Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Orphan Articles: The Dark Matter of Wikipedia

Akhil Arora[1], **Robert West**[1*], **Martin Gerlach**[2]
[1]EPFL
[2]Wikimedia Foundation
akhil.arora@epfl.ch, robert.west@epfl.ch, mgerlach@wikimedia.org

Abstract

With 60M articles in more than 300 language versions, Wikipedia is the largest platform for open and freely accessible knowledge. While the available content has been growing continuously at a rate of around 200K new articles each month, very little attention has been paid to the discoverability of the content. One crucial aspect of discoverability is the integration of hyperlinks into the network so the articles are visible to readers navigating Wikipedia. To understand this phenomenon, we conduct the first systematic study of *orphan articles*, which are articles without any incoming links from other Wikipedia articles, across 319 different language versions of Wikipedia. We find that a surprisingly large extent of content, roughly 15% (8.8M) of all articles, is de facto invisible to readers navigating Wikipedia, and thus, rightfully term orphan articles as the *dark matter* of Wikipedia. We also provide causal evidence through a *quasi-experiment* that adding new incoming links to orphans (de-orphanization) leads to a statistically significant increase in their visibility in terms of the number of pageviews. We further highlight the challenges faced by editors for de-orphanizing articles, demonstrate the need to support them in addressing this issue, and provide potential solutions for developing automated tools based on cross-lingual approaches. Overall, our work not only unravels a key limitation in the link structure of Wikipedia and quantitatively assesses its impact but also provides a new perspective on the challenges of maintenance associated with content creation at scale in Wikipedia.

1 Introduction

Wikipedia is the largest multi-lingual platform on the Internet for open and freely accessible knowledge. As of November 2022, Wikipedia comprised 60M articles across 319 different language versions, and it has since been growing at a rapid rate of around 200K articles per month. In fact, in order to bridge knowledge gaps (Redi et al. 2021), there have been a plethora of efforts to systematically add content that is currently absent, e.g., formation of organized groups such as Wiki Women in Red (WMF 2015c) to add articles about women (Vitulli 2018), development of automatic tools such as Project Quicksilver (Primer.ai 2020) to surface missing articles by generating a list of people who are missing from Wikipedia based on news, or translation of existing articles into other languages (Wulczyn et al. 2016). These initiatives have been extremely successful—for example, Wikipedia's content translation tool (WMF 2014) has helped to create more than 1M new articles (Ozurumba 2021). As a result, one of the main challenges is how to maintain this ever-increasing volume of content. Specifically, it is crucial to properly integrate new articles into the existing network structure. In fact, hyperlinks play a crucial role in the encyclopedia and editors have developed a dedicated guideline to "build the web" in English Wikipedia's manual of style (WMF 2004b), primarily to enable readers to access relevant information on other Wikipedia pages easily. While the largest share of traffic to Wikipedia comes from search engines, a substantial fraction (38%) of pageviews result from traffic via internal hyperlinks (Piccardi et al. 2023).

Thus, it is problematic if existing articles are not integrated into the network structure because they will suffer from a lack of visibility to readers. In addition, the lack of visibility reflects structural biases such as the gender gap (Beytía and Wagner 2022). For example, the visibility of biographies about women is systematically lower than for biographies about men (Wagner et al. 2015, 2016). Different community-driven campaigns, which have been successfully adding and improving content about women, have been shown to be less successful at addressing structural biases that limit their visibility (Langrock and González-Bailón 2022). Previous research demonstrated in a quasi-experiment a spill-over effect of attention in Wikipedia (Kummer 2014; Zhu, Walker, and Muchnik 2020) suggesting a causal relation that visibility can be improved by adding relevant incoming links to articles. This provides evidence that the lack of visibility can be improved by suitable interventions.

In this work, we explore the question of the lack of visibility of articles in more than 300 language versions of Wikipedia. We specifically focus on so-called *orphan articles*, which are defined as articles that do not have any incoming links from other articles in the main namespace of Wikipedia.[1] These articles are of particular interest since they are de facto invisible for readers navigating hyperlinks in Wiki-

[*]Robert West is a Wikimedia Foundation Research Fellow.

[1]https://en.wikipedia.org/wiki/Wikipedia:Orphan

pedia. Specifically, we aim to address the following research questions:

- **RQ1**: What are the key characteristics of orphan articles? (Sec. 4)
- **RQ2**: Does adding incoming links (de-orphanization) increase the visibility of orphan articles? (Sec. 5)
- **RQ3**: What is the current state of de-orphanization and what are the potential ways to improve it? (Sec. 6)

To answer the aforementioned research questions, we conduct the first systematic study on orphan articles in Wikipedia and show that orphans make up a surprisingly large fraction of articles. We also establish causal evidence through quasi-experiments that orphan articles are significantly less visible than non-orphan articles. We then describe the challenges faced by editors in addressing this issue and sketch potential solutions to develop models to support their efforts, demonstrating the opportunities for using our insights in future works. Together, these results provide a new perspective on maintenance costs associated with content creation and challenges in making existing knowledge discoverable.

2 Related Work

In this section, we review existing works that overlap closely with our study. For additional related work, please see Appendix A.

Orphans articles in Wikipedia. The English Wikipedia contains an information page about orphan articles (WMF 2003), which states that "these pages can still be found by searching Wikipedia, but it is preferable that they can also be reachable by links from related pages; it is therefore helpful to add links from other suitable pages with similar or related information." Moreover, according to the manual of style (WMF 2004b), de-orphanizing articles is an important aspect of "building the web", as hyperlinks are crucial for helping readers in conveniently finding related information while reading Wikipedia. Even the guide for creating new pages suggests to "Provide internal links to the article from other pages".[2] Editors use a maintenance template (WMF 2004a) to mark orphan articles with a note visible at the top of the article and organize them in specific categories. A group of editors from WikiProject Orphanage (WMF 2007b) are "dedicated to clearing up the immense backlog of orphaned articles" and provide suggestions for how to de-orphanize articles. For this, they have a set of tools at their hand such as *findlink* (mentioned in the hat-note of each orphan), which suggests new links from where to link an article based on string matches of the page-title (Betts 2008). However, despite the organized efforts, in English Wikipedia, the number of articles tagged with an orphan template has been decreasing very slowly from a peak of 140K in 2017 to around 80K in 2023 (WMF 2006a). In opposition to orphan articles (no incoming links), there are the so-called dead-end articles, which are articles that contain no outgoing links to other Wikipedia articles. Similarly, these articles are marked with a maintenance template (WMF 2007a) but for English Wikipedia, the number of affected articles is in the low single digits.

Spillover effect. Different recent studies have demonstrated a so-called spillover effect in Wikipedia (Kummer 2014, 2018; Zhu, Walker, and Muchnik 2020). These are based on quasi-experiments suggesting a causal effect of newly added incoming links to the attention received by articles. For example, (Zhu, Walker, and Muchnik 2020) compared a "treatment" group of articles edited through organized campaigns with a "control" group of articles that did not receive any edits, finding a significant increase in the number of pageviews for articles that were newly linked from the treated articles but that themselves were neither in the control or treatment group. However, none of the aforementioned studies investigated orphan articles specifically.

Knowledge gaps and visibility. Wikipedia and its sister projects such as Wikimedia Commons, Wikidata, or Wiktionary, suffer from a wide range of knowledge gaps (Redi et al. 2021). For example, the content gender gap refers to the fact that only 15–20% of biography articles in Wikipedia are about women (Konieczny and Klein 2018). This gap has been confirmed in countless studies also taking into account more nuanced metrics such as notability (Tripodi 2021) and has also been confirmed beyond content for the population of editors (Hill and Shaw 2013; Ford and Wajcman 2017) and readers (Johnson et al. 2021). One often overlooked aspect, however, has been raised about biases in the visibility of already existing content (Beytía and Wagner 2022). Anecdotally, it has been reported that deletion of biography articles is more common if the subject is not yet mentioned on other Wikipedia articles (Vitulli 2018). Several studies documented how articles about women are systematically less visible than articles about men using proxies such as PageRank centrality (Wagner et al. 2015, 2016). This is especially interesting in view of studies showing that organized campaigns are successful at adding content about women that would otherwise be missing, but are less successful at addressing structural biases that limit the visibility of women-focused content as addressing these biases directly is non-trivial (Langrock and González-Bailón 2022). Also, there is evidence that using existing tools for recommending links to support editors in addressing this problem could actually reinforce those biases (Ferrara et al. 2022).

Complementary to all the aforementioned works, with the primary focus on visibility, orphan articles provide a more nuanced approach to measuring knowledge gaps in Wikimedia projects (Redi et al. 2021).

Cross-lingual approaches in Wikipedia. With more than 300 active language versions, Wikipedia is an intrinsically multilingual project. While there are community-created lists of vital articles,[3], i.e., articles that every Wikipedia should have, it has been found that there are substantial differences between different language versions (Hecht and Gergle 2010; Bao et al. 2012). Thanks to the efforts of,

[2] https://en.wikipedia.org/wiki/Help:Drawing_attention_to_new_pages

[3] https://meta.wikimedia.org/wiki/List_of_articles_every_Wikipedia_should_have

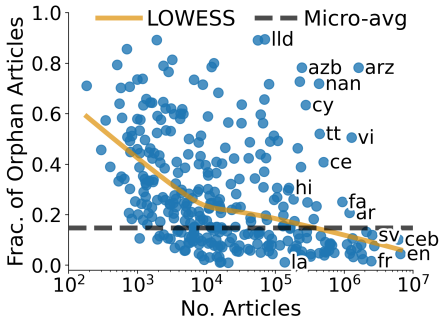


Figure 1: Analyzing the extent of orphan articles across all Wikipedia language versions.

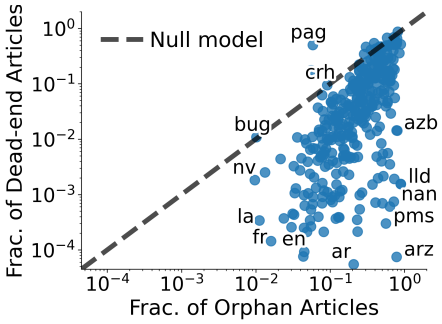


Figure 2: Comparing the extent of orphans with that of of dead-end articles across all Wikipedia language versions.

among others, multilingual editors (Hale 2014), content has been shown to propagate from one language version to another (Valentim et al. 2021; Yoon et al. 2022). This has also been the motivation for leveraging content translation systematically in order to grow the different language versions of Wikipedia on the level of articles (Wulczyn et al. 2016), sections (Piccardi et al. 2018), or section titles (Aslam and Sáez-Trumper 2022). Specifically, the model for recommending articles for translation has been developed into a tool by the Wikimedia Foundation (Laxström, Giner, and Thottingal 2015). This tool supports editors to translate articles from one language into another by providing a first draft of the article using automatic translation (WMF 2014). This approach has been extremely successful, with over 1M translated articles as of 2021 (Ozurumba 2021). Along similar lines, one recent study proposed to improve the overall inter-connectivity among articles by taking advantage of existing links in other language versions (Lotkowski 2017); however, the work considers only one specific case translating any possible link from English to Scots Wikipedia.

3 Data and Resources

In this section, we describe in detail the datasets used for studying orphan articles in Wikipedia. We consider 319 different language versions of Wikipedia and collect data spanning 7 monthly snapshots ranging from August 2022 to February 2023. Unless stated otherwise, the results presented in this paper are based on the monthly snapshot of November 2022. For other snapshots, the results portrayed similar trends, and are therefore omitted. All the publicly accessible resources (data, descriptive statistics, and code) required to reproduce the analyses in this paper are available at https://github.com/epfl-dlab/wikipedia-orphans.

Wikipedia hyperlink network. For each language version, we construct its 'directed' hyperlink network by leveraging the `pagelinks` table, which tracks all internal links among Wikipedia articles and is available as a SQL dump (WMF 2015a) released by Wikipedia on a monthly basis. Note that we resolve redirects (Hill and Shaw 2014) and only consider links between articles in the main namespace[4] of Wikipedia. Specifically, we use the dumps released on the first of each month, e.g., for November 2022, we use the dump dated '2022-11-01' to extract a total of 60M articles and 3.5B links across 319 language versions. Additionally, using the Wikidata dump (WMF 2015b) dated '2023-02-26', each article was appropriately mapped to its corresponding unique language-agnostic Wikidata identifier (QID), which further facilitates matching articles across languages.

Orphan-articles data. This data consists of orphan articles in Wikipedia, which are articles with no incoming links from any other main namespace articles in the same language version of Wikipedia. This data is used primarily in Sec. 4.

De-orphanizing-links data. This data consists of new incoming links added to orphan articles. Specifically, for a given month, e.g., November 2022, we obtain the added links by computing the set difference between links existent in Wikipedia in December and November 2022, respectively. Next, to obtain the de-orphanizing links we restrict ourselves to added links with orphan articles from November 2022 as the target. This data is used primarily in Sec. 5.

Wikipedia article features. For each article, we extract the following features: topic, quality, time since creation (age), whether it was created by a bot, the gender (for biography articles), and pageviews.

- **Topics**: We use the language-agnostic topic model developed by (Johnson, Gerlach, and Sáez-Trumper 2021), which assigns topic labels to articles based on the taxonomy (Halfaker and Johnson 2019) developed with inputs from the Wikipedia editor community (WMF 2006b). We use the 4 top-level topic labels from the taxonomy, namely 'Culture', 'Geography', 'History and Society', and 'STEM'. For each topic label, the model predicts the probability for an article to be assigned to that topic, and the label is assigned only if the predicted probability is > 0.5.
- **Quality**: Article quality is computed using the language-agnostic quality model by (Johnson 2021), which uses fea-

[4] Articles in Wikipedia are grouped into collections called 'namespaces', which differentiate between their purpose at a high level. For details please see https://w.wiki/6hoy.

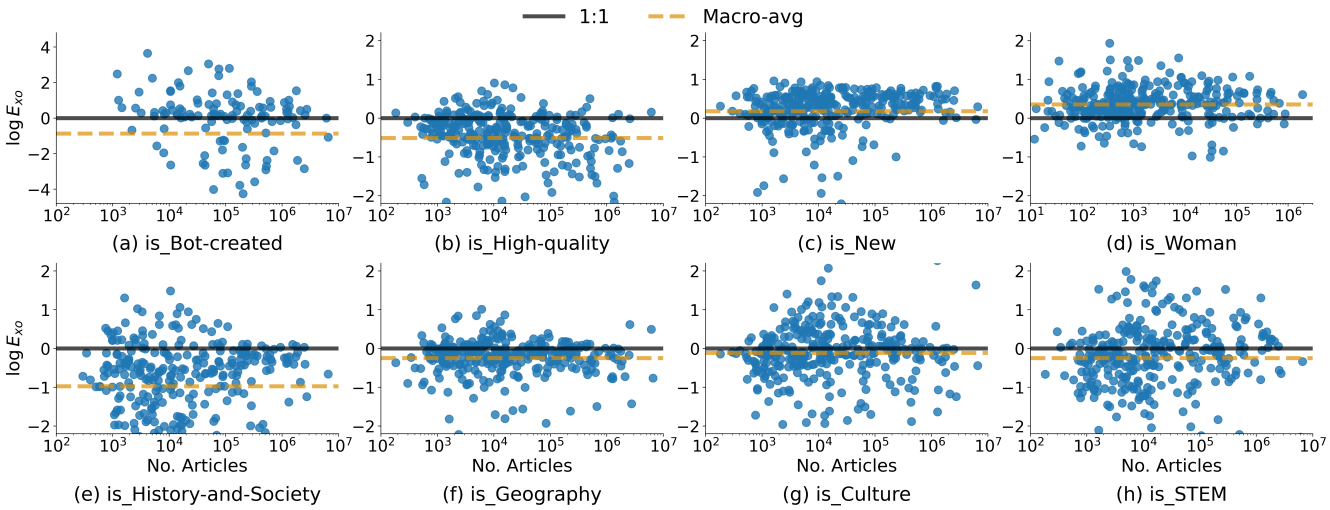


Figure 3: Characterizing orphans based on article features in all Wikipedia language versions. For a given feature and language, points above the $\mathbf{y} = \mathbf{0}$ line indicate an over-representation of the feature among orphans in that language.

tures such as article-length, number of links, sections, references, etc., to obtain a score between $[0, 1]$.

- **Age**: For each article, we extract its creation timestamp from its revision history, and represent its age using the UNIX timestamp format.

4 Characterizing Orphan Articles

In this section, we assess the extent of orphan articles, contrast it with the extent of dead-end articles, and characterize orphan articles based on different article features in all language versions of Wikipedia.

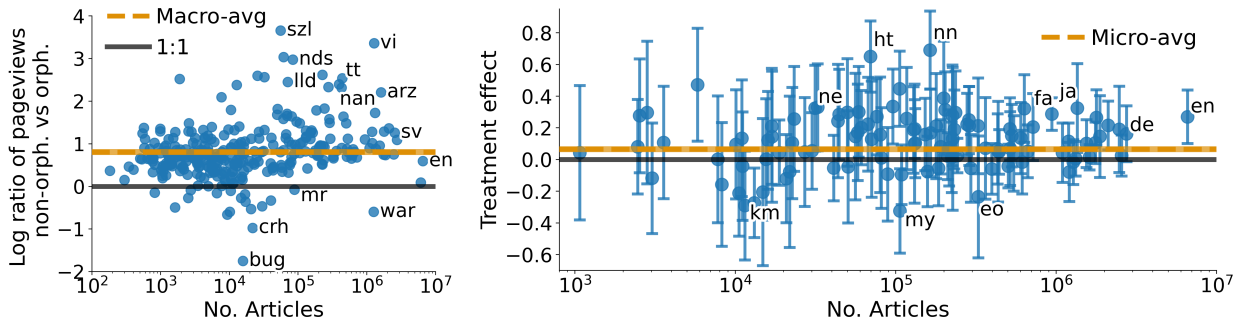
Extent of orphans. Counting the number of orphan articles, we find that the fraction of orphan articles is surprisingly large (Fig. 1): out of the total 60M articles across all the 319 Wikipedia language versions, 8.8M (14.7%) are orphan articles. This observation is not driven by only a few outliers but is consistent across (almost) all language versions of Wikipedia: there are more than 100 Wikipedia language versions with at least 30% orphan articles. As portrayed by the LOWESS regression fit (Cleveland and Devlin 1988), smaller Wikipedia language versions tend to have a higher fraction of orphans; yet larger Wikipedia language versions can also have above-average orphan-rates. For example, among the 20 largest Wikipedia language versions, we find that Egyptian Arabic (arz, 78%), Vietnamese (vi, 50%), Persian (fa, 25%), and Arabic (ar, 21%) portray high orphan-rates. In relative terms, English Wikipedia is an outlier with only 5% orphans, however, this still corresponds to more than 300K articles.

Comparison with dead-end articles. In comparison to orphan articles (no incoming links), at 300K ($\sim 0.5\%$ of all articles), dead-end articles (no outgoing links) can be considered virtually non-existent (Fig. 2). For almost all Wikipedia language versions, we find that the fraction of dead-ends is often (at least) an order of magnitude lower than the fraction of orphans. For example, Egyptian Arabic (arz) possesses 1.25M orphans (78%) but only 121 dead-ends (0.007%). We thus find that the problem of orphan articles is very *distinct from and of much larger scope* than the problem of dead-end articles. This is perhaps intuitive: while the issue of dead-end articles can be addressed by editing the respective article itself, orphan articles can only be addressed by (identifying and) editing other articles.

Characterizing orphans. To better understand which types of articles are found more commonly among orphans (such as whether the article is about a specific topic), we perform a characterization of orphan articles based on the article features described in Sec. 3. For a given feature (x), we calculate how many of the orphan articles (o) have that feature, i.e., the conditional probability $P(x|o)$. By comparing $P(x|o)$ with the overall propensity of the feature x among all articles ($P(x)$), the ratio $\log E_{xo} = P(x|o)/P(x)$ shows whether feature x is over-represented ($\log E_{xo} > 0$) or under-represented ($\log E_{xo} < 0$) among orphan articles.

Note that for the purpose of this analysis we require binary article features. While most article features are binary by construction, we binarize the numeric features 'age' and 'quality' by partitioning the set of articles in each language into two groups—old vs. new and high vs. low quality, respectively—using their median value. We investigated the following article features and whether they are over- or under-represented among orphans (Fig. 3).

Bot-created articles are under-represented among orphans. However, the variation is large and there is a considerable number of Wikipedia language versions for which bot-created articles are substantially over-represented. For example, among the 20 largest Wikipedia language versions, we find that in Italian (it) Wikipedia (similar trends observed for Chinese (zh) and Portugese (pt)), $P(\text{bot}|o) = 0.19$, which is much larger than the overall fraction of bot-created articles in Italian, $P(\text{bot}) = 0.06$. Moreover, for Bulgar-



(a) Correlation: Non-orphans vs. Orphans (b) Quasi-experiment: Treated (orph. → de-orph.) vs. Control (orph. → orph.)

Figure 4: Comparing the pageviews received by orphan and non-orphan articles across all Wikipedia language versions. The error bars denote 95% CIs, and have been omitted from (a) as they were small and therefore impacting readability.

ian (bg) Wikipedia (similar trends observed for Malay (ms) and Afrikaans (af)), $P(\text{bot}|o) = 0.51$, which is more than 4 times the overall fraction of bot-created articles in Bulgarian, $P(\text{bot}) = 0.12$. This finding could point to undesired artifacts emanating from the use of semi-automatic tools for content creation.

Considering the **gender** of biography articles, we find that articles about women are over-represented among orphans. Overall, we know that between 15-20% of biographies are about women[5]. However, among orphan articles, we find a much higher percentage of biographies about women. For example, in English (en) Wikipedia $P(\text{woman}|o) = 0.29$, which is much larger than the overall fraction of women biographies in English, $P(\text{woman}) = 0.19$. An even more extreme example is Catalan (ca) Wikipedia with $P(\text{woman}|o) = 0.42$, while $P(\text{woman}) = 0.20$. This shows that biography articles about women are disproportionately more likely to be orphan articles.

Next, **high-quality** articles are under-represented among orphans, and thus, articles with lower quality are more likely to be orphans across almost all Wikipedia language versions. Moreover, while newer articles (**age**) are slightly over-represented among orphans, the effect size is relatively small. Finally, all **topics** are equally represented among orphans; the only exception is 'History and Society' which is, on average, substantially underrepresented among orphans.

At this juncture, it is important to note that establishing causality is neither the focus nor the intent of the aforementioned analysis, which solely reveals correlations between different article features and their existence among orphans.

5 Visibility of Orphan Articles

In this section, we investigate the visibility of orphan articles to readers navigating Wikipedia. By definition, we know that *structurally* they are less visible within Wikipedia because there are no incoming links pointing to orphans from other articles. Here, we want to assess to which degree this also holds *functionally*, i.e., is this also reflected in orphan articles receiving fewer pageviews?

[5]https://humaniki.wmcloud.org/

5.1 Analyzing Correlations

As a first step, we compare the pageviews received by orphan articles with that of non-orphan articles, and find that orphans receive substantially fewer pageviews than non-orphans (Fig. 4a). Specifically, due to the long tail in the distribution of pageviews for articles, we compare the mean of the logarithm of the pageviews between the two groups. Averaged across all Wikipedia language versions, we find that the mean for non-orphans is twice as high as the mean for orphans. This indicates that orphans are, on average, less visible and less visited than non-orphan articles. However, this observation is only a correlation, i.e., we cannot conclude that the number of pageviews is lower *because* the articles are orphans.

5.2 Establishing Causality

In order to establish a causal link that fewer pageviews are a result of an article being an orphan, we conduct a quasi-experiment (Rosenbaum 2017) and use difference-in-differences (Angrist and Pischke 2008), a widely used causal inference method, to study the change in the number of pageviews for those orphans that were de-orphanized.

Setting up treatment-control groups. To setup the quasi-experiment, we follow the process portrayed in Fig. 5a. As treatment group, we consider all articles that were orphans in the monthly snapshot of October 2022 but at some point in the following month, i.e., November 2022, received a de-orphanizing incoming link (cf. *de-orphanizing-links data* in Sec. 3) so they were not orphans anymore. For each de-orphanized (treated) article, we consider the same article, albeit in a different Wikipedia language version in which it remained an orphan, as control. In this way, a given orphan article gets de-orphanized (treated) in one language but remains an orphan (control) in another. The motivation to match on the same article is to construct a control group that is as similar to the treatment group as possible, thereby accounting for potential confounding effects due to often fast shifts in attention to specific topics or current events. As a potential limitation, this setup assumes an absence of

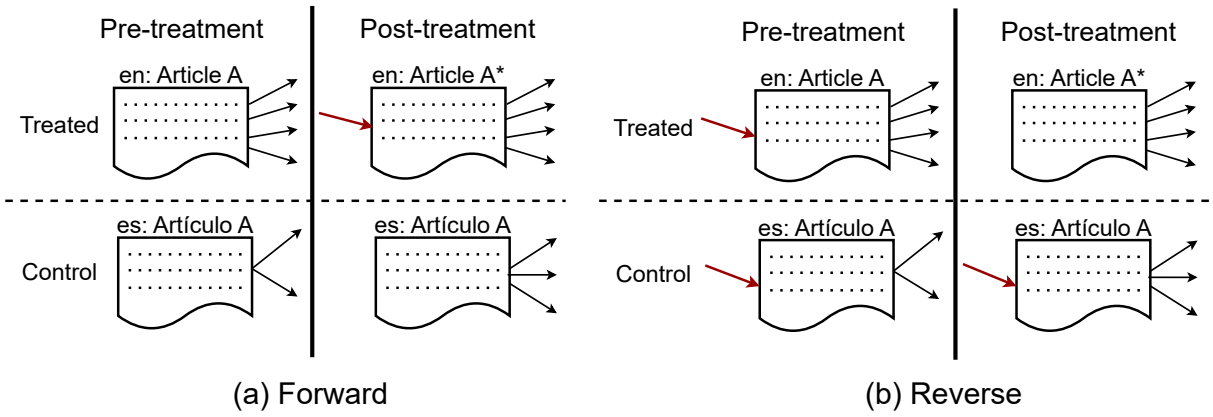


Figure 5: A pictorial representation of the quasi-experiment: (a) Forward: an article that receives a new incoming link (denoted in red font) is considered as treated, whereas the same article in another language that does not receive any new incoming links is considered as control; (b) Reverse: an article that loses an incoming link is considered as treated, whereas the same article in another language that does not lose any incoming links is considered as control.

language-specific shifts in attention, which we discuss in detail in Sec.7.3.

The aforementioned process yields 36,707 treated-control article pairs across 192 language versions (no article was de-orphanized in the remainder 127 versions).

Difference-in-Differences (DiD). We use the following DiD model to estimate the effect of de-orphanization on article visibility by comparing the aforementioned treatment-control groups three months before (August–October 2022) and after (December 2022–February 2023) the treatment.

$$Y_{it} = \beta_0 + \beta_1 \text{deorph}_i + \beta_2 \text{after}_t + \beta_3 \text{deorph}_i \text{after}_t + \varepsilon_{it}, \quad (1)$$

where Y_{it} is the logarithm of the number of pageviews received by article i in the month t, deorph_i indicates whether article i was de-orphanized or not, after_t indicates whether the month t is before or after the treatment month, and ε_{it} is the error term. The coefficient β_3 denotes the causal effect of article de-orphanization on its visibility measured by the number of pageviews. We extend the aforementioned DiD model to (1) estimate language-specific treatment effect by adding language as a categorical variable into the model, and (2) estimate month-specific treatment effect by transforming after_t from a binary to a categorical variable.

Results. The DiD model described in Eq. 1 yields a statistically significant overall increase of 6.5% ($p < 10^{-10}$) in the number of pageviews for articles de-orphanized in November 2022. Next, we estimate the treatment effect for 120 language versions in which at least 30 articles were de-orphanized. While the treatment effect differs across Wikipedia language versions, we find a statistically significant ($p < 0.05$) increase for 25 whereas a decrease for 8 language versions (the effects were not significant for the remainder 87 languages), respectively (Fig. 4b). The largest increase can be observed for Norwegian (nn) and Haitian Creole (ht), whereas the largest decrease is observed for Cebuano (ceb).

Moving ahead, we estimate the month-specific treatment effect (Fig. 6a). It is important to point out the following: (1) we observe a statistically significant ($p < 0.001$) positive DiD effect (7.8%) immediately after de-orphanization, (2) the positive effect is persistent for the entire post-treatment duration, and (3) the pre-treatment difference is statistically indistinguishable from 0, suggesting that our quasi-experimental setup generates treatment and control groups that portray similar behavior prior to de-orphanization, while also providing some evidence in favor of the existence of parallel-trends pre-treatment. Moreover, we obtain qualitatively similar findings if other months are chosen as treatment months (Supplementary Fig. S1). Overall, the aforementioned points highlight the robustness of our findings.

Finally, we find that the increase in pageviews for the treatment group is, indeed, mostly driven by readers using the newly added incoming links (Fig. 6b). For this, we stratify the number of pageviews with respect to their referrer (internal: via a Wikipedia link, external: from an external website or external search engine, and unknown: missing referrer). Fitting a separate DiD model for each case yields statistically significant ($p < 0.001$) effect sizes only for pageviews with internal referrer.

Note that the vast majority (80%) of de-orphanized articles received exactly one incoming link (Supplementary Fig. S2). Thus, the reported treatment effects in all the aforementioned analyses can be approximately attributed to be emanating from a single link.

Inverting the treatment. The previous analysis provides causal evidence that adding incoming links to orphans leads to an increase in the number of pageviews. An alternative explanation could argue that causality works in the opposite direction, i.e., an increase in the number of pageviews could have lead to the added incoming link because the increase in attention will make it more likely that an editor will encounter and make edits related to the articles. In order to rule out this alternative explanation, we analyze the inverse process: the treatment group comprises articles that are orphanized whereas the control group comprises articles that

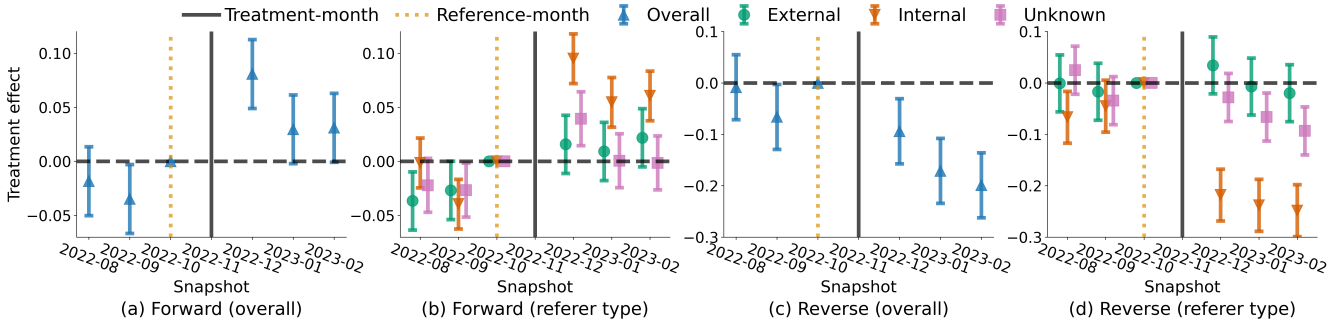


Figure 6: Per-month DiD treatment effect with 95% CIs for the (a)-(b) forward and (c)-(d) reverse setup considering November 2022 as the treatment month.

remain non-orphans (Fig. 5b). Considering November 2022 as the treatment month, this process yields 12,560 treated-control article pairs across 121 language versions (no article was orphanized in the remainder 198 versions).

Our DiD model reveals that treated articles experience a statistically significant reduction of 13% ($p < 10^{-10}$) in the number of pageviews. Further, similar to the forward setup, (1) the pre-treatment difference is statistically indistinguishable from 0, (2) the reduction is persistent for the entire post-treatment duration (Fig. 6c), and (3) the reduction is prevalent only for pageviews with internal referrer (Fig. 6d).

It is important to highlight that the causal direction is less contentious in this setup: it is very unlikely that a decrease in pageviews would cause editor activity involving the removal of the corresponding link we considered as treatment. Overall, this analysis provides further confidence for establishing the causal direction that adding incoming links to orphan articles leads to an increase in the number of pageviews.

6 De-orphanization in Practice

In this section, we assess the current state of organic de-orphanization, demonstrate the challenges faced by Wikipedia editors in de-orphanizing articles, and provide potential solutions for developing automated de-orphanizing tools.

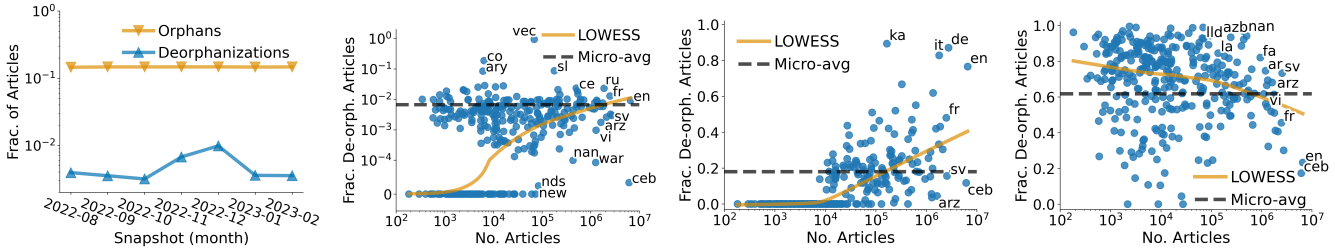
Current state. Wikipedia editors have developed different approaches for de-orphanizing articles such as through marking orphan articles via maintenance templates or coordination via WikiProjects. While these efforts, on average, facilitate de-orphanization of around 35K articles per month, in comparison to the overall fraction of orphans, the rate of de-orphanization is more than an order of magnitude smaller at around 0.5% per month (Fig. 7a). With this rate, it would take more than 20 years to work through the backlog of currently existing orphans. However, with the addition of new content, new orphans are also created such that the overall fraction of orphans remains approximately constant despite the continuous efforts by editors.

We observe some variation in the rate of de-orphanization across Wikipedia language versions (Fig. 7b). There are only few Wikipedia languages that exceed a rate of 1% in de-orphanization. In contrast, there are 94 languages with no de-orphanizations at all. Taking into account that the latter is more common for smaller language versions, using a LOWESS regression fit we find an overall positive correlation between the size of the language version and the rate of de-orphanization.

Challenges. These observations raise the question about potential reasons for the low rates of de-orphanization. In Sec. 4, we showed that the number of orphan articles is typically much larger than the number of dead-end articles. This suggests that adding new incoming links is a more difficult task than adding new outgoing links to articles. While the latter can be easily added by editing the respective article itself, for adding new incoming links the workflow for editors is more complex because one has to first identify other articles where a link to the orphan articles can be inserted.

To help editors in this task, the maintenance templates to mark orphan articles suggests the use of the findlink tool (Betts 2008) to identify suitable candidates. Findlink is a community-developed tool that tries to locate unlinked mentions of the specified article title in other articles via a relatively simple text-based search. However, this approach often yields very few candidates for orphans because the title of the orphan article does not yet appear as a potential mention in the text of any other article. Moreover, findlink works well only for very large language versions, such as English (en), German (de), and Italian (it). In fact, the performance is substantially low for smaller language versions (Fig. 7c) with no candidates returned whatsoever for 190 language versions. Overall, this approach yields at least one candidate only for 1.6M (18%) out of 8.8M orphans. From this, we conclude that available tools such as findlink are struggling to support editors in identifying candidates for de-orphanization.

Potential solution. Inspired by the success of content translation approaches in Wikipedia (Wulczyn et al. 2016), we test whether cross-lingual approaches could be a useful signal to identify candidates for de-orphanization. The hypothesis is that for an orphan article a in a Wikipedia language w, the same article might not be an orphan in another Wikipedia language version $w' \neq w$, thereby possessing an incoming link from article s to a in w'. If such an article s already exists in the Wikipedia language version w, we have identified a natural candidate for a new link from s to a



(a) Orphans vs. De-orphanization (b) De-orphanizations per Wiki (c) Findlink (d) Cross-lingual

Figure 7: Analyzing (a)-(b) the current state of de-orphanization, and the fraction of orphans that can be potentially de-orphanized using (c) Findlink, and (d) Cross-lingual approaches across all Wikipedia language versions.

to de-orphanize a in w. These link candidates can generally be considered of high precision because they have already been vetted by one or more communities of editors. We find that this approach could provide suggestions for a vast majority of orphan articles (Fig. 7d). Overall, for 5.5M (62%) out of 8.8M orphans, this approach yields at least one link candidate for de-orphanization, which already exists in at least one other Wikipedia language version. In many cases (3.2M), we could actually identify not just one but 10 or more different (incoming) link candidates per orphan article. While this heuristic is similarly effective across almost all Wikipedia languages, where other languages generally contain link candidates for more than half of the orphan articles, the effectiveness is slightly better for smaller Wikipedia languages. In fact, an outlier seems to be English Wikipedia with only 23% but even this amounts to link candidates for more than 68K orphan articles. We have developed a tool based on this heuristic, which is publicly available at https://linkrec.toolforge.org/.

7 Discussion

7.1 Summary of Findings

Many orphans. The number of orphan articles is surprisingly large: 8.8M (14.7%) out of 60M articles do not have any incoming links. This observation is not limited to only a few or small Wikipedia language versions, rather for more than 100 Wikipedia language versions the percentage of orphans is above 30%, including Egyptian Arabic (78%) and Vietnamese (50%), which are among the 20 largest Wikipedia language versions. In comparison, the number of dead-end articles, i.e., articles without any outgoing links, is very low across all languages (less than 0.5%). We find that orphan articles are negatively correlated with being: (1) of higher quality and (2) being about the topic of history and society, while possessing a slight positive association with being newer. More importantly, we showed that orphan articles encode structural biases: biography articles about women are substantially more common among orphans than expected from their overall frequency.

Lack of visibility. Orphan articles have, in general, fewer pageviews than non-orphan articles. We find causal evidence that orphan articles that were de-orphanized by editors receive a statistically significant increase in the number of pageviews. Specifically, we found that this increase is mainly driven by internally-referred pageviews from other Wikipedia articles which contain a link to the de-orphanized article.

Challenges for editors. The rate of organic de-orphanization is alarmingly low. For the snapshots we considered, editors added new incoming links to ∼35K orphan articles. While this constitutes an impressive effort by the community, at that rate it would take approximately 20 years to de-orphanize all orphan articles (assuming no newly created orphan articles). One hypothesis is that existing tools do not support editors in addressing this issue effectively. For example, FindLink (the tool suggested to editors in the orphans maintenance template) generally does not yield many results for orphan articles, especially for smaller languages. However, our results show that an orphan article in one language is not always an orphan in other languages. This suggests that we can develop an approach for identifying articles from which to link to orphans via link translation. Our results shows that this could be effective for 5.5M (62%) orphan articles.

7.2 Implications and Broader Impact

Maintenance vs content creation. While there exists a plethora of efforts to build methods and tools for mitigating content gaps: content translation (WMF 2014) to create new content, entity linking (Arora, García-Durán, and West 2021; Gerlach et al. 2021; Culjak et al. 2022; García-Durán, Arora, and West 2022) to ground concepts in knowledge bases, entity alignment (Leone et al. 2022; Sun et al. 2020) to enrich knowledge graphs, etc., there exists very little support for maintaining the created content. An important aspect of maintenance work is to integrate new articles into the hyperlink network of Wikipedia. While it does not necessarily add new content, it is crucial for the visibility of these articles. Adding incoming links to articles is also more difficult than adding content to (or creating) the article itself, since it requires editing other articles. In fact, it has been shown that community-organized campaigns such as Art+Feminism are very successful at improving the content of articles about women; but are less successful at increasing the structural visibility of articles by, e.g., adding new

inlinks (Langrock and González-Bailón 2022). To this end, this work focuses on improving the structural visibility of articles by adding inlinks to orphan articles.

Supporting editors. Our insights demonstrate the need to support editors to address the issue of orphan articles. This could be achieved by developing machine-learning models that could generate suggested edits in a machine-in-the-loop approach (Gerlach et al. 2021). Such approaches have been shown to be effective for generating outgoing links in the context of structured tasks for newcomer editors (WMF 2021). While the main focus for the latter was the action of the edit itself, extending this framework to adding new inlinks would, therefore, increase the value of the added links.

Cross-lingual approaches. Our analysis demonstrates the potential of cross-lingual approaches for building well-founded solutions to address the issue of orphan articles. These cross-lingual approaches not only yield a scalable and robust signal but also have the main advantage that derived models are easily interpretable for editors (e.g. editor communities from other Wikipedia languages have already vetted the information). This is in line with previous work on content translation in Wikipedia.

Knowledge gaps. Orphan articles provide a more nuanced approach to measuring knowledge gaps in Wikimedia projects (Redi et al. 2021). While the most common approach is to count the number of articles for, e.g., biography articles of different genders, it has been pointed out that this should be complemented by other aspects; most notably the quality and the visibility of articles (Miquel, Gerlach, and Johnson 2021). The current work provides a starting point to systematically operationalize knowledge gaps in terms of their visibility through orphan articles.

7.3 Limitations and Future work

While constructing treatment-control groups by matching on the same article (albeit in a different language version) is a powerful way of accounting for most potential confounders, as acknowledged in Sec. 5.2, one subtle limitation in this setup is the assumption that different languages portray similar trends in attention shifts for a fixed article. Specifically, one or more language versions may portray an increase in attention for an article (or a topic), which eventually could lead to that article being de-orphanized (treated) and thus, acting as a confounder. Moreover, for a given article (or topic), the expertise of the editor community may also greatly vary across language versions, and thus, an article could be de-orphanized in a given language primarily because of the existence of editor expertise. That said, the aforementioned limitations are unlikely to simultaneously impact both the forward (de-orphanization) and reverse (orphanization) setups (Sec. 5.2). Overall, considering the two setups in conjunction alleviates most limitations that could impact causal inference.

We showed that pageviews to de-orphanized articles increase significantly. One open question is whether these are "additional" pageviews or whether this simply corresponds to a shift of pageviews from other articles (i.e. "cannibalization"). Similarly, does the number of different readers who access these articles also increase? These questions are difficult to assess not only due to privacy restrictions of the data on readership to Wikipedia articles, but also the fluctuations in overall access volume to Wikipedia in order to disentangle potentially competing articles. Note that we showed an increase in pageviews when considering "organic" de-orphanziations performed by editors. It remains an open question whether "artificial" de-orphanizations resulting from suggestions to editors via link translation would result in a similar impact, however, it is an interesting question in its own right and constitutes as future work.

While cross-lingual approaches to Wikipedia content contain a rich signal to extend content coverage, there are some caveats. One challenge is that translations might not meet certain quality standards. In an extreme example, it was recently found that nearly half of Scots Wikipedia was created by someone who does not speak Scots (Harrison 2020). One additional concern is to avoid "language imperialism" and take into account the cultural context (Miquel-Ribé and Laniado 2018), i.e., different Wikipedia versions cover the same topics differently (Hecht and Gergle 2010).

The preferential attachment model is one of the best-known models to understand observed properties of real-world networks (including Wikipedia (Capocci et al. 2006)), e.g., with respect to their degree distributions (Newman 2018). However, within this framework networks typically yield a large fraction of nodes with in-degree 0 (i.e. orphans). While this might be natural in many contexts such as the World Wide Web, it is an undesirable property for Wikipedia due to the lack of visibility for much of its existing content. This leads to the question of alternative network formation processes that do not lead to large number nodes with in-degree 0; and how to apply this in the context of Wikipedia's communities.

7.4 Ethical Considerations

In our opinion, this work has no major ethical considerations. All the datasets and resources used in this work are publicly available and do not contain any private or sensitive information about Wikipedia readers. Moreover, all the findings are based on analyses conducted at an aggregate level, and thus, no individual-level inferences can be drawn. Finally, we took utmost care to distinguish claims establishing causality from those that present non-causal findings, and thus, we do not foresee any negative media impact, especially around misrepresentation of findings, emanating from this research. We confirm that we have read and abide by the AAAI code of conduct.

8 Conclusion

Our work constitutes the first characterization of orphan articles as the dark matter of Wikipedia: a surprisingly large fraction of articles across all 319 language versions of Wikipedia is de-facto invisible to readers of Wikipedia when navigating the hyperlink network. Our analysis not only reveals the existence of a causal link between the addition of incoming links to orphans and an increase in their visibility

in terms of the number of pageviews, but also demonstrates the need to develop automated tools to support editors in addressing this issue, e.g., via cross-lingual approaches. The latter would further help address structural biases related to (the lack of) visibility of articles about, e.g., women in the context of the gender gap. Overall, our results provide a new perspective on the challenges and costs of maintenance associated with the constant creation of new content.

Acknowledgements

We would like to thank Leila Zia, Miriam Redi, Isaac Johnson, Marko Čuljak, Veniamin Veselovsky, Guiseppe Russo, and Manoel Horta Ribeiro for insightful discussions as well reviewing an initial draft of this paper. West's lab is partly supported by grants from Swiss National Science Foundation (200021_185043), Swiss Data Science Center (P22_08), H2020 (952215), Microsoft Swiss Joint Research Center, and Google. We also gratefully acknowledge generous gifts from Facebook, Google, and Microsoft.

References

Anderka, M.; and Stein, B. 2012. A breakdown of quality flaws in Wikipedia. In *WICOW/AIRWeb Workshop on Web Quality*, 11–18.

Angrist, J. D.; and Pischke, J.-S. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arora, A.; García-Durán, A.; and West, R. 2021. Low-Rank Subspaces for Unsupervised Entity Linking. In *EMNLP*, 8037–8054.

Arora, A.; Gerlach, M.; Piccardi, T.; García-Durán, A.; and West, R. 2022. Wikipedia Reader Navigation: When Synthetic Data Is Enough. In *WSDM*, 16–26.

Aslam, M.; and Sáez-Trumper, D. 2022. Section Alignment at Large Scale. https://meta.wikimedia.org/wiki/Research:Expanding_Wikipedia_articles_across_languages/Inter_language_approach/Section_Alignment_at_Large_Scale. Accessed: 2023-05-01.

Bao, P.; Hecht, B.; Carton, S.; Quaderi, M.; Horn, M.; and Gergle, D. 2012. Omnipedia: bridging the wikipedia language gap. In *CHI*, 1075–1084.

Betts, E. 2008. The Findlink tool. https://edwardbetts.com/find_link/. Accessed: 2023-05-01.

Beytía, P.; and Wagner, C. 2022. Visibility layers: a framework for systematising the gender gap in Wikipedia content. *Internet policy review*, 11(1).

Capocci, A.; Servedio, V.; Colaiori, F.; Buriol, L.; Donato, D.; Leonardi, S.; and Caldarelli, G. 2006. Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E*, 74(3): 1–6.

Cleveland, W. S.; and Devlin, S. J. 1988. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83: 596–610.

Culjak, M.; Spitz, A.; West, R.; and Arora, A. 2022. Strong Heuristics for Named Entity Linking. In *NAACL SRW*, 235–246.

De Ruvo, G.; and Santone, A. 2015. Analysing Wiki Quality Using Probabilistic Model Checking. In *WETICE*, 224–229.

Ferrara, A.; Espin-Noboa, L.; Karimi, F.; and Wagner, C. 2022. Link recommendations: Their impact on network structure and minorities. In *WebSci*, 228–238.

Ford, H.; and Wajcman, J. 2017. 'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap. *Social studies of science*, 47(4): 511–527.

García-Durán, A.; Arora, A.; and West, R. 2022. Efficient Entity Candidate Generation for Low-Resource Languages. In *LREC*, 6429–6438.

Gerlach, M.; Miller, M.; Ho, R.; Harlan, K.; and Difallah, D. 2021. Multilingual Entity Linking System for Wikipedia with a Machine-in-the-Loop Approach. In *CIKM*, 3818–3827.

Hale, S. A. 2014. Multilinguals and Wikipedia editing. In *WebSci*, 99–108.

Halfaker, A.; and Geiger, R. S. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. *CSCW*, 4: 1–37.

Halfaker, A.; and Johnson, I. 2019. The WikiTax Taxonomy. https://github.com/wikimedia/wikitax. Accessed: 2023-05-01.

Harrison, S. 2020. What happens to Scots Wikipedia now? https://slate.com/technology/2020/09/scots-wikipedia-language-american-teenager.html. Accessed: 2023-05-01.

Hasan Dalip, D.; André Gonçalves, M.; Cristo, M.; and Calado, P. 2009. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *JCDL*, 295–304.

Hecht, B.; and Gergle, D. 2010. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. In *CHI*, 291–300.

Hill, B. M.; and Shaw, A. 2013. The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation. *PloS one*, 8(6).

Hill, B. M.; and Shaw, A. 2014. Consider the Redirect: A Missing Dimension of Wikipedia Research. In *Proceedings of The International Symposium on Open Collaboration*, 28. ACM. ISBN 9781450330169.

Johnson, I. 2021. Language-Agnostic Quality. https://meta.wikimedia.org/wiki/Research:Prioritization_of_Wikipedia_Articles/Language-Agnostic_Quality. Accessed: 2023-05-01.

Johnson, I.; Gerlach, M.; and Sáez-Trumper, D. 2021. Language-Agnostic Topic Classification for Wikipedia. In *WWW (Companion)*, 594–601.

Johnson, I.; Lemmerich, F.; Sáez-Trumper, D.; West, R.; Strohmaier, M.; and Zia, L. 2021. Global gender differences in Wikipedia readership. In *ICWSM*.

Konieczny, P.; and Klein, M. 2018. Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator. *New Media & Society*, 20(12): 4608–4633.

Kummer, M. E. 2014. Spillovers in Networks of User Generated Content: Pseudo-Experimental Evidence on Wikipedia. Available at SSRN: https://ssrn.com/abstract=2567179.

Kummer, M. E. 2018. Attention in the Peer Production of User Generated Content - Evidence from 93 Pseudo-Experiments on Wikipedia. Available at SSRN: https://ssrn.com/abstract=3431249.

Langrock, I.; and González-Bailón, S. 2022. The Gender Divide in Wikipedia: Quantifying and Assessing the Impact of Two Feminist Interventions. *Journal of Communication*, 72(3): 297–321.

Laxström, N.; Giner, P.; and Thottingal, S. 2015. Content Translation: Computer assisted translation tool for Wikipedia articles. In *EAMT*, 194–197.

Leone, M.; Huber, S.; Arora, A.; García-Durán, A.; and West, R. 2022. A Critical Re-evaluation of Neural Methods for Entity Alignment. *PVLDB*, 15(8): 1712–1725.

Lewoniewski, W.; Wecel, K.; and Abramowicz, W. 2019. Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics. *Computers*, 8(3).

Lotkowski, M. 2017. Automatic Wikipedia Link Generation Based On Interlanguage Links. arXiv:1701.01858.

McMahon, C.; Johnson, I.; and Hecht, B. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. *ICWSM*, 11(1).

Miquel, M.; Gerlach, M.; and Johnson, I. 2021. Developing Metrics for Content Gaps (Knowledge Gaps Taxonomy). https://meta.wikimedia.org/wiki/Research:Developing_Metrics_for_Content_Gaps_(Knowledge_Gaps_Taxonomy). Accessed: 2023-05-01.

Miquel-Ribé, M.; and Laniado, D. 2018. Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. *Frontiers of physics*, 6.

Newman, M. E. J. 2018. *Networks: An introduction*. Oxford University Press, 2nd edition.

Ozurumba, U. 2021. Content translation tool helps create one million Wikipedia articles. https://diff.wikimedia.org/2021/11/16/content-translation-tool-helps-create-one-million-wikipedia-articles/. Accessed: 2022-12-9.

Piccardi, T.; Catasta, M.; Zia, L.; and West, R. 2018. Structuring Wikipedia Articles with Section Recommendations. In *SIGIR*, 665–674.

Piccardi, T.; Gerlach, M.; Arora, A.; and West, R. 2023. A Large-Scale Characterization of How Readers Browse Wikipedia. *ACM Trans. Web*.

Piccardi, T.; Gerlach, M.; and West, R. 2022. Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions. In *WWW (Companion)*.

Primer.ai. 2020. Project Quicksilver. https://quicksilver.primer.ai/. Accessed: 2023-05-01.

Redi, M.; Gerlach, M.; Johnson, I.; Morgan, J.; and Zia, L. 2021. A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft). arXiv:2008.12314.

Rosenbaum, P. 2017. *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press.

Sun, Z.; Zhang, Q.; Hu, W.; Wang, C.-M.; Chen, M.; Akrami, F.; and Li, C. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment*, 13: 2326 – 2340.

Tripodi, F. 2021. Ms. Categorized: Gender, notability, and inequality on Wikipedia. *New Media & Society*.

Valentim, R. V.; Comarela, G.; Park, S.; and Sáez-Trumper, D. 2021. Tracking Knowledge Propagation Across Wikipedia Languages. *ICWSM*, 15: 1046–1052.

Vincent, N.; Johnson, I.; and Hecht, B. 2018. Examining Wikipedia With a Broader Lens: Quantifying the Value of Wikipedia's Relationships with Other Large-Scale Online Communities. In *CHI*, 1–13.

Vincent, N.; Johnson, I.; Sheehan, P.; and Hecht, B. 2019. Measuring the Importance of User-Generated Content to Search Engines. *ICWSM*, 505–516.

Vitulli, M. A. 2018. Writing Women in Mathematics into Wikipedia. *Notices of the AMs*, 65(3).

Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *ICWSM*.

Wagner, C.; Graells-Garrido, E.; Garcia, D.; and Menczer, F. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science*, 5(1): 1–24.

Warncke-Wang, M.; Cosley, D.; and Riedl, J. 2013. Tell me more: an actionable quality model for Wikipedia. In *WikiSym*, 1–10.

WMF. 2003. Wikipedia:Orphan. https://en.wikipedia.org/wiki/Wikipedia:Orphan. Accessed: 2023-05-01.

WMF. 2004a. Template:Orphan. https://en.wikipedia.org/wiki/Template:Orphan. Accessed: 2023-05-01.

WMF. 2004b. Wikipedia Manual of Style (Linking). https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking. Accessed: 2023-05-01.

WMF. 2006a. Category:Orphaned articles. https://en.wikipedia.org/wiki/Category:Orphaned_articles. Accessed: 2023-05-01.

WMF. 2006b. WikiProject Council/Directory. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Directory. Accessed: 2023-05-01.

WMF. 2007a. Template:Dead end. https://en.wikipedia.org/wiki/Template:Dead_end. Accessed: 2023-05-01.

WMF. 2007b. Wikipedia:WikiProject Orphanage. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Orphanage. Accessed: 2023-05-01.

WMF. 2014. MediaWiki Content Translation Tool. https://www.mediawiki.org/wiki/Content_translation. Accessed: 2023-05-01.

WMF. 2015a. Wikimedia Downloads. https://dumps.wikimedia.org/backup-index.html. Accessed: 2023-05-01.

WMF. 2015b. Wikimedia Downloads (Wikidata). https://dumps.wikimedia.org/wikidatawiki/entities/. Accessed: 2023-05-01.

WMF. 2015c. WikiProject Women in Red. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red. Accessed: 2023-05-01.

WMF. 2016. Wikipedia ORES: Article quality. https://www.mediawiki.org/wiki/ORES#Article_quality. Accessed: 2023-05-01.

WMF. 2021. Add-a-Link Experiment Analysis. https://www.mediawiki.org/wiki/Growth/Personalized_first_day/Structured_tasks/Add_a_link/Experiment_analysis, _December_2021. Accessed: 2023-05-01.

Wulczyn, E.; West, R.; Zia, L.; and Leskovec, J. 2016. Growing Wikipedia Across Languages via Recommendation. In *WWW*, 975–985.

Yoon, J.; Park, J.; Yun, J.; and Jung, W.-S. 2022. Quantifying knowledge synchronisation in the 21st century. arXiv:2202.01466.

Zhu, K.; Walker, D.; and Muchnik, L. 2020. Content Growth and Attention Contagion in Information Networks: Addressing Information Poverty on Wikipedia. *Information Systems Research*, 31(2): 491–509.

ICWSM Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. Sections 4, 5, and 6 introduce the methods and the relative explanations.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. Please see Data and Resources (Section 3).
   (e) Did you describe the limitations of your work? Yes. Please see Section 7.3.
   (f) Did you discuss any potential negative societal impacts of your work? Yes. Please see Section 7.4. Overall, we do not foresee any significant negative impact.
   (g) Did you discuss any potential misuse of your work? Yes. Please see Section 7.4. Overall, we do not foresee any significant negative impact.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. Please see Data and Resources (Section 3) and Discussion (Section 7).
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? N/A
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? N/A
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? N/A
   (b) Did you mention the license of the assets? N/A
   (c) Did you include any new assets in the supplemental material or as a URL? N/A
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? N/A.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? N/A.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? N/A.

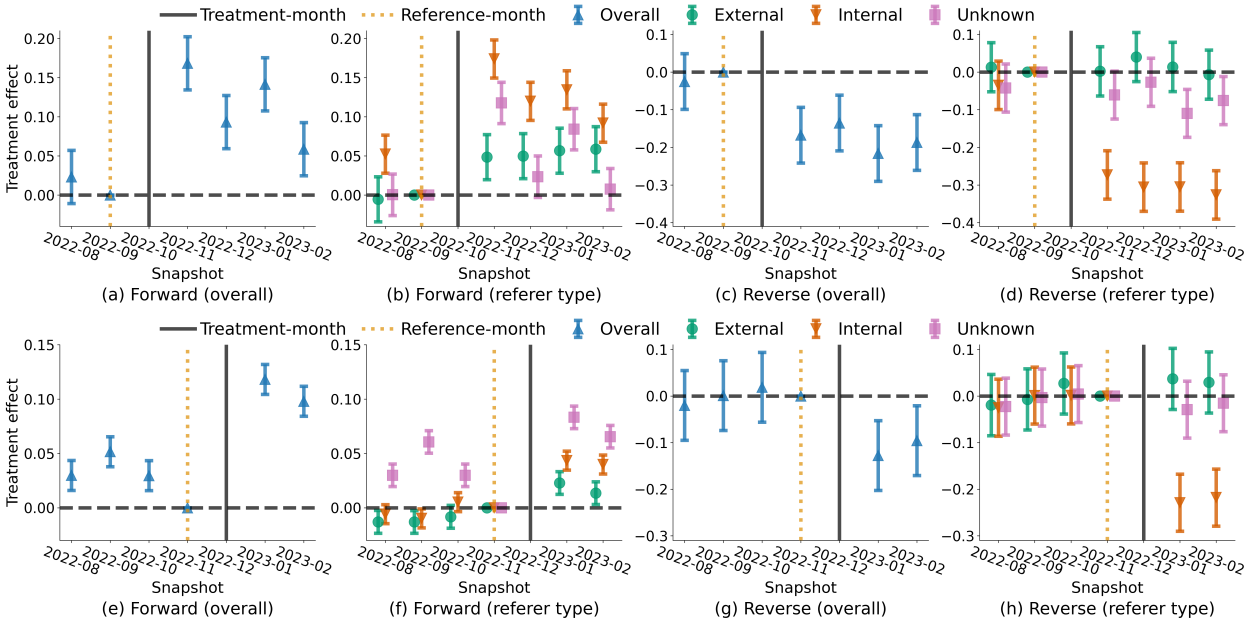


Figure S1: Per-month DiD treatment effect with 95% CIs for the forward and reverse setup considering October 2022 (top) and December 2022 (bottom) as the treatment month.

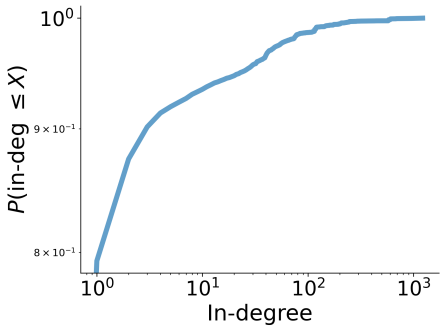


Figure S2: Cumulative distribution function of the number of added incoming links via organic de-orphanization across all Wikipedia language versions in November 2022.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset)? N/A

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? This is an observational study. Participant recruitment was not required.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A. This study did not require IRB approval.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A

(d) Did you discuss how data is stored, shared, and deidentified? Yes. Please see Data and Resources (Section 3).

A Additional Related Work

Article quality models in Wikipedia. There are several publicly available models that aim to automatically assess the quality of articles in Wikipedia such as ORES (WMF 2016). These models assess the quality based on features extracted from the content available in the specific articles. Typically they do not consider the number of incoming links (Halfaker and Geiger 2020; Lewoniewski, Wecel, and Abramowicz 2019; Warncke-Wang, Cosley, and Riedl 2013), though this has been suggested in some works (Hasan Dalip et al. 2009; Anderka and Stein 2012; De Ruvo and Santone 2015).

Reader navigation. Wikipedia's hyperlinks are crucial for readers' navigation between articles (Arora et al. 2022). While the majority of pageviews originate from an external search engine such as Google (48%), 38% of pageviews are referred from other Wikipedia articles (i.e. an internal referrer) (Piccardi et al. 2023). When considering reading sessions (i.e. combining sequentially visited articles by the same reader), most readers visit only a single article (68–72%) and thus do not use hyperlinks. However, the distribution of the number of visited articles shows a long tail with tens of millions of reading sessions consisting of 10 or more pageviews (Piccardi, Gerlach, and West 2022). Interestingly, it was found that a substantial fraction of readers use an external search engine to navigate between articles despite the availability of a corresponding hyperlink in Wikipedia. Such phenomena have been the subject of different efforts to sketch the interdependence between Wikipedia and search engines (McMahon, Johnson, and Hecht 2017; Vincent et al. 2019) as well as other online platforms more generally (Vincent, Johnson, and Hecht 2018).