Drawing 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
Qatar Computing Research Institute
Doha, Qatar
chato@acm.org

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 online 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 online lives beyond Wikipedia. We broadly characterize editors by investigating how their online 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 popular culture. Then we inspect how editors' general interests relate to the articles to which they contribute; e.g., we confirm the intuition that editors are more familiar with 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. Our results are relevant as they illuminate novel aspects of what has become many Web users' prevalent source of information.

Categories and Subject Descriptors: H.1.2 **[Models and Principles]**: User/Machine Systems—*Human factors*.
General Terms: Experimentation, Human Factors, Measurement.
Keywords: Wikipedia, editors, Web usage, expertise.

1. INTRODUCTION

Wikipedia is one of the technological and sociological 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 make more than 100 edits every month.[1]

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 data obtained by the Yahoo! Toolbar and look for specific URLs corresponding to Wikipedia edits. This way we can identify Wikipedia editors and obtain insights into their browsing behavior 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:

First, we find that on broad average Wikipedia editors seem, on the one hand, more 'hungry for information' 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 popular 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.

Finally, we introduce a notion of interest in, or familiarity with, a topic based on users' search query histories and show that across topical domains Wikipedia editors show significantly more familiarity with the edited articles than average users. We also refine 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 demonstrate that more substantial contributions tend to come from editors more familiar with the edited topic, and that editors with a Wikipedia account expose more familiarity than other editors.

Apart from being interesting in its own right, characterizing Wikipedia editors may be useful from a practical perspective. One of the most pressing challenges Wikipedia is currently facing is to combat a decline in the number of active editors [20]. To fill the resulting void, readers need to be converted into editors, and in order to target the most promising readers, it can help a lot to know what a typical editor is like. In this respect, organizations such as the Wikimedia Foundation can directly profit from the results of our research.

The rest of this 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 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 editors' familiarity with the edited topics and how it correlates with quantities

*Part of this work was done while all authors where at Yahoo! Research Barcelona (R.W. as an intern). A compressed version of this paper was previously published [17].

[1] `http://en.wikipedia.org/w/index.php?title=Wikipedia:Wikipedians&oldid=457802123`

WikiSym '12, August 27–29, 2012, Linz, Austria.

such as the size of edits or the topics of the edited articles. Finally, Section 6 discusses future work and concludes the paper.

2. RELATED WORK

A significant amount of research has been done on Wikipedia, both because of its significance (it is the seventh most visited site on the Web[2]) 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 April 2011 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 a desire to volunteer 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. The use of toolbar data for studying user behavior on the Web is a well established paradigm; for a recent study including over 50 million pageviews 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 pageviews that Wikipedia serves are related to editing actions.

Expertise. In small-scale analyses, experts can be identified by using surveys or looking for academic or technical qualifications. In a large-scale analysis, however, proxies for expertise need to be used. The preferred method used by expert-finding methods that rely on traffic analysis has been to identify expertise with being *familiar* with a topic, more than being *proficient* at topic-related tasks; 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). In this work we adopt a similar methodology.

3. DATA SET DESCRIPTION

Since early 2008, users of the Yahoo! Toolbar[3] have the option to allow Yahoo! to collect information about the websites they visit. In accordance with our privacy policy,[4] we employ these data for research purposes without using or accessing personally identifiable information about the toolbar user at any time.

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

We use toolbar ids as anonymous user ids. Though it is possible that Wikipedia editors use several distinct computers—with or without a 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 (cf. Section 4) and the observations concerning familiarity (cf. Section 5) are only more pronounced, and the trends we identify are expected to remain 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 (0.089% of all users), readers-only of the English Wikipedia (58%), and those that do not read any language version of Wikipedia (41%). Note that users reading or editing only non-English Wikipedia versions are excluded from our analysis.

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 per group. Note that this implies that Wikipedia users (and editors in particular) generally spend more time online. Later, in Section 4.1, we analyze in detail *how* they spend it.

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 (users often click on the 'edit' link but make no changes [14]) 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[5] and keep only those for which we find a match. Not all edits are equal, though; e.g., a revert edit might appear to have introduced significant new content (if a delete edit is being reverted), while in fact it merely reintroduces content previously added by a potentially different editor. However, we are interested in edits that are likely to correspond to novel content creation on behalf of the acting editor, and hence we ignore the 113 revert edits we could identify in our data set by checking the edit summary for specific substrings such as 'revert' or 'rv'.[6] For the same reason we will sometimes (where noted) also restrict ourselves to edits of a minimum size.

Finally, we removed all edits that were immediately (i.e., within a couple of seconds) followed by a bot,[7] as in those cases it is not

[2] http://www.alexa.com/siteinfo/wikipedia.org
[3] http://toolbar.yahoo.com/
[4] http://info.yahoo.com/privacy/

[5] We use the Wikipedia API at http://www.mediawiki.org/wiki/API:Properties.
[6] The most important patterns are listed at http://en.wikipedia.org/w/index.php?title=Wikipedia:Edit_summary_legend&oldid=458695721. Our regular-expression approach detects most of the reverts according to our inspection of the result.
[7] An automatic agent that does maintenance tasks on Wikipedia, identified via the official bot registry at http://en.wikipedia.org/w/index.php?title=Wikipedia:

possible to reliably identify which of the two (toolbar user or bot) performed the edit.

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 weight 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 one user made to one article.

We define the *edit size* of an EAP as the maximum edit size[8] 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 at the same time added 101, we will count an edit size of 1. The distribution of edit sizes is heavy-tailed for both positive as well as negative edit sizes: as observed in previous work [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 zero.

3.4 Edit topic distribution

Wikipedia is a general encyclopedia spanning many realms of knowledge that vary widely along several dimensions: some topical areas are large, others are small; some change rapidly, others contain information that is updated only rarely.

It will be useful for our analysis to understand how the edits in our sample are distributed over topical areas. To estimate this distribution we proceed as follows. First 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[9] 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 [15]. For instance, ANTHOLOGY gets the category ARTS/HUMANITIES/LITERATURE, and CARNIVOROUS PLANT is classified as SCIENCE/BIOLOGY/BOTANY.

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 work; e.g., Holloway *et al.* [9] show that seven of the ten largest categories by number of articles in Wikipedia are related to music, films, or television. This entertainment bias will be a recurring theme in many places throughout our analysis.

3.5 Potential sampling biases

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 typical Wikipedia editors. For instance, it is conceivable that the aforementioned bias towards the entertainment domain might not hold for Wikipedia editors in general but rather be a peculiarity of those editors that also use our toolbar.

We investigate potential biases by comparing our toolbar sample to a representative sample of 74K recent Wikipedia edits, collected on Wikipedia over a period of 6 days, by fetching 500 recent edits every hour[10] (here we deal with single edits rather than EAPs).

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.

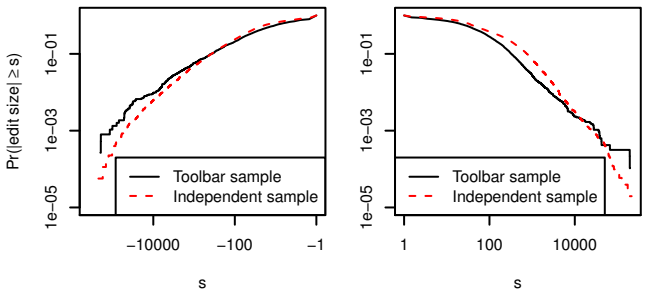

Figure 1: Log-log plots of the complementary cumulative distribution function (ccdf) of edit size. *Left:* edits of negative size. *Right:* edits of positive size.

In summary we find that the editors in our toolbar sample seem somewhat less involved in Wikipedia than those in the independent sample. This is hinted at by the following numbers (the null hypothesis of the pairs of means being equal is rejected by t-tests with Bonferroni-corrected $p < 0.05$): (1) Our sample has fewer edits in namespaces other than the main namespace[11] (5.1% vs. 12%), and we expect those non-standard namespaces to be edited by more involved users. (2) The toolbar sample comprises slightly more edits of negative size (34% vs. 30%) but (3) significantly fewer by users logged in to Wikipedia (41% vs. 77%). (4) Among the edits made by logged-in users, the percentage made by administrators is much smaller in the toolbar sample (0.39% vs. 12%).

The distributions of edit sizes are qualitatively similar in the two samples (cf. Fig. 1), with the toolbar sample containing smaller edits on average (median positive edit size: 36 vs. 52 bytes; median negative edit size: −15 vs. −26 bytes).

The distributions of editor age, measured in the number of days between the edit and the time the respective editor account was established, are also qualitatively similar (cf. Fig. 2; every editor is counted only once). The main difference is that the independent sample of live Wikipedia has a longer tail of very old users, which results in a much higher average age compared to the toolbar sample (mean 967 vs. 475 days; median 837 vs. 66).

All these numbers suggest that on average the editors in the toolbar sample are newer to, and less involved in, Wikipedia than the contributors of typical edits. While it is important to be aware of this bias, we think it does not compromise the relevance of our work, for two reasons: First, the group of newer editors is of particular interest to the Wikipedia community, an issue we discuss in Section 6. Second, the aforementioned differences notwithstand-

List_of_bots_by_number_of_edits&oldid=447262313.

[8] In practice, the largest edit is typically the first one in a series of micro-edits, followed by small corrections.

[9] http://dir.yahoo.com

[10] http://en.wikipedia.org/wiki/Special:RecentChanges

[11] File, File_talk, Format, Help, Special, Template, Template_talk, User, User_talk, Wikipedia, Wikipedia_talk

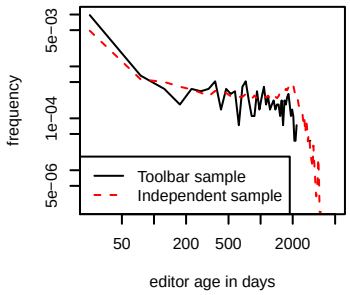

Figure 2: Log-log plot of the distribution of editor age.

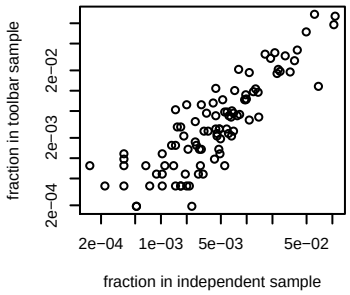

Figure 3: Log-log scatter plot of the distributions of categories of edited articles. There is high correlation between the toolbar sample used in this paper and an independent sample of recent edits on live Wikipedia. The outlier to the right is RECREATION/SPORTS/SOCCER, which is due to a tournament that was ongoing at sampling time.

ing, the toolbar and independent samples are very similar in terms of the articles edited, which becomes clear when we correlate the distributions over categories of edited articles for the two samples: Articles were again mapped to Yahoo! Directory categories as described in Section 3.4, and Fig. 3 shows a scatter plot of the two category distributions. A diagonal line of slope 1 would imply that the two distributions are identical, and we see that this is indeed nearly the case, with a Pearson correlation coefficient of $r = 0.88$. We also note that, in particular, the entertainment bias of Section 3.4 is a Wikipedia-wide phenomenon not specific to our sample.

As a final remark, it would also be desirable to know whether our sample of non-editors is biased. However, unfortunately we do not have access to ground-truth Web usage logs for non-editors that could parallel the public Wikipedia logs for the case of editors.

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 three times as many pageviews as readers-only, and nine 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.[12] 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 Yahoo! Directory for the respective category (e.g., ENTERTAINMENT/GAMES). Details are provided in Table 2.

Fig. 4 contains a summary of the differences between the three groups with respect to these 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 website for readers-only is in a middle ground between editors and non-readers. 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 Fig. 4 include the following. Wikipedia editors seem more 'information-hungry' (more news, more educational domains, more reference lookups, and more searches) but also more interested in popular culture (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 most salient 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 five video views. For each view, we use the YouTube API[13] 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 (Bonferroni-corrected $p < 0.05$) differences for the following categories: editors watch more ENTERTAINMENT (19% vs. 17%), FILM (6.5% vs. 4.9%), and GAMES (4.6% vs. 2.2%—i.e., over twice as much), whereas non-editors watch more videos from the categories PEOPLE and AUTOS.

These numbers lend further support to the hypothesis that Wikipedia editors are more immersed in popular 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 You-

[12]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.)

[13]`https://developers.google.com/youtube`

Domain class	Description
REFERENCE	From Yahoo! Directory; this class contains dictionaries, Q&A sites, and encyclopedias; phone books, 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 Yahoo!, Google, Bing, etc., but also site search, product search, and the like. To this 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 Brin's [5], starting from a seed list of such patterns for the major Web search engines. This mechanism initialized with, say, the pattern `p=` used on `http://search.yahoo.com`, might 'learn' that a common value for this parameter is `howto%20install%20latex`. 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.
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 were classified into the Yahoo! Directory category ENTERTAINMENT/GAMES by an in-house machine-learned classifier. The same classifier was used for several of the following categories.
PROGRAMMING	All domains classified into the category COMPUTERS&INTERNET/PROGRAMMING&DEVELOPMENT.
SPORTS	All domains classified into the category 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	`youtube.com`
`.org` (non-Wiki)	All domains under the `.org` top-level domain (non-profit organizations), except `wikipedia.org`.
ADULT	All domains from the category SOCIETY&CULTURE/SEXUALITY, combined with those listed on `http://www.tblop.com`.
SOCIAL NETWORK	`facebook.com`, `myspace.com`, `hi5.com`, `orkut.com`, `friendster.com`

Table 2: Description of the domain classes referred to in Fig. 4.

Tube, 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 made in 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: 0.85% vs. 0.28%, all domains in the 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: 0.28% vs. 0.035%, all domains in the class MOVIES & TV: 1.9% vs. 1.4%); t-tests for checking for a difference in means yielded $p < 0.01$ 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.

Recall from the opening paragraph of Section 4 that Wikipedia usage is correlated with Web usage in general, with editors having on average three times as many pageviews as readers-only, and the latter three times as many as non-readers. Therefore, it is conceivable that many of the differences shown by Fig. 4 might be caused by a user's overall number of pageviews as a single latent factor: it would be intuitive to expect users with more overall online time to spend a larger fraction of that time on entertainment-related sites (e.g., while procrastinating). We rule out this hypothesis in the following experiment: in each of the three user groups, we take each user as a data point and compute Pearson's correlation coefficient between the *absolute* overall number of pageviews and the *relative* frequency of each domain category of Fig. 4. In no user group and for no category do we find a large correlation; in particular, relative YouTube usage is even slightly negatively correlated with absolute Web usage, with Pearson's $r \approx -0.04$.

Let us draw a quick summary of the emerging picture: editors spend more time online; they seem more 'information-hungry' than average users, in the sense that they read more news, search more, and look up more things on reference and academic sites; and 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.

11.239	**`facebook.com`**	−1.681	`google.com`
0.022	`picnik.com`	−1.493	`wikipedia.org`
0.014	`farmville.com`	−1.249	`youtube.com`
0.011	`google.lk`	−0.222	`google.co.in`
0.009	`formspring.me`	−0.168	`ebay.com`

Table 3: Entries of the first principal component of the user–domain matrix with the largest absolute values. *Left:* top positive domains. *Right:* top negative domains.

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?

To this end, we perform principal component analysis (PCA) on users' relative domain frequency vectors.[14] 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 3. The entries with the largest absolute weights tell us with respect to which domains there is most variation among editors; e.g., Facebook's high weight—by far the largest—implies that there are many editors with very high and many with very low Facebook activity. In other words, Facebook seems to be the main line of divide within the group of editors. Furthermore, entries of the same sign are correlated and those of opposite signs anticorrelated: A positive correlation between using Wikipedia on the one hand and Google (or search in general) and YouTube on the other was already hinted at by Fig. 4, and now we also see that editors with heavy Facebook usage tend to frequent Google, YouTube, and Wikipedia[15] less.

[14] We ignore pageviews on `*.yahoo.com`, to minimize bias, but the result is nearly exactly the same when we keep it.

[15] Even when the special domain `wikipedia.org` is removed from the data matrix, the PCA results remain otherwise unchanged.

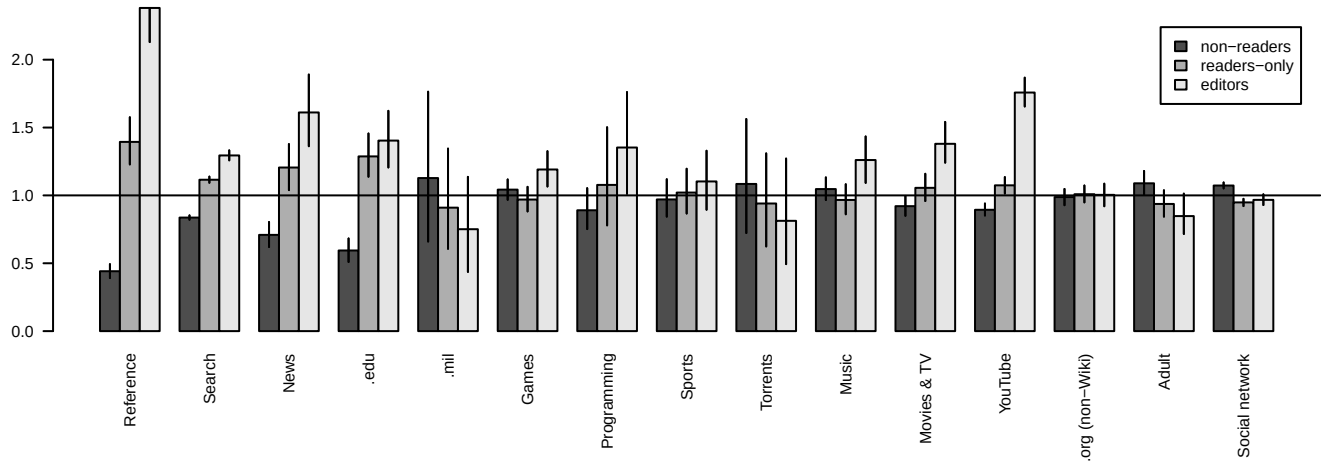

Figure 4: Category and domain frequencies for the three user groups, each macro-averaged over its users. The horizontal bar at 1.0 corresponds to the overall average for each category for general Web users computed by weighting the three groups by their relative sizes (41%, 58%, 0.089%). All fractions are normalized by this global average so we can plot everything in one figure. Error bars are 95% confidence intervals obtained by bootstrap sampling.

We obtain further evidence for a dichotomy 'much Facebook vs. much Wikipedia' among editors by looking at how the level of Facebook usage correlates with editing behavior. To this end, we first cluster all editors using the k-medoids algorithm,[16] which yields two fairly balanced clusters (the optimal number according to average silhouette width) and, due to the strong influence of the first principal component, essentially groups the editors according to their loadings with respect to that component: 47% of editors fall into the Facebook cluster, 53% into the less-Facebook cluster. Now, to see how the editing behavior of less-Facebook editors differs from that of those in the Facebook cluster, we compute the means of certain editing-related properties for each cluster. In order to allow for easier interpretability, we group the large number (93) of Yahoo! Directory categories into the 12 high-level categories ENTERTAINMENT, NEWS & MEDIA, BUSINESS, HUMANITIES LAW & SOCIAL SCIENCE, HEALTH, HOBBIES, SCIENCE, ARTS, SPORTS, ADULT, SHOPPING, and TECHNOLOGY. Then we compute for both clusters a distribution over these categories, with respect to the edits made, and find two significant (Bonferroni-corrected $p < 0.1$) differences: editors in the Facebook cluster have more edits in ENTERTAINMENT (47% vs. 40%), while in the other cluster we see more edits related to NEWS & MEDIA (4.7% vs. 2.5%).

Editors from the less-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 Wikipedia edits based on revision history and author reputation features. We find that the average trust value attributed to edits in the less-Facebook cluster is 0.22, while it is only 0.086 in the Facebook cluster. (All reported differences are significant with Bonferroni-corrected $p < 0.05$). 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 less-Facebook users comes from the fact that in the less-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 vs. 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 NEWS & MEDIA edits. Users from the less-Facebook cluster are more involved in Wikipedia as signified by (1) larger edits, (2) a higher chance of being logged in to Wikipedia, (3) more edits per user, and (4) higher edit trust scores.

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 above approach (PCA and investigating the first principal component) also for the sets of readers-only and non-readers. In both cases, the result looks very similar to Table 3, and hence the distinction 'Facebook vs. less-Facebook' spans across user groups.

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 how familiar they are with the areas in which they make edits.

5.1 Defining interest and familiarity

Ideally, we would like to quantify whether editors are experts in their domains. Unfortunately, this is a subtle notion capturing the proficiency at topic-related tasks, which is hard to measure, particularly in a large-scale analysis such as ours. Hence, we consider a more amenable notion of *familiarity*. Following White *et al.* [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.

[16] 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.

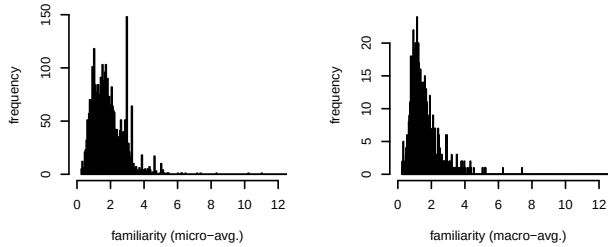

Figure 5: Histograms of familiarity, both micro-averaged (each EAP contributes equally) and macro-averaged (each user contributes equally).

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):

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

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

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

Some query categories are more common than others [15], so high interest in a certain topic alone is not necessarily very informative. We define someone to be familiar with a topic if their interest is significantly above average. Formally, we define e's *familiarity* with an article a as

$$F_e(a) := I_e(a)/I_{\text{avg}}(a). \tag{3}$$

A familiarity greater than 1 implies above-average interest in the given topic, and conversely for a familiarity less than 1. While it may be a somewhat unfamiliar use of the word, we will speak of the 'familiarity of an edit' (measuring how familiar the editor is with the edited article) and the 'familiarity of an editor' (meaning the average familiarity of an editor over all articles he/she edited).

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 (cf. Table 2) is included, with these exceptions: navigational queries (defined using a click-entropy threshold [15]), 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 characters and contain no whitespace character.

5.2 How familiar are editors with their edited topics?

Using the definition of familiarity from the previous section, we now characterize the familiarity 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 sizeable edits that can be achieved with a mere click rather than through novel content creation. This leaves us with 83% of all EAPs.

Fig. 5 shows histograms of familiarity over all EAPs. The micro-averaged familiarity is 1.85 (the 95% confidence interval computed via bootstrap resampling is $[1.82, 1.88]$), and macro-averaged—i.e., first averaging all EAPs for each user—it is 1.52 (with 95% confidence interval $[1.46, 1.57]$). 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 familiarity histogram (left part of Fig. 5). 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 are more familiar with the topics they edit than the average Internet user.

This bias towards familiarity 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 he/she edits the page, regardless of article-specific familiarity. 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. According to this argument, the results above would not necessarily be edit-specific but would 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 familiarity obtained this way is 1.62 (with 95% confidence interval $[1.59, 1.65]$) and macro-averaged it is 1.41 (confidence interval $[1.36, 1.46]$). So, since these numbers are significantly lower than when an actual edit rather than a mere article view takes place, the observed familiarity cannot be solely explained by the simple interest-only model described above.

4.16 SPORTS	1.80 HEALTH
2.70 TECHNOLOGY	1.68 SCIENCE
2.66 NEWS & MEDIA	1.65 ENTERTAINMENT
2.16 HUMANITIES LAW & SOCIAL SCIENCE	1.54 ARTS
2.03 SHOPPING	1.53 HOBBIES
1.92 BUSINESS	0.87 ADULT

Table 4: List of categories in order of decreasing category-specific familiarity.

5.3 Is there more familiarity in some domains than others?

The previous section indicates that, on average, Wikipedia edits are made by people with above-average interest in a particular topic. But do 'experts' exist to the same extent across all topics? To answer this, we average the familiarity 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 familiarity. We find that familiarity 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 familiarity of 0.87. However, only 11 edits were contributed in this category. Table 4 contains a listing in order of decreasing category-specific familiarity.

What about other categories? Can we make a statement about whether editors of certain categories are also familiar with other domains? We answer this question quantitatively using the notion of *co-familiarity,* proceeding as follows: For each edit category c_1, we compute a co-familiarity profile. The profile has one entry per category c_2, the value being the mean familiarity with 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. 6. In this *co-familiarity graph,* one partition (the upper one in Fig. 6) represents edit categories, the other (the lower one in the figure) familiarity categories. Edges are drawn from c_1 to c_2 if, on average, editors that edit c_1 have familiarity greater than 1 in category c_2. Additionally, an edge's gray tone represents familiarity strength.[17]

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

The co-familiarity 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 popular 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 familiarity instead of raw domain frequencies: familiarity 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-familiarity graph originates from editors of ENTERTAINMENT articles. On the flip side, editors of ENTERTAINMENT have no other areas of strong familiarity, visualized by the lack of strong 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. Note that, on the contrary, editors of SCIENCE and BUSINESS seem to be more versatile: they are familiar with several areas beyond what they edit.

5.4 What are the correlates of familiarity?

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

Fig. 7 summarizes our findings graphically. The x-axes show the respective properties, the y-axes familiarity. 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 two plots in the upper row relate features of EAPs to familiarity: First, long edits (notably the very long ones) come from editors with more familiarity; this is a good sign: small edits are often minor corrections such as typo fixes, while the large ones are the proper content contributions, which we would hope to come from real 'experts'. Second, articles with greater edit trust [1] come from editors with more familiarity.

The third plot in the upper row 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 the editors of those articles are slightly more familiar with the edited content (we chose three comments as the threshold to have two roughly balanced sets to compare).

The remaining properties describe the editors whose familiarity we are evaluating. Most notably, editors who use a Wikipedia account show more familiarity (second row, first plot). This is good, as the more involved users are more familiar with what they edit.

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',[18] a Wikipedia-internal award given to deserving editors by other Wikipedia editors (or even by themselves). Those that do have one also have significantly more familiarity (second row, third plot). Next, editors that have ever made a comment on any article are more familiar with the edited article (second row, fourth plot). We also correlate familiarity with the number of edits the user has made overall (first row, fourth plot) and with the time for which he/she has been registered with Wikipedia (first row, last plot). While the error bars are too large to make a strong statement, it seems that the 'newbies' have less familiarity: considering only the editors with exactly one edit and those that have been registered for at most one day (i.e., they probably registered to make the edit in our data set), we see that these users are least familiar with what they edit.

Finally, we conjectured that users who view talk pages of articles are more involved and show greater familiarity 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 familiarity is systematically higher for certain editors, articles, and edits. The notion of interest and familiarity we adopted was based on a random sample of 1K queries from the editors' entire browsing histories. Now we are interested in investigating interest 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 than her/his usual queries are.

For this purpose, we analyze our data from a temporal perspective: for each edit, we extract the queries issued by the editor in the time span 30 minutes immediately before the edit. In doing so, we ignore immediate duplicates. 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 it against the number of searches in a non-edit situation. To this end, 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 familiarity* 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 familiarity. Let us call this the *familiarity boost*. We also compute the same ratio after removing the last query before

[17] The graph looks basically identical when we use median rather than mean familiarity per category, indicating that the results are not dominated by outliers.

[18] http://en.wikipedia.org/wiki/Wikipedia:Barnstars

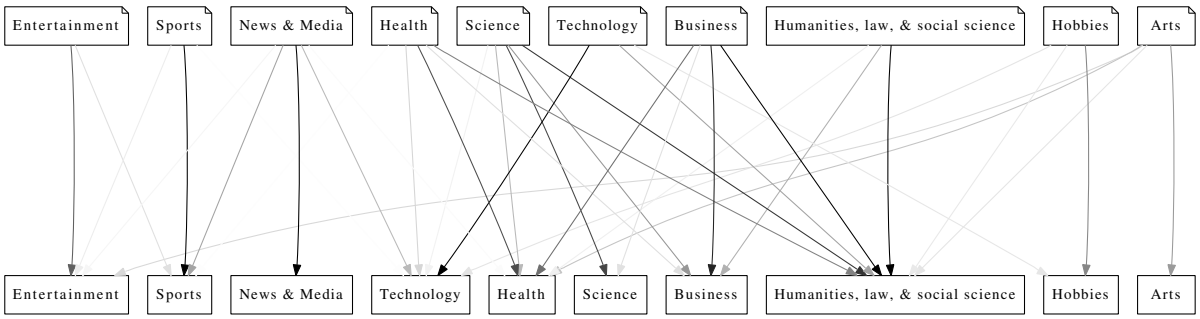

Figure 6: The co-familiarity graph: a bipartite graph connecting edit categories (on top) with familiarity categories (on the bottom). An edge's gray tone represents familiarity strength. People editing science articles are familiar with many categories. Similarly, editors from many domains are familiar with the humanities, law, and social science.

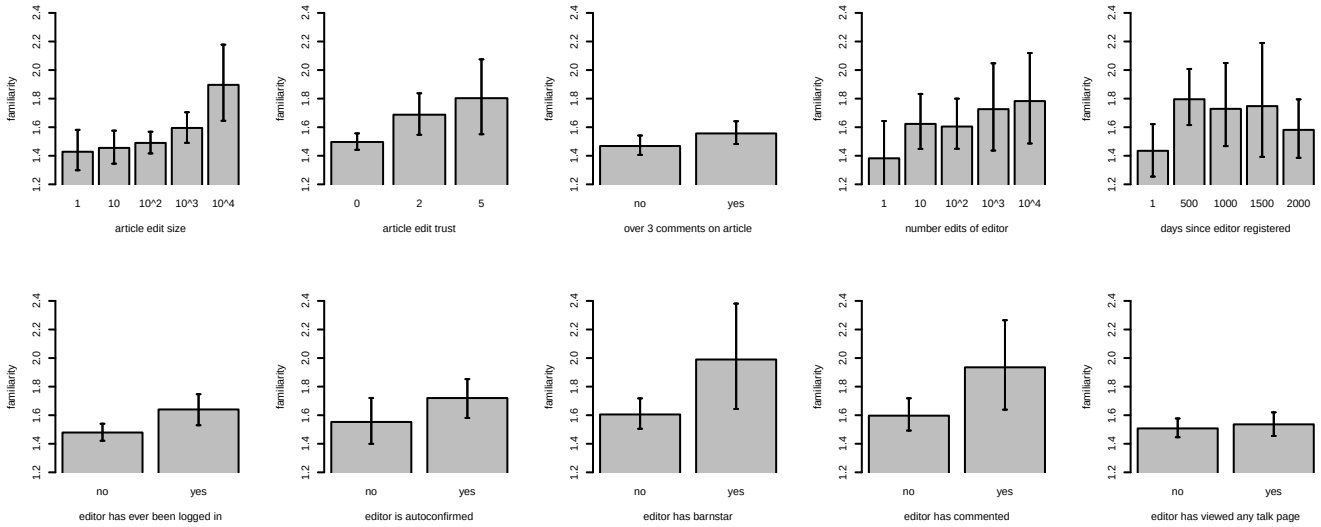

Figure 7: A breakdown of micro-averaged familiarity (on the y-axis) according to various features. EAPs were binned into bins of roughly equal size. Noteworthy observations are: (1) longer edits tend to be done by users with more familiarity, (2) the notion of familiarity correlates with an existing measure of edit trust, (3) editors with more edits tend to have more familiarity, and (4) users with a barnstar tend to have more familiarity.

the edit, suspecting that this might often be a navigational query taking the editor to the edited page.

The micro-averaged familiarity 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, which confirms the intuition that editors seem to show increased interest—likely for research purposes—in the topic of the article in question immediately before editing it.

6. CONCLUSIONS

Concerns about the quality of Wikipedia have been present since the beginning of the project, and systematic approaches trying to provide enough evidence to close this debate [7] seem to fuel it instead. Hence, the efforts made to understand the processes that underlie content creation in Wikipedia continue. 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 contributing to it?

At first glance, Wikipedians seem to share two common traits according to our analysis: they are entertainment-loving and information-hungry. The entertainment bias may be explained by the fact that seven of the ten largest categories of article topics are entertainment-related [9], which in turn might be the case because the world produces information of public interest in terms of movies, TV, music, etc. at a much faster rate than in term of, say, the sciences or humanities. The 'information hunger', too, might have been expected, for the knowledge that editors pour into Wikipedia must be gathered somewhere, resulting in more online time in general, and more searches, reference lookups, and news-reading in particular. But not only do editors search more than average Internet users, they do so especially in the areas to which they contribute. This is not surprising but nonetheless an analysis like ours is necessary to confirm the intuition (and the hope) that editors are familiar with their domains at a level significantly above the average.

It is thus tempting to broadly sketch editors as being 'smart but fun', but the full picture needs to be drawn in a more faceted manner. In previous work, Welser *et al.* [16] identified four key social

roles of Wikipedia editors: substantive experts, technical editors, vandal fighters, and social networkers. Regarding the latter role, our analysis additionally shows that social-network activity outside Wikipedia—particularly on Facebook—has echoes inside Wikipedia, too: editors who are heavily active on Facebook tend to be less involved in Wikipedia. One possible explanation could be that being arguably more social takes a lot of time, which could otherwise be dedicated to more extensive editing.

On the flip side, the editors who are more involved in Wikipedia exhibit more familiarity with their active areas—a sign of a healthy and competent encyclopedic community. Furthermore, some areas seem to draw better 'experts' than others; e.g., search queries by SPORTS editors are four times more related to sports than the average editor's query, whereas those few who edit articles from the ADULT domain have fewer than average adult-related search queries. Also, the large volume of entertainment-related edits seems to come mostly from editors immersed primarily in popular culture, and conversely these editors contribute primarily to entertainment-related articles. On the contrary, those working on science, business, or the humanities seem to be more 'generalist'. Considering that there is more than just a single type of editor in Wikipedia, we conclude that the community should continue to foster diversity, e.g., through specialized 'WikiProjects' that cater to particular subgroups of users (e.g., those interested in improving articles about video games[19]).

An interesting avenue for future research along the lines of this work is to understand what makes readers become first-time editors. Knowing the answer to this question is important in order to combat the current decline in the number of active editors, a daunting challenge that has led the Wikimedia Foundation to make the 'recruitment and acculturation of newer editors' one of its key goals [20]. In this light, it might even be an advantage that our data set is biased towards newer editors (cf. Section 3.5). Our work provides some useful first pointers, e.g., by showing that new editors are often less familiar with the edited articles than more senior ones, complementing the survey finding that 64% of Wikipedians start editing because of minor issues such as typo fixes [19]. Given the different motivations and editing skills, it might be a worthwhile strategy to personalize the viewing and editing interfaces for specific types of users in order to lower the threshold of becoming an editor.

To conclude, while many spots of our portrait of Wikipedia editors remain to be filled in, we hope that it will eventually help to inform focused strategies for convincing the aptest and most promising readers to become active members of one of the most intriguing—and useful—phenomena of the Web.

7. ACKNOWLEDGEMENTS

This research has received funding from the Spanish Ministry of Science and Innovation through project CEN-20101037 CENIT Social Media, the EC's Seventh Framework Programme (FP7/2007-2013) under grant agreement 288024 (LiMoSINe project), and the Torres Quevedo program. We also thank Andreas Kaltenbrunner for contributing data.

8. REFERENCES

[1] B. T. Adler, K. Chatterjee, L. De Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *WikiSym*, 2008.

[2] B. T. Adler, L. de Alfaro, I. Pye, and V. Raman. Measuring author contributions to the Wikipedia. In *WikiSym*, 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, 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. In *The World Wide Web and Databases*, volume 1590 of *Lecture Notes in Computer Science*. 1999.

[6] U. Cress and J. Kimmerle. A systemic and cognitive view on collaborative knowledge building with wikis. *Intern. J. Computer-Supported Collaborative Learning*, 3(2), 2008.

[7] J. Giles. Internet encyclopedias go head to head. *Nature*, 438, 2005.

[8] Y. A. Hamburger, N. Lamdan, R. Madiel, and T. Hayat. Personality characteristics of Wikipedia members. *CyberPsychology & Behavior*, 11(6), 2008.

[9] T. Holloway, M. Božičević, and K. Börner. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity*, 12, 2007.

[10] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *WWW*, 2010.

[11] K. K. Lee and G. G. Karuga. The role of cognitive conflict in open-content collaboration. In *AMCIS*, 2010.

[12] O. Nov. What motivates Wikipedians? *CACM*, 50, 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. González-Barahona, and G. Robles. A quantitative approach to the use of the Wikipedia. In *ISCC*, 2009.

[15] I. Weber and A. Jaimes. Who uses Web search for what? And how? In *WSDM*, 2011.

[16] H. Welser, D. Cosley, G. Kossinets, A. Lin, F. Dokshin, G. Gay, and M. Smith. Finding social roles in Wikipedia. In *iConference*, 2011.

[17] R. West, I. Weber, and C. Castillo. A data-driven sketch of Wikipedia editors. In *WWW*, 2012.

[18] R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on Web search behavior. In *WSDM*, 2009.

[19] Wikimedia Foundation. Editor survey. Online, April 2011. `http://meta.wikimedia.org/w/index.php?title=Editor_Survey_2011&oldid=2872379`.

[20] Wikimedia Foundation. Wikimedia strategic plan: A collaborative vision for the movement through 2015. Online, February 2011. `http://strategy.wikimedia.org/wiki/Wikimedia_Movement_Strategic_Plan_Summary`.

[19] `http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Video_games`

APPENDIX

A. ARTICLE–QUERY SIMILARITY

First note that 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 ten results. Each such result comes with a classification into 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 $\mathrm{sim}'(a,q)$. Using this measure as described, a query could in general have a similarity of less than 1 with itself unless the categories of its 10 results are all the same. We account for this by considering the ratios $\mathrm{sim}'(a,q)/\mathrm{sim}'(a,a)$ and $\mathrm{sim}'(a,q)/\mathrm{sim}'(q,q)$; the final similarity $\mathrm{sim}(a,q)$ is then defined by the harmonic mean of these two ratios (harmonic instead of arithmetic because the numerators rather than the denominators are the same).