UvA-DARE (Digital Academic Repository)

On the Value of Wikipedia as a Gateway to the Web

Piccardi, T.; Redi, M.; Colavizza, G.; West, R.

Citation for published version (APA):

On the Value of Wikipedia as a Gateway to the Web

Tiziano Piccardi
EPFL
Switzerland
tiziano.piccardi@epfl.ch

Miriam Redi
Wikimedia Foundation
United Kingdom
miriam@wikimedia.org

Giovanni Colavizza
University of Amsterdam
Netherlands
g.colavizza@uva.nl

Robert West*
EPFL
Switzerland
robert.west@epfl.ch

ABSTRACT

By linking to external websites, Wikipedia can act as a gateway to the Web. To date, however, little is known about the amount of traffic generated by Wikipedia's external links. We fill this gap in a detailed analysis of usage logs gathered from Wikipedia users' client devices. Our analysis proceeds in three steps: First, we quantify the level of engagement with external links, finding that, in one month, English Wikipedia generated 43M clicks to external websites, in roughly even parts via links in infoboxes, cited references, and article bodies. Official links listed in infoboxes have by far the highest click-through rate (CTR), 2.47% on average. In particular, official links associated with articles about businesses, educational institutions, and websites have the highest CTR, whereas official links associated with articles about geographical content, television, and music have the lowest CTR. Second, we investigate patterns of engagement with external links, finding that Wikipedia frequently serves as a stepping stone between search engines and third-party websites, effectively fulfilling information needs that search engines do not meet. Third, we quantify the hypothetical economic value of the clicks received by external websites from English Wikipedia, by estimating that the respective website owners would need to pay a total of $7–13 million per month to obtain the same volume of traffic via sponsored search. Overall, these findings shed light on Wikipedia's role not only as an important source of information, but also as a high-traffic gateway to the broader Web ecosystem.

1 INTRODUCTION

Thanks to the collaborative effort of a community of volunteer editors, Wikipedia is the world's largest encyclopedia and an important source of information for millions of people. Wikipedia serves its content as a regular website, allowing editors to add hyperlinks in

order to enable readers to more easily find additional content, both internal and external. Internal links help readers locate relevant encyclopedic content by navigating from article to article. In contrast, external links enrich articles with additional content

---

*Robert West is a Wikimedia Foundation Research Fellow.

that should not or cannot be included in Wikipedia itself. There are various reasons to add external links,[1] with linked content ranging from official websites, to news articles used as references,[2] to copyrighted material.

In this paper, we are interested in quantifying and characterizing the outgoing traffic generated by Wikipedia towards external content. Given Wikipedia's crucial societal role and global reach, it is essential to understand how it interacts with the broader Web by driving traffic to external websites. The resulting insights can inform the platform's future design and thus allow it to better cater to readers' information needs around external content. As Web traffic has monetary value—in particular when the traffic goes to commercial websites—

Figure 1: Example of an official link, in infobox of Wikipedia article about Internet Archive's Wayback Machine.

an investigation of the external traffic generated by Wikipedia also sheds new light on the poorly understood role it has as a provider not only of information, but also of economic wealth.

Research questions. We approach the question of Wikipedia's value as a gateway to the Web from two angles: informational and economic. Concretely, we pose three research questions:

RQ1 Level of engagement with external links: What total volume of traffic does Wikipedia drive to third-party websites? What is the click-through rate of external links, and how does it vary across types of linked content? (Sec. 4)

RQ2 Patterns of engagement with external links: How do users interact with external links? Do they click through fast or slow, and how does this vary across types of linked content? In what navigational situations do clicks to external websites occur? (Sec. 5)

RQ3 Economic value of external links: What is the monetary value of the traffic from Wikipedia to external websites? If website owners had to pay for an equivalent amount of traffic via sponsored search, how much would this cost? (Sec. 6)

---

[1]https://en.wikipedia.org/wiki/Wikipedia:External_links
[2]https://en.wikipedia.org/wiki/Wikipedia:Citing_sources

Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West

Summary of findings. Based on usage logs gathered over a one-month period from English Wikipedia users' client devices (Sec. 3), we quantified the level of engagement with external links **(RQ1)**, determining that English Wikipedia generated 43 million clicks to external websites during the month we studied, despite the fact that, on average, the click-through rate (CTR) of external links was only 0.08%. While most external links (95.5%) occurred in article bodies and cited references (accounting for about two thirds of the external traffic), a disproportionately large fraction (23%) of the total traffic came from a relatively small fraction (0.8%) of all external links, namely from *official links* to the website of the entity covered in the respective article. Such official links are regularly listed in so-called *infoboxes,* short tabular summaries of key facts about the covered entity (see Fig. 1 for an example). Since official links witnessed a vastly increased CTR of 2.47% (vs. 0.08% over all external links), we focused our analysis on official links. In a topical analysis, we found that official links associated with articles about businesses, educational institutions, and websites had the highest CTR (a first indicator of the economic value of Wikipedia's external links), whereas official links associated with articles about geographical content, television, and music had the lowest CTR.

By analyzing patterns of engagement with external links **(RQ2),** we observed that Wikipedia frequently serves as a stepping stone between search engines and third-party websites. We captured this effect quantitatively as well as in a manual analysis, where we found that URLs that are down-ranked or censored by search engines, and thus not retrievable via search, can often be found in Wikipedia infoboxes, which leads search users to take a detour via Wikipedia. We conclude that Wikipedia regularly and systematically meets information needs that search engines do not meet, which further confirms Wikipedia's central role in the Web ecosystem.

Finally, we aimed to quantify the hypothetical economic value of the clicks received by external websites from English Wikipedia **(RQ3)**. Wikipedia is, of course, free, and it runs thanks to the donations of thousands of people. We thus cannot ask how much money Wikipedia could earn by charging a fee for external clicks—this hypothetical scenario is simply too far from reality—but we may approach the question from a different angle, asking how much money external-website owners would have to pay in order to obtain an equivalent number of clicks by other means, such as paid ads. In this spirit, we applied the Google Ads API to the content of official websites linked from Wikipedia in order to generate keywords for sponsored search and estimated their cost per click at market price. We conclude that the owners of external websites linked from English Wikipedia's infoboxes would need to collectively pay a total of around $7–13 million per month (or $84–156 million per year) for sponsored search in order to obtain the same volume of traffic that they receive from Wikipedia for free.

These numbers exceed even the ballpark guess given in a bullish 2013 analysis [15] that, unlike ours, was not based on real click logs, but on generic rates commonly assumed in the online ad industry, and estimated that Wikipedia could earn $2.5 million monthly via affiliate links. Although our analysis of monetary value should mostly be taken as an indicative "back-of-the-envelope" calculation, it highlights the importance of Wikipedia not only as a source of information, but also as a gratuitous provider of economic wealth.

2 RELATED WORK

User engagement on the Web. Being able to quantify user engagement is crucial for websites, especially for those with an advertising-based business model [2]. Researchers from various fields have investigated ways to measure users' attention, interest, and engagement with websites and their ads. Several works have tried to predict engagement with content in social media based on social interest metrics, such as the number of post comments or likes [3, 6, 13]. Researchers in information retrieval have also investigated methods to estimate users' satisfaction and engagement with textual and visual Web search engines [14, 29, 36]. In computational advertising, existing works have tried to improve ad serving based on target engagement metrics [4, 35], or to directly predict ad click-through rates [17].

User engagement on Wikipedia. Users' reasons for engaging with Wikipedia are varied, and consequently bring about different usage behaviors [16, 28], ranging from in-depth information seeking to using Wikipedia as a stopover towards other locations on the Web [30]. The most relevant prior work is a study by Piccardi et al. [24], who used the same dataset as the present paper, but focused exclusively on external links in cited references, whereas the present study considers all external links, with a focus on official links listed in infoboxes. Piccardi et al. [24] found that users engage little with the external links included in references, and that they do so more frequently from short, possibly unsatisfactory Wikipedia articles and in order to visit recent content, open access sources, and references about life events (births, deaths, marriages, etc.). A similar negative relationship between Wikipedia article quality and user engagement with external sources via references was found specifically for medical content [18].

Wikipedia's own traffic is influenced by its connections to the larger Web ecosystem, and its interdependence with search engines, and Google in particular, has real consequences. On the one hand, Wikipedia's content improves Google search results (e.g., via content snippets); on the other hand, this might keep users from visiting Wikipedia itself [20]. In the present paper, we take a novel angle by focusing not on traffic from the rest of the Web to Wikipedia, but on traffic from Wikipedia to the rest of the Web.

The economic value of Wikipedia. The value of Wikipedia to the world is not only high, but also difficult, if not impossible, to fully quantify in purely economic terms. It has been shown that Wikipedia is essential—or has the potential to be—in a variety of spillovers with substantial economic impact. For instance, it is of critical importance to Web search engines, such as Google [20], and has also been shown to be useful to improve, or even predict, financial markets [22, 23, 34]. Wikipedia can be used to inform economic development policies [27], improve the visibility of places, with direct positive consequences on tourism [12], and even predict and monitor global health and diseases [9, 11]. Furthermore, Wikipedia has been shown to influence the very development of science [31]. Nevertheless, and perhaps surprisingly, to the best of our knowledge Wikipedia's economic value as an information gateway to the Web has rarely been discussed in previous work. In a rare exception, researchers have considered the value of Wikipedia in providing traffic to Reddit and Stack Overflow [32].

Table 1: Click statistics for external links embedded in Wikipedia articles.

Link location	Total links		Clicks		Total articles	Mean CTR ± SD*	Median click time†
	Number	Perc. of total	Number	Perc. of total			
Infobox	2.8M	4.5%	12.5M	29.1%	1.3M	0.90% ± 2.2%	18.6s (45.6s)
Official links	*506K*	*0.8%*	*9.8M*	*22.7%*	*506K*	*2.47% ± 3.0%*	*20.7s (47.8s)*
Body	24.9M	39.5%	16.2M	37.8%	4.0M	0.14% ± 0.7%	35.4s (90.9s)
References	35.3M	56.0%	14.2M	33.1%	3.9M	0.03% ± 0.2%	51.8s (131.4s)
All links	63.1M	100%	43.1M	100%		0.08% ± 0.5%	32.9s (87.1s)

*CTR = click-through rate; considering only links with at least 300 impressions during the one-month study period. †Inter-quartile range in parentheses.

3 DATA AND DEFINITIONS

3.1 Wikipedia client-side logs

In order to analyze user engagement with Wikipedia's external links, we made use of the dataset collected by Piccardi et al. [24]. This dataset consists of logs of all reader interactions with external links in English Wikipedia articles over the one-month period from 24 March to 21 April 2019. The data was captured by the browser on the client side and includes all clicks on external links and a uniformly random sample (33%) of all pageview events, organized into *sessions,* i.e., sequences of events from the same user in the same browser tab. This article reports results using the full dataset when describing external-click events. Whenever pageview counts are involved, we extrapolate from the 33% sample.

The data was collected in accordance with Wikimedia's privacy policy[3] and processed exclusively on Wikimedia computing machinery. Although the data does not contain personally indentifiable information beyond what is implicit in browsing behavior, it cannot be shared publicly. For transparency, we publish our data analysis code at https://github.com/epfl-dlab/WikipediaAsWebGateway.

3.2 Article characteristics

At the time of data collection, Wikipedia had around 5.8M articles, which were loaded by readers more than 4.5 billion times during the month studied. We characterized each article by popularity (pageviews during the month studied) and length (number of characters). The popularity distribution was very skewed: 50% of the articles had fewer than 42 views, 90% had fewer than 894 views. In contrast, the average number of pageviews in one month was 700. The most visited 1,550 articles, which represented 0.02% of all articles, accounted for 10% of all pageviews. The most visited pages were articles about topics that were trending in April 2019, such as NIPSEY HUSSLE (5.7M views), NOTRE-DAME DE PARIS (4.7M), BONNIE AND CLYDE (3.5M), or GAME OF THRONES (SEASON 8) (2.6M). Most of the articles were short, and similar to popularity, the number of characters showed a skewed distribution, with a median of 3,888, and an average of 7,793 characters.

ORES topics. ORES[4] is a toolkit offered by Wikimedia that, among other things, includes functionality for labeling articles with topics, based on a manually curated taxonomy of 64 topics [10] derived from WikiProjects.[5] Based on this categorization, ORES offers a

classifier that predicts, for a given article, its probability of belonging to each of the 64 topics. Since a single article may belong to multiple topics, the 64 probabilities generally do not sum to 1. We used topic labels in binarized form, considering an article to belong to a topic if the corresponding probability is greater than 50%. Note that, although the taxonomy is hierarchically organized in two levels, in this work we only considered the 57 lower-level topics (listed along the x-axis of Fig. 3). Having run the classifier on all articles in the dataset, we observed that overall the most common topic was BIOGRAPHY (1.7M articles), followed by SPORTS (1.4M) and NORTH AMERICA (950K). The least common topics were EASTERN AFRICA (11K) and LIBRARIES & INFORMATION (14K).

3.3 External links

External links form our central object of study, so we extracted detailed information about them from Wikipedia articles. Since parsing content from articles in wikitext format might result in missing hyperlinks [21], we extracted the external links from the articles in HTML format instead. As this paper focuses specifically on those Wikipedia links that lead to external websites, we adopt the convention that, whenever we simply say "link", we implicitly mean "external link".

We partitioned the external links in the dataset into three classes, according to their position on the page: *infobox,* article *body,* and *references. Infoboxes* are tables (typically rendered by the browser on the right-hand side of the page on desktop devices, or at the top of the page on mobile devices) that summarize key information by adding semi-structured content (see Fig. 1). In addition to images and textual properties, this area can contain—potentially many— external links pointing to external geolocation services, official registries, or official websites. The links in an article *body* appear inline within the main textual content of the page or in dedicated sections such as "External links" or "See also". Links in article bodies are more heterogeneous, including links to social media pages, PDF documents, or related external material. Finally, we considered as *references* all links used to cite external content in support of a statement. Typically they appear at the bottom of the page, reachable from the article body via numbered link anchors.

During the period considered, Wikipedia had 5.3M articles that contained at least one of 63.1M external links (totaling 49.8M unique target URLs). Table 1 (column "Total links") summarizes these values. In total, 35.3M (56.0%) of these links appeared in references, 24.9M (39.5%) in article bodies, and 2.8M (4.5%) in infoboxes. Around 1.3M articles in English Wikipedia had an infobox with links, and the average number of links per infobox in these articles was 2.08.

Links spanned from official company links (e.g., schlenkerla.de) to geocoordinates on geohack.toolforge.org to institutional registries (e.g., National Register of Historic Places).

3.4 Official links

To further qualify the infobox links, we designed a binary classifier that can distinguish between *official links* and other types of link. It was trained on a random sample of 2,000 infobox links, manually annotated as "official" or "other". This resulted in a training set of 387 official links and 1,613 other links. To characterize each link, we then computed the following features:

- **URL length:** Number of characters in the URL path (guided by the intuition that official-link URLs tend to be short).
- **Similarity of URL with article title:** Motivated by the usefulness of character n-grams for URL-based topic classification [5], we computed the character n-grams ($n = 1, \ldots, 4$) of the title of the article where the link was placed, the link's anchor text (if non-empty), the domain of the link URL, and the path of the link URL. We then computed, and used as features, the Jaccard similarity of sets of n-grams for three pairs: title/URL-domain, anchor-text/URL-domain, title/URL-path.
- **Similarity of context with marker words:** Jaccard similarities between the character n-gram sets ($n = 1, \ldots, 4$) of high-precision marker words ("official", "website", "homepage", "URL") and of the link's anchor text and context (i.e., the text within the same <TR> tag as the link).

We used these 10 features and the manual labels in a random forest classifier, achieving 5-fold cross-validated precision 0.980 (SD 0.009), recall 0.983 (SD 0.005), and F1 score 0.982 (SD 0.007).

Applying this classifier to all the links in the dataset, we found that 506K of the 63.1M external links corresponded to the official website of the entity described in the respective article. Broken down by article topic, the largest number of official links was associated with NORTH AMERICA (27.2%), EUROPE (25.0%), MEDIA (20.2%), BIOGRAPHY (18.9%), ASIA (17.1%), EDUCATION (10.5%), and BUSINESS AND ECONOMICS (9.5%), with the latter containing mainly articles about companies. (The percentages sum to more than 100% because articles may belong to multiple topics.)

As expected, according to the classifier, the vast majority (98.1%) of articles with an official link had exactly one official link, and conversely, the vast majority (97.1%) of official links appeared as such in exactly one article's infobox. This has the added advantage that official links can be characterized by features derived from their corresponding Wikipedia articles (e.g., topics, content words).

3.5 Definitions

Definition: click-through rate (CTR). Our main metric for measuring engagement with external links is the *click-through rate (CTR)*, which, intuitively, is simply the number of times a link was clicked, divided by the number of times the link was displayed (by virtue of being contained in an article that contained the link). In practice, care must be taken, as it frequently happens that the same article is viewed multiple times in the same session, e.g., because the user refreshes the page or clicks the back button. To guard against overcounting such multiple views, we grouped all pageviews of the
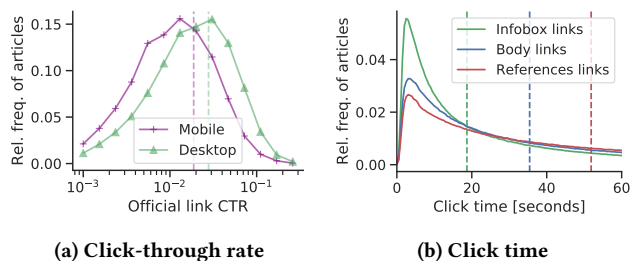

(a) Click-through rate (b) Click time

Figure 2: Usage of external links. (a) Distribution of click-through rate of official links by device type (vertical lines: means). (b) Distribution of click time by link type (vertical lines: medians).

same article a that occurred during the same session s and call the unique pair (a, s) one visit of a.

With N_l as the number of visits (i.e., distinct (a, s) pairs) upon which link l was displayed, and C_l as the number of visits upon which link l was clicked, we define the CTR of link l as C_l/N_l, i.e., the fraction of visits upon which l was clicked, out of all visits upon which l was displayed. Since each official link belongs to exactly one article (with extremely rare exceptions; cf. Sec. 3.4), we may also, in a slight abuse of terminology, speak of the "CTR of an article", implying the CTR of the official link associated with the article.

In order to reliably estimate CTRs, we need to avoid small denominators, so we restricted our analyses to links that were displayed upon at least 300 pageview events. In the case of official links, this resulted in a set of 160K links (and their corresponding articles).

Definition: click time. In order to capture how long users dwell on an article before they click an external link, we define the notion of *click time,* which measures the number of seconds between the pageview event on which the link was displayed and the click on the link itself. If the same external link was clicked multiple times in the same session, we only considered the first pageview that was accompanied by an external click.

Since click times are unbounded above and follow a heavy-tailed distribution, we used medians, rather than means, for aggregation.

4 RQ1: LEVEL OF ENGAGEMENT WITH EXTERNAL LINKS

We start our analysis by quantifying the level of engagement with external links, both overall (Sec. 4.1) and by article topic (Sec. 4.2).

4.1 Overall click statistics

Overall click statistics are summarized in Table 1. During our one-month data collection period, there was a total of around 4.5 billion Wikipedia pageviews, which led to around 43.1M clicks on external links. The total volume of external clicks was distributed roughly evenly over the three classes of external links: those in infoboxes (12.5M), those in references (14.2M), and those in article bodies (16.2M). As the vast majority of external links was located in references (56.0%) and article bodies (39.5%), the CTR of infobox links (0.90%) vastly exceeded that of links in references (0.03%) and article bodies (0.14%). To ascertain that this was not simply caused by the fact that infobox links appear higher up on the page, we also
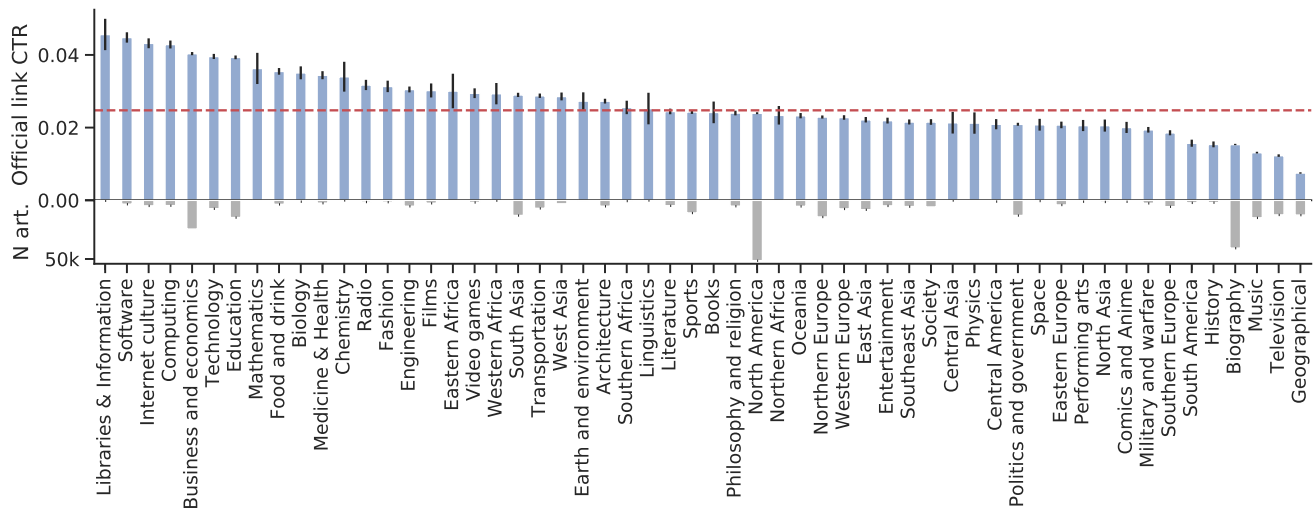
Figure 3: Official-link click-through rate by article topic. *Blue bars:* means with bootstrapped 95% confidence intervals. *Gray bars:* number of articles with official links. *Red dashed line:* global mean.

computed the CTR of article-body links appearing in the top 20% of the page. This yielded a CTR of 0.20%, much closer to the 0.14% of article-body links overall than to the 0.90% of infobox links.

Official links. Official links play a key role. Although they constituted only 18% of the 2.8M infobox links, they accounted for 78% of the 12.5M clicks on infobox links, with a CTR of 2.47%, nearly 3 times as high as that of infobox links overall. The average CTR was even higher on desktop devices, where it reached 2.78%, vs. 1.87% on mobile (Fig. 2a). Given their prominence, we shall focus mostly on official links from here on, and unless stated otherwise, we henceforth refer to official links when simply writing "links".

Geographical differences. The top 5 countries by pageview volume were, in this order, the United States, the United Kingdom, India, Canada, and Australia. They generated 71.6% of the total traffic. Among these countries, the CTR on official links was highest in the U.S. (2.36%), followed by India (2.14%), the U.K. (1.53%), Canada (1.38%), and Australia (1.11%).

4.2 Click-through rates by topic

Next, we aim to understand how the click-through rates of official links vary by topic as defined by the ORES classifier introduced in Sec. 3.2. Since, in the vast majority of cases, official links are associated with exactly one article, we may label each official link with the topics of that article. (Throughout the discussion that follows, keep in mind that each article, and thus each official link, may be labeled with multiple topics.)

Fig. 3 visualizes the mean CTR (in blue), as well as the number of articles with official links (in gray), by topic. We see that official links relating to LIBRARIES & INFORMATION, SOFTWARE, and INTERNET CULTURE had the highest click-through rates, whereas geographical content, media-related content, and biographies on average saw the lowest engagement.
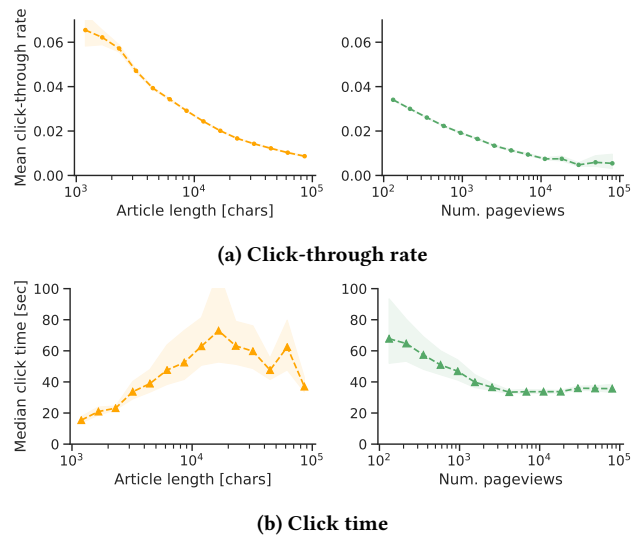

(a) Click-through rate

(b) Click time

Figure 4: (a) Click-through rate and (b) click time of official links as functions of article length (left) and popularity (right), with 95% CIs. Official links on longer pages are clicked more rarely and more slowly; those on more popular pages are clicked more rarely and more quickly.

Controlling for article length and popularity. The length and popularity of a Wikipedia article correlated strongly and negatively with the CTR of the official link contained in its infobox (Fig. 4a), possibly because longer articles, by offering more information, reduce the user's need to gather additional information from external links, and because more popular articles are more likely to appear in shallower information-seeking sessions [24]. Since length and popularity also vary by topic, they might act as confounds that could potentially explain an observed variation of CTR by topic.

To tease these two confounds apart from the impact of topics alone, we controlled for length and popularity in a matched analysis, as follows. We split the set of articles at the median CTR into high- vs. low-CTR articles, and we split the length and popularity ranges into 1,000 equally sized bins each. We then defined a bipartite graph with edges between articles that fell in different CTR halves, but in the same bins with respect to length and popularity. Using the Euclidean distance in the space defined by the logarithmic length and popularity as edge weights, we found a minimum matching and retained only the 112K matched articles (out of originally 160K). This procedure successfully balanced the dataset.[6]

In the balanced dataset, we binarized the CTR by splitting at the median and fit a logistic regression to model whether an article belonged to the high- or low-CTR group, with topic indicators as predictors (pseudo $R^2 = 0.20$, $p < 10^{-307}$). (The advantage of performing regression modeling rather than a simple comparison of per-topic average CTRs is that topics are correlated, which is accounted for by the regression model.)
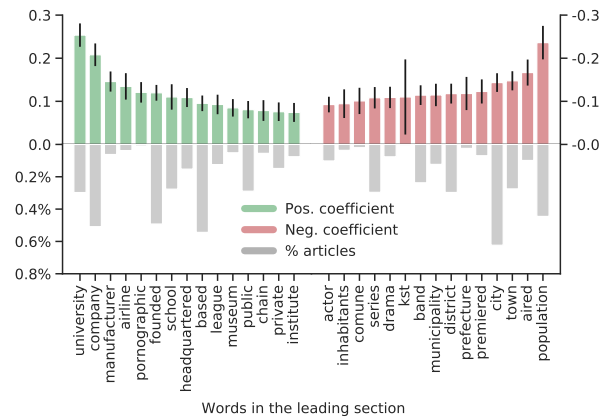
The 15 largest positive and negative coefficients, plotted in Fig. 5a, revealed a slightly different ranking than Fig. 3, with BUSINESS AND ECONOMICS and EDUCATION emerging as the strongest predictors of a high CTR, whereas GEOGRAPHICAL and TELEVISION remained the strongest predictors of a low CTR.

Top of the CTR ranking. While manually screening the data, we realized that, among the articles with the highest official-link CTR, there was a disproportionate fraction of articles about websites (which are generally classified by ORES under the topic INTERNET CULTURE), and in particular websites related to file sharing and pornography, some with CTRs of 40% or more, e.g., Library Genesis (47%), RARBG (45%), or The Pirate Bay (43%). To determine whether official links of Wikipedia articles about websites dominated the top of the CTR ranking in general, we repeated the above regression analysis with a small modification: instead of predicting the top half vs. the bottom half of the article ranking with respect to CTR, we now predicted the top L articles (an absolute, rather than relative number) vs. the same number of samples from the bottom half, matched on length and popularity. This way, plotting the fitted coefficients for a given topic as a function of L reveals whether the topic is particularly over-represented among the highest-CTR official links (manifested in a sharply decreasing curve). The results, presented in Fig. 6, clearly show that INTERNET CULTURE—a topic held by most articles about websites—is indeed particularly over-represented among the articles with the very highest official-link CTR. Similar effects were observed for SOCIETY (a loose mix of articles), SPORTS, SOFTWARE, and ENTERTAINMENT, among others. On the contrary, we observed that GEOGRAPHICAL, BIOGRAPHY, and TELEVISION, among others, were particularly under-represented among the highest-CTR official links.

Fine-grained topical analysis. The topics from the ORES classifier used above are rather broad. In order to obtain more fine-grained insights, we conducted a word- rather than topic-level analysis, where we represented an official link by the words contained in the lead paragraph of the article in whose infobox the link

---
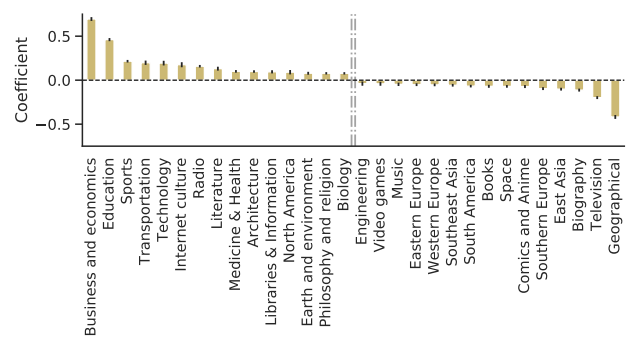
(a) Topic-based analysis

(b) Word-based analysis

Figure 5: Association of click-through rate of official links with article properties, captured via 15 largest positive and negative coefficients (with 95% CIs) from binary logistic regression models that predict above- vs. below-median CTR, using as predictors (a) article topics or (b) words from lead paragraphs (controlling for article length and popularity). Gray bars in (b): percentage of articles whose lead paragraph contains the word.

appeared as an official link (via z-score-standardized TF-IDF vectors restricted to the 3,000 most frequent words across all articles with official links). Mirroring the above regression analysis, but now using words rather than topics as predictors (pseudo $R^2 = 0.31$, $p < 10^{-307}$), this analysis revealed words associated with high- vs. low-CTR links. Fig. 5b, which shows the 15 words with the largest positive and negative coefficients, confirms our previous findings while adding nuance. We see that EDUCATION specifically marks high-CTR links about universities, schools, institutes, and museums; BUSINESS AND ECONOMICS, about companies, manufacturers, chains, and airlines; and INTERNET CULTURE, about adult websites.

Summary. Taking stock of the findings so far, we reiterate that official links play a key role among Wikipedia's external links, with CTRs far above those of other types of external link. We observed a large amount of variation depending on article topics, with official links associated with articles about websites, software, businesses, education, and sports seeing particularly high engagement.
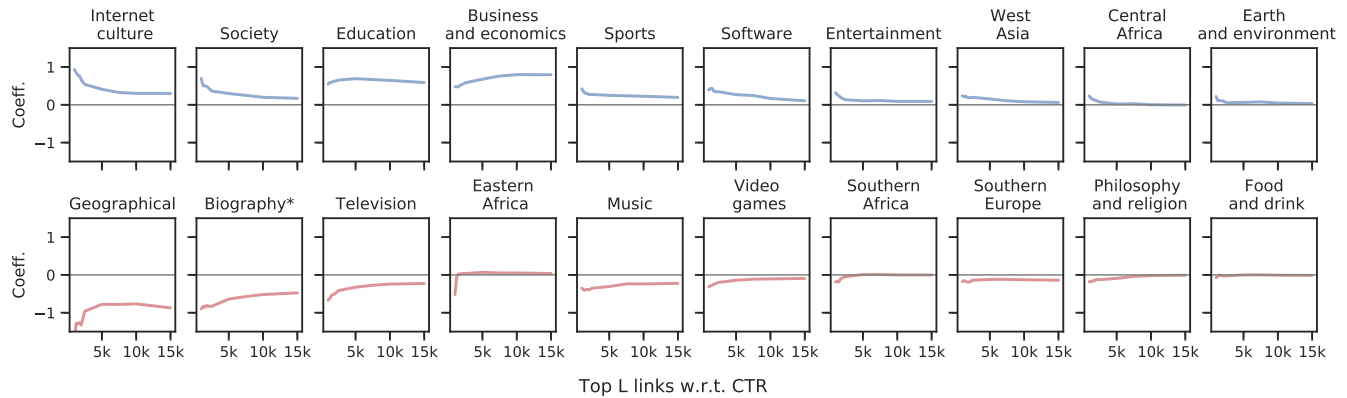
Figure 6: Prevalence of topics among most frequently clicked official links. We fitted binary logistic regression models that used article topics as predictors to predict if an article's official link is among the top L highest-CTR links. Plots show regression coefficients for individual topics (predictors) as functions of L (for values of L between 1K and 15K). Topics are sorted by the leftmost values of their curves. Sharply decreasing [increasing] curves correspond to topics that are particularly over-represented [under-represented] among the links with the most extreme CTR. (More details: Sec. 4.2, "Top of the CTR ranking".)

5 RQ2: PATTERNS OF ENGAGEMENT WITH EXTERNAL LINKS

Above, we established which types of external link have a particularly high CTR. Next, we investigate more closely the patterns by which users engage with external links.

5.1 Click time

We start by analyzing the click time (cf. Sec. 3.5), which captures how long users dwell on an article before leaving it toward an external website via an external link.

Click time statistics are summarized in Table 1, and click time distributions are plotted in Fig. 2b, for the three types of link: those appearing in infoboxes, article bodies, and references, respectively. (Note that, although we consider only official links in most of this paper—and indeed in the rest of this section—we nevertheless report this basic statistic for all types of external link beyond official links only.) The global median click time was 32.9 seconds (31.8 seconds for desktop, 34.4 seconds for mobile), with a much lower value for infobox links (18.7 seconds; 20.1 seconds for official links), and larger values for the article-body links (35.4 seconds) and reference links (51.8 seconds). The short click time of infobox links, however, seems to be due to their prominent position within articles: when approximately controlling for position by considering only article-body links in the top 20% of the page, the median click time dropped to 22.2 seconds, only 10% longer than for infobox links.

After this general characterization, we from here on focus on official links in infoboxes. Similar to CTR (Fig. 4a), the click time of official links was correlated with article length and popularity (Fig. 4b), such that clicks took more time on longer and on less popular articles. When analyzing click time by topic, we thus again controlled for these two factors via matching, as they could otherwise confound the analysis. Analogously to the setup of Sec. 4.2, we split the set of articles at the median click time into articles with "slow" vs. "fast" clicks on official links, and subsequently found a bipartite matching in order to ensure that the distribution of length

and popularity was nearly identical in articles with "slow" vs. "fast" official-link clicks.[7] We then fitted a linear regression on the resulting balanced dataset in order to model the logarithmic click time as the outcome variable, using (as in Sec. 4.2) topic indicators as predictors ($R^2 = 0.090$, $p < 10^{-307}$).

The 15 largest positive and negative coefficients are plotted in Fig. 7a, revealing that clicks on official links to entertainment-related websites occurred faster, whereas links to websites on more classic encyclopedic topics, such as biographies, geographical content, history, etc., occurred more slowly.

To gain more granular insights than at the coarse topic level, we again ran an analysis at the word level, parallel to the word-level analysis of Sec. 4.2, but this time in the linear regression setup just described ($R^2 = 0.17$, $p < 10^{-307}$). The words most indicative of slow and fast clicks (Fig. 7b) mirror the findings from the topic-level analysis (Fig. 7a), but we now also see that official links on articles about websites and universities were clicked particularly fast.

5.2 Wikipedia as a stepping stone

We just noted that official links on articles about websites typically have short click times. Moreover, in Sec. 4.2 we had observed that such links are also strongly over-represented among the highest-CTR links. This observations led us to hypothesize that the main interest of users interacting with these links might be to find these links to begin with, rather than to find the content that surrounds the links in their respective Wikipedia articles. In other words, we hypothesized that Wikipedia serves as a "stepping stone" toward external websites, whereupon users barely set foot onto Wikipedia before leaving again towards different content that they actually were intent on finding. To further investigate this hypothesis, we considered the relationship between CTR and click time, visualized in Fig. 8, which presents a scatter plot of the coefficients obtained from the previously described regressions for CTR (x-axis;

---

[7] The standardized mean differences in logarithmic article length and logarithmic popularity dropped from 0.076 to 0.0001, and from 0.30 to 0.00005, respectively. 138K of the original 160K data points were retained after matching.

(a) Topic-based analysis
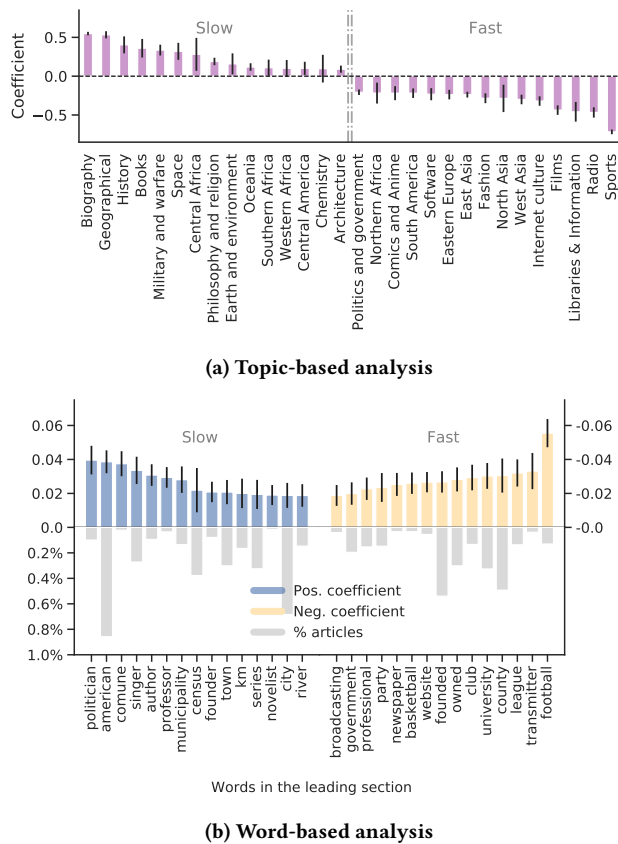

(b) Word-based analysis

Figure 7: Association of click time of official links with article properties, captured via 15 largest positive and negative coefficients (with 95% CIs) from linear regression models that predict logarithmic click time, using as predictors (a) article topics or (b) words from lead paragraphs (controlling for article length and popularity). Gray bars in (b): percentage of articles whose lead paragraph contains the word.

cf. Sec. 4.2) and click time (y-axis; cf. Sec. 5.1). As the plot shows, topics with a high CTR tended to have low click times (lower right quadrant), notably SPORTS, RADIO, and INTERNET CULTURE (a topic that, as mentioned, primarily tags articles about websites). We take this as another indicator of the existence of a class of articles from which users leave Wikipedia frequently and fast.

If indeed a distinct class of articles on Wikipedia are used heavily as stepping stones, then we should be able to identify many articles that were visited primarily from outside of Wikipedia (e.g., from search engines), rather than from other Wikipedia articles (via internal navigation), and that had a high CTR on the external links they contain. With the goal of finding such articles, we define the *external-referrer frequency (ERF)* of an article as the fraction of all visits made to the article via a click from a referral page external to Wikipedia. Note that external referrals almost exclusively stemmed from search engines: about 70% of them specified a search engine referrer URL in the logs, and about 29% did not specify any referrer URL, but it is suspected that a majority of these visits also come from
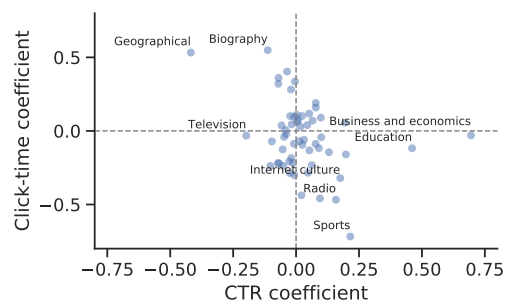

Figure 8: Click-through rate (x-axis) vs. click time (y-axis) of official links. Each point represents one topic. CTR and click time of a topic captured in terms of the topic's coefficient in the regressions summarized in Fig. 5a and 7a, respectively.

search [19]. Hence, we included empty referrers in our analysis, although the conclusions were identical when excluding them.

The ERF histogram, plotted in Fig. 9a, shows that most articles were primarily visited from outside of Wikipedia, but only few articles had a very high ERF close to 1. Although extreme-ERF articles were in total visited less than medium-ERF articles (Fig. 9b), they generated a total number of external clicks comparable to that generated by medium-ERF articles (Fig. 9c). The most important piece of evidence, however, comes from Fig. 9d, which plots, on the y-axis, the mean CTR for situations where the respective article was reached from an external referrer. The plot shows that the articles that were nearly exclusively reached from external referrers (with an ERF close to 1) are precisely those articles that also had the highest CTR after being reached from an external referrer.[8]

Taken together, these facts provide evidence of a class of articles that serve as mere "stepping stones", "revolving doors", or "in-and-outs": users come from elsewhere in order to find a particular link and immediately leave Wikipedia by clicking that link.

But why, then, would users go through Wikipedia in the first place, if all they want is to go to a website linked from Wikipedia? We shall discuss potential reasons for this behavior in Sec. 7.

6 RQ3: ECONOMIC VALUE OF EXTERNAL LINKS

Our final set of analyses aims to estimate the monetary value of the traffic that Wikipedia drives to external websites.

The idea behind our calculations is the following. Search engines generally charge website owners money in exchange for driving clicks to their sites from sponsored search results (e.g., Google Ads), whereas Wikipedia conveys a large volume of traffic to official websites for free. We therefore ask, "How much money would a search engine want from website owners to obtain, via ads, the same number of clicks they obtain from Wikipedia for free?"

While we could be asking, "How much money could Wikipedia earn by charging a fee for external clicks?", we consider this counterfactual scenario too far from reality: Wikipedia is open and free by design, and it functions rather differently from platforms

---

[8] This analysis only considered articles with at least 100 visits from external referrers, in order to avoid noise due to division by small numbers.

(a)   (b)   (c)   (d)
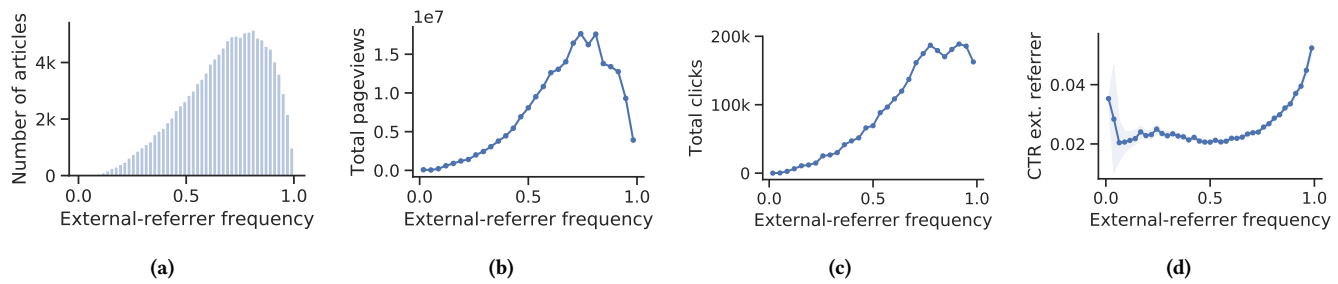
Figure 9: Quantification of Wikipedia's role as a stepping stone toward external websites. (a) Histogram of *external-referrer frequency (ERF)* of Wikipedia articles, where ERF is defined as the fraction of times the article was visited via a referral page external to Wikipedia. (b) Total number of pageviews of articles within each ERF bin. (c) Total number of official-link clicks of articles within each ERF bin. (d) Official-link CTR upon pageviews with an external referrer (most likely a search engine), with 95% CIs. Articles with an extreme ERF close to 1 are rare (a), but generate a disproportionately large number of official-link clicks (c vs. b), especially when reached from search engines (d).

driven mostly by advertising. Moreover, we could not estimate the "price paid" to Wikipedia in this scenario, as we do not know whether website owners would be willing or able to actually pay for any hypothetical click fees. On the contrary, estimating the "price asked"—the cost of online ads—is entirely feasible.

6.1 Methodology

Google Ads. To estimate the price of achieving a certain set of URL clicks via online ads, we used the Google Ads API, with "sponsored search" as the ads network. Google Ads is one of the most prolific advertising platforms, and the primary source of revenue for Google's parent company, Alphabet [1]. The Google Ads API allows advertisers to create campaigns for promoting a URL by placing bids on campaign-related search keywords. The bid expresses the maximum amount that the website owner is willing to pay each time the promoted URL is clicked when shown on the search result page for the respective keyword. When a user searches for a keyword specified by the campaign, an auction system determines which sponsored URL to show among all the candidates competing for the keyword.[9] When the user clicks a promoted link, the campaign owner pays the auction value. Note that the paid price is not necessarily equal to, but only bounded by, the website owner's bid.

From URLs to keywords. Our intended analysis started from clicks on official links (URLs) observed in the Wikipedia logs and aimed to estimate how much these clicks would cost when obtained via Google Ads instead of Wikipedia. The Google Ads API, on the contrary, requires keywords, not URLs, as input. Thus, in order to leverage Google Ads for our analysis, we had to work our way backwards and determine appropriate keywords that a website owner might use to advertise a given URL. Since the choice of the right keywords is critical to increase a website's discoverability while keeping ad costs down [26], Google Ads offers a tool called *Keyword Planner*, which, given a website URL, generates a set of relevant keywords, alongside information on the historical bidding range and search volume for those keywords. Using the Keyword Planner, we generated 11 keywords for each official link: the title

---

[9] With the "broad search" option, the number of matching searches increases by also considering substrings and permutations of the tokens in campaign keywords.

Table 2: Keywords, alongside estimated average cost per click (CPC), for two example websites.

	Coursera	American Airlines
From title	COURSERA	AMERICAN AIRLINES
Keywords recommended by Google Ads Keyword Planner	ONLINE COURSES	AA
	ONLINE COLLEGES	AIRLINE FLIGHTS
	ONLINE CLASSES	AIRLINE TICKETS
	MOOC	AIRLINES
	ONLINE LEARNING	AMERICAN
	FREE ONLINE COURSES	AMERICAN AIRLINES FLIGHTS
	ONLINE DEGREES	CHEAP AIR TICKETS
	OPEN UNIVERSITY COURSES	CHEAP AIRLINE TICKETS
	ONLINE EDUCATION	FLIGHT TICKETS
	ONLINE UNIVERSITIES	US AIRWAYS
Est. avg. CPC	$0.79	$1.10

of the corresponding Wikipedia article, plus the top 10 keywords returned by the Keyword Planner. As examples, Table 2 summarizes the most relevant keywords generated for two different websites: Coursera (a popular online course platform) and American Airlines.

Cost-per-click (CPC) forecasting. Once the set of keywords and the bids have been set, Google Ads can make a prediction about the cost of the campaign through its forecasting tool. The prediction model uses historical data to simulate the auction system and provides an estimated number of clicks and the average price for every keyword. The tool predicts the campaign's average *cost per click (CPC)* by combining the keyword costs with their expected click-through rates. In practice, the forecasting tool simulates campaigns for specific target countries. We used the top 5 English-speaking countries (U.S., U.K., India, Canada, and Australia; cf. Sec. 4.1) as target countries, since they accounted for a large portion (71.6%) of all external-link clicks in the Wikipedia logs studied here.

Estimating the value of official links. Leveraging the above tools, we estimated how much a website owner would need to pay to Google Ads for a single click to their website as follows:

(1) Obtain keywords for the website via the Keyword Planner.
(2) For each keyword, set the bid to the 80th percentile of the keyword's historical auction price.
(3) Estimate the cost (CPC) for one click on the website link by feeding the keywords and their bids to the forecasting tool (using "broad search", cf. footnote 9).
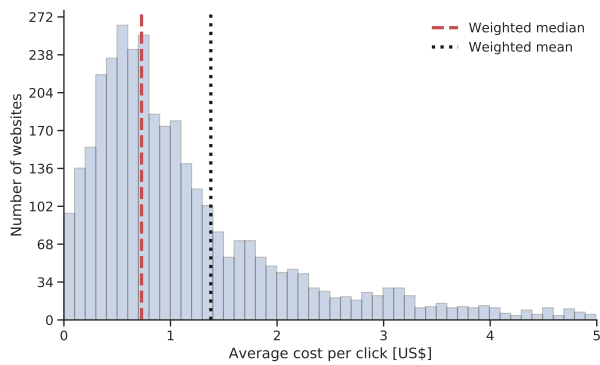
Figure 10: Distribution of official-link CPC, estimated via Google Ads API. Vertical lines represent weighted versions of median and mean, respectively, where each link was weighted according to its click volume in Wikipedia logs.

We emphasize that setting a high bid (step 2) does not automatically entail a high CPC. Indeed, as we shall see, the winning price was usually much lower than the bid. A high bid ensures that we are likely to win the simulated auction and that the promoted link is actually displayed to the user, which is required in order to obtain clicks—the event whose cost we are aiming to estimate.

6.2 Results

Cost per click (CPC). We applied the above-described procedure to a total of 3,600 official links from Wikipedia infoboxes, obtained by randomly sampling an equal number of articles from each of the 57 topics. During the one-month study period, these 3,600 official links were clicked 2.73M times in total.

As mentioned, the bid passed to the forecasting tool is not necessarily the winning price of the simulated auction; it merely caps expenses. In practice, the auction price reached our bid in only 8.9% of cases; on average, the auction price was 58% of the bid.

The CPC distribution is shown in Fig. 10. When weighting all links evenly (i.e., macro-averaging), the mean and median CPC is $1.64 and $0.90, respectively. As not all links are equally popular, a more reasonable estimate of the overall CPC may be obtained by weighting each link according to the number of clicks it received (i.e., micro-averaging). The resulting weighted CPC is slightly lower, with a mean and median of $1.37 and $0.73, respectively (vertical lines in Fig. 10).

Investigating the weighted mean CPC for individual topics, we found considerable variation, with the highest CPCs for MATHEMATICS, MEDICINE & HEALTH, BOOKS, and ARCHITECTURE, and the lowest CPCs for MUSIC, SPORTS, FASHION, and FILMS (omitting from the list topics that mark geographical regions, such as NORTH AMERICA).

Monthly value of traffic to official websites. Multiplying the weighted CPC with the overall number of 9.8M clicks on official links during the one-month study period, we estimate the total monthly value generated by the traffic from Wikipedia to official websites as **$13.4 million** when using the mean CPC estimate, or as **$7.2 million** when using the median CPC estimate.

Broken down by topic, we obtained the (mean-based) estimates of Fig. 11. The topic with the highest total monthly value ($1.9M) is NORTH AMERICA, a macro category assigned to a large set of articles, including U.S. companies and people. It is followed by BUSINESS AND ECONOMICS ($1.3M), BIOGRAPHY ($1.3M), TECHNOLOGY ($1.0M), and SOFTWARE ($0.9M).

7 DISCUSSION

Wikipedia as a gateway to the Web. While the value of Wikipedia's knowledge is fairly well known, less known was the hidden and significant additional value of Wikipedia as a gateway. Building on top of existing work related to Wikipedia readers' behavior analysis, we have uncovered a new perspective on the role of Wikipedia in the broader Web ecosystem. We offered a description of the value of Wikipedia as a gateway under multiple levels of analysis. Overall, we found that a substantial fraction of Wikipedia readers use the encyclopedia as a gateway to the broader Web: readers engage more and faster with official links in the article's infobox than with links in the article body or in the reference section. We found that Wikipedia's role as a gateway to external content is particularly pronounced when users visit articles about websites, software, businesses, education, and sports, among others, where the click-through rate of official links is the highest.

Wikipedia as a stepping stone. We found an inverse relation between the time it takes to click on an external link and its average click-through rate, showing clusters of topics, such as SPORTS and INTERNET CULTURE where engagement with links was high and fast, or conversely, where engagement with link was low and slow, such as BIOGRAPHY and GEOGRAPHICAL. We also observed that articles that were visited particularly frequently from external referrers (mostly search engines) also had a particularly high probability of an official-link click after being reached from external referrers. These results indicate that a certain distinct set of Wikipedia articles is leveraged by users in the spirit of "stepping stones" or "revolving doors", which are reached nearly exclusively from external referrers (mostly search) and from which the user leaves Wikipedia immediately again by clicking an official external website link. This begs the question: If users visit Wikipedia from a search engine result page only to leave it immediately towards a third website, why would they not simply use the search engine to locate the third website to begin with?

We conjecture that the reason is that the search engine cannot fulfill the user's information need in such situations, whereas Wikipedia can. When manually screening the data (focusing on popular articles with at least 30K pageviews during the one-month study period), we found that, among the articles with the highest CTR, there was a disproportionate fraction of file-sharing (5 of the top 6) and pornographic (5 of the top 15) websites. Such search results are frequently censored or down-ranked by search engines, depending on the search engine's corporate policy as well as legislation in the user's country. Indeed, manually searching Google for the names of the 15 articles with the highest CTR (and more than 30K monthly pageviews) from two locations (U.S. and Germany) revealed that 5 file-sharing websites and 1 pornographic website were not listed by Google among the top 10 search results in at least one of the two
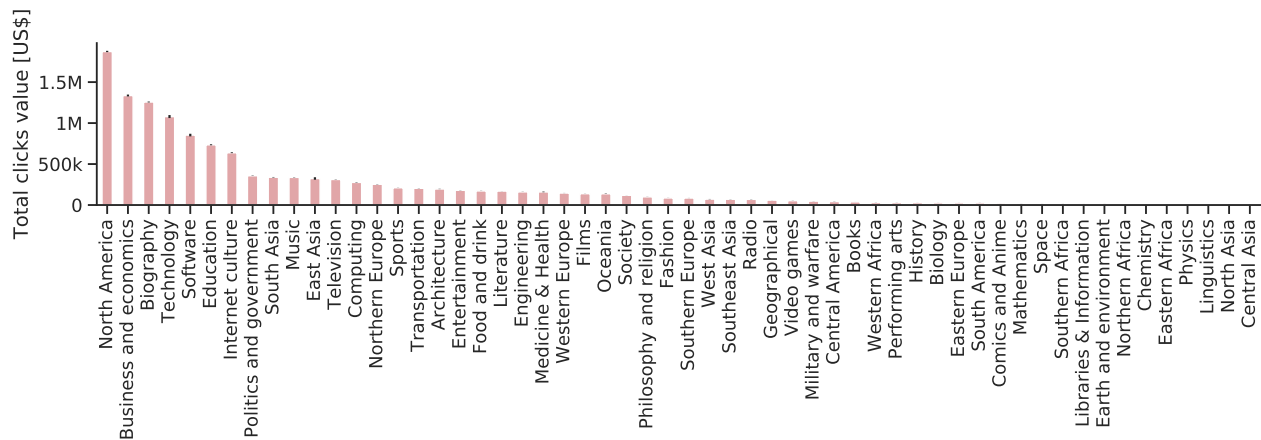
Figure 11: Estimated total monthly value of official links in Wikipedia infoboxes by topic, obtained by multiplying the mean cost per click (CPC) of links from the respective topic with the total number of clicks on those links in the Wikipedia logs.

locations. (Additionally, two controversial websites were not online anymore at the time of research, about 18 months after data collection.) While these findings remain small-scale and anecdotal, they suggest that Wikipedia fills a functional gap, as a workaround for content suppressed by search engines (sometimes for valid reasons, e.g., when the linked material is copyrighted or illegal).

Wikipedia's role as a transitory stepping stone towards external content can have important implications for Web user studies. For example, researchers working on disinformation diffusion might want to take into account the function of Wikipedia as a short yet crucial stopover in Web users' information-seeking journeys.

The economic value of traffic generated by Wikipedia. Finally, we set out to estimate the monetary value of Wikipedia as a gateway to the broader Web. The infoboxes contained in English Wikipedia articles collectively list over half a million official-website links, which were clicked 9.8M times during our one-month study period. These clicks were generated by Wikipedia for free, amidst a Web ecosystem that is majorly powered by paid ads. We asked, "If the respective website owners wanted to achieve the same number of clicks via sponsored search results, what would be the price?" We estimate that achieving the 9.8M monthly clicks on official links would cost a total of $7–13 million using Google Ads. Extrapolating to 12 months, the yearly cost would amount to $84–156 million. This is a remarkably high number, considering that the annual revenue of the Wikimedia Foundation, the non-for-profit organization that operates Wikipedia and its sister projects, stands at around $110 million,[10] coming entirely from donations and voluntary contributions. We also emphasize that the estimated economic value of $84–156 million pertains to English Wikipedia only, whereas Wikimedia's annual revenue of $110 million needs to support all Wikimedia projects across languages.

We showed that, when buying clicks from Google Ads instead of obtaining them from Wikipedia for free, the types of businesses that would have to pay the most would be North American companies, as well as software and technology businesses. While the

narrative about tech companies' donations to Wikipedia has often been around their massive usage of the free encyclopedic content for products and algorithms [7, 33], these findings might provide yet another perspective on how these companies benefit from the hard work of hundreds of thousands of volunteer editors.

More broadly, our work expands the small body of literature on measuring the value of Wikipedia to the Web [20, 32]. While previous work focused on the value of content *production,* for example estimating that Wikipedia generates $1.7 million of Reddit and Stack Overflow's revenue, based on the amount of Wikipedia-linked posts on those platforms [32], we focused here on the value of Wikipedia *traffic.* We provided, for the first time, an estimation of the monetary value offered—for free—by Wikipedia to the broader Web ecosystem by means of link navigation.

Limitations and future work. This study should be considered in the light of its limitations. Most notably, it was constrained to data collected during one month from English Wikipedia only, and as such provides a limited view of readers' general behavior. Future work should replicate the study for different time periods and language versions in order to paint a more complete and inclusive picture of Wikipedia readers' engagement with external links.

Besides broadening the scope, future work should also go deeper by more closely investigating why certain types of official link see particularly high or low CTRs (e.g., the CTR of links to geographical and biographical content was particularly low; cf. Fig. 5a). Also, considering that official links related to BUSINESS AND ECONOMICS saw the highest CTRs, it will be interesting to analyze which businesses benefit most from the free traffic provided by Wikipedia.

Whereas our investigation of the volume (RQ1) and patterns (RQ2) of engagement with external links on Wikipedia was primarily measurement-based, our estimation of the economic value of Wikipedia as a gateway to the Web (RQ3) was more speculative. On the one hand, we operated under the assumption that our methodology for obtaining costs per click via the Google Ads API is sound and provides accurate estimates, despite the fact that we relied on keyword suggestions of unknown quality from the Google Ads Keyword Planner and on uncertain auction simulations

---

[10]https://wikimediafoundation.org/about/financial-reports

from the Google Ads forecasting tool. On the other hand, and more fundamentally, estimating the economic value of the Web traffic generated by Wikipedia necessarily requires arguing about a hypothetical, counterfactual ("what if") situation, in our case, "What if website owners were to pay for the same number of clicks via Google Ads instead of obtaining them from Wikipedia for free?"

Although similar reasoning has been applied to estimate the value of images from Wikimedia Commons [8], it remains open how realistic that "what if" is: as a matter of fact, Wikipedia *is* providing those clicks for free, so why would website owners ever decide to pay for them instead? As one concrete example, one could imagine a situation where a website owner would want to increase traffic to their site, in which case our estimates indicate how much it would cost them to double the traffic they already receive from Wikipedia for free. Alternatively, one could imagine scenarios where Wikipedia were to be blocked by censorship or ceded to exist entirely, in which case our estimated economic value of traffic from Wikipedia would correspond to the loss on behalf of website owners due to the lack of that traffic. Finally, and more boldly, one could imagine a setting where Wikipedia decided to introduce a fee for clicks on official website links, in which case our estimate would upper-bound the amount of extra revenue Wikipedia could possibly earn from such a fee. Although the latter setting is highly unrealistic, we consider it a useful thought experiment that can help emphasize Wikipedia's importance as a provider of free traffic.

As a final remark, we argue that all notions of economic value are fundamentally counterfactual at heart, as they always consider "what if" scenarios ("If A were to give X to B, how much money would B give to A in return?"), which is also the reason why business valuations of companies are routinely criticized as absurdly off [25].

To conclude, we hope this work will offer ideas and methods to those interested in exploring Wikipedia's role in the larger Web ecosystem in more depth, and that it will help quantify the true value of the largest encyclopedic knowledge repository on the Web.

REFERENCES

[1] Alphabet Inc. 2019. Alphabet announces fourth quarter and fiscal year 2019 results. https://web.archive.org/web/20210104101629/https://abc.xyz/investor/static/pdf/2019Q4_alphabet_earnings_release.pdf

[2] Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. 2011. Towards a science of user engagement (position paper). In *Proc. UMWA*.

[3] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on Instagram. In *Proc. CHI*.

[4] Nicola Barbieri, Fabrizio Silvestri, and Mounia Lalmas. 2016. Improving post-click user engagement on native ads via survival analysis. In *Proc. WWW*.

[5] Eda Baykan, Monika Henzinger, Ludmila Marian, and Ingmar Weber. 2009. Purely URL-based topic classification. In *Proc. WWW*.

[6] Jörg Claussen, Tobias Kretschmer, and Philip Mayrhofer. 2013. The effects of rewarding user engagement: The case of Facebook apps. *Information Systems Research* 24, 1 (2013), 186–200.

[7] Megan Rose Dickey. 2019. Google.org donates $2 million to Wikipedia's parent org. *TechCrunch* (22 January 2019). https://web.archive.org/web/20210213130411/https://techcrunch.com/2019/01/22/google-org-donates-2-million-to-wikipedias-parent-org/

[8] Kristofer Erickson, Felix Rodriguez Perez, and Jesus Rodriguez Perez. 2018. What is the Commons worth? Estimating the value of Wikimedia imagery by observing downstream use. In *Proc. OpenSym*.

[9] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. 2014. Global disease monitoring and forecasting with Wikipedia. *PLoS Computational Biology* 10, 11 (2014), e1003892.

[10] Aaron Halfaker and R.Stuart Geiger. 2019. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *Proc. ACM Human-Computer Interaction*.

[11] Kyle S. Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M. Hyman, Alina Deshpande, and Sara Y. Del Valle. 2015. Forecasting the 2013–2014 influenza season using Wikipedia. *PLOS Computational Biology* 11, 5 (2015), e1004239.

[12] Marit Hinnosaar, Toomas Hinnosaar, Michael E Kummer, and Olga Slivko. 2019. Wikipedia matters. *SSRN* (2019).

[13] Yuheng Hu, Shelly Farnham, and Kartik Talamadupula. 2015. Predicting user engagement on Twitter with real-world events. In *Proc. ICWSM*.

[14] Yushi Jing, David Liu, Dmitry Kislyuk, Andrew Zhai, Jiajing Xu, Jeff Donahue, and Sarah Tavel. 2015. Visual search at Pinterest. In *Proc. KDD*.

[15] Michael Johnston. 2013. Wikipedia revenue analysis: How a wiki could make $2.3B a year. *MonetizePros* (25 June 2013). https://web.archive.org/web/20201112024633/https://monetizepros.com/features/analysis-how-wikipedia-could-make-2-8-billion-in-annual-revenue/

[16] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the world reads Wikipedia: Beyond English speakers. In *Proc. WSDM*.

[17] Xiaoliang Ling, Weiwei Deng, Chen Gu, Hucheng Zhou, Cui Li, and Feng Sun. 2017. Model ensemble for click prediction in Bing search ads. In *Proc. WWW*.

[18] Lauren A Maggio, Ryan M Steinberg, Tiziano Piccardi, and John M Willinsky. 2020. Meta-Research: Reader engagement with medical content on Wikipedia. *eLife* 9 (2020), e52426.

[19] Gene McKenna. 2014. Experiment shows up to 60% of "direct" traffic is actually organic search. *Search Engine Land* (8 July 2014). https://web.archive.org/web/20201020025709/https://searchengineland.com/60-direct-traffic-actually-seo-195415

[20] Connor McMahon, Isaac Johnson, and Brent Hecht. 2017. The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In *Proc. ICWSM*.

[21] Blagoj Mitrevski, Tiziano Piccardi, and Robert West. 2020. WikiHist.html: English Wikipedia's full revision history in HTML format. *Proc. ICWSM*.

[22] Helen Susannah Moat, Chester Curme, Adam Avakian, Dror Y. Kenett, H. Eugene Stanley, and Tobias Preis. 2013. Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports* 3, 1 (2013), 1–5.

[23] Helen Susannah Moat, Chester Curme, H. Eugene Stanley, and Tobias Preis. 2014. Anticipating stock market movements with Google and Wikipedia. *NATO Science for Peace and Security Series C: Environmental Security* (2014), 47–59.

[24] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2020. Quantifying engagement with citations on Wikipedia. In *Proc. WWW*.

[25] Jamie Powell. 2020. Is GMO's Montier right on 'absurd' US stocks? *Financial Times* (18 August 2020). https://www.ft.com/content/6ad966ed-17c8-484f-85b3-f685c8d28b13

[26] Michael Scholz, Christoph Brenner, and Oliver Hinz. 2019. AKEGIS: Automatic keyword generation for sponsored search advertising in online retailing. *Decision Support Systems* 119 (2019), 96–106.

[27] Evan Sheehan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. 2019. Predicting economic development using geolocated Wikipedia articles. In *Proc. KDD*.

[28] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read Wikipedia. In *Proc. WWW*.

[29] Yang Song, Xiaolin Shi, and Xin Fu. 2013. Evaluating and predicting user engagement change with degraded search relevance. In *Proc. WWW*.

[30] Nathan TeBlunthuis, Tilman Bayer, and Olga Vasileva. 2019. Dwelling on Wikipedia: Investigating time spent by global encyclopedia readers. In *Proc. OpenSym*.

[31] Neil Thompson and Douglas Hanley. 2017. Science is shaped by Wikipedia: Evidence from a randomized control trial. *SSRN Electronic Journal* (2017).

[32] Nicholas Vincent, Isaac Johnson, and Brent Hecht. 2018. Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities. In *Proc. CHI*.

[33] Rachel Withers. 2018. Amazon owes Wikipedia big-time. *Slate* (11 October 2018). https://web.archive.org/web/20210202152028/https://slate.com/technology/2018/10/amazon-echo-wikipedia-wikimedia-donation.html

[34] Sean Xin Xu and Xiaoquan (Michael) Zhang. 2013. Impact of Wikipedia on market information environment: Evidence on management disclosure and investor reaction. *MIS Quarterly* 37, 4 (2013), 1043–1068.

[35] Xing Yi. 2015. Dwell time based advertising in a scrollable content stream. US Patent App. 13/975,157.

[36] Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. How well do offline and online evaluation metrics measure user satisfaction in Web image search?. In *Proc. SIGIR*.