9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Orphan articles: The dark matter of Wikipedia

Keywords: Wikipedia, knowledge gaps, multilinguality, visibility, quasi-experiment, network optimization, link recommendation

Introduction. With 60M articles in more than 300 language versions, Wikipedia is the largest platform for open and freely accessible knowledge of the Internet and it is continuously growing at a rate of 200K newly added articles each month. In fact, there are many efforts to systematically add content that is absent [1] through different initiatives such as organized groups and campaigns, or translating articles across languages [2]. As a result, one of the main challenges is how to maintain this ever-increasing volume of content. For example, it is crucial to properly integrate new articles into the existing network structure so that readers can find these articles through hyperlinks, which is one of the main ways to access content on Wikipedia. While the largest share of traffic to Wikipedia comes from search engines, a substantial fraction (38%) of pageviews result from traffic via internal hyperlinks [3]. In this work, we explore the question of the lack of visibility of articles by focusing on so-called orphans – articles that do not have any incoming links.

Data. We consider 305 language versions of Wikipedia from the monthly snapshots of January and February 2022. We extract the link networks among articles and match links across languages using the articles' Wikidata item ids. In addition, for each article, we extract the following features: topic, quality, time since creation, whether it was created by a bot, whether it is a disambiguation page, the gender (for biography articles). In total, we have 58M articles and 870M links.

Results. We find a surprisingly large number of orphan articles in Wikipedia: roughly 8.4M of all articles do not have any incoming links, which corresponds to 14.5% of the total of 58M articles. The extent of orphan articles varies across the different language versions (wikis): For large wikis such as English, the fraction of orphans is below 10% but in absolute terms this often corresponds to hundreds of thousands of articles. Among the 20 largest wikis (in terms of number of articles), we find the highest fraction of orphans for Egyptian Arabic (81.0%), Vietnamese (51.5%), Persian (22.3%), Arabic (20.9%), and Swedish (17.2%). For many of the smaller wikis, the fraction of orphans is consistently above 30%. For comparison, less than 0.1% of articles do not have any outgoing links showing that orphans (no incoming links) are a distinct and much more severe problem.

We find that orphans are associated with different article characteristics. While orphans are much more likely to be disambiguation pages (e.g., 50% for English) and thus might not need any improvement, we see that, for some languages, a substantial number of orphans are articles that were created by bots; e.g., Cebuano: 632K (99.6%), Swedish: 435K (86.8%), and Vietnamese: 495K (75.9%). This suggests an unintended negative consequence of automatic article creation. More importantly, the prevalence of orphans mirrors existing biases such as the gender gap. Overall, 15–20% of all biography articles are about women. Only considering orphans, the share of women biographies is much higher (e.g., around 40% for German). This means that women biographies are much more likely to be orphans than expected from their overall propensity.

We next observe that adding incoming links to orphan articles (i.e. de-orphanization) leads to a significant increase in their visibility. In a quasi-experimental setup, we consider all orphans that were de-orphanized by editors in a wiki in a given month as our treatment group. In order to rule out potential confounders (such as general increase in interest in the specific

topic), we consider a control group comprised of the *same* article in another wiki in which it remained an orphan. Considering the difference-in-difference of the number of pageviews as a proxy for the effective visiblity of each article, we find, on average, a more than 50% increase for the treatment group in comparison to the control group (cf. Fig. 1).

However, editors de-orphanized only 50K orphan articles in the month we considered—with this rate, it would take 14 years to de-orphanize all orphans (even if no new orphans were to be created). In order to support editors in this task, we propose a simple approach to surface new incoming links for articles based on link translation. This is based on the observation that, for an orphan in a given wiki w, we can often find existing incoming links for the *same* article in another wiki $w' \neq w$ (provided the source article also exists in w). The main advantage of suggesting new incoming links in this way is that we can reasonably assume that the translated links are good since they have already been vetted by editors in one or more communities. In addition, they are easily interpretable for users and contain information about where in article (e.g., which section) they would best fit. We find that this approach could, in principle, work at scale—link translation can generate suggestions for new incoming links for 4.9M (59%) orphan articles across all wikis.

Evaluating the quality of these suggestions for a small set of orphan articles that were actually de-orphanized by editors, we find that our link translation approach works well, substantially outperforming existing tools for editors or other heuristics. Interestingly, we observe that link recommendation algorithms based on sophisticated graph neural networks (such as GraphSage [4])–while showing exceptional link prediction performance on randomly chosen links—provide substantially worse recommendations for orphans. This suggests that in the regime of data sparsity (e.g. articles with few links) state-of-the-art approaches for link recommendation struggle (cf. Fig. 2) and could be significantly improved by taking into account reliable domain-specific signals such as link translation.

Discussion. Studying orphan articles in more than 300 languages in Wikipedia, we characterized the surprisingly large extent of content that is de-facto invisible for readers navigating hyperlinks in Wikipedia. We proposed a simple and interpretable approach to suggest new incoming links to orphan articles at scale based on link translation, which outperforms existing tools available to editors. This approach can improve the visibility of orphan articles, in line with previous natural experiments demonstrating a spill-over effect of attention in Wikipedia [5]. Furthermore, we believe this can help address structural biases such as the gender gap in Wikipedia. For example, visibility of biographies of women is systematically lower than for biographies on men [6]. While community-driven campaigns are successful at adding and improving the content about women, they are less successful at addressing structural biases that limit their visibility [7]. Existing link recommendation algorithms are prone to reinforcing those biases and can reduce the visibility of minorities [8].

[1] Redi et al. *arXiv preprint*, 2020. arXiv: 2008.12314

[2] Wulczyn et al. In *WWW*, 2016. doi: 10.1145/2872427.2883077

[3] Piccardi et al. *ACM Trans. Web*, 2023. doi: 10.1145/3580318

[4] Hamilton et al. In *NeurIPS*, 2017. doi: 10.5555/3294771.3294869

[5] Zhu et al. *Information Systems Research*, 2020. doi: 10.1287/isre.2019.0899

[6] Wagner et al. *EPJ Data Science*, 2016. doi: 10.1140/epjds/s13688-016-0066-4

[7] Langrock et al. *The Journal of communication*, 2022. doi: 10.1093/joc/jqac004

[8] Ferrara et al. In *WebSci*, 2022. doi: 10.1145/3501247.3531583

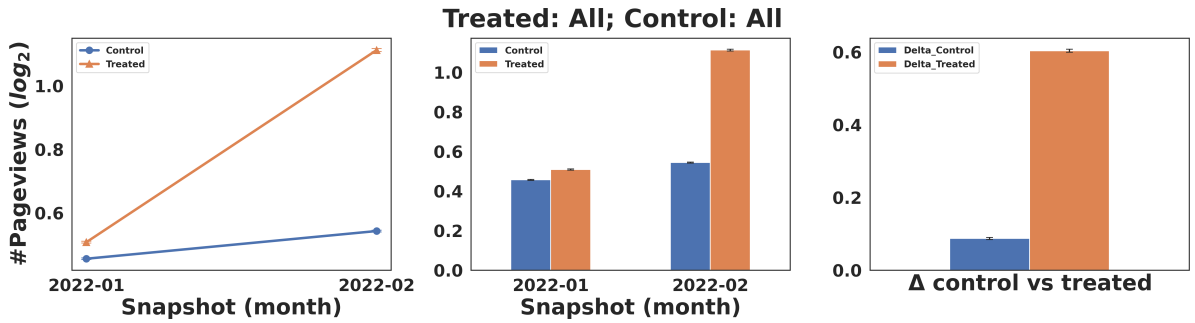

Figure 1: Differences-in-differences analysis to showcase the treatment effect of de-orphanizing an article on its visibility measured via pageviews.

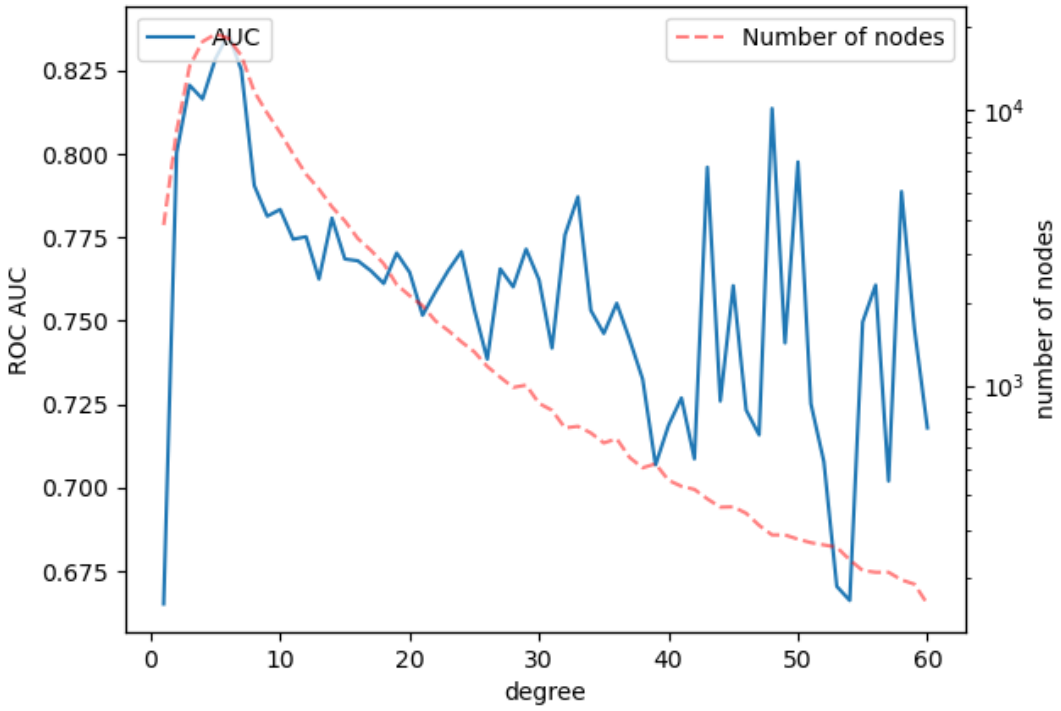

Figure 2: Analyzing the impact of node degree on the link prediction performance of GraphSage on Simple English Wikipedia.