Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Cross-Platform and Cross-Lingual Dynamics of Wikipedia Sharing and Contribution

Akira Matsui

Center for Computational Social Science, Kobe, Japan
amatsui@rieb.kobe-u.ac.jp

Abstract

Wikipedia is a key knowledge source, but the literature suggests that attention from social media platforms rarely leads to edits. This research examines this relationship using a 15-year dataset of links from three major social media platforms (Twitter, Reddit, and Discord) that contains the link-sharing history of multiple languages and edits of Wikipedia articles. We first characterize the shared articles in terms of their language co-occurrence and editing history. Then, we apply a staggered difference-in-differences design with unshared language editions as controls. The present work reveals several distinctive platform-specific patterns. Reddit favors developed articles in a Eurocentric sphere, while Twitter is more dynamic and engages earlier. Our analysis also confirms the findings of prior work, finding that the association between sharing and editing activity is not strong. However, we also observed that high-intensity sharing correlates with measurable article growth. This study thus contributes to the literature by refining the one-way street view with comprehensive characteristics using multi-platform and multi-lingual analysis and showing that attention and contribution are conditionally connected.

Introduction

Wikipedia provides a unique environment where people collaboratively create and accumulate knowledge. On this platform, editors primarily contribute as volunteers, providing the essential data that fuels modern AI, such as large language models (LLMs), constituting a primary source of high-quality, diverse text for major pre-training datasets (Gao et al. 2020; Brown et al. 2020). Beyond its role in model training, Wikipedia serves as a dominant authoritative source within the citation patterns of prominent AI answer engines (Wikimedia Foundation 2025; Profound 2025). Beyond knowledge, the Wikipedia platform also accumulates records of human behavior, with its edit histories documenting how people cooperate to construct knowledge. Therefore, understanding how to attract and maintain editorial contributions to preserve the Wikipedia ecosystem has been a central question in Wikipedia research.

Recently, the literature on this topic has turned its attention to how external platforms consume Wikipedia articles to understand how Wikipedia contributes to the internet in general (Meier 2022; Gildersleve et al. 2025; Veselovsky et al. 2024). This line of research has largely converged on the paradigm that the Wikipedia platform serves as a reliable source for social media platforms. However, Wikipedia does not receive reciprocal attention from these platforms in the form of article editing; the attention garnered on other platforms does not translate into meaningful editorial contributions (Vincent, Johnson, and Hecht 2018; Piccardi and West 2025; Gildersleve et al. 2025). This *"one-way street"* finding posits a relationship where platforms derive substantial value by linking to Wikipedia, but the attention they direct back rarely translates into meaningful editorial contributions to the Wikipedia ecosystem.

While the implications of this one-way street finding are powerful and pose a certain degree of risk to the sustainability of the ecosystem, several areas remain for investigation. Our goal in this paper is to provide a more granular characterization of this relationship by asking: under what specific, observable conditions is external attention associated with any change in editorial activity, however subtle? First, not all social media platforms are the same. The degree of this "one-wayness" can depend on the nature of the platform where article links are shared.

In addition, the literature is often constrained not only to a single platform but also to a specific language, especially the English-language Wikipedia, given its dominance on the internet. However, investigating multi-language editing behavior can generate useful implications, as shown by prior work (Piller, Zhang, and Li 2022; Samoilenko et al. 2016; Hale 2014; Kim et al. 2015; Matsui et al. 2025).

Also, the fact that an article is shared does not necessitate that it be edited. If an article is already developed with little room for further edits, it is unsurprising that a social media share would not lead to editing activity. Thus, one should characterize the maturity of Wikipedia articles shared on social media platforms, for instance, in terms of article length. Conversely, it is plausible that minor editorial changes could be correlated with social media shares. For example, a link to a Wikipedia article about a public figure might be shared on the internet in response to a breaking news event. While existing scholarship provides a foundation, it is often constrained by data limitations. Specifically, Vincent, Johnson, and Hecht (2018) focus exclusively on English Wikipedia,

which underscores the need for larger-scale analyses to account for the confounding factors inherent in such restricted samples. This poses a key challenge: disentangling the statistical association of a share from the influence of the underlying real-world event that motivates both the share and the need for a Wikipedia update.

To address this challenge, we employ a staggered difference-in-differences (DiD) framework. It is crucial to note that our primary aim is not to establish a strict causal relationship but rather to use the DiD design as a robust method for controlling for unobserved confounders. By comparing a shared article to its unshared counterparts in other languages, our approach allows us to mitigate the influence of external events that would likely affect all language editions simultaneously. This helps us to more accurately characterize the specific changes in editing activity that are temporally associated with the sharing event itself. For this, we utilize the multilingual aspect of the Wikipedia platform to detect differences for the same entity or topic across languages (Kim et al. 2015; Hale 2014; Park et al. 2015). Using this constructed multi-platform and multilingual dataset, we investigate the following three central research questions.

- **RQ1: Cross-Lingual Attention Across Platforms**
  How do patterns of Wikipedia sharing differ across platforms and language editions?
- **RQ2: Characteristics of Shared Articles**
  What are the intrinsic characteristics of articles shared on different platforms, particularly concerning their maturity within the edit lifecycle?
- **RQ3: The Attention-Contribution Link**
  How do platform, share intensity, and timing affect the link between social media sharing and subsequent editorial activity?

Findings Our findings clarify the aforementioned one-way street hypothesis by utilizing a robust method and comprehensive data, filling gaps in the literature. The present work confirms prior research showing that the average effect of a share on editorial activity is negligible. However, we reveal a subtle but statistically significant positive association between sharing and article growth, conditional on the platform and the intensity of attention (RQ3). In addition, our analysis suggests that different platforms exhibit distinct Wikipedia article-sharing behaviors in terms of language co-occurrence (RQ1) and the maturity of articles (RQ2). This study refines our understanding of the digital ecosystem, moving from a monolithic model to a more nuanced, conditional account of when and how attention and contribution are intertwined.

Contributions This paper makes three primary contributions. First, we construct and analyze a large-scale, fifteen-year longitudinal dataset tracking Wikipedia links shared on Twitter, Reddit, and Discord to facilitate supplementary analyses across numerous languages. Second, using this unique dataset, we provide the first comparative, multi-platform, and multilingual analysis of the relationship between social media sharing and Wikipedia editing, characterizing distinct cross-platform behaviors and filling a key gap in the literature. Third, we advance the methodology for studying this phenomenon by employing a DiD framework that compares shared articles with their counterparts in other languages. This approach serves to control for confounding real-world events, thereby yielding a more robust estimate of how external attention relates to content growth on Wikipedia and offering a generalizable framework for studying online knowledge systems.

Related Research

This section discusses the background of the presented study, focusing mainly on research related to the Wikipedia platform (Piccardi and West 2025). Thanks to its nature, the Wikipedia platform is a unique arena for understanding human behavior in cooperative environments where the "wisdom of the crowd" operates (Arazy, Morgan, and Patterson 2006; Kittur et al. 2007). In particular, this paper focuses on the editing history of Wikipedia articles to understand their links to social media platforms. While the literature reveals important aspects of human behavior from article-viewing data (Zhou et al. 2024; Piccardi et al. 2024), mining the revision history of articles can contribute to our understanding of article quality (Raman et al. 2020; Shenoy et al. 2021).

Understanding Editing Behavior Although the present study does not thoroughly cover the discussion among editors, the literature on this interaction reveals that such communication plays a pivotal role in maintaining the quality of articles (Viegas et al. 2007; Bryant, Forte, and Bruckman 2005; Yang et al. 2016). Therefore, researchers have proposed user modeling to capture the essential nature of editing behavior, for example, by using network models (Adler and De Alfaro 2007; Flöck and Acosta 2014; Javanmardi, Lopes, and Baldi 2010) or opinion dynamics (Ciampaglia, Flammini, and Menczer 2015; Tasnim Huq and Ciampaglia 2021). Similarly, the literature contributes to understanding the factors associated with editors' engagement (Halfaker, Kittur, and Riedl 2011; Halfaker et al. 2013) and their roles on the platform (Kittur et al. 2007). Importantly, research sheds light on the challenges the Wikipedia platform faces regarding its sustainability. For example, Umarova and Mustafaraj (2019) recently detected partisan relationships in Wikipedia articles related to news media. Also, Matsui, Miyazaki, and Murayama (2024) point out the socioeconomic disparities in editing politicians' Wikipedia pages.

Multilingualism The literature also points out the importance of studying the multilingual aspect of the Wikipedia platform because the structure of knowledge in a given language depends on its linguistic characteristics associated with cultural or geopolitical contexts (Kramsch 2014; Li, Haider, and Callison-Burch 2024; Matsui et al. 2025). It also suggests that the interaction of multilingual editing demonstrates the importance of cross-language connections within Wikipedia (Hale 2014; Kim et al. 2015). Samoilenko et al. (2016) also reveal the cross-lingual and cross-cultural aspects spanning languages by studying the co-occurrence of knowledge among language pairs.

Article Consumption Outside of Wikipedia Not only within the Wikipedia platforms, research on Wikipedia also examines the consumption patterns of its articles from the internet. Modeling such consumption behavior can depict how the platform contributes to our society as a knowledge provider. There are multiple studies that examine article consumption data on social media platforms (Meier 2022; Gildersleve et al. 2025). Recently, Veselovsky et al. (2024) investigated links across the World Wide Web, collecting more than 90 million links. They demonstrate that most of them appear on news and science websites, usually for "explanatory references rather than evidence."

Relation to the Presented Work Based on the findings discussed in this section, this paper contributes to the literature by analyzing the one-way street hypothesis suggested by Vincent, Johnson, and Hecht (2018) from multiple perspectives. First, to address the literature's suggestion that multilingual analysis is necessary, we study the language co-occurrence of Wikipedia article sharing (RQ1). We then study the maturity of Wikipedia articles when they are shared. Prior research suggests that the primary motivation for sharing Wikipedia articles is for reference (Veselovsky et al. 2024). Then, we compare what kind of articles are shared on social media platforms and study the differences among platforms (RQ2). Lastly, our paper employs robust identification strategies to capture editing behavior around article sharing by leveraging the multilingual features of our dataset (RQ3).

Data & Methods

Data

In this paper, we construct a dataset covering multiple social media platforms to capture the multi-faceted aspects of Wikipedia article-sharing behavior. We use two open datasets from Reddit and Twitter. Additionally, we analyze a dataset of online chat channels on Discord to extract posts containing Wikipedia article links.

For Reddit data, we use the WikiReddit dataset from Gildersleve et al. (2025). The dataset contains a comprehensive collection of Wikipedia links shared on Reddit over a multi-year period (2020–2024). That paper analyzes the effect of article sharing on Reddit on editing behavior, reporting a negative or negligible effect on the number of edits. We also use the TWikiL dataset (Meier 2022), which collects tweets with Wikipedia article links. Although Twitter has since been rebranded as X, the platform was named Twitter at the time of data collection. Therefore, in this paper, we refer to the platform as Twitter and use the term "tweet" instead of "post." The third dataset, from Aquino et al. (2025), was not originally published for Wikipedia research; we use this Discord open server dataset to explore article-sharing behavior on a chatting platform, which has distinct characteristics from the first two. In summary, our dataset comprises data from Reddit (discussion) and Twitter (micro-blogging), with Discord (chatting) included as a supplementary analysis in the Appendix.

Using the Wikipedia article links from these three platforms, we retrieve the complete editing history of the corresponding articles via the Wikimedia API. For this stage, we also use the Wikimedia Pageviews API to retrieve view count data (MediaWikiAPI 2022). Importantly, we collect all language versions of an article using its Wikidata ID for our analysis. For example, if a user posts a link to an English Wikipedia article, we retrieve the complete editing histories of all available language versions of that article. This comprehensive, multilingual data is particularly important for addressing RQ3.

We then combine the datasets and identify the sharing of a Wikipedia article link on any of the three platforms as the primary "treatment" event. We summarize this data at a daily resolution. Since the same Wikipedia page can be shared multiple times, we record the date and platform for each sharing event for every unique page. This temporal information is essential for our analyses in RQ2 and RQ3, and we provide further details when reporting the corresponding results.

Method

We next describe the methods to address the three research questions proposed in the Introduction.

RQ1 and RQ2 For RQ1 and RQ2, we employ a simple method to describe Wikipedia article-sharing behavior by aligning it with the editing history of the shared articles. RQ1 studies cross-lingual co-occurrence (i.e., instances where the same article is shared across different language editions within the same platform) to understand the multilingual aspect of each platform. Our analysis first calculates the proportion (i.e., share) of co-occurrences at each time point. By computing and comparing this share for each platform, we analyze how the co-occurrence rate changes over the course of our data collection period. However, this co-occurrence analysis is not performed for the Discord dataset, as links to the English Wikipedia constitute the vast majority of its data. For RQ2, we characterize the identified shared articles. We first examine each article's revision history to determine at what relative point the sharing event occurred. This normalization allows us to compare different articles using a uniform standard.

RQ3 To investigate the changes in a Wikipedia article after a sharing event on social media, we employ a DiD framework. The DiD framework is a quasi-experimental method often used to estimate the impact of an intervention. While the strength of DiD lies in controlling for time-invariant unit characteristics and time trends common to all units, our matching of shared articles to their unshared counterparts in other languages addresses the concern about unobserved, time-varying confounders (i.e., the influence of underlying real-world events). In this study, we adapt this framework not to claim causality, but to leverage this robust design.

To quantify the magnitude of the association, we use this framework to address a key methodological challenge: the simultaneity bias that arises when an external event both motivates an article share and necessitates an update to the article itself. For example, when a prominent researcher wins a Nobel Prize, their Wikipedia article is likely shared on social media, while concurrently, editors update the article

to include this new information.[1] To disentangle these effects, our design leverages the unique multilingual structure of Wikipedia. We use other language editions of the same article as a control group, which allows us to characterize the changes associated with the share while controlling for edits driven by the underlying exogenous event. Specifically, we identify unshared counterparts of the same article in other language editions that were not shared on the platform during the observation window. This ensures that the control group represents the same conceptual entity but without the specific social media exposure in the dataset. By comparing the shared article against this linguistically diverse control set, we can distinguish platform-specific attention effects from global interest spikes that would affect all language versions simultaneously. This approach is powerful because a major real-world event would likely spur edits across multiple language editions, not just the one shared on a specific platform.

We analyze the relationship from two perspectives: Share Order and Intensity. Share Order distinguishes between the first and subsequent shares of an article on different days. Intensity refers to the number of times an article is shared on the same day, which we use as a proxy for a convergence of attention.

To estimate the effects, we use an event study model. The model is adapted to incorporate intensity as a measure of treatment strength, an approach used in other contexts of online platform events.

The intensity-adjusted model is specified as:

$$Y_{ilt} = \alpha_{il} + \gamma_t + \sum_{k=-K}^{L} \beta_k \cdot D_{ilt}^k \cdot \text{Intensity}_{il} + \epsilon_{ilt} \quad (1)$$

For our standard model, the intensity term is omitted:

$$Y_{ilt} = \alpha_{il} + \gamma_t + \sum_{k=-K}^{L} \beta_k \cdot D_{ilt}^k + \epsilon_{ilt} \quad (2)$$

Where Y_{ilt} is the outcome (article size in bytes) for article i in language l during time point t; α_{il} and γ_t are article-language and time point fixed effects, respectively. These control for time-invariant article characteristics and any time-based trends common to all articles; D_{ilt}^k is an indicator variable equal to 1 if time point t is k periods relative to the initial sharing event; Intensity_{il} is the number of same-day shares for article i in language l; ϵ_{ilt} is the error term.

Our parameter of interest is β_k, which captures the average difference in the trajectory of article size following a sharing event, relative to the trajectory of unshared control articles. Standard errors are clustered at the article level to account for potential serial correlation. We employ two temporal resolutions in our analysis: daily and weekly. These different time scales allow us to examine the research question from complementary perspectives. Our primary analysis operates at the daily level, capturing changes in article size in the days immediately before and after a share event. This fine-grained approach enables us to detect how much editing activity increases following a share. Moreover, by estimating changes in the days immediately preceding the share event, we can assess whether the treatment and control groups exhibit parallel trends. In addition, we conduct a weekly-level analysis to capture broader patterns of change as a supplementary analysis. This analysis focuses on reporting $\beta_{k=\text{week }0}$, which represents the average difference in article size during the seven days including and following the share event, compared to the preceding seven days. We apply this weekly analysis specifically to estimate effects when articles experience multiple share events in the dataset.

[1] To aid conceptual understanding, we present an illustrative example of article size differences following the 2020 Nobel Prizes in Figure 9 in the Appendix. Note that this figure is provided solely for illustrative purposes to clarify the underlying mechanism, not as part of our formal analysis.

Results

RQ1: Cross-Lingual Attention Across Platforms

To investigate the one-way street hypothesis in different linguistic contexts, we first map the baseline patterns of cross-lingual information sharing, revealing the linguistic and cultural borders that shape how Wikipedia articles are shared across platforms. Figures 1 and 2 present the co-occurrence patterns for the top 10 most frequently shared language co-occurrences in each platform. This restriction yields a sufficient sample size to understand the differences in co-occurrence that we observe over 288,000 occurrences on Reddit and 5 million occurrences on Twitter. On the other hand, we did not find a meaningful number of co-occurrences on Discord and therefore focused only on the other two in this analysis. Additionally, we present the overall time-series trends of shares by language in Figure 10 in the Appendix.

Our analysis reveals that co-occurrence is primarily a short-term phenomenon. Varying the time window from one to three or seven days did not significantly alter the rankings, indicating that most co-sharing events happen within a 24-hour period, likely as reactions to the same immediate real-world event.

More importantly, the patterns reveal distinct linguistic ecosystems on each platform. On Reddit, the landscape is dominated by English paired with other major European languages, with German-English being the most frequent pair (Figure 1). This suggests Reddit functions as a hub for English-speaking communities that frequently reference content from other European languages, reflecting established patterns of knowledge exchange. These co-occurrence rankings have also remained remarkably stable over time.

In stark contrast, Twitter exhibits a more dynamic linguistic structure. While high-frequency pairs like English-French and English-Spanish are common, Twitter shows a significantly higher prevalence of co-occurrences involving Japanese Wikipedia, such as English-Japanese. Furthermore, the share of certain pairs, like English-French, has trended upwards, suggesting that Twitter's linguistic dynamics are more fluid than Reddit's stable hierarchy.

The analysis of lower-ranked pairs (Figure 2) reinforces this divergence. The platforms share few common pairs in this tier, but within each platform, the remaining, less fre-

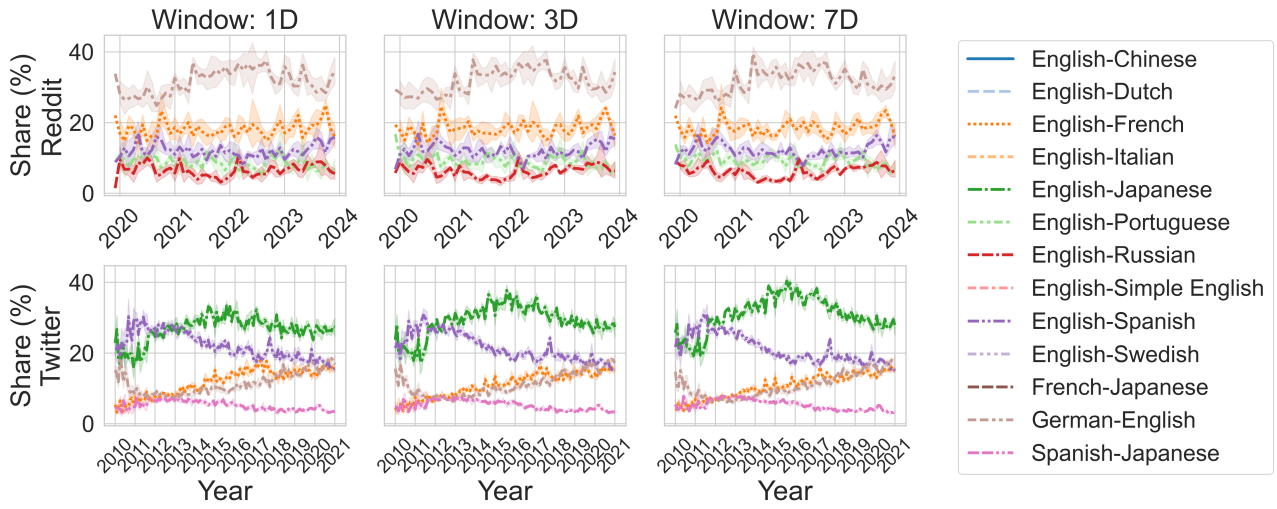


Figure 1: Language co-occurrence rankings for the top 5 most frequently shared Wikipedia language editions across social media platforms. The figure shows the frequency of articles on the same conceptual topic being shared across different language Wikipedias. The legend is shared with Figure 2 for comparison. Shaded areas represent 95% CIs.

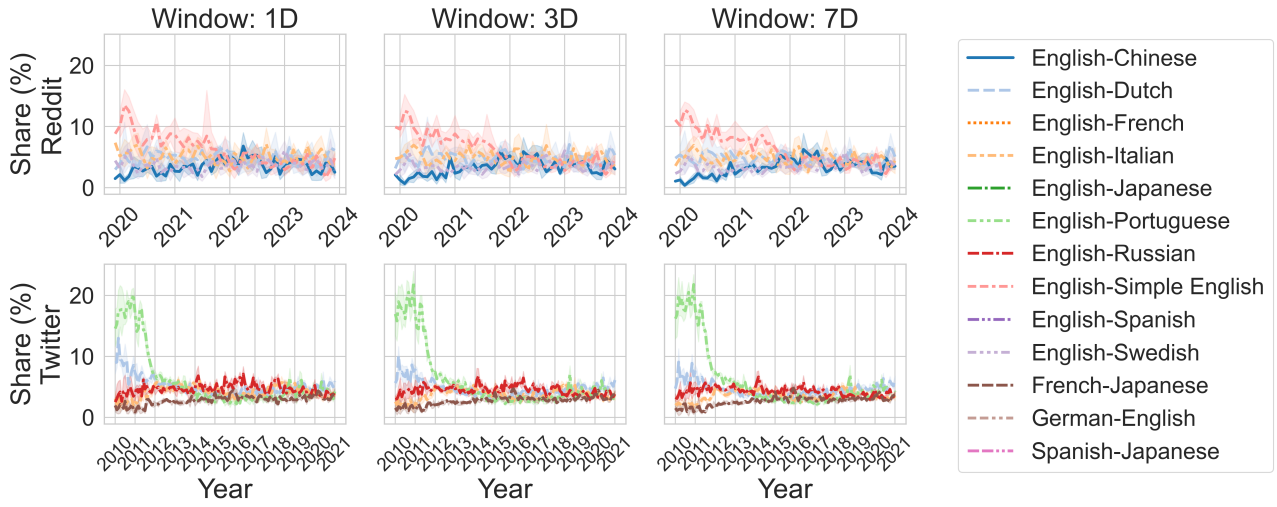


Figure 2: Language co-occurrence rankings for Wikipedia language editions, ranked 6th–10th by sharing frequency. This extends the analysis to show secondary language clusters and their cross-platform sharing patterns. The legend is shared with Figure 1 for comparison. Shaded areas represent 95% CIs.

quent pairs hold similar shares. This implies that while a few dominant language pairs form a clear hierarchy, a collection of secondary languages co-occur at a comparable, lower frequency. In summary, the landscape of cross-lingual information sharing is not uniform but highly platform-dependent. Reddit reflects a more traditional, Eurocentric linguistic sphere, while Twitter hosts a more multipolar environment where non-European languages like Japanese play a significant role. This diversity challenges a monolithic view of the literature, indicating that the contrast between Reddit's Eurocentric focus and Twitter's multipolar nature offers distinct avenues for future research.

Having established these distinct sharing patterns, we now turn to the articles themselves. We next examine the characteristics of the articles shared within these ecosystems (RQ2) before investigating whether this attention translates into the editorial contributions that sustain Wikipedia (RQ3).

RQ2: Characteristics of Shared Articles

Building on prior research showing that social media links to Wikipedia serve a referential function, we ask at what stage of an article's development it is shared. To this aim, we investigate whether users share articles that are almost complete or those that are still evolving, which clarifies whether Wikipedia is used as a static reference or a dynamic source.

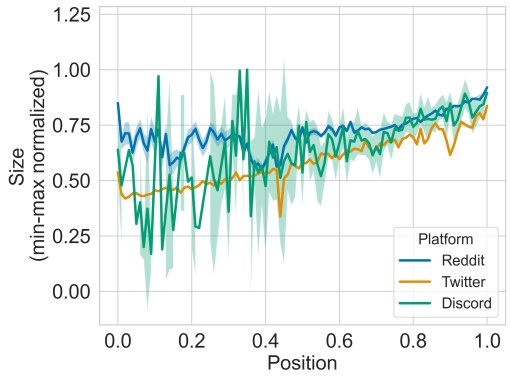


Figure 3: Trajectory of average article size (normalized per article) for articles shared on social media platforms over time, indicating whether sharing behavior favors longer, more comprehensive articles during specific periods. Shaded areas represent 95% CIs.

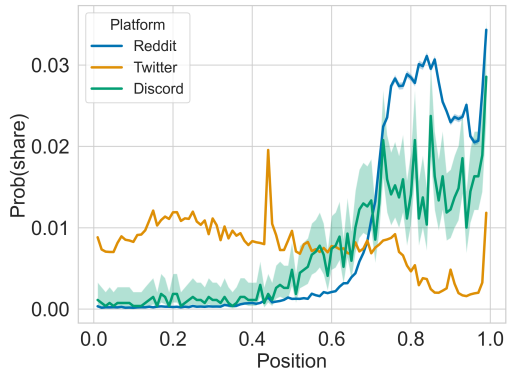


Figure 4: Overall trajectory of sharing probability over time, showing how the likelihood of Wikipedia articles being shared on social media platforms changes across different time periods. Shaded areas represent 95% CIs.

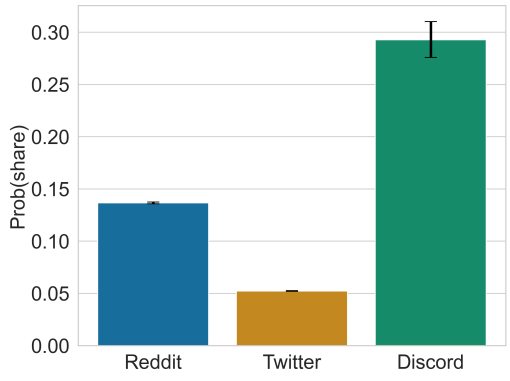


Figure 5: The ratio of articles that have no edits at or after the time of a sharing event. The bars represent 95% CIs.

Because an article's "maturity" is difficult to quantify directly, we use a proxy: the relative "position" of a sharing event within an article's edit history, normalized from its first edit (0) to its most recent edit (1). To validate this proxy, we first analyze the trajectory of article size (Figure 3). Article size tends to fluctuate in the earlier stages of the edit history before showing a clearer upward trend in the later stages. In this analysis, we observe over 300,000 Wikidata IDs in 198 languages on Twitter, over 72,000 Wikidata IDs in 62 languages on Reddit, and over 2,260 Wikidata IDs in 14 languages on Discord.

Our analysis of sharing timings reveals distinct patterns across platforms (Figure 4). On Twitter, articles are shared at a relatively high rate during the earlier stages of their edit history. Conversely, on Reddit and Discord, articles are shared more frequently during the later stages of their edit history. This difference likely reflects the distinct use cases of each platform: on Twitter, users share articles earlier in their lifecycle, which often continue to receive substantial revisions after sharing, whereas Reddit and Discord users tend to share articles that are further along in their edit history. The localized spike near position 0.4 in the Twitter trajectory appears to reflect a binning-related irregularity arising from articles with few total edits, whose discrete normalized positions concentrate at specific points within the interval; similar patterns are visible in both panels of Figure 6.

A common trend across all platforms, however, is that sharing frequency increases as articles develop. Focusing on the proportion of articles shared at or after their last edit reveals significant differences across platforms, with Twitter showing the lowest value, followed by Reddit and Discord (Figure 5) ($p < 0.01$). We also confirm that these platform-specific patterns are not largely driven by language differences. Figure 6 shows no significant differences in sharing position trends among the dominant language editions. This finding suggests that platform-specific norms, rather than language characteristics, primarily drive sharing behavior.

RQ3: The Attention-Contribution Link

We employ a DiD framework to analyze whether an observable change occurs in a Wikipedia article after it is shared on social media. While prior studies found that sharing did not increase the number of editors (Vincent, Johnson, and Hecht 2018; Gildersleve et al. 2025), they did not investigate whether the article's content itself changed (Vincent, Johnson, and Hecht 2018; Gildersleve et al. 2025). We therefore measure the extent to which information is added by estimating the change in article size (in bytes).

It is important to note that this analysis is not intended to establish a causal effect. As discussed in the Introduction, our goal is to characterize the changes in Wikipedia articles that are shared on social media, particularly because prior work suggests the effect of such shares on editing activity is small.

To this end, we use the DiD framework to analyze the magnitude of change in article size following a social media share. In this framework, we compare articles that were shared on social media with a matched control group of unshared articles from the same period. More specifically, the control group consists of unshared counterparts of the

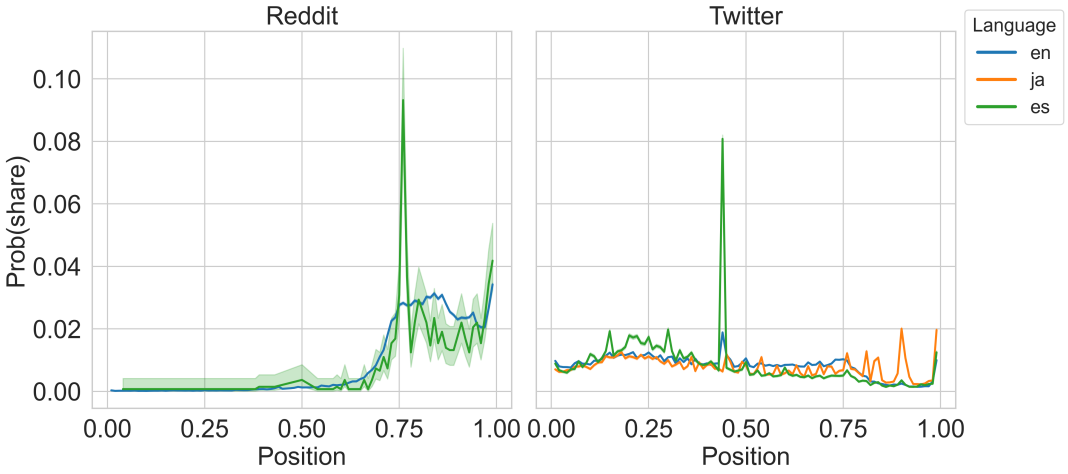


Figure 6: Sharing probability trajectories broken down by language edition demonstrate distinct temporal patterns across social media platforms. Trajectories for the same Wikipedia language edition vary between Reddit and Twitter. Shaded areas represent 95% CIs. A complementary analysis of Discord data is provided in Figure 11 in the Appendix.

same article in other language editions. This approach allows us to account for endogenous events that might affect both Wikipedia article sharing and editing, and to isolate the change associated with the share.

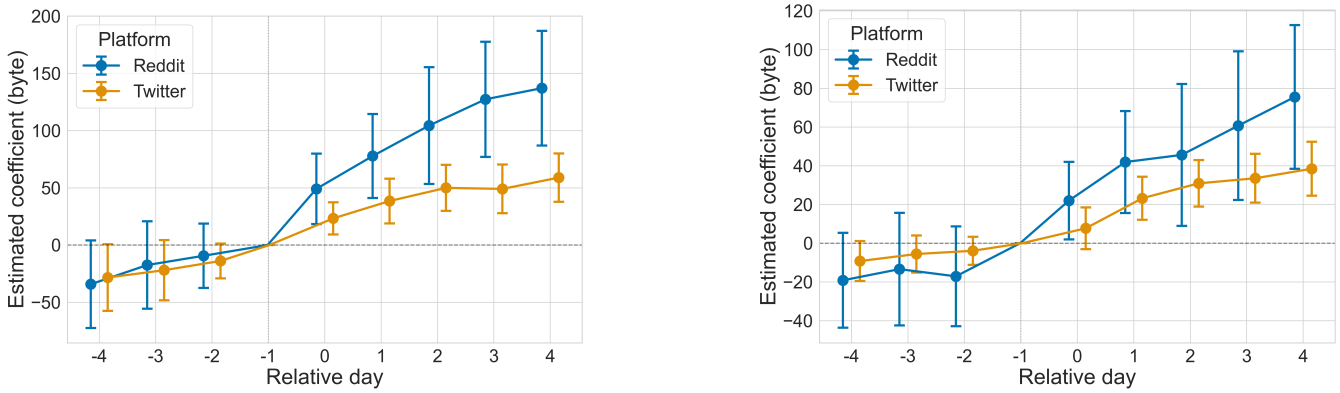
We also examine this relationship from a perspective of intensity. Intensity refers to the number of times an article is shared on the same day, indicating a convergence of attention. We adapt the DiD framework to incorporate intensity by using the number of same-day shares as a measure of treatment strength, an approach used in other contexts of online platform analysis (Zeng, Danaher, and Smith 2022).

In addition, this analysis focuses on the top seven language editions by frequency to ensure a valid control group; including inactive editions would inflate the estimated differences due to their lack of regular editing activity. These languages are German, English, Spanish, French, Russian, Japanese, and Italian. The dataset used for the estimation contains over 280,000 articles on Reddit, 200,000 on Twitter, and 12,000 on Discord. As with the other analyses, we focus on the Reddit and Twitter datasets, reporting the Discord results to complement the main findings.

Figure 7 presents the results of the daily-level DiD analysis. We use the day before the share event (day $t = -1$) as the reference period. In the pre-treatment period (days $t < 0$), the 95% confidence intervals include zero, indicating no statistically significant difference ($p > 0.05$) relative to the reference. This result supports the parallel trends assumption between the treatment and control groups. Note that our outcome variable is the article size in bytes, representing the cumulative stock of content, rather than the size of marginal edits. Consequently, any divergence that emerges on day $t = 0$ persists and is incorporated into the differences observed on subsequent days (e.g., day $t = 1$).

In contrast, we observe statistically significant differences from the share day onwards, indicating an average increase in article size. Specifically, the estimated magnitude of this increase ranges from approximately 50 to 200 bytes in our datasets, which corresponds to roughly one or two sentences of text. A comparison between the Standard DiD model and the Intensity-adjusted DiD model shows that the coefficient estimates are smaller in the latter. This difference reflects the positive association between share intensity represented by the number of same-day shares on the share day and the magnitude of the change in article size. Furthermore, adjusting for intensity refines our interpretation of platform-specific differences. While the point estimates from the Standard DiD model suggest that Reddit is associated with larger editorial changes than Twitter, particularly after day 3, this gap narrows in the Intensity-adjusted model. These findings suggest that accounting for the magnitude of attention is crucial for characterizing the connections between article share events on social media platforms and article editing on Wikipedia. In addition, we estimate the same models using the Discord dataset and present them in Figure 12 in the Appendix. We observe a statistically significant difference from the control group at day $t = -4$ in both models, which casts some doubt on the validity of the parallel trends assumption for this platform. Therefore, while these results should be interpreted with caution, they are broadly consistent with our main findings: the group of shared Wikipedia articles exhibits a greater increase in size compared to the unshared group.

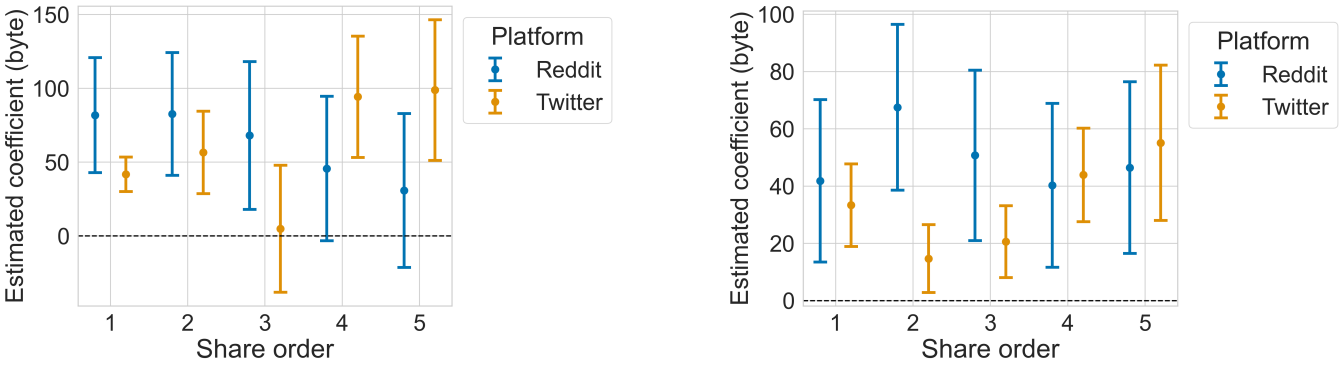
A natural follow-up question concerns the dynamics of repeated exposure: how does the association with article size change when a Wikipedia article is shared on multiple occasions across different days? Specifically, we ask whether the relationship between sharing and editing becomes stronger or weaker with repetition. One hypothesis posits that the magnitude of the effect diminishes as the number of shares accumulates. Alternatively, repeated mentions might cumulatively build user interest, thereby strengthening the association with each successive share. To study this, we define "Share Order," which categorizes share events by sequence, distinguishing the initial share from subsequent ones.



(a) Standard DiD model.

(b) Intensity-adjusted DiD model.

Figure 7: Daily-level difference-in-differences analysis, focusing on the first share in the datasets. Day 0 represents the share day, and day -1 serves as the reference period. Figure 7a presents the coefficients from the standard DiD model (Equation 2), while Figure 7b shows results from the intensity-adjusted model (Equation 1). These plots display the estimated difference in article size (in bytes) for each day relative to the share day, comparing shared articles to their unshared counterparts in other language editions. The bars represent 95% CIs.



(a) Standard DiD model.

(b) Intensity-adjusted DiD model.

Figure 8: Difference-in-differences event study plots by share order. When the same article was shared on multiple occasions, we estimated the model separately for each share event. Figure 8a presents the coefficients from the standard DiD model, while Figure 8b shows results from the intensity-adjusted model. These plots display the estimated change in article size (in bytes) in the weeks before and after sharing, comparing shared articles to their unshared counterparts in other language editions. The bars represent 95% CIs.

Figure 8 illustrates the results from our DiD framework. The results for Reddit and Twitter show statistically significant associations. The coefficient estimates are significant ($p < 0.02$) in both the standard and intensity-adjusted models. However, the point estimates are small, aligning with prior research indicating that the impact of sharing on an article's content is not large. Based on the point estimates, sharing on Reddit is associated with a larger increase in article size for the first three share orders. For subsequent shares, this difference diminishes or reverses. Furthermore, the intensity-adjusted analysis reveals that the association is significant for the first share order across all three platforms. For later shares, the association remains significant only for Reddit and Twitter. This finding indicates a correlation between the number of times a link is shared on the same day and the change in the article's size; a higher number of same-day shares is associated with a larger increase in size. On the other hand, the results of the Discord dataset presented in Figure 13 in the Appendix depict large standard errors; we therefore exclude them from this main plot for clarity.

Discussion and Conclusion

Our study examined the relationship between the sharing of Wikipedia articles on multiple social media platforms and subsequent editing activity. We revealed that while the general premise of minimal inflow from these platforms holds, the relationship is more nuanced and conditional than previously understood. Our analyses for RQ1 and RQ2 suggested that sharing patterns and the characteristics of the shared articles are highly heterogeneous across platforms.

Reddit, with its focus on relatively developed articles, serves as an English-centric hub with other Eurocentric linguistic clusters. On the other hand, we find that Twitter, with its more fluid and multipolar linguistic dynamics, serves distinct functions in the knowledge ecosystem. This nature makes Twitter a crucial field for studying non-Western knowledge dynamics, such as those in the Japanese community, in contrast to the Western-centric environment of Reddit. Furthermore, these findings point to a potential functional division of labor within the web knowledge ecosystem that warrants future investigation: Twitter may facilitate real-time debate on developing topics, whereas Reddit could serve as a reference hub for established ones, with Wikipedia functioning as the central archive consolidating this information.

Building on these findings, our analysis using the DiD framework for RQ3 confirmed that sharing on social media is associated with a subtle but statistically significant increase in article size, after controlling for the confounding effects of the external events that prompted the shares. The daily-level DiD analysis demonstrated that this increase emerges immediately on the day of sharing and persists thereafter. This finding largely corroborates previous results suggesting a negligible effect of sharing events on Wikipedia editing behavior, showing that external attention, on average, does not translate into tangible changes (Vincent, Johnson, and Hecht 2018; Piccardi and West 2025; Gildersleve et al. 2025). However, our analysis also revealed that this relationship is platform-dependent. Our estimates suggest that the change in article size after a sharing event is positively associated with intensity, measured by the number of shares on the share day. Furthermore, this association is correlated with share order, which represents the number of different days an article was shared. This implies that for Twitter, repeated sharing across different days is associated with a larger increase in article size. In contrast, the point estimates for Reddit suggest roughly the opposite effect.

This statistically significant yet limited effect size identified in this study resolves the debate surrounding the "one-way street hypothesis." Previous research struggled to distinguish whether the lack of observed editing was due to insufficient statistical power or the actual absence of the phenomenon. Our analysis of a large-scale dataset demonstrates that this relationship is not absent, but not strong. This implies that while social media platforms direct attention to Wikipedia, the pathway to editorial contribution remains narrow.

While this study utilizes a rigorous method, it has limitations. Our reliance on observational data means we cannot entirely rule out unobserved confounders. As discussed, we did not aim to establish a causal relationship between article sharing and editing behavior; however, future research should aim to establish causal links more definitively. Another promising research direction could be to analyze specific links to understand intentional sharing behaviors, such as coordinated information manipulation.

Our dataset does not have equal sample sizes for the three platforms. In particular, the Discord dataset covers a smaller set of Wikipedia articles than those for Reddit and Twitter, resulting in large standard errors that make it difficult to draw firm conclusions for that platform. Also, we use article size to measure information growth, rather than analyzing the specific text differences between versions. We rely on this metric because comparing text changes across different languages is technically challenging. In addition, assessing the meaning of each edit on a large scale is difficult, as a change could range from a simple grammar correction to a major addition of facts. Therefore, while article size is a useful proxy for content accumulation, our analysis does not capture the qualitative nature of the editorial changes. A promising future direction would be a qualitative analysis to capture these differences at the time of sharing events and to report how increased attention from social media platforms alters Wikipedia content. The methodology presented in this study, which utilizes cross-lingual comparisons to control for external shocks, extends to other domains. For instance, future research could apply this framework to examine how social media exposure or referencing by AI influences contributions to open-source software or global movie and product reviews. Such work can thereby generalize the link between attention and contribution across the web.

Despite these limitations, the central finding of a narrow path carries important implications. Rather than viewing this merely as a negative result that editing does not increase, we can interpret it as evidence that the consumption (viewing) and supply (editing) of information are functionally but loosely connected. The fact that social media attention is not immediately associated with major article revisions may serve as a protective barrier, maintaining the stability and neutrality of information. If external attention were easily converted into editing activity, the risk of vandalism or biased contributions could increase. Future research should investigate whether this limited relationship effectively serves to preserve the stability of the encyclopedia.

This structural resistance suggests that sporadic viewing rarely translates into editorial action. This aligns with prior research indicating that typical Wikipedia editors are driven by intrinsic motivation (Nov 2007) and that frequent contributors often start editing early in their engagement (Panciera, Halfaker, and Terveen 2009). Taken together with our findings, this implies that simply increasing social media exposure is insufficient to accelerate the recruitment of new editors or the update of information. This is consistent with the observation that edits related to breaking news events are often performed by experienced editors (Keegan, Gergle, and Contractor 2013). However, since viewing is a first step toward becoming an editor, such exposure on social media may still contribute to the long-term growth of the editor population (Antin and Cheshire 2010).

Ethical Statement

This research was conducted in adherence with the highest ethical standards. All data used is from publicly available datasets and APIs, originally collected with user consent where applicable. Our analysis is performed on aggregated, anonymized data at the article level, and no personally identifiable information of social media users or Wikipedia editors is used or stored. We confirm that all text in this paper

was written by the authors, but we use Gemini for grammar and spelling checks and to improve the clarity of the author-written text. The content and intellectual contributions remain entirely those of the human author.

Acknowledgements

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research Number 24K16359. Also, This work was supported by Research Institute of Science and Technology for Society, Japan, Grant Number JPMJRS23L4.

References

Adler, B. T.; and De Alfaro, L. 2007. A content-driven reputation system for the Wikipedia. In *Proceedings of the World Wide Web Conference*.

Antin, J.; and Cheshire, C. 2010. Readers are Not Free-Riders: Reading as a Form of Participation on Wikipedia. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*.

Aquino, Y.; Bento, P.; Buzelin, A.; Dayrell, L.; Malaquias, S.; Santana, C.; Estanislau, V.; Dutenhefner, P.; Evangelista, G. H.; Porfírio, L. G.; et al. 2025. Discord Unveiled: A Comprehensive Dataset of Public Communication (2015-2024). *arXiv preprint arXiv:2502.00627*. Dataset https://huggingface.co/datasets/SaisExperiments/Discord-Unveiled-Compressed.

Arazy, O.; Morgan, W.; and Patterson, R. 2006. Wisdom of the crowds: Decentralized knowledge construction in Wikipedia. In *Annual Workshop on Information Technologies & Systems*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bryant, S. L.; Forte, A.; and Bruckman, A. 2005. Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the international ACM SIGGROUP conference on Supporting group work*.

Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2015. The production of information in the attention economy. *Scientific Reports*, 5(1): 9452.

Flöck, F.; and Acosta, M. 2014. WikiWho: Precise and efficient attribution of authorship of revisioned content. In *Proceedings of the World Wide Web Conference*.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Gildersleve, P.; Beers, A.; Ito, V.; Orozco, A.; and Tripodi, F. B. 2025. WikiReddit: Tracing Information and Attention Flows Between Online Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Hale, S. A. 2014. Multilinguals and Wikipedia editing. In *Proceedings of the 2014 ACM Conference on Web Science*.

Halfaker, A.; Geiger, R. S.; Morgan, J. T.; and Riedl, J. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5): 664–688.

Halfaker, A.; Kittur, A.; and Riedl, J. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the International Symposium on Wikis and Open Collaboration*.

Javanmardi, S.; Lopes, C.; and Baldi, P. 2010. Modeling user reputation in wikis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(2): 126–139.

Keegan, B.; Gergle, D.; and Contractor, N. 2013. Hot off the wiki: Structures and dynamics of Wikipedia's coverage of breaking news events. *American Behavioral Scientist*, 57(5): 595–622.

Kim, S.; Park, S.; Hale, S. A.; Kim, S.; Byun, J.; and Oh, A. H. 2015. Understanding Editing Behaviors in Multilingual Wikipedia. *PLoS ONE*, 11.

Kittur, A.; Chi, E.; Pendleton, B. A.; Suh, B.; and Mytkowicz, T. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Alt.CHI*.

Kramsch, C. 2014. Language and Culture. *AILA review*, 27(1): 30–55.

Li, B.; Haider, S.; and Callison-Burch, C. 2024. This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Matsui, A.; Miyazaki, K.; and Murayama, T. 2024. Throw Your Hat in the Ring (of Wikipedia): Exploring Urban-Rural Disparities in Local Politicians' Information Supply. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Matsui, A.; Toriumi, F.; Yoshida, M.; Murayama, T.; and Hironaka, S. 2025. Global Patterns of Knowledge: Language, Genre, and the Geography of Knowledge. *arXiv preprint arXiv:2507.22271*.

MediaWikiAPI. 2022. MediaWiki API. https://www.mediawiki.org/wiki/API.

Meier, F. 2022. TWikiL–the Twitter Wikipedia Link Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Nov, O. 2007. What motivates wikipedians? *Communications of the ACM*, 50(11): 60–64.

Panciera, K.; Halfaker, A.; and Terveen, L. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the 2009 ACM International Conference on Supporting Group Work*.

Park, S.; Kim, S.; Hale, S. A.; Kim, S.; Byun, J.; and Oh, A. H. 2015. Multilingual Wikipedia: Editors of Primary Language Contribute to More Complex Articles. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Piccardi, T.; Gerlach, M.; and West, R. 2024. Curious rhythms: Temporal regularities of wikipedia consumption.

In *Proceedings of the International AAAI Conference on Web and Social Media*.

Piccardi, T.; and West, R. 2025. Navigating Knowledge: Patterns and Insights from Wikipedia Consumption. In Yasseri, T., ed., *Handbook of Computational Social Science*. Edward Elgar Publishing Ltd.

Piller, I.; Zhang, J.; and Li, J. 2022. Peripheral multilingual scholars confronting epistemic exclusion in global academic knowledge production: a positive case study. *Multilingua*.

Profound. 2025. AI Platform Citation Patterns: How ChatGPT, Google AI Overviews, and Perplexity Source Information. https://www.tryprofound.com/blog/ai-platform-citation-patterns. Accessed: 2026-1-13.

Raman, N.; Sauerberg, N.; Fisher, J.; and Narayan, S. 2020. Classifying Wikipedia article quality with revision history networks. In *Proceedings of the International Symposium on Open Collaboration*.

Samoilenko, A.; Karimi, F.; Edler, D.; Kunegis, J.; and Strohmaier, M. 2016. Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity. *EPJ Data Science*, 5(1): 9.

Shenoy, K.; Ilievski, F.; Garijo, D.; Schwabe, D.; and Szekely, P. 2021. A Study of the Quality of Wikidata. *Journal of Web Semantics*, 72: 100679.

Tasnim Huq, K.; and Ciampaglia, G. L. 2021. Characterizing Opinion Dynamics and Group Decision Making in Wikipedia Content Discussions. In *Companion Proceedings of the World Wide Web Conference*.

Umarova, K.; and Mustafaraj, E. 2019. How partisanship and perceived political bias affect wikipedia entries of news sources. In *Companion Proceedings of The World Wide Web Conference*.

Veselovsky, V.; Piccardi, T.; Anderson, A.; West, R.; and Arora, A. 2024. Web2Wiki: Characterizing Wikipedia Linking Across the Web. *arXiv preprint arXiv:2405.09491*.

Viegas, F. B.; Wattenberg, M.; Kriss, J.; and van Ham, F. 2007. Talk Before You Type: Coordination in Wikipedia. In *Proceedings of the Annual Hawaii International Conference on System Sciences*.

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 *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Wikimedia Foundation. 2025. In the AI era, Wikipedia has never been more valuable. https://wikimediafoundation.org/news/2025/11/10/in-the-ai-era-wikipedia-has-never-been-more-valuable/. Accessed: 2026-1-13.

Yang, D.; Halfaker, A.; Kraut, R.; and Hovy, E. 2016. Who did what: Editor role identification in Wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Zeng, H. S.; Danaher, B.; and Smith, M. D. 2022. Internet governance through site shutdowns: The impact of shutting down two major commercial sex advertising sites. *Management Science*, 68(11): 8234–8248.

Zhou, D.; Patankar, S.; Lydon-Staley, D. M.; Zurn, P.; Gerlach, M.; and Bassett, D. S. 2024. Architectural styles of curiosity in global Wikipedia mobile app readership. *Science Advances*, 10(43): eadn3268.

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
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? YES
   (e) Did you describe the limitations of your work? YES
   (f) Did you discuss any potential negative societal impacts of your work? YES
   (g) Did you discuss any potential misuse of your work? YES
   (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
   (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? YES
   (b) Have you provided justifications for all theoretical results? YES
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? YES
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? YES
   (e) Did you address potential biases or limitations in your theoretical framework? YES
   (f) Have you related your theoretical results to the existing literature in social science? YES
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? YES
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? YES
   (b) Did you include complete proofs of all theoretical results? YES

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)? YES
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? YES
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? YES
   (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)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? YES
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? YES
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? YES
   (b) Did you mention the license of the assets? YES
   (c) Did you include any new assets in the supplemental material or as a URL? NO
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? YES
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? YES
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

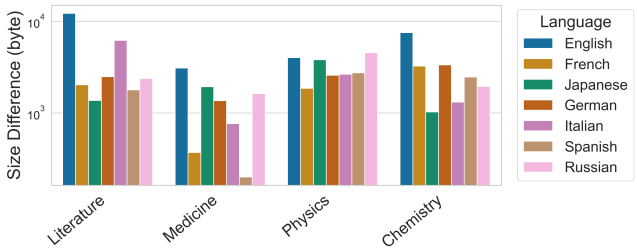


Figure 9: Illustrative example of article size changes following the 2020 Nobel Prize announcements. This figure shows how Wikipedia articles about Nobel laureates in 2020 evolved in size after the prize announcements.

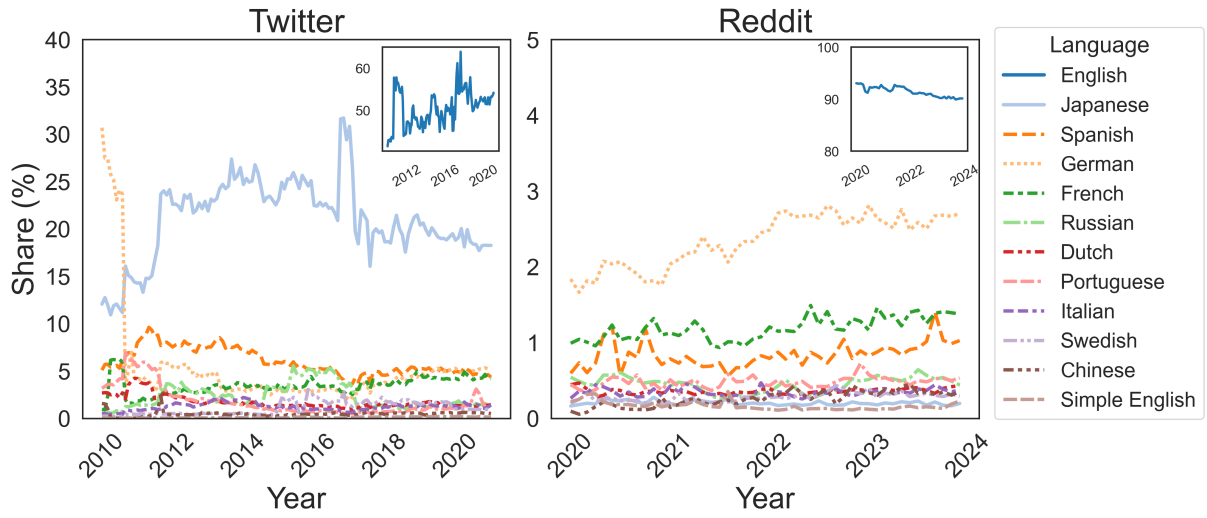


Figure 10: Share of shared Wikipedia articles by language edition on Reddit and Twitter. Due to the disproportionately high volume of shares for the English Wikipedia compared to other languages, its trend is presented in an inset to allow for a clearer visualization of the other language editions. The x and y-axis labels in the inset are identical to those of the main plot.

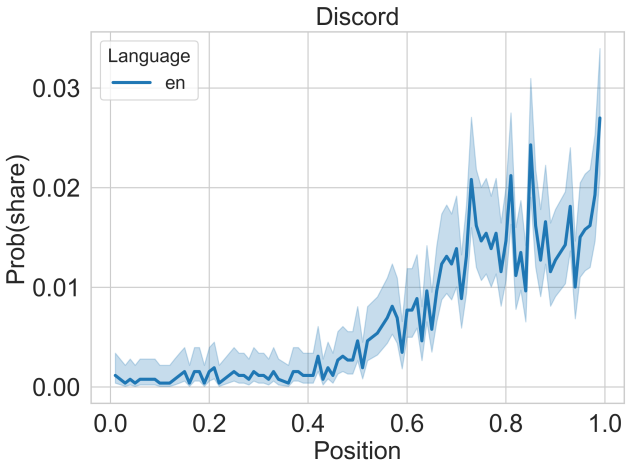
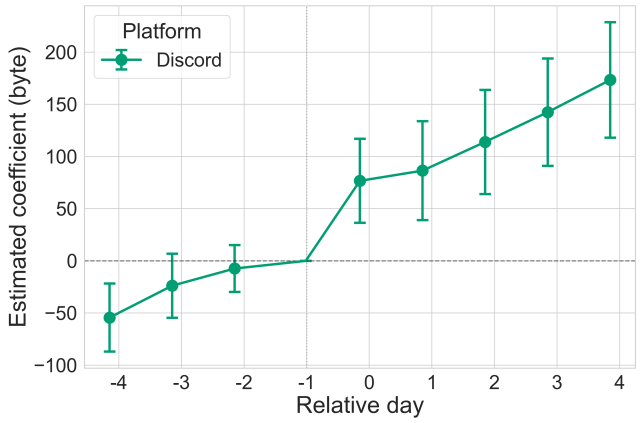
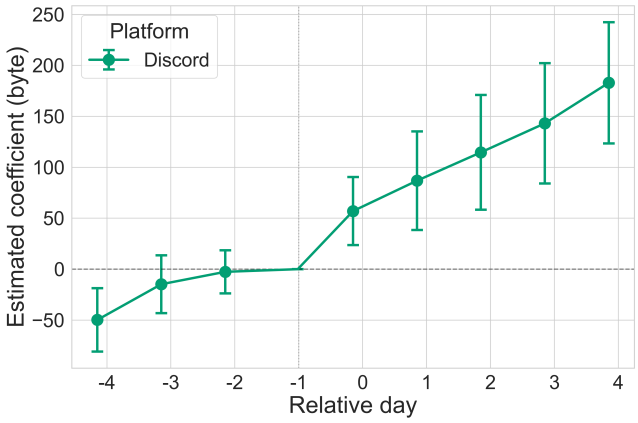


Figure 11: Sharing probability trajectories on Discord. This figure presents a complementary analysis to the Reddit and Twitter results shown in Figure 6. Shaded areas represent 95% CIs.

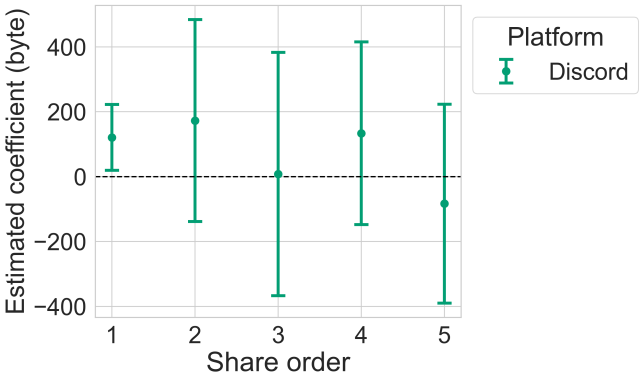


(a) Standard DiD model.

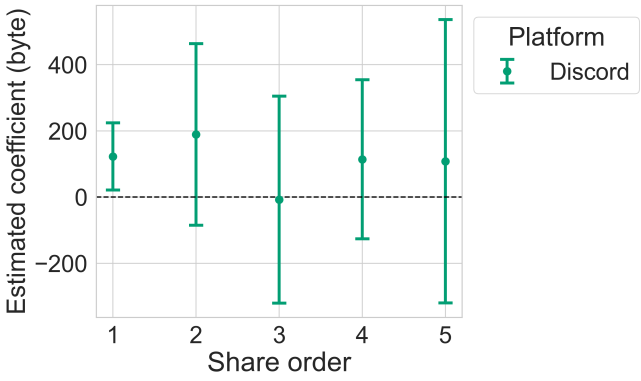


(b) Intensity-adjusted DiD model.

Figure 12: Daily-level difference-in-differences analysis, focusing on the first share in the datasets. Day 0 represents the share day, and day -1 serves as the reference period. Figure 12a presents the coefficients from the standard DiD model (Equation 2), while Figure 12b shows results from the intensity-adjusted model (Equation 1). These plots display the estimated difference in article size (in bytes) for each day relative to the share day, comparing shared articles to their unshared counterparts in other language editions. This figure presents a complementary analysis to the Reddit and Twitter results shown in Figure 7. Shaded areas represent 95% CIs.



(a) Standard DiD model.



(b) Intensity-adjusted DiD model.

Figure 13: Difference-in-differences event study plots for the change in article size by share order. When the same article was shared on multiple occasions, we estimated the model separately for each share event. Figure 13a presents the coefficients from the standard DiD model. These plots display the estimated change in article size (in bytes) in the weeks before and after sharing, comparing shared articles to their unshared counterparts in other language editions. Figure 13b shows results from the intensity-adjusted model. The bars represent 95% CIs. This figure presents a complementary analysis to the Reddit and Twitter results shown in Figure 8.