviews on game websites than the average Web user. Also, the same survey states that a typical editor is ‘computer savvy but not necessarily a programmer’; indeed, we find that editors have significantly more than average pageviews on programming websites.

Given that the fraction of visits to YouTube is one of the differentiating factors, we decided to compare editors to the other two groups with respect to what they watch on the site. For this comparison, we sample five YouTube views for each user, ignoring users with less than 5 video views. For each view, we use the YouTube API to get additional information, in particular the category the video was posted under. For this analysis, we compare editors to non-editors, i.e., we lump readers-only and non-readers together. We compute category distributions and compare them for the two groups. A *t*-test yields significant ( $p < .05$ ) differences for the following categories: editors watch more entertainment (19% vs. 17%), film (6.5% vs. 4.9%), shows (1.3% vs. 0.89%) and games (4.6% vs. 2.2%—i.e., over twice as much!). Non-editors watch more videos from the categories people, howto, autos and animals. We also checked other video features such as length, age, number comments, average rating but did not find any significant differences.

These numbers lend further support to the hypothesis that Wikipedia editors are more immersed in pop culture and that they play more games. This analysis also allows us to make an additional statement: one might have argued that the lower interest in entertainment-related domains among non-editors stems from the hypothetical fact that the non-editor group is less familiar with entertainment-focused media platforms such as YouTube. But, as we can see, even conditioning on users being familiar with YouTube the increased level of interest in entertainment and games among editors persists.

The entertainment bias is in tune with the fact that most Wikipedia edits are from the entertainment domain (cf. Section 3). It is therefore an interesting question whether the entertainment bias is characteristic of all editors or just of those that edit the many entertainment articles.

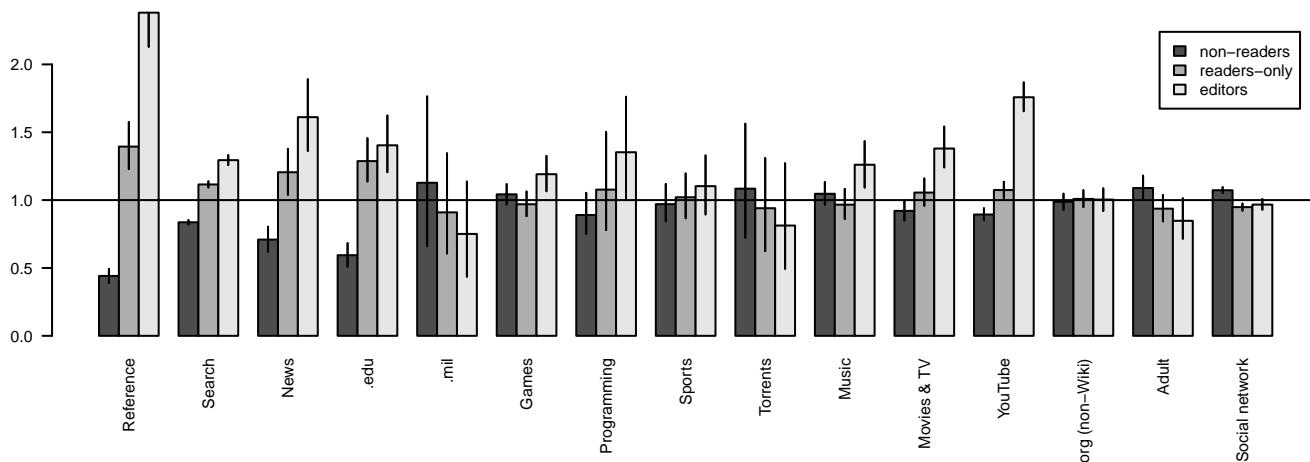
To answer this question, we note that editors of entertainment articles have a significantly higher level of entertainment pageviews than those editors that do not edit entertainment articles (YouTube: 9.2% vs. 7.7%, IMDb: .85% vs. .28%, all domains in class ‘Movies & TV’: 2.8% vs. 1.9%). But also those editors that never edit entertainment articles have a significantly higher fraction of pageviews on entertainment domains than non-editors (YouTube: 7.7% vs. 3.8%, IMDb: .28% vs. .035%, all domains in class ‘Movies & TV’: 1.9% vs. 1.4%). *t*-tests for checking for a difference in means yielded  $p < .01$  for all reported numbers. In Section 5.3 we will further investigate the question whether the focus on entertainment is pervasive in the entire editor community or just in parts of it.

Let us draw a quick summary of the emerging picture: editors spend more time online; they seem more sophisticated than average users, in the sense that they read more news, search more, look up more things on reference and academic sites; they are more computer savvy, reading more programming sites. But by no means are they mere bookworms: they are also more interested in music, movies, and TV, and play more online games.

## 4.2 Are there different classes of editors?

In the previous section we have compared editors to non-editors. Now we want to see how homogeneous the group of editors is. Are all editors the same? If not, how do they differ?

<sup>11</sup>This mechanism initialized with, say, the pattern `p=` used on <http://search.yahoo.com>, might ‘learn’ that a common value for this parameter is `britney%20spears`. It would then take this value and try to find it in other patterns. This way it would automatically discover `search=` as a new pattern.



**Figure 2: Category and domain frequencies for the three user groups, each macro-averaged over its users. The horizontal bar at 1.0 corresponds to the overall average for each category for general Web users computed by weighting the three groups by their relative frequencies (41%, 58%, 0.089%). All fractions are normalized by this global average so we can plot everything in one figure. Error bars are 95% confidence intervals obtained by bootstrap sampling.**

<p><b>11.239 facebook.com</b>  0.022 picnic.com  0.014 farmville.com  0.011 google.lk  0.009 formspring.me</p>	<p>−1.681 google.com  −1.493 wikipedia.org  −1.249 youtube.com  −0.222 google.co.in  −0.168 ebay.com</p>
--	--

**Table 2: Entries of the first principal component of the user-domain matrix with the largest absolute values. Left: top positive domains. Right: top negative domains.**

To this end, we perform principal component analysis (PCA) on users’ relative domain frequency vectors.<sup>12</sup> The first principal component captures 47% of the total variance and tells us a lot about the main differences between editors. The most important entries of the first principal component are listed in Table 2. The entries with the largest positive and those with the largest negative weights tell us with respect to which domains editors differ most.

We see that the main line of divide is Facebook. Its weight is nearly seven times as large as that of the next largest entry (google.com) by absolute value. The domain wikipedia.org also gets a relatively large weight, but even when this special domain is removed from the data matrix, the result is unchanged.

To see if the differences in domain frequencies are specific to the set of Wikipedia editors or apply to Web users in general, we repeated the same approach (PCA and investigating the first principal component) also for the sets of readers-only and non-readers. In both cases, facebook.com was by far the strongest dimension, with google.com and live.com being relatively strong and of opposite sign. This indicates that the distinction ‘Facebook vs. non-Facebook’ (where Facebook is the most prominent,

but stands along other social media sites including Farmville and formspring.me) spans across user groups.

Whereas Table 2 indicates differences in domain frequencies, we are also interested in the corresponding differences in terms of editing behavior. To this end, we cluster all editors using the  $k$ -medoids algorithm<sup>13</sup> and compute the means of certain properties for each cluster. The best clustering (according to average silhouette width) has two fairly balanced clusters: 47% of editors fall into the ‘Facebook’ cluster and 53% into the ‘non-Facebook’ cluster. Note that, due to the strong influence of the first principal component, the clustering essentially groups the editors according to their loadings with respect to the first principal component.

We note that for the sake of easier interpretability we first grouped the large number (93) of Yahoo! Directory categories into the 12 high-level categories ENTERTAINMENT, BUSINESS, HUMANITIES LAW & SOCIAL SCIENCE, HEALTH, NEWS & MEDIA, HOBBIES, SCIENCE, ARTS, SPORTS, ADULT, TECHNOLOGY, SHOPPING. Then we computed for both clusters a distribution over these categories, with respect to the edits made, and found 6 significant ( $p < .05$ ) differences: editors in the Facebook cluster have more edits in entertainment (47% vs. 40%) and shopping (1.7% vs. 0.48%), while in the other cluster we see more edits related to business (3.9% vs. 2.2%), news & media (4.7% vs. 2.5%), hobbies (including games, travel, autos, pets) (13% vs. 10%), and science (6.3% vs. 4.2%).

Editors in the non-Facebook cluster also make significantly longer edits (mean/median edit size 200/45 vs. 123/33) and are more likely to be logged in (26% vs. 16%). Not only are these editors more involved with the Wikipedia community, they also create higher-quality content. To quantify this notion, we use the ‘WikiTrust’ metric [1], which assigns trust values (ranging from 0 to 9) to

<sup>12</sup>We again ignore pageviews to \*.yahoo.com to minimize toolbar bias, but the result is nearly exactly the same when we keep it.

<sup>13</sup>As the feature space of relative domain frequencies is very sparse and as many dimension are correlated with each other, we operate in the dimensionality-reduced space resulting from PCA. We find a good dimensionality by looking for an ‘elbow’ in the plot of eigenvalues: a dimensionality of 60 (out of 10K) explains 92% of the variance.

wiki edits based on revision history and author reputation features. We find that the average trust value attributed to edits in the non-Facebook cluster is 0.22, while it is only 0.086 in the Facebook cluster. (All reported differences are significant with  $p < .05$  and with non-overlapping confidence intervals). This is in tune with previous work that has found that contributions by logged-in users are of higher quality than those that do not register [3].

More support for the larger involvement of non-Facebook users comes from the fact that in the non-Facebook cluster we have over twice as many edits per user (3.9 vs. 1.8). To check whether this is caused by only a few ‘power editors’, we exclude the top 5% and the bottom 5% users in each cluster (in terms of number of edits) before computing means, but find that even then the numbers of edits per user are still significantly different, at 1.8 and 1.2.

In summary, the major difference between editors is their use of Facebook. Users from the cluster with more Facebook activity produce more entertainment edits, whereas the other cluster produces more edits in SCIENCE and NEWS & MEDIA. Users from the non-Facebook cluster are more involved in Wikipedia as signified by (i) larger edits, (ii) a higher chance of being logged in to Wikipedia, (iii) more edits per user, and (iv) a higher edit trust score.

## 5. DO WIKIPEDIA EDITORS KNOW THEIR DOMAIN?

So far, all our analyses were based on the domain frequency representation of users, and on derived categories. We now concentrate on the group of editors and investigate whether they are experts in the areas in which they make edits.

### 5.1 Defining expertise

We use expertise in the sense of familiarity (not necessarily proficiency). Following [18], we consider experts on a topic those who have seen more information on that topic than regular users, in our case those users who have issued many search queries related to a topic.

For each editor  $e$ , we sample 1K search queries uniformly at random without replacement. Call this the editor’s *query history*  $Q_e$ . We also sample 1K random queries from the set of all queries issued by all editors. Call this the *average query history*  $Q_{avg}$ .

Now define an editor  $e$ ’s *interest* in a Wikipedia article  $a$  as the mean similarity of their queries with the article (the definition of the article–query similarity  $\text{sim}(a, q)$  is rather technical and is given in Appendix A):

$$I_e(a) := \frac{1}{|Q_e|} \sum_{q \in Q_e} \text{sim}(a, q) \quad (1)$$

Similarly, we define the *average interest* in an article

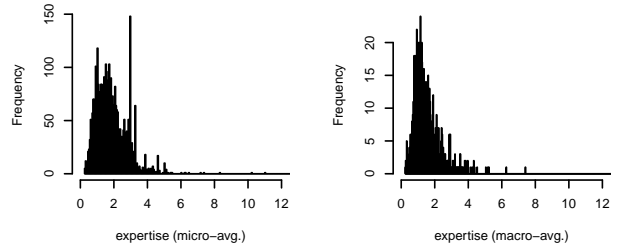
$$I_{avg}(a) := \frac{1}{|Q_{avg}|} \sum_{q \in Q_{avg}} \text{sim}(a, q) \quad (2)$$

Certain query categories are much more common than others (see e.g. [16]), so high interest in a certain topic alone is not necessarily very informative. We define that someone is an expert in a topic if their interest is significantly above average. Formally, we define  $e$ ’s expertise in an article  $a$ :

$$E_e(a) := I_e(a)/I_{avg}(a) \quad (3)$$

An expertise greater than 1.0 implies above-average interest in the given topic, and conversely for an expertise less than 1.0.

Since we want searches to capture the user’s interests as well as possible, we define search broadly, not only as queries to search engines. Everything matching a bootstrapped regular expression



**Figure 3: Histograms of expertise, both micro-averaged (each EAP contributes equally) and macro-averaged (each user contributes equally).**

(cf. Section 4.1) is included, with these exceptions: navigational queries (defined using a click entropy threshold, see [16]), queries issued on Facebook (they relate to a user’s personal circle of friends and do not reveal their interests) and queries that are longer than 30 chars and contain no whitespace character.

As a validation of our notion of ‘expertise’ we will see later (c.f. Fig. 5) that expertise correlates both with edit trust [1] and with the possession of a ‘barnstar’, a type of award/badge given to deserving editors by other editors.

### 5.2 Are editors experts in their edited topics?

Using the definition of expertise from the previous section, we now characterize the expertise distribution of Wikipedia edits. In these experiments we neglect all EAPs of negative size difference for the same reason we neglect revert-only EAPs: to not consider sizable edits that can be achieved with a click rather than through novel content creation. This leaves us with 83% of all EAPs.

Fig. 3 shows a histogram of expertise over all EAPs. The micro-averaged expertise is 1.85 (with 95% confidence intervals of [1.82, 1.88] computed via bootstrap resampling) and macro-averaged, first averaging all EAPs for each user, it is 1.52 ([1.46, 1.57]). The micro-averaged value is higher because there is one user making 134 music-related edits (nearly all of them about one specific TV show), pertaining to the spike in the micro-averaged expertise histogram (left part of Fig. 3). Summarizing we can say that an edit is on average more than 1.5 times as related to the editor’s personal query history as it is to a random sample of queries, indicating that editors know more about the topics they edit than the average Internet user.

This bias towards expertise could, however, also be explained by a simpler model as follows: for every Wikipedia page visited by a potential editor, there is a fixed probability  $p$  with which they edit the page, regardless of article-specific expertise. Now, if the user visits Wikipedia pages according to their general interest, the constant fraction of articles edited will, of course, be more similar to the user’s personal history than to a random history. So the results above are not necessarily edit-specific but simply confirm the intuition that users visit Wikipedia pages similar to their general interests.

To refute this counter-argument, we reran the same experiment, with edits replaced by non-edited yet visited articles. For a user who edited  $n$  different articles we now sample  $n$  different viewed yet not edited articles. The micro-averaged expertise obtained this way is 1.62 ([1.59 1.65]) and macro-averaged it is 1.41 ([1.36 1.46]). So, since these numbers are significantly lower than when an actual edit rather than a mere article view takes place, the observed exper-

4.16 SPORTS  
 2.70 TECHNOLOGY  
 2.66 NEWS & MEDIA  
 2.16 HUMANITIES LAW & SOCIAL SCIENCE  
 2.03 SHOPPING  
 1.92 BUSINESS

1.80 HEALTH  
 1.68 SCIENCE  
 1.65 ENTERTAINMENT  
 1.54 ARTS  
 1.53 HOBBIES

**Table 3: List of categories in order of decreasing category-specific expertise.**

tise cannot be solely explained by the simple interest-only model described above.

### 5.3 Is there more expertise in some domains than others?

The previous section indicates that, on average, Wikipedia edits are made by people with above-average knowledge about a particular topic. But do experts exist to the same extent across all topics? To answer this, we average the expertise of the edit’s author with respect to the edited article over all edits in a given category (again, we consider grouped high-level categories, cf. discussion in Section 4.2). This gives us the category-specific expertise. We find that expertise differs across categories, but also that it is significantly greater than 1 everywhere, with the only exception of the ADULT category, which has a micro-averaged expertise of 0.87. However, only 11 edits were contributed in this category. Table 3 contains a listing in order of decreasing category-specific expertise.

What about other categories? Can we make a statement about whether editors of certain categories are also experts in other domains? We answer this question quantitatively using the notion of *co-expertise*, proceeding as follows: For each edit category  $c_1$ , we compute a co-expertise profile. The profile has one entry per category  $c_2$ , the value being the mean expertise in category  $c_2$  of all users editing articles from  $c_1$  (we consider micro-averages, such that users are weighted by how many different articles they have edited in category  $c_1$ ).

The results can be represented as a bipartite graph, as shown in Fig. 4. In this *co-expertise graph*, one partition (the upper one in Fig. 4) represents edit categories, the other (the lower one in the figure) expertise categories. Edges are drawn from  $c_1$  to  $c_2$  if, on average, editors that edit  $c_1$  have expertise greater than 1 in category  $c_2$ . Additionally, an edge’s gray tone represents expertise strength.<sup>14</sup>

We have already seen that in all categories (besides the ADULT category), there is significant expertise on behalf of the people who edit articles in that category; in the co-expertise graph, this is manifested as strong vertical arrows.

The co-expertise graph can also be used to shed light on the following question: Is editors’ overall focus on entertainment (a fact that has re-occurred as the result of many of our experiments) caused by all editors equally or by a subgroup of editors that is deeply immersed in pop culture? A first indicator that the latter might be the case is the fact that the number of pageviews on entertainment-related domains is higher for editors of entertainment-related articles than for other editors (cf. Section 4.1). This is now confirmed using the more sophisticated notion of expertise instead of raw domain frequencies: expertise in the ENTERTAINMENT domain resides mostly in the group of editors of that category, as visualized by the fact that the only strong arrow leading into the bottom ENTERTAINMENT node of the co-expertise graph originates from

editors of ENTERTAINMENT articles. On the flip side, editors of ENTERTAINMENT have no other areas of strong expertise, visualized by the lack of outgoing arrows from the upper ENTERTAINMENT node. Hence, the simplistic image of all Wikipedia editors being entertainment-loving has to be faceted: rather, the overall focus on entertainment may be attributed to a group of entertainment-only specialists that contribute many edits.

Also note that, on the contrary, editors of SCIENCE and BUSINESS seem to be more versatile: they are experts in several areas beyond what they edit.

### 5.4 What are the correlates expertise?

Next we look into the question whether some quantities correlate with the expertise of an edit. We consider properties of the edit (e.g., edit size), the edited article (e.g., whether it has many received many comments), and the editor (e.g., whether he/she has been logged into Wikipedia etc.).

Fig. 5 summarizes our findings graphically. The  $x$ -axes show the respective properties, the  $y$ -axis expertise. The  $x$ -axes often had to be binned in unequally sized intervals to give roughly comparable sample sizes in each bucket. The  $x$ -labels show the upper ends of the bin intervals. We include all EAPs of an edit size of at least 0. Error bars indicate 95% confidence intervals, obtained by bootstrap resampling. Note that the confidence intervals are often large.

The first 2 plots in the upper row relate features of EAPs to expertise:

1. Long edits (notably the very long ones) come from editors with more expertise; this is a good sign: small edits are often minor corrections such as typo fixes, while the large ones are the real content contributions, which we would hope to come from real experts.
2. Articles with greater edit trust [1] come from editors with more expertise.

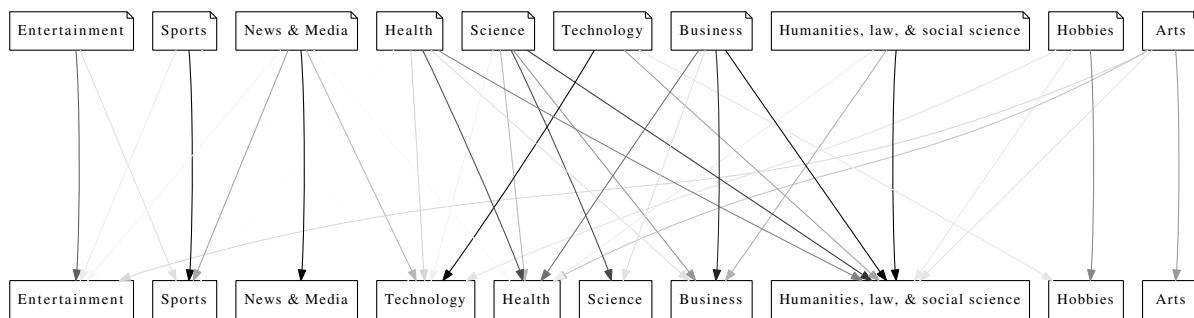
The third plot refers to the number of article comments, a property of the article that is edited: articles typically do not have many comments (median 2), but when they do, they are the more debated ones. It seems there is slightly more expertise on those articles (we chose three comments as the threshold to have two roughly balanced sets to compare).

The remaining properties describe the editors whose expertise we are evaluating:

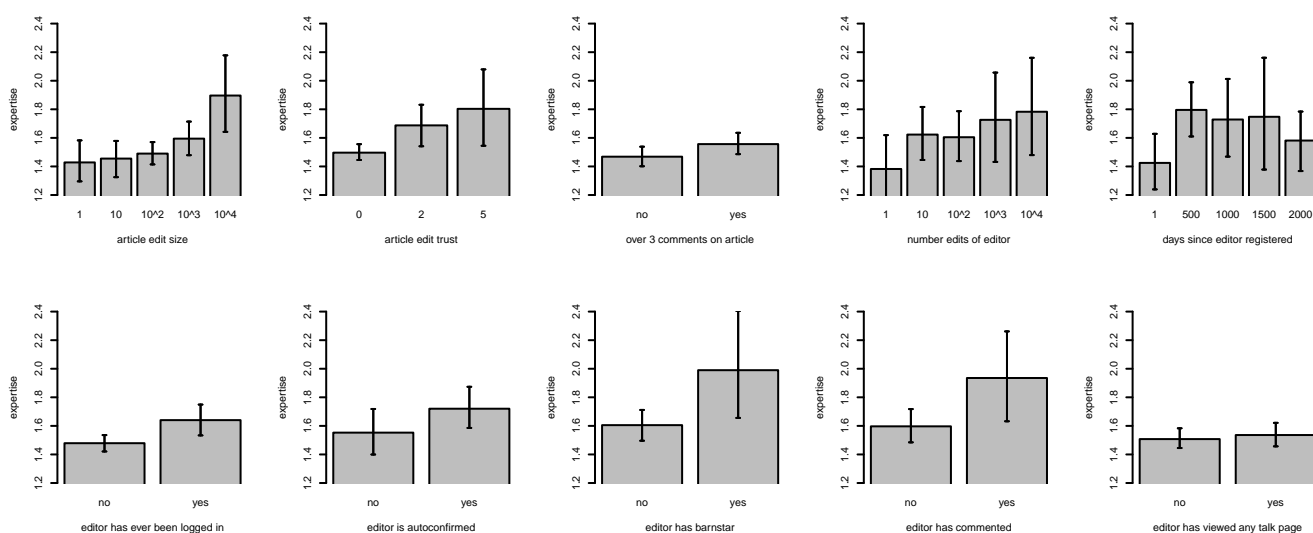
1. Most notably, editors that use a Wikipedia account show more expertise. That is good as the more involved users are better experts.
2. This is confirmed by further findings (where we only consider editors that have ever been logged in to Wikipedia because only for them can we find the respective properties in the Wikipedia logs): among those logged in, we check if they have a barnstar;<sup>15</sup> those that do have one also have significantly more expertise. Note that this fact may be interpreted as supporting the validity of our definition of expertise.
3. Editors that have ever made a comment on any article are more experts.
4. We also correlate expertise with the number of edits the user has made overall and with the time they been registered with

<sup>14</sup>The graph looks basically identical when we use median rather than mean expertise per category, indicating that the results are not dominated by outliers.

<sup>15</sup>A barnstar is a Wikipedia-internal award given by other Wikipedia editors. See <http://en.wikipedia.org/wiki/Wikipedia:Barnstars> for details.



**Figure 4: The co-expertise graph: a bipartite graph connecting edit categories (on top) with expertise categories (on the bottom). An edge’s gray tone represents expertise strength. People editing science articles have expertise in many categories. Similarly, editors from many domains have expertise in humanities, law and social science.**



**Figure 5: A breakdown of micro-averaged expertise (on the y-axis) according to various features. EAPs were binned into bins of roughly equal size. Noteworthy observations are: (i) longer edits tend to be done by users with more expertise, (ii) the notion of expertise correlates with an existing measure of edit trust, (iii) editors with more edits tend to have more expertise, and (iv) users with a barnstar tend to have more expertise.**

Wikipedia for; while the error bars are too large to make a strong statement, it seems that the ‘newbies’ have less expertise: considering only the editors with exactly one edit and those that have been registered for at most 1 day (i.e., they probably registered to make the edit in our toolbar logs) we see that these users have the lowest expertise from among all bins.

- Also, we conjectured that users that view talk pages of articles are more involved and show greater expertise, but could not confirm this (rightmost plot in lower row).

### 5.5 Do editors do research just before an edit?

In the previous subsection, we have seen that expertise is systematically higher for certain editors, articles, and edits. The notion of expertise we adopted was based on a random sample of 1K queries from the editors’ entire browsing histories. Now we are interested in investigating expertise from a temporal perspective. Specifically,

we would like to find out if the editor’s queries just before the edit are more related to the edit.

For this purpose, we define the notion of *temporal expertise*: for each edit, we extract the queries issued by the editor in the time span 30 min immediately before the edit. While doing so, we ignore immediate duplicate queries. For instance, the query sequence  $(q, q, q, r, q, s)$  (6 queries) will be taken as  $(q, r, q, s)$  (4 queries).

The mean number of searches in the time window is 3.75 (median: 2.75). To get a sense of whether this number is high, let us compare this against the number of searches in a non-edit situation. For this purpose, we replace each unique edited article by a unique viewed yet not edited article and count the searches within 30 minutes before the view. Surprisingly, there are *more* searches before non-edit views than before edits (mean: 4.91, median: 3). Possible explanations could be that the edit itself takes time away from the 30 minutes—time that could be used for searching, or that ‘research’ might not be in the form of search engine queries, but rather in that of Wikipedia views.

Editor  $e$ 's temporal expertise with respect to an edit  $a$  is defined as in Equation 3, with the difference that the interest is now not computed from a random sample of 1K of the editor's queries but based on the queries in the 30-minute pre-edit window. To see if the searches immediately before the edit are more related to the edited article than average we look at the ratio of the editors' temporal with their general expertise. Let us call this the 'expertise boost'. We also compute the same ratio after removing the last query before the edit suspecting that this might often be a navigational query taking the editor to the edited page.

The micro-averaged expertise boost (i.e., each EAP is a data point) is 2.96, or 2.35 when excluding the last query. Macro-averaged (i.e., each user is a data point) these numbers become 4.15 and 3.06 respectively.

We conclude that editors seem to show increased interest in the topic of the article in question immediately before editing it. In the next section we will investigate in more detail how exactly they arrive on the article page, and whether they do so in a targeted fashion.

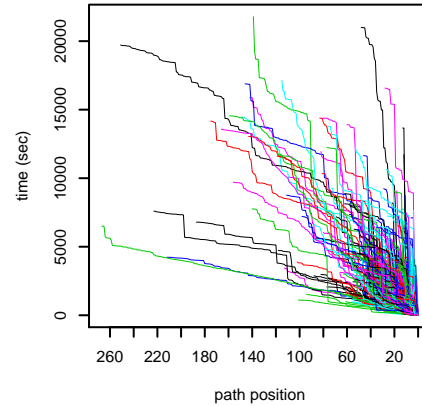
## 6. HOW DO EDITORS ARRIVE AT THE EDITED ARTICLE?

To address this question we look at the referrer chains ending in a article edit. Such a referrer chain is constructed by retracing the sequence of referrers in reverse chronological order. In total, we considered 2,652 referrer chains. Note that a single chain can and often does contain several edits.

We observe two main types of referral chains: 47% of the chains (1,251) originate from a search and 49% (1,287) start on Wikipedia with the user typing the URL directly or having bookmarked it. Only 114 (4.3%) of chains originate from other pages. Of the chains with a search origin, 67.7% (871) stem from a query that is highly related to the edited article title (we define this as  $\text{sim}(a, q) > .5$ , see Appendix A). A surprisingly high 30.1% (388) of search openings go directly from the search engine result page to the edit, making up 14.6% of all edit chains.

The mean chain length is 14.5 pageviews, with a median of 9. There's a long tail in terms of chain length and this explains the fact that the mean is larger than the median. Fig. 6 visualizes these chains as follows. On the  $x$ -axis we plot the position along the click path in terms of clicks to go till the final edit (at 0); we see that there are a few very long chains, although most are short. The  $y$ -axis has the time till the final edit. We see that the longest click chains span more than 5 hours of browsing activity. The slope of the wires gives an impression of the click chain's speed: a steep slope corresponds to lots of time spent on few page views and a flat slope corresponds to a fast 'clicker'. On average, chains have a click rate of 97 seconds per click, but there is a lot of variability. Chains that start with a highly related search also have shorter length, with a mean of 10.0 and a median of 7. On the other hand, chains starting on the Wikipedia home page have a mean length 18.3 and a median 12. We therefore conclude that edits that use a search engine as an entry point typically imply less pre-edit Wikipedia reading on the editor's behalf, compared to edits that start on Wikipedia's main page. We conjecture that these might often be the edits the editor stumbled across, rather than had in mind to begin with.

To study correlation with expertise, we divided the referrer chains into four sets according to quartiles of the chain length. In increasing order of length the average expertise values are 1.74, 1.79, 1.92 and 2.00, indicating that users who read more Wikipedia articles before committing an edit tend to have more expertise. The longer referral chains might be a manifestation of pre-edit research, while



**Figure 6:** Each line in this graph corresponds to a chain of referrers ending in a edit event. The number of pageviews of such a chain is plotted on the  $x$ -axis and the time duration on the  $y$ -axis. A flat line corresponds to a rapid succession of clicks and a steep line to a slow succession.

the edits occurring in short click chains possibly correspond to minor changes such as typo fixes.

## 7. CONCLUSIONS AND FUTURE WORK

Concerns about the quality of Wikipedia have been present since the beginning of the project, and systematic approaches to try to provide enough evidence to close this debate [7] seem to fuel it instead.

This paper attempts to shed light on yet another question that is of importance in order to understand the phenomenon of Wikipedia: Who are the people that contribute to it? In particular, we try to draw the portrait of Wikipedia editors in a data-driven fashion, and approach the question from many different angles, characterizing their typical Web usage patterns and their levels of expertise.

While editors as a whole seem to expose certain traits, such as being more sophisticated than average Web users, it is important to note that there is no such things as the quintessential Wikipedia editor. Rather, there seem to be specific classes of editors heavily specialized in entertainment, sports, or news media, and in general in popular culture. On the other end of the spectrum, there is another class of editors that works in topics such as science, technology, business, humanities, etc.; given that they seem to be experts in more than one topic, we may think that there is also a large population of 'generalists' in Wikipedia, whose knowledge may be really encyclopedic (pun intended).

Considering that there is not a single type of editor in Wikipedia, it should—as a collaborative system—continue striving to accommodate diversity. In this light, specialized *WikiProjects* that cater to particular subgroups of users (e.g. 'WikiProject: Video Games'<sup>16</sup>) should be encouraged.

From a methodological viewpoint, our method is more general than what we have presented here: it can be applied to any site in which content can be produced and where we can reliably identify actions that constitute productive acts; and we could even think

<sup>16</sup>[http://en.wikipedia.org/wiki/Wikipedia:WikiProject\\_Video\\_games](http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Video_games)

about linking the explicit behavior we observe to psychological variables in order to confirm or refute existing hypotheses.

## 8. REFERENCES

- [1] B. Adler, K. Chatterjee, L. De Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *Proceedings of the 4th International Symposium on Wikis*, pages 26:1–26:12, 2008.
- [2] B. T. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contributions to the wikipedia. In *Proceedings of the 4th International Symposium on Wikis (WikiSym)*, pages 15:1–15:10, 2008.
- [3] D. Anthony, S. W. Smith, and T. Williamson. The quality of open source production: Zealots and good samaritans in the case of Wikipedia. Technical report, Dartmouth College, Computer Science, 2007.
- [4] Y. Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, 2006.
- [5] S. Brin. Extracting patterns and relations from the World Wide Web. *The World Wide Web and Databases*, pages 172–183, 1999.
- [6] U. Cress and J. Kimmerle. A systemic and cognitive view on collaborative knowledge building with wikis. *I. J. Computer-Supported Collaborative Learning*, 3(2):105–122, 2008.
- [7] J. Giles. Internet encyclopedias go head to head. *Nature*, 438:900–901, 2005.
- [8] Y. A. Hamburger, N. Lamdan, R. Madiel, and T. Hayat. Personality Characteristics of Wikipedia Members. *CyberPsychology & Behavior*, 11(6):679–681, 2008.
- [9] T. Holloway, M. Bozicevic, and K. Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors. *Complexity*, 12:30–40, January 2007.
- [10] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th international conference on World wide web (WWW)*, pages 561–570, 2010.
- [11] K. K. Lee and G. G. Karuga. The role of cognitive conflict in open-content collaboration. In *Proc. of AMCIS 2010*, 2010.
- [12] O. Nov. What motivates Wikipedians? *Communications of the ACM (CACM)*, 50:60–64, November 2007.
- [13] F. Ortega. *Wikipedia: A quantitative analysis*. PhD thesis, Universidad Rey Juan Carlos, Madrid, Spain, 2009.
- [14] A. J. Reinoso, F. Ortega, J. M. Gonzalez-Barahona, and G. Robles. A quantitative approach to the use of the wikipedia. In *IEEE Symposium on Computers and Communications (ISCC)*, pages 56–61, 2009.
- [15] M. Wattenberg, F. Viégas, and K. Hollenbach. Visualizing activity on Wikipedia with chromograms. *Human-Computer Interaction–INTERACT 2007*, pages 272–287, 2007.
- [16] I. Weber and A. Jaimes. Who uses web search for what: and how. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM)*, pages 15–24, 2011.
- [17] H. Welsler, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, pages 122–129, 2011.
- [18] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on*

*Web Search and Data Mining (WSDM)*, pages 132–141, 2009.

- [19] Wikimedia Foundation. Editor survey, April 2011. [Online; accessed September 13, 2011].

## APPENDIX

### A. ARTICLE–QUERY SIMILARITY

First of all, we treat an article  $a$  as a query, too, simply by using its title as a proxy. This way, we only have to define the similarity between two queries. For this purpose, we issue both queries to the Yahoo! search engine and obtain the top 10 results. Each such result comes with a classification into the Yahoo! Directory according to a machine-learned classifier. Categories are hierarchical and an example is ENTERTAINMENT / SPORTS / TENNIS (length 3). We then compute the weighted average pairwise *category similarity* between the two result lists: the similarity between two categories is the length of their longest common prefix, divided by the length of the shorter category. The weight for pair  $(i, j)$  is  $(10 - i) + (10 - j)$  (normalized such that all weights sum to 1). Call this weighted average category similarity  $\text{sim}'(a, q)$ . Using this measure as described, a query could in general have a similarity of less than 1.0 with itself unless the categories of its 10 results are all the same. We account for this by considering the ratios  $\text{sim}'(a, q) / \text{sim}'(a, a)$  and  $\text{sim}'(a, q) / \text{sim}'(q, q)$ ; the final similarity  $\text{sim}(a, q)$  is then defined by the harmonic mean of these 2 ratios (harmonic instead of arithmetic because the numerators rather than the denominators are the same).