UNIVERSITY OF AMSTERDAM
Institute for Logic, Language and Computation

Bringing Science to the Public: The Role of Wikipedia in Scientific Communication
Puyu Yang

Bringing Science to the Public

The Role of Wikipedia in Scientific Communication

Puyu Yang

Bringing Science to the Public: The Role of Wikipedia in Scientific Communication

Puyu Yang

Bringing Science to the Public: The Role of Wikipedia in Scientific Communication

ILLC Dissertation Series DS-2026-04

Institute for Logic, Language and Computation

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Cover design by Puyu Yang.
Printed and bound by Ipskamp Printing.

ISBN: 978-94-6536-036-2

Bringing Science to the Public: The Role of Wikipedia in Scientific Communication

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op maandag 23 februari 2026, te 13.00 uur

door Puyu Yang
geboren te Ningxia

Promotiecommissie

Promotor:	prof. dr. R. Fernández Rovira	Universiteit van Amsterdam
Copromotor:	prof. dr. G. Colavizza	Københavns Universitet
Overige leden:	prof. dr. ing. R.A.M. van Rooij	Universiteit van Amsterdam
	dr. ir. J. Kamps	Universiteit van Amsterdam
	prof. dr. K. Sima'an	Universiteit van Amsterdam
	prof. dr. L.R. Waltman	Leiden University
	dr. S. Haustein	University of Ottawa
	dr. N. Robinson García	University of Granada

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Contents

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Giovanni Colavizza. Your patience, guidance, and encouragement have been the most precious support throughout my PhD journey. Working with you has not only shaped my doctoral research but has also profoundly influenced my broader academic and personal growth. It has been both a joy and an honor to be your PhD student, and I am truly grateful for the opportunity you gave me. Although most of our meetings took place online, you were always present when I needed advice, offering timely and insightful support. From our very first paper together, you guided me step by step with care and honesty, helping me gradually find my own direction. I cannot thank you enough for the trust, generosity, and inspiration you have provided during these years.

I am equally indebted to my co-supervisor, Raquel Fernández. Although our meetings were less frequent, you were always there at the most important moments, and your perspective was consistently enlightening and constructive. I will always value the warm and supportive discussions we had during my annual reviews, which showed your sincere care for my work and growth. Your careful comments and invaluable suggestions during the writing of this thesis have been a tremendous help. I am very grateful for your support and very happy to have had the chance to work with you.

Thanks to the understanding and encouragement of Giovanni and Raquel, I was able to see light at the end of the tunnel during my PhD. For this, I am profoundly grateful.

I would also like to express my sincere thanks to my committee members: Ludo Waltman, Stefanie Haustein, Nicolas Robinson-Garcia, Robert van Rooij, Jaap Kamps, and Khalil Sima'an. It is a privilege to have you serve on my committee, and I am grateful for the time, effort, and thoughtful engagement you have invested in reading and evaluating my thesis.

My gratitude also goes to the ILLC. This institute is not only an outstanding academic environment but also a truly warm and supportive community. I am

thankful for the friendly atmosphere in LAB42, which made my daily work both enjoyable and inspiring. I would especially like to thank my office mates and colleagues: Ece, for being such a wonderful deskmate, for your kindness, help, and our many enjoyable conversations; Evgenia, for your company and the many lunches we shared; and Joris, for the many pleasant moments in our office. I am also grateful to current and former PhD colleagues—Alina, Marco, Oskar, Julian, Michael, Jaap, and Zhi—who made my time at ILLC fun and rewarding. Special thanks also to Ewout and Peter for their assistance and care with many practical matters during my PhD. I also want to acknowledge all other ILLC members with whom I shared lunches and met at events. Your companionship and kindness made this journey much lighter.

I am also thankful to Liangping and Congfeng. It was a pleasure to meet you and work alongside you, and I wish you both every success in your research and future paths.

I owe a special debt of gratitude to the CWTS in Leiden. Visiting CWTS was a truly inspiring experience, and I am deeply grateful to my hosts and co-authors, Vincent and Rodrigo, for your invaluable guidance both within and beyond my research. I learned a great deal from working with you. I would also like to thank my colleagues at CWTS—Dongyi, Qianqian, Juan, Huilin, Biegzat, Karin, Bram, Leyan, Huizhen, Wei Quan, and many others—for the wonderful time we spent together. I am grateful for your support, collaboration, and friendship.

Finally, I would like to thank my family. To my parents, thank you for your endless support and encouragement. I could never have come this far without your love and trust. Most importantly, I want to thank my wife, Chengdi. Thank you for being by my side throughout these years, for your care, patience, and faith in me. You have always been the light in moments of confusion and doubt, giving me strength when I needed it most. These years have been a journey of growth for both of us, and I have learned so much from you. I am deeply grateful for your unwavering support, your belief in me, and your love.

Amsterdam Puyu Yang
September, 2025.

Chapter 1

Introduction

In the contemporary information environment, scientific knowledge is more abundant and accessible than ever before. From peer-reviewed publications to institutional datasets and preprints, the volume and diversity of scientific outputs have expanded rapidly (Piwowar et al., 2018; Fraser et al., 2019; Lin et al., 2020). While this proliferation holds promise for democratizing access to science, it also introduces new complexities. Members of the public must navigate an increasingly fragmented and unevenly accessible information landscape (Brossard, 2013; Nisbet and Scheufele, 2009). In this context, questions surrounding how scientific knowledge is filtered, translated, and made publicly legible have become ever more pressing.

Wikipedia occupies a central position in this shifting landscape of public knowledge. As the largest open and collaboratively edited encyclopedia, it is a dominant information source worldwide. Its articles consistently rank among the top results for science-related web searches, and its multilingual content structure supports a global readership that includes students, educators, journalists, and policymakers alike (Nielsen, 2007; Heilman and West, 2015; Jemielniak, 2020; Kittur et al., 2007a). Wikipedia does not produce original research. Instead, it curates and organizes information derived from external sources, both academic and non-academic, and makes this knowledge publicly available through a distinctive editorial process rooted in community norms[1]. These norms include an emphasis on verifiability, neutrality, and reliance on reliable sources (Mesgari et al., 2015). Together, they reflect Wikipedia's epistemological commitment to serving as a trustworthy public knowledge infrastructure, even though it operates outside traditional academic gatekeeping mechanisms (Pavalanathan et al., 2018; Arazy et al., 2006; Forte et al., 2009).

As an intermediary between the scientific community and the broader public, Wikipedia serves as a powerful interface for science communication (Mesgari et al., 2015). Its editors face the task of evaluating sources, resolving disputes,

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

contextualizing technical content, and presenting complex ideas in accessible ways (Forte et al., 2009). These challenges are particularly acute in articles dealing with rapidly evolving topics, interdisciplinary subjects, or politically contested issues (Keegan et al., 2013; Greenstein and Zhu, 2012b). At stake in these editorial decisions is not merely the accuracy of individual claims, but the broader public framing of what counts as credible scientific knowledge, who gets to speak for science, and how knowledge is structured and legitimized in open digital platforms (Hu, 2024; Avieson, 2022).

This thesis investigates how Wikipedia supports the public communication of science through its practices of citation and knowledge organization. Scientific journal articles and news media sources are central to how information is sourced and substantiated on the platform. References serve not only to bolster claims, but also to signal reliability, highlight specific viewpoints, and shape the perceived authority of content. Citation practices thus play a crucial role in structuring Wikipedia's knowledge base and, by extension, influence how science is represented in public discourse. Despite their significance, the selection and use of references on Wikipedia remain relatively underexplored from a systematic, empirical perspective.

The overarching aim of this thesis is to examine how Wikipedia incorporates, filters, and recontextualizes external scientific and media sources, and how these citation practices reflect and shape the platform's function as a public science communication tool. This investigation is situated at the intersection of science communication, information infrastructure studies, and platform governance. It engages with key questions about how epistemic authority is constructed in digitally mediated environments, how disciplinary hierarchies and accessibility barriers shape knowledge visibility, and how editorial decisions are made in collaborative online contexts.

A central focus of this inquiry is the knowledge structure of Wikipedia's science-related content. Wikipedia articles are highly heterogeneous in scope and depth, ranging from concise overviews to densely referenced entries. This raises the research question of which scientific sources are cited and how different areas of Wikipedia rely on them. Prior research shows that only a small fraction of Wikipedia's total citations point to peer-reviewed scientific literature, and that these citations tend to cluster within specific domains, particularly biology and medicine, often drawn from a narrow set of high-impact journals (Nielsen, 2007). Building on this, our analysis of citation patterns in this thesis finds that STEM literature indeed dominates Wikipedia references, and that biographies act as important bridges connecting STEM fields to the humanities. These findings suggest that Wikipedia not only reflects the hierarchies of academic science but also reshapes them, privileging certain forms of expertise while potentially reconfiguring disciplinary boundaries to respond to public-facing concerns.

Another important dimension of Wikipedia's knowledge infrastructure is its reliance on news media (Fetahu et al., 2015). News sources are frequently used

to document recent developments and sociopolitical contexts surrounding scientific topics. However, media outlets differ widely in factual reliability and political orientation, raising concerns about potential biases in knowledge representation (Saez-Trumper, 2019; Morgan, 2019). This leads to the research question of whether political bias and factual reliability of news sources shape their selection on Wikipedia. Our analysis indicates a moderate yet systematic liberal bias in the selection of news media sources, largely independent of factual reliability. This suggests that editorial choices may be influenced by factors beyond mere accuracy. Understanding how media sources are selected and cited is therefore essential for assessing the platform's epistemic practices and its broader impact on public discourse. These findings highlight the challenges of implementing Wikipedia's guidelines for reliability and neutrality in a dynamic, volunteer-driven environment (Aragón and Saez-Trumper, 2021), and underscore the role of editorial decisions in shaping public perceptions of science.

The question of accessibility is also central to the thesis. As the open access (OA) movement gains momentum in scholarly publishing, concerns have arisen about how freely available research influences the visibility, credibility, and contestation of scientific information on public platforms (Tattersall et al., 2022). In particular, it is unclear to what extent OA status affects the likelihood of a publication being cited on Wikipedia and its involvement in editorial disputes. To address this, we examine how the open access status of scientific publications shapes their integration into Wikipedia and their use in contested content. Our analysis shows that OA articles are cited more frequently than paywalled articles, with the effect especially pronounced for recent and highly cited works. We also find that OA articles are more likely to be involved in disputes, particularly in social sciences and humanities, and that contested claims appear sooner after publication. These findings highlight that accessibility does more than facilitate inclusion; it also accelerates engagement and scrutiny in collaborative editorial processes. Openly available research is easier to mobilize in Wikipedia editing, influencing both the speed and content of scientific knowledge dissemination. This underscores the strategic as well as logistical implications of OA for public communication of science, revealing a complex interplay between accessibility, visibility, and contestation (Hara and Doney, 2015; Wyatt et al., 2016; Steiert, 2025).

Each of these dimensions, including knowledge structure, media source reliability, open access, and editorial dynamics, offers a distinct perspective for understanding how Wikipedia operates as a platform for scientific communication. This thesis approaches Wikipedia not as a passive repository of information but as an active knowledge infrastructure. It is a space where decisions about what constitutes knowledge are continuously negotiated, where credibility is actively constructed, and where various publics interact with science in ways shaped by technological, social, and political contexts.

By investigating citation patterns across scientific and media sources, and by

analyzing how editorial norms and accessibility constraints shape knowledge representation, this thesis contributes to a growing body of interdisciplinary research on digital epistemology, public engagement with science, and open knowledge platforms. It seeks to illuminate the evolving relationship between science and society, as mediated by one of the most influential public repositories of knowledge in the digital age.

The findings are of relevance to multiple stakeholder communities. For science communicators and educators, they highlight the importance of curating and contextualizing scientific content for general audiences. For platform designers and policy-makers, the results underscore the need for infrastructures that support equitable, reliable, and transparent knowledge integration. For researchers in science and technology studies, communication, and information science, this thesis offers empirical insights into how epistemic norms are enacted in one of the world's most impactful digital knowledge platforms.

1.1 Thesis Overview

This thesis consists of two parts with the overarching goal of understanding how Wikipedia serves as a platform for scientific communication. A background chapter sets the stage for both parts by outlining the broader context relevant to the thesis as a whole.

Background In this chapter, I first focus on the role of scientific knowledge within Wikipedia, exploring how Wikipedia integrates and manages scientific sources. It covers Wikipedia's editorial principles and policies, the characteristics of scientific and news media citations, and the challenges posed by political polarization and reliability of news sources. This sets the stage for understanding Wikipedia as a dynamic platform for scientific communication and knowledge curation.

The second part addresses the interaction between open access publishing and the dissemination of scientific knowledge through Wikipedia. It examines how open access influences Wikipedia's citation practices and its impact on public understanding of science. Furthermore, it considers Wikipedia as a site of scientific disputes where open access publications play a significant role. Together, these two parts offer complementary perspectives on how Wikipedia shapes and reflects the broader ecosystem of scientific communication.

Part One: The Science in Wikipedia

The first part of this thesis investigates how Wikipedia engages with both scientific literature and news media by analyzing the *Wikipedia Citations* (Singh et al., 2021). It focuses on two key aspects: the distribution of scientific knowledge within Wikipedia and the reliability and political orientation of the news media sources it cites.

Chapter 4 presents a comprehensive analysis of Wikipedia's use of scientific

publications. Drawing on *Wikipedia Citations*, the study reveals that the majority of journal articles cited in Wikipedia originate from STEM disciplines, particularly biology and medicine. Furthermore, Wikipedia's biographical articles help bridge scientific domains with the humanities, especially history. These findings enhance our understanding of Wikipedia's function as a broker of scientific knowledge and its role in facilitating public access to science.

Chapter 5 shifts attention to the citation of news media sources. While Wikipedia relies on a wide range of external references, news outlets account for nearly one-third of all citations, prompting concerns about their impact on content neutrality and credibility. This chapter systematically examines the political polarization and factual reliability of these sources. Through quantitative analysis, it explores how Wikipedia's open editorial model influences media source selection and how media bias may shape perceptions of trustworthiness. The results underscore the complex interplay between user-generated knowledge and external media ecosystems, offering insights into how Wikipedia mediates between scientific content and broader public discourse.

Part Two: The Role of Open Access in the Dissemination of Science

The second part of this thesis focuses on the role of open access (OA) publishing in shaping the flow of scientific knowledge into Wikipedia and, by extension, into broader public discourse. It examines how the accessibility of scientific literature influences citation patterns in Wikipedia and how open access affects knowledge representation in both routine article development and in editorial conflicts.

Chapter 7 investigates whether open access facilitates the integration of scientific knowledge into Wikipedia. Using *Wikipedia Citations*, this study analyzes how the open access status of scientific articles affects their likelihood of being cited. The results show that Wikipedia cites open access articles at significantly higher rates than would be expected based on their overall presence in the scientific publishing landscape. This effect remains robust even when controlling for other factors such as citation counts and publication year. Open access articles are more likely to be cited, particularly when they are recent or highly cited, indicating that accessibility plays a key role in shaping which scientific knowledge becomes part of the public record. These findings emphasize the importance of open access in enabling timely and equitable dissemination of research through widely used public platforms like Wikipedia.

Chapter 8 extends the analysis by exploring the role of open access in contexts of editorial controversy. Focusing on disputed or contested Wikipedia articles, the study combines citation data with edit histories to examine which types of scientific sources are invoked during content disputes. The analysis reveals that open access articles are more likely to be cited in controversial articles and tend to be introduced shortly after publication. This pattern suggests that ease of access not only supports citation, but also accelerates the incorporation of scientific findings into contested spaces where knowledge is actively negotiated. The chapter also shows that disputes involving open access sources are particularly common

in the social sciences and humanities, reflecting differences in how disciplines are represented and debated in Wikipedia.

Together, these chapters highlight the dual function of open access: increasing the visibility and use of scientific knowledge, and shaping how that knowledge is mobilized in collaborative and contested environments. This part contributes to broader discussions around open science, knowledge equity, and the governance of digital public knowledge infrastructures.

Conclusion (Chapter 9) In this chapter, I summarize the key findings and contributions of the thesis, reflect on its limitations, and outline directions for future research. This chapter also considers broader implications for open science, platform governance, and public engagement with scientific knowledge.

Overall, the research presented in this thesis contributes to our understanding of how scientific information flows into Wikipedia, how different types of sources—particularly open access articles and news media—shape the encyclopedia's representation of science, and how these dynamics influence public access to and engagement with reliable knowledge. The findings shed light on the informational structures underpinning one of the most widely used knowledge platforms in the world, highlighting both the opportunities and challenges involved in making science more open, visible, and democratically negotiated in the digital age.

1.2 List of Publications

The contents of this thesis are based on the following publications, listed in the order they appear in the chapters:

- Yang, P., & Colavizza, G. (2022). *A Map of Science in Wikipedia.* In *Companion Proceedings of the Web Conference 2022* (pp. 1289–1300).
- Yang, P., & Colavizza, G. (2024). *Polarization and Reliability of News Sources in Wikipedia. Online Information Review*, *48*(5), 908–925.
- Yang, P., Shoaib, A., West, R., & Colavizza, G. (2024). *Open Access Improves the Dissemination of Science: Insights from Wikipedia. Scientometrics*, *129*(11), 7083–7106.
- Yang, P., Traag, V. A., Costas, R., & Colavizza, G. (2025). *Weaponized Citations: The Role of Open Access Publications in Wikipedia's Scientific Disputes.* arXiv preprint arXiv:2510.14071.

I list below the other works that I contributed to during my PhD:

- Yang, P., & Colavizza, G. (2025). *Research Data in Scientific Publications: A Cross-Field Analysis.* arXiv preprint arXiv:2502.01407.

During my PhD, I have presented my work in the following international events:

- **2025** International Conference on Science, Technology and Innovation Indicators (STI), Bristol, United Kingdom.
- **2025** International Society for Scientometrics and Informetrics (ISSI), Yerevan, Armenia.
- **2023** CZI Hackathon: Mapping the Impact of Research Software in Science, San Francisco, United States.
- **2023** Workshop on Open Citations and Open Scholarly Metadata, Bologna, Italy.
- **2023** International Society for Scientometrics and Informetrics (ISSI), Bloomington, United States.
- **2023** CWTS Scientometrics Summer School (CS3), Leiden, Netherlands.
- **2022** The 26th International Conference on Science, Technology and Innovation Indicators (STI), Granada, Spain.
- **2022** The Web Conference 2022 (WWW), Lyon, France.

Chapter 2

Background

2.1 Wikipedia as a Knowledge Infrastructure

Wikipedia is the world's largest free online encyclopedia, with over 65 million articles in more than 300 languages and about one billion unique monthly visitors[1]. Unlike traditional expert-authored encyclopedias, Wikipedia is created and maintained through a decentralized editorial model driven by volunteers who continuously create, revise, and update content in real time. This allows Wikipedia to remain responsive to emerging events and to offer dynamic, up-to-date knowledge (Mesgari et al., 2015; Jemielniak and Aibar, 2016; Black, 2008).

As a complex network of interlinked articles representing concepts connected through semantic hyperlinks, Wikipedia facilitates nonlinear exploration and contextual navigation of knowledge (Kim et al., 2019). This structure reflects its nature as a collaborative and evolving knowledge infrastructure.

Wikipedia's credibility is upheld by three fundamental content policies: Neutral Point of View (NPOV), Verifiability, and No Original Research (Mesgari et al., 2015; Borra et al., 2014). These require fair representation of significant viewpoints and that all content be supported by reliable, published sources, emphasizing verifiability over absolute truth (Kaffee and Elsahar, 2021; Lewoniewski et al., 2023). Editors must avoid personal interpretations and cite external, verifiable publications.

The platform embodies a broader epistemological shift from traditional expert-driven knowledge production toward socially constructed, collaborative modes (Fallis, 2008; Fuchs, 2007). Its decentralized governance and focus on verifiability exemplify this transformation.

Wikipedia relies on collective intelligence, where contributions from diverse users aggregate into high-quality content (Surowiecki, 2005). The editorial community is heterogeneous, comprising a core of highly active, experienced editors

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

alongside many occasional contributors and newcomers, supporting both content stability and growth (Halfaker et al., 2013).

The community has developed complex editorial norms and collaborative workflows that function as a self-governed regulatory system. Quality is maintained through transparency, public deliberation, and procedural legitimacy. A key practice is citing "reliable, independent, and published" secondary sources to ensure content reflects established knowledge rather than original interpretation (Teplitskiy et al., 2017).

The definition of "reliable sources" on Wikipedia is context-sensitive rather than fixed. Peer-reviewed academic publications are generally prioritized in scientific topics, while reputable journalistic or institutional sources may be accepted for contemporary or interdisciplinary subjects, especially when scholarly sources are unavailable or outdated. Editorial consensus determines source reliability, considering independence, publication status, and recency.

Academic literature plays a significant role in Wikipedia's citation ecosystem. Approximately 8.3% of references are to scholarly sources, underscoring peer-reviewed research's importance in establishing credibility (Singh et al., 2021). Editors act as gatekeepers who select scientific findings to include, shaping the public dissemination of authoritative knowledge (Jemielniak and Aibar, 2016). However, editor expertise varies widely, with only a minority holding doctoral degrees (Glott et al., 2010). This can affect source evaluation, especially when distinguishing reliable academic work is complex (Murray et al., 2018). Wikipedia favors research from prestigious journals, with recent and open-access publications more likely to be cited (Arroyo-Machado et al., 2020; Benjakob et al., 2022).

Beyond scholarly references, Wikipedia extensively cites news media, particularly for breaking news and topics with limited academic coverage. News sources typically constitute 20–30% of external references in English Wikipedia (Fetahu et al., 2015). The editorial community applies quality filters, favoring reputable, centrist outlets such as Reuters, BBC, and The New York Times, while avoiding hyperpartisan or low-credibility media (Greenstein and Zhu, 2012b). This selective sourcing acts as an informal gatekeeping mechanism, influencing how scientific and political narratives are presented.

The platform's decentralized governance model offers flexibility but also leads to conflicts, especially on politically sensitive topics. "Edit wars" may arise when contributors repeatedly change content to reflect opposing views (Yasseri et al., 2012). In these disputes, citations to scientific or high-impact sources often serve as rhetorical tools to assert authority or challenge alternative perspectives (West et al., 2010).

Wikipedia functions as a socio-technical infrastructure that blends user participation, transparent rules, and automated tools. Bots and algorithms help detect vandalism, flag citation needs, and evaluate source reliability, while human editors retain central judgment in content disputes and neutrality enforcement (Zagorova et al., 2021; Lewoniewski et al., 2020).

Research into Wikipedia's citation graph structures reveals patterns in how scientific domains interconnect, highlighting emergent knowledge flows (Silva et al., 2011). Comparative studies between Wikipedia's category taxonomy and formal classification systems like the Universal Decimal Classification expose differences between emergent online taxonomies and traditional curated knowledge organization (Salah et al., 2012).

Wikipedia's role extends beyond knowledge dissemination to influencing scholarly visibility. Although only a small fraction of academic articles—such as 4% of PLOS articles—are cited on Wikipedia, such citations significantly amplify public exposure and contribute to altmetric measures (Lin and Fenner, 2014; Kousha and Thelwall, 2017). Articles cited tend to come from high-impact, often open-access journals, reinforcing specific forms of scientific authority (Nielsen, 2007; Teplitskiy et al., 2017).

The content on Wikipedia shows a temporal bias towards recent events and newer sources, reflecting both its dynamic editing culture and editors' preferences (Jemielniak et al., 2019; Sundin, 2011). Maintaining long-term accuracy amid ongoing updates remains a challenge.

Wikipedia's public scientific influence has grown with its integration into broader technology ecosystems. Its content is indexed by search engines, used by digital assistants, and incorporated into educational and media workflows (Yang and Colavizza, 2022a; Maggio et al., 2017). The COVID-19 pandemic illustrated Wikipedia's capacity for rapid, collaborative updating based on evolving scientific knowledge, balancing timeliness with reliance on peer-reviewed and open-access sources (Colavizza, 2020).

While news media sources contribute to Wikipedia's timely coverage, they also introduce epistemic challenges. News outlets vary in reliability and editorial standards, which can subtly affect content quality. The risk of polarization, especially in politically sensitive topics, highlights the importance of carefully evaluating news sources for both factual accuracy and potential bias (Lazer et al., 2018; Patterson, 2011).

2.2 Quantitative Research on Wikipedia and Scientific Citations

Wikipedia has long been a subject of interest for researchers studying large-scale collaborative knowledge production. Its transparent and publicly available revision history provides a uniquely rich dataset for quantitative inquiry. Scholars have used a variety of computational and statistical methods to examine editorial behavior, content evolution, citation dynamics, and platform infrastructure.

One of the earliest systematic analyses was conducted by Voss (Voss, 2005), who studied the German-language edition of Wikipedia. His work modeled exponential

growth in article counts, database size, user participation, and hyperlink density, noting that article and author contributions follow Lotka-like distributions, similar to those found in scientific publishing.

Visualization tools have helped illuminate editorial patterns. Viegas et al. (Viégas et al., 2007) introduced the "History Flow" tool to trace article development and visualize individual editors' contributions. Chromograms and edit timelines were used to distinguish between original content creation, maintenance, and vandalism reversion.

Subsequent research has examined the relationship between editor activity and article quality. Wilkinson and Huberman (Wilkinson and Huberman, 2007) observed that articles with higher visibility, measured by network centrality, tend to attract more edits and exhibit greater stability. Ortega et al. extended this work across multiple language editions, identifying differences in editing patterns between short articles and longer, more mature entries. They found that a small core group of contributors is responsible for the majority of edits, reflecting a long-tail distribution in participation (Kittur et al., 2007a).

Temporal analyses have revealed bursty and self-similar editing behavior, deviating from the random Poisson processes that characterize general web activity (Almeida et al., 2007). These findings suggest that editorial attention on Wikipedia is highly clustered in time, particularly around controversial or high-interest topics.

Citation practices have also received significant attention. Articles in domains such as medicine, biochemistry, and physics are cited more frequently than those in the humanities or social sciences (Teplitskiy et al., 2017). Biomedical entries, in particular, often rely heavily on academic sources and are widely consulted by both the public and professionals (Heilman et al., 2011; Maggio et al., 2017). Editors tend to prefer peer-reviewed, high-impact, and open-access sources (Yang and Colavizza, 2022a), reinforcing established hierarchies of scientific authority. Recent studies using bibliometric datasets such as OpenAlex and Unpaywall show that open-access status significantly increases the likelihood of citation, especially during high-urgency situations like the COVID-19 pandemic (Yang et al., 2024; Piwowar et al., 2018).

Analyses of citation timing suggest that most references are added once an article reaches a stable phase of development and receives frequent contributions (Kaffee and Elsahar, 2021; Chen and Roth, 2012).

To study the relationship between Wikipedia and the broader scientific ecosystem, researchers have integrated Wikipedia citation data with external bibliographic databases such as Crossref, PubMed, and Dimensions (Teplitskiy et al., 2017; Arroyo-Machado et al., 2020; Torres-Salinas et al., 2019). These connections allow for analyses of journal prestige, funding sources, and disciplinary representation. Co-citation and bibliographic coupling studies show that citation patterns on Wikipedia often mirror those found in the scholarly literature (Nicholson et al., 2021), with medicine, molecular biology, and environmental sciences being

especially prominent (Arroyo-Machado et al., 2020; Torres-Salinas et al., 2019).

Editor expertise and community structure significantly shape how citations are selected and maintained. While Wikipedia is open to contributions from anyone, studies show that citation decisions are often made by a relatively small group of experienced editors whose choices reflect implicit hierarchies of knowledge and trust (Yarovoy et al., 2020).

Despite Wikipedia's extensive readership, user engagement with its references remains limited. Click-through data suggest that external citations are rarely accessed (Piccardi et al., 2020). However, references are more likely to be consulted in shorter or lower-quality articles, indicating that they serve a more functional role when article content is sparse or ambiguous (Piccardi et al., 2023). User behavior varies by context, with browsing patterns differing between casual readers and those seeking academic information (Singer et al., 2017).

Controversies over citation practices have also drawn scholarly attention. Disputes over source selection frequently arise in politically sensitive topics, where opposing ideological positions clash. Tools such as revert tracking, edit-distance metrics, and temporal clustering have been used to study the dynamics of "edit wars" and content stabilization (Yasseri et al., 2013; Brandes and Lerner, 2008; Wagner et al., 2015).

To support editorial practices, researchers have developed automated tools that assist in citation management. Machine learning and natural language processing have been used to detect biased or unsupported claims (Hube and Fetahu, 2018), identify citation gaps (Fetahu et al., 2016), and recommend relevant sources based on article content (Przybyla et al., 2022). These systems help reinforce Wikipedia's core principles of verifiability and neutrality at scale.

Despite the breadth of existing research, notable gaps remain. Most studies focus on the English-language Wikipedia, with limited comparative work across other language editions. In addition, many quantitative approaches emphasize citation frequency without considering the rhetorical or epistemic roles of references in context (Benjakob et al., 2022; Esteves Gonçalves da Costa and Cukierman, 2019). As a result, scholars increasingly call for interdisciplinary approaches that combine large-scale data analysis with qualitative methods to better understand how knowledge legitimacy is constructed on Wikipedia.

2.3 Wikipedia, Open Access, and Scientific Visibility

Open Access (OA) refers to the unrestricted and cost-free availability of scholarly publications, aiming to enhance the visibility, accessibility, and impact of scientific knowledge regardless of readers' financial or institutional barriers (Tennant et al., 2016; Redalyc et al., 2003). Various OA models exist, including gold OA (fully

open and publisher-hosted), green OA (archived in institutional repositories), hybrid OA (subscription journals offering OA options), and bronze OA (freely available without explicit licensing) (Piwowar et al., 2018). These models reflect evolving norms and practices in scholarly communication.

Empirical research indicates that OA significantly improves access to journal articles (Björk et al., 2010), though estimates of OA prevalence vary depending on data sources and methodologies. For example, studies based on Crossref data report approximately 27.9% of scholarly articles as openly accessible (Piwowar et al., 2018), while analyses using Google Scholar data suggest over 50% OA availability for articles published since 2007 (Martín-Martín et al., 2018; Archambault et al., 2014). Bronze OA, which generally denotes articles made freely available on publisher websites without clear licensing, appears to be the most common OA type (Piwowar et al., 2018). There are also clear disciplinary differences in OA coverage, with general science, technology, and biomedical fields exhibiting higher OA rates compared to engineering, arts, and humanities (Archambault et al., 2014; Martín-Martín et al., 2018).

The so-called Open Access Citation Advantage (OACA), where OA articles receive more citations than paywalled ones, has been widely studied but remains controversial. Reported citation advantages range from 18% to 40% depending on data sources such as Web of Science and Scopus (Piwowar et al., 2018; Archambault et al., 2014). Some disciplines, including philosophy, political science, electrical engineering, and mathematics, show more pronounced citation benefits associated with OA (Antelman, 2004). Green OA articles in institutional repositories often attract more citations than gold OA or subscription-only articles (Young and Brandes, 2020). However, a recent systematic review of 134 empirical studies found mixed results: 47.8% confirmed a positive citation advantage, 27.6% found no effect, and 23.9% reported effects only under specific conditions (Langham-Putrow et al., 2021). These findings suggest that the citation advantage of OA varies according to discipline, timing, and access pathway.

Wikipedia, as the largest online encyclopedia, provides a unique context to study the role of OA in scientific visibility. It aims to disseminate knowledge globally and relies heavily on peer-reviewed research to support its content (Black, 2008). Research shows that articles cited in Wikipedia are often from high-impact journals and are more likely to be openly accessible than average scholarly articles (Nielsen, 2007; Teplitskiy et al., 2017). In fields such as medicine and psychology, Wikipedia serves both as a public information source and a gateway to academic literature (Maggio et al., 2017; Schweitzer, 2008).

Since 2015, altmetric indicators have tracked Wikipedia citations as measures of research impact, although their significance remains debated (Lin and Fenner, 2014; Kousha and Thelwall, 2017). For instance, only 4% of PLOS articles have been found cited on Wikipedia (Lin and Fenner, 2014), with citation coverage varying widely across disciplines. Nonetheless, Wikipedia's engagement with OA sources is growing. One study reported that 13.44% of Wikipedia references come

from OA journals (Arroyo-Machado et al., 2020), and recent analyses indicate that over 30% of cited sources are OA, with this proportion increasing over time (Pooladian and Borrego, 2017).

The reciprocal relationship between scientific literature and Wikipedia highlights the platform's potential as an amplifier of scholarly visibility. By citing peer-reviewed research, Wikipedia can increase public exposure to scientific findings and potentially influence their impact within academic communities (Nielsen, 2007; Teplitskiy et al., 2017). However, the mechanisms through which Wikipedia integrates and highlights different forms of scholarly communication, such as Open Access, are complex (Arroyo-Machado et al., 2020). The diversity of OA types and their varying accessibility levels may affect how editors select and use sources (Piwowar et al., 2018).

Moreover, the dynamic and collaborative nature of Wikipedia editing means that citations are not static endorsements but may be subject to contestation, revision, and negotiation over time (Yasseri et al., 2012).

Despite these insights, the interaction between OA and Wikipedia remains underexplored. Previous studies often analyze citations at the journal level and rely on manual matching methods, potentially overlooking distinctions among OA types (green, gold, hybrid, bronze) and OA versions hosted outside traditional publishers (ElSabry, 2017). Additionally, most research treats citations as static, ignoring the editorial dynamics of how OA articles are introduced, contested, or maintained within Wikipedia's evolving content.

Part One: The Science in Wikipedia

Chapter 3

Overview

This part investigates how Wikipedia functions as a knowledge infrastructure through the lens of its sourcing practices. It examines the nature and structure of scientific and news citations across Wikipedia, and what they reveal about the encyclopedia's representation of scientific knowledge, source reliability, and editorial norms. Chapters 4 and 5 address two complementary dimensions: the use of scholarly citations and the political and factual character of news media sources. Together, these studies contribute to our understanding of how Wikipedia curates, organizes, and legitimizes knowledge from external sources. They also provide empirical insight into Wikipedia's citation practices and their implications for knowledge equity and integrity. I investigate the following research questions:

- **What types of scientific articles are cited in Wikipedia, and what do they reveal about the coverage and topical structure of scientific knowledge in the encyclopedia?**
  In Chapter 4, I explore Wikipedia's use of academic sources by analyzing large-scale citation data from Wikipedia to scientific journals. I identify which scientific domains are most prominently represented, and how citation patterns reflect the organization of scientific knowledge across Wikipedia articles. Network analysis is used to map relationships between articles and journals, providing a structural view of how science is embedded in Wikipedia.

- **Are news media sources cited in Wikipedia politically polarized, and how does their factual reliability relate to that polarization?**
  Chapter 5 focuses on Wikipedia's citation of news media sources, with an emphasis on the political orientation and factual accuracy of those outlets. Drawing on third-party metadata and a large citation dataset, I assess the extent of polarization in news citations and whether reliability and political leaning are systematically related. Regression modeling is used to uncover

the factors that shape how politically or factually slanted media sources are cited in different areas of Wikipedia.

By studying both scientific and media citations, these chapters provide a foundational map of the sources that sustain Wikipedia's epistemic infrastructure. They offer a descriptive but essential account of how credibility, coverage, and bias are distributed in one of the world's most influential public knowledge platforms.

Chapter 4

A Map of Science in Wikipedia

4.1 Introduction

Wikipedia is the largest, free and collaborative encyclopedia to date. Its importance cannot be overstated as Wikipedia provides reliable access to information worldwide. Wikipedia editors use primary and secondary sources in support of the statements they make in Wikipedia. These sources are quoted, cited and added to the list of references in any Wikipedia article. While previous work has focused on the contents of Wikipedia and its collaborative editing, it is only recently that scholars have begun to systematically investigate Wikipedia's sources.

Understanding which scientific sources are cited in Wikipedia is a necessary first step in situating the platform within broader systems of knowledge production. By mapping the distribution of scientific literature within Wikipedia, this chapter establishes a foundation for the rest of the thesis. Later chapters will build on this baseline by examining other types of sources such as news media (Chapter 5), exploring the role of open access in shaping visibility (Chapter 7), and analyzing how citations are mobilized in editorial disputes (Chapter 8).

Developing a better understanding of the scientific sources Wikipedia relies on is important to assess its coverage, reliability and representation of human knowledge, including in view of informing its future development. Expanding upon previous work (Teplitskiy et al., 2017; Torres-Salinas et al., 2019; Colavizza, 2020; Arroyo-Machado et al., 2020), we pose here the following two descriptive research questions (RQs):

1. RQ1: Which scientific sources are cited from Wikipedia and what are their characteristics?

2. RQ2: Which areas of Wikipedia rely on scientific sources and how do they relate?

In order to contribute towards answering them, in the present contribution we focus on citations given from Wikipedia articles to scientific journal articles.

We rely on a recently published dataset as a source of such citation data (Singh et al., 2021). Firstly, we provide a descriptive overview of the journal articles cited from Wikipedia (RQ1); secondly, we make use of network analysis and study the bibliographic coupling network of Wikipedia articles, respectively the co-citation network of journal articles (RQ2).

4.2 Previous Work

References in Wikipedia Wikipedia is a tertiary source which strives to provide reliable contents in a neutral way (Mesgari et al., 2015). To this end, Wikipedia editors follow established standards, guidelines and workflows to expand Wikipedia articles, and add references to them (Kaffee and Elsahar, 2021). The bulk of the referencing activity seems to occur when an article has reached a certain level of maturity and number of edits. Furthermore, references "also tend to be contributed by editors who have contributed more frequently and more substantially to an article, suggesting that a subset of more qualified or committed editors may exist for each article" (Chen and Roth, 2012). Despite this approach, Wikipedia's contents vary greatly in quality, including across languages (Roy et al., 2021).

The automatic improvement of Wikipedia's contents is an area of active research. While bots already patrol and improve the quality of Wikipedia's references (Zagorova et al., 2021), recent work also focused on automatically flagging sentences in need for a citation (Redi et al., 2019) and on assessing a source's reliability (Lewoniewski et al., 2020).

Using Wikipedia Given its broad scope in contents, the usage of Wikipedia varies greatly too: "for instance, we observe long and fast-paced page sequences across topics for users who are bored or exploring randomly, whereas those using Wikipedia for work or school spend more time on individual articles focused on topics such as science"(Singer et al., 2017). The usage of Wikipedia is even more significant when considering countries with varied languages and socio-economic characteristics (Lemmerich et al., 2019). Wikipedia also fulfills a specific role within the broader Web: it serves as a stepping stone between search engines and third-party websites (Piccardi et al., 2021). The complementarity of Wikipedia and search engines, such as Google, is particularly significant for scientific information seeking (Mesgari et al., 2015).

Previous studies have explored the use and usability of Wikipedia as a source of biomedical information. A randomized controlled trial focused on placing evidence of the effects of treatments (in this case, for schizophrenia) within Wikipedia pages finds no significant changes in the full-text accesses of the treated pages, but an effect in their altmetric scores (Adams et al., 2020). The readability of the most viewed Wikipedia articles on diseases is of varying quality, with many articles still

too difficult to read for a general readership (Brezar and Heilman, 2019). More generally, while Wikipedia is a prominent health information source in terms of views and visibility (e.g., often top in Google searches), the study of its impact in this respect is still too limited to draw any general conclusion (Smith, 2020).

A distinct set of recent studies explored the use of references in Wikipedia, and in particular references to external sources. Work on WikiProject Medicine shows that its readers appear to use links to external sources to verify and authorize Wikipedia content, rather than to examine the sources themselves (Maggio et al., 2020). A Wikipedia-wide study of engagement with external references found that users click on them only very rarely (once for every 300 page views, on average) (Piccardi et al., 2020). Crucially, users more often look for more when reading lower quality and shorter articles, which possibly do not contain what they seek (Piccardi et al., 2020). These findings further underline the importance of providing high-quality contents within Wikipedia itself.

Science and Wikipedia Wikipedia strives to convey information grounded in scientific results. It thus provides visibility to scientific research, and is considered an altmetric source in this respect (Sugimoto et al., 2017), yet the influence goes both ways. In fact, previous work has established that being cited from Wikipedia can increase the citation impact of an article (Thompson and Hanley, 2018). Understanding and monitoring which scientific results underpin Wikipedia's contents, and why, is therefore of critical importance.

Recent work is gradually improving our understanding of the matter. The open release of datasets of citations from Wikipedia to its sources is helping in broadening access to the essential data for tackling the question at hand (Singh et al., 2021; Zagorova et al., 2021). Furthermore, several previous studies have been able to rely on altmetric data. Some trends clearly emerge from this literature. Journal articles cited from Wikipedia are more likely than average published in high-impact journals (e.g., by impact factor), and in open access (Nielsen, 2007; Teplitskiy et al., 2017). Articles cited from Wikipedia are 'uncited' and untested by subsequent studies in rates proportional to the rest of the scientific literature, nevertheless they also receive a higher rate of supporting citations (Nicholson et al., 2021).

Wikipedia's capacity to rapidly and reliably integrate novel scientific results to respond to ongoing public events or crises has also been assessed. The COVID-19 pandemic provides a recent example (Colavizza, 2020). Most notably, the areas of Wikipedia where the editors are highly organized and include domain experts, for example several WikiProjects, appear to fare better in this respect. Indeed, the scope of expert involvement in editing Wikipedia is substantial. A recent study found that approximately 10%–30% of Wikipedia's contributors have substantial subject-matter expertise in the topics that they edit (Yarovoy et al., 2020).

Recent results by Arroyo-Machado et al. (2020), extending previous work by

the same team (Torres-Salinas et al., 2019), directly relate to our study. The authors perform a co-citation analysis of Wikipedia's sources, relying on a dataset of "847,512 references made by 193,802 Wikipedia articles to 598,746 scientific articles belonging to 14,149 journals indexed in Scopus." They use Altmetrics data to retrieve Wikipedia citations to journal articles. They study the co-citation network of journals and Scopus main field categories, as referenced by Wikipedia. The most significant scientific domains cited by Wikipedia include Medicine, Biochemistry, Genetics, and Molecular Biology. They confirm that the most important journals are multidisciplinary, and include prominent venues such as Nature, Science, PNAS, and that articles from high-impact factor journals are more likely to be cited from Wikipedia. Lastly, they find that only 13.44% of Wikipedia citations are to open access journals. Directly expanding upon Arroyo-Machado et al. (2020), we use here a larger and more recent dataset, considering the more granular level of analysis of journal articles, and we deepen the analysis by comparing the bibliographic coupling network of Wikipedia articles with the co-citation network of journal articles.

In conclusion, it is also worth noting that understanding how Wikipedia is structured according to how it uses scientific publications complements work relying on its internal link network. For example, previous results have found that different scientific domains possess distinct internal link network organizations, with modular structures for Biology and Medicine, but a sparse structure for Mathematics and a dense core for Physics (Silva et al., 2011).

4.3 Data and Methods

We assemble data from a variety of sources to perform our study.

Wikipedia Citations We use *Wikipedia Citations* as our main dataset (Singh et al., 2021). It consists of more than 29M citations extracted from the over 6M articles composing the English Wikipedia as of May 2020. In *Wikipedia Citations*, each citation is automatically classified as being to a book, journal article or Web content. Approximately 2.5M citations are classified as to a journal article, of which 1,705,085 are equipped with a DOI, either from Wikipedia itself or retrieved from Crossref. These citations to journal articles come from 405,358 distinct Wikipedia article pages and refer to 1,157,571 distinct DOIs. Citations to journal articles clearly comprise a relatively limited share of all citations contained in Wikipedia, and thus likely serve a specific purpose (Singh et al., 2021). A large share of these citations are given to articles published over the past 20 years, and the most cited journals include Nature, Science, the Journal of Biological Chemistry and PNAS. We use this set of citations in what follows.

ORES Topics, WikiProjects, Dimensions In order to perform our analysis, we enrich the *Wikipedia Citations* dataset with other sources of information. Firstly, we equip Wikipedia articles with information about their topics and the WikiProject they belong to, if any. The topics of Wikipedia articles are retrieved using the ORES Web service[1], which exposes a topic model of Wikipedia trained using Language-Agnostic Topic Classification (LATC) (Johnson et al., 2021) and assigns each Wikipedia article to a taxonomy rooted into the four categories of: Geography, Culture, History and Society, and STEM.[2] Through the ORES API, we could extract topics for Wikipedia articles covering 99.7% of the citations in *Wikipedia Citations*.

Further, we equip Wikipedia articles with information on their WikiProject. "A WikiProject is a group of contributors who want to work together as a team to improve Wikipedia"[3], for example to focus on a specific topic area such as WikiProject Mathematics or WikiProject India, or curate a specific aspect of the encyclopedia, for instance WikiProject Disambiguation. The English Wikipedia currently includes over 2,000 WikiProjects. Using public data (Johnson and Halfaker, 2020), we could equip with WikiProjects information Wikipedia articles comprising 96.3% of the citations in *Wikipedia Citations*.

Finally, we use Dimensions (Herzog et al., 2020) to retrieve metadata for all the journal articles with a DOI cited from Wikipedia. While no ideal single bibliographic data source exists yet, Dimensions provides for broad source coverage and relies in substantial part on the open Crossref repository, making it a meaningful choice for our study (Visser et al., 2021). What is more, we are further interested in an article's Field of Research (FOR) classification.[4] The Fields of Research are organized into hierarchies, with divisions (top, largest), groups and fields (bottom, smallest). Dimensions exposes divisions as *major fields*, and groups as *minor fields*; we adopt this naming convention in what follows. By querying the Dimensions' API we were able to match 96% of all the unique DOIs from *Wikipedia Citations*. All these data were retrieved in June 2021.

4.3.1 Methods

For our study we make a comparison between two undirected networks in turn extracted from the directed citation network of Wikipedia to journal articles: the co-citation network among scientific journal articles, and the bibliographic coupling network among Wikipedia articles. Using the conceptual framework

[1]We made use of the second API at the following address: `https://wiki-topic.toolforge.org/#lang-agnostic-model`.

[2]`https://www.mediawiki.org/wiki/ORES/Articletopic#Taxonomy`.

[3]`https://en.wikipedia.org/wiki/Wikipedia:WikiProject`.

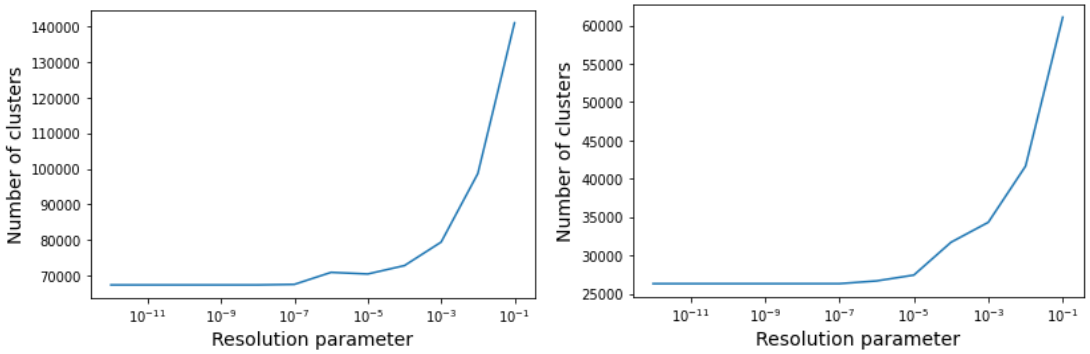
[4]The Fields of Research classification follows the research areas defined in the Australian and New Zealand Standard Research Classification (ANZSRC). See: `https://app.dimensions.ai/browse/categories/publication/for`.

of Costas et al. (2021), we can say that the co-citation network is made of co-Wikipedia citations whereby two journal articles are cited by the same Wikipedia article(s), and the bibliographic coupling network is composed of Wikipedia articles connected when they cite the same journal articles. Both networks are clustered using the Leiden algorithm (Traag et al., 2019), and their modular structures are compared. In this way, we aim to clarify which areas of Wikipedia rely on scientific sources. What is more, we seek to show how similar clusters of Wikipedia articles (bibliographic coupling) and similar clusters of co-cited scientific articles (co-citation) are internally related (RQ2).

Citation networks While constructing the co-citation network, we remove nodes (journal articles) that are cited only once from Wikipedia, as they would be isolated in the co-citation network since they are never co-cited. This results in a network of 1,050,686 nodes (91% of 1,157,571) and 17,916,861 edges. Similarly, we remove nodes (Wikipedia articles) that cite only one journal article, as they would be isolated in the bibliographic network. This gives a network of 257,452 nodes (64% of 405,358) and 27,473,262 edges. This bibliographic coupling network is thus not only denser, but it also does not include a higher share of isolated nodes than the co-citation network. Nodes in both networks are equipped with relevant metadata: WikiProject and ORES topics for the bibliographic coupling network, Fields of Research for the co-citation network. Furthermore, in both networks every edge is weighted according to how many times any two nodes are cited together (co-citation) or jointly cite the same items (bibliographic coupling).

Network clustering In order to equip our networks with a clustering solution, we use the Leiden algorithm (Traag et al., 2019), a popular choice that is fast and provides specific guarantees over the resulting clustering.[5] The Leiden algorithm uses a resolution (hyper)parameter to control the balance between the number of clusters and their size. A higher resolution parameter leads to more, smaller clusters. We find a reasonable value for the resolution parameter empirically, by inspecting the number of clusters at varying values of the parameter. Figure 4.1a shows the number of clusters at the varying size of the resolution parameter for the co-citation network, and Figure 4.1b does the same for the bibliographic coupling network. We avoid choosing extreme values, and settle for reasonable elbows which can be found, in both cases, at a value of the resolution parameter of $1e-4$.

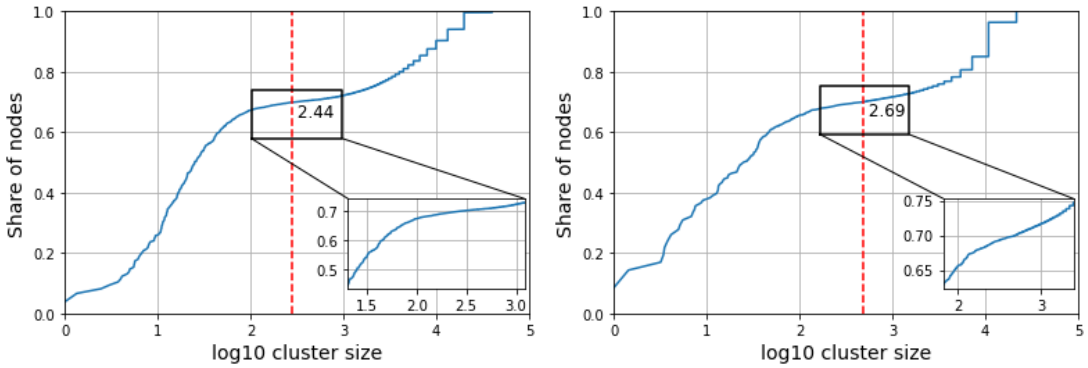
[5]For our analyses we relied on *igraph* 0.9.6 and *leidenalg* 0.8.3.



(a) Co-citation. (b) Bibliographic coupling.

Figure 4.1: Number of clusters at varying values of the resolution parameter.

Supernetworks In order to further coarsen our networks, we also construct 'supernetworks' of clusters. In a supernetwork, each node is a cluster of nodes from an underlying network. A supernetwork is therefore a network of clusters. A supernetwork is often easier to visualize and inspect, as it contains a much smaller number of nodes than the original network. We construct supernetworks by weighting each node with the number of nodes that a cluster contains, and by weighting each edge by summing the weights of the edges between the nodes of any two clusters. The use of aggregated metadata from the nodes of a cluster is detailed below, when discussing results. In order to identify and focus on the largest and most representative clusters, we further trim the supernetworks at a given cluster size threshold. We do so empirically, by inspecting the cumulative share of nodes that would be included in a supernetwork, at varying cluster size thresholds. Results are shown in Figure 4.2a (co-citation) and Figure 4.2b (bibliographic coupling). In both cases, reasonable cutoffs can be found at elbows corresponding to a cumulative share of included nodes of .7. This cutoff corresponds in turn to cluster sizes of 98 (co-citation) and 48 (bibliographic coupling), below which clusters are removed from the supernetworks.



(a) Co-citation. (b) Bibliographic coupling.

Figure 4.2: Cumulative share of nodes included in the supernetworks, per cluster size threshold. Clusters are ordered from largest (left) to smallest (right).

4.4 Results

We organise our results into three parts, addressing RQ1 in the first, and RQ2 in the second and third sections. First, we provide an overview of the scientific sources cited from Wikipedia. Next we analyse how Wikipedia articles are organised accordingly using bibliographic coupling networks. Lastly, we analyse how journal articles cited from Wikipedia are organised in turn, using co-citation networks.

4.4.1 Science in Wikipedia

Table 4.1: Number of citations and of journal articles per Field of Research (major fields), using fractional counting.

Field of Research (major field)	**Citations**	**Journal articles**
06 Biological Sciences	440,029.8 (25.8%)	255,382.0 (22.1%)
11 Medical and Health Sciences	371,304.7 (21.8%)	278,289.3 (24.0%)
04 Earth Sciences	69,835.7 (4.1%)	42,865.6 (3.7%)
02 Physical Sciences	69,678.3 (4.1%)	42,405.1 (3.7%)
03 Chemical Sciences	67,931.0 (4.0%)	53,396.7 (4.6%)
17 Psychology and Cognitive Sciences	64,092.0 (3.8%)	48,072.9 (4.2%)
21 History and Archaeology	63,292.9 (3.7%)	40,855.7 (3.5%)
16 Studies in Human Society	57,414.7 (3.4%)	40,640.1 (3.5%)
09 Engineering	42,994.5 (2.5%)	34,626.0 (3.0%)
01 Mathematical Sciences	41,408.8 (2.4%)	31,755.2 (2.7%)
08 Information and Computing Sciences	33,286.9 (2.0%)	25,300.0 (2.2%)
20 Language, Communication and Culture	31,969.4 (1.9%)	23,256.9 (2.0%)
05 Environmental Sciences	22,993.5 (1.3%)	15,069.7 (1.3%)
22 Philosophy and Religious Studies	21,448.8 (1.3%)	15,654.1 (1.4%)
14 Economics	18,230.9 (1.1%)	13,720.5 (1.2%)
07 Agricultural and Veterinary Sciences	15,525.6 (0.9%)	12,146.2 (1.0%)
15 Commerce, Management, Tourism and Services	12,691.6 (0.7%)	10,146.7 (0.9%)
13 Education	10,534.9 (0.6%)	8,714.2 (0.8%)
19 Studies in Creative Arts and Writing	10,304.9 (0.6%)	8,085.8 (0.7%)
18 Law and Legal Studies	9,971.6 (0.6%)	7,189.6 (0.6%)
10 Technology	5,740.8 (0.3%)	4,433.4 (0.4%)
12 Built Environment and Design	2,886.5 (0.2%)	2,274.6 (0.2%)
Missing	221,517.0 (13.0%)	143,291.0 (12.4%)
Total	**1,705,085 (100%)**	**1,157,571 (100%)**

Fields of Research Table 4.1 shows the number of citations and the number of journal articles cited from Wikipedia per Dimension's major Field of Research. Given that journal articles (and, later on, Wikipedia articles) are our units of analysis, we use fractional counting to account for an article belonging to multiple categories (Perianes-Rodriguez et al., 2016), unless otherwise specified. It can be seen that almost half of citations are given to articles in the fields of biological, medical and health sciences, which in turn make up about 46% of all journal articles cited from Wikipedia. More in general, STEM fields make up for a large part of both citations and articles, while the most represented areas outside of STEM are history and sociology. Looking at minor Fields of Research, in Table 4.2 (only the top 10 are shown), we can appreciate how the most cited fields include genetics and cell biology, evolutionary biology and ecology in biology; clinical science and public health, psychology and neurosciences in medicine; history in the humanities. These fields capture over a fifth of all citations from Wikipedia.

Table 4.2: Number of citations and of journal articles per Field of Research (minor fields), using fractional counting.

Field of Research (minor fields)	**Fractional counting**	**Unique DOIs number**
0604 Genetics	76,299.8 (4.5%)	39,393.3 (3.4%)
0601 Biochemistry and Cell Biology	71,505.5 (4.2%)	45,098.7 (3.9%)
1103 Clinical Sciences	49,352.5 (2.9%)	38,700.8 (3.3%)
1117 Public Health and Health Services	32,066.8 (1.9%)	23,282.4 (2.0%)
1701 Psychology	27,375.5 (1.6%)	20,623.2 (1.8%)
1109 Neurosciences	25,030.1 (1.5%)	19,117.9 (1.7%)
2103 Historical Studies	24,134.6 (1.4%)	15,873.0 (1.4%)
0403 Geology	23,114.6 (1.4%)	13,464.6 (1.2%)
0602 Ecology	23,033.6 (1.4%)	14,862.2 (1.3%)
0603 Evolutionary Biology	18,246.4 (1.1%)	8,667.2 (0.7%)

Citation counts We assess the distribution of (journal article) citations given to articles cited from Wikipedia, using data from Dimensions. There is a wide variation in citations counts, with many articles with no or few citations, and some with a high number. The maximum number of received citations is 214,886 and the minimum is 0, with a mean at 189 and a median at 33. When considering recent citations (over the past two years), the maximum is 34,845, the minimum is 0, with a mean at 36.4 and a median at 5. Furthermore, 60% of articles cited from Wikipedia received fewer than 10 citations in the past two years. Articles cited fewer than 100 times account for 70% of the total cited articles, and only about 3% of articles are cited 1,000 times or more. It is worth noting that 10.4% of journal articles are never cited and 22.5% have no recent citations either, according to Dimensions. Finally, we note that 75,248 (4.4%) articles were missing this information from Dimensions.

Most cited articles and journals Among the top journal articles by a number of received citations (See Appendix, Section A.1), only two are open access, six are authored by scholars based in the United States, and only one is published after 2000. The most cited article is "Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4", published in Nature in 1970, and cited 214,886 times according to Dimensions. Instead, the most cited article in terms of recent citations (past two years) is "Deep Residual Learning for Image Recognition", published at the 2016 Conference on Computer Vision and Pattern Recognition, with 34,845 recent citations. The top citing Wikipedia articles, instead, frequently include reference or survey articles. For example, the series "this year in" is often a source of highly citing Wikipedia articles (e.g., "2018 in paleontology" includes 580 citations to journal articles). In Table 4.3, we list the most cited journals in Wikipedia. Nature, PNAS and Science top the list, confirming findings from previous work (Arroyo-Machado et al., 2020).

Table 4.3: Most cited journals

Journal name	**Citations**
Nature	37,287
PNAS	31,801
Science	26,903
Journal of Biological Chemistry	24,518
PLOS ONE	12,997
Zootaxa	10,006
Cell	9,318
Genome Research	8,961
The Astrophysical Journal	8,882
Astronomy & Astrophysics	7,292

Age of articles and open access In Figure A.1, we show the distribution of the publication years for journal articles cited from Wikipedia. The trend is clear: most cited articles were published in the past two decades. We also assess the open access availability of these journal articles using Dimensions data. We find that 686,952 (41%) articles are available in some form of open access, while a majority number of 942,885 (55%) remain closed access, and for 75,248 (4%) this information is missing. This constitutes a considerably higher fraction of cited OA articles than what previous work has found (Arroyo-Machado et al., 2020).

Citation flows In conclusion of this first overview section, we analyse the flow of citations from Wikipedia to journal articles, on the one hand, grouping Wikipedia articles by ORES topics and WikiProjects, and on the other hand, grouping journal articles by major FOR categories. We show the river plots of citations from Wikipedia articles by ORES topics in Figure 4.3 and by (top-10) WikiProjects in Figure 4.4. The flow of citations from STEM Wikipedia articles confirms the importance of the biological, medical and health sciences in Wikipedia, while other topics are more evenly distributed across fields of research. When considering WikiProjects, we only kept the top-10 by (fractional) number of Wikipedia articles, which have a strong STEM focus. Partially as a consequence, the flow of citations again favours the biological, medical and health sciences, but also distributes across other STEM fields of research. It is worth highlighting the role of the WikiProject Biography, which focuses on the biographies of notable persons. This project spans across fields of research, for example by covering the lives of prominent scientists, and connects them with historical fields in turn.

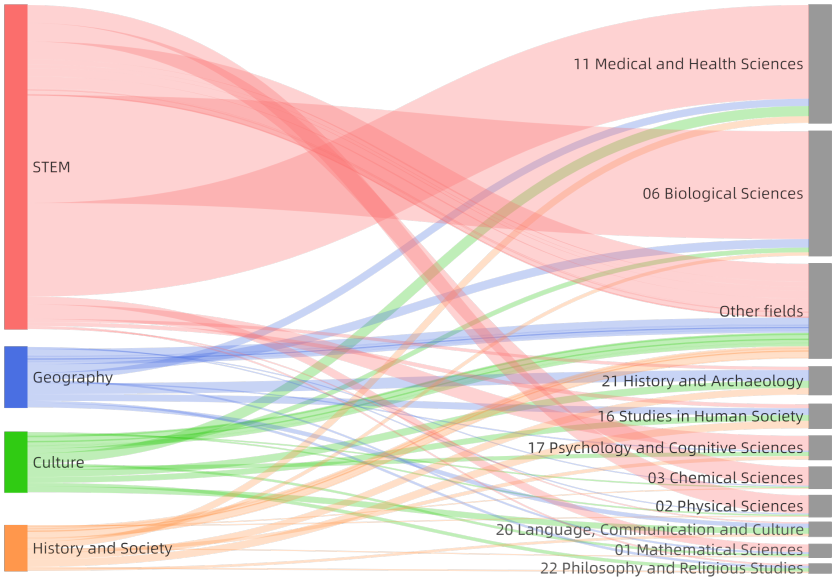


Figure 4.3: Citation flow from ORES topics to major fields of research.

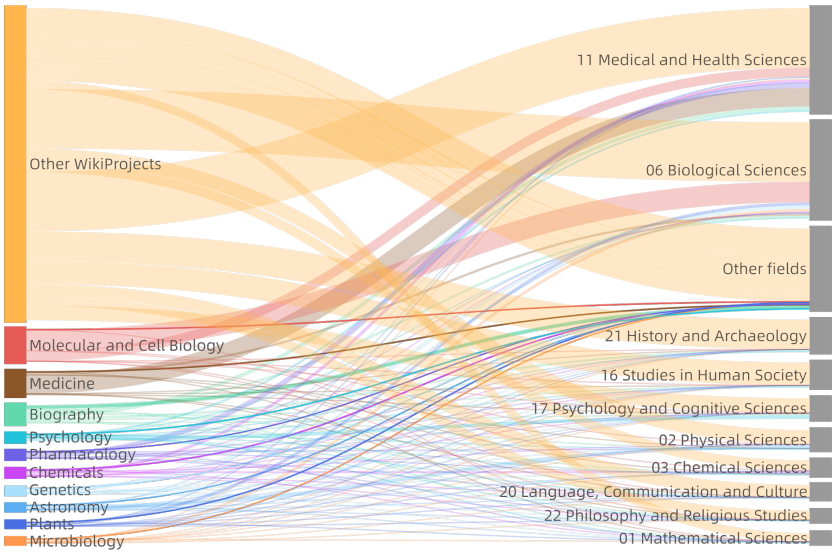


Figure 4.4: Citation flow from the top-10 WikiProjects to major fields of research.

4.4.2 Wikipedia from a Science Perspective

Bibliographic coupling network The bibliographic coupling network of Wikipedia articles includes 257,452 nodes and 27,473,262 edges, with a density of 0.0008. The corresponding supernetwork has 31,642 nodes (clusters) and 9,158 edges, for a density of 0.000018. We further aggregate information about

ORES topics and WikiProjects to the supernetwork using fractional counting. In Figure A.2 we show the distribution of the cluster sizes of the underlying bibliographic coupling network, distinguishing between isolated and connected clusters. A large number of small, isolated clusters exists, while at higher cluster sizes the clusters become increasingly connected.

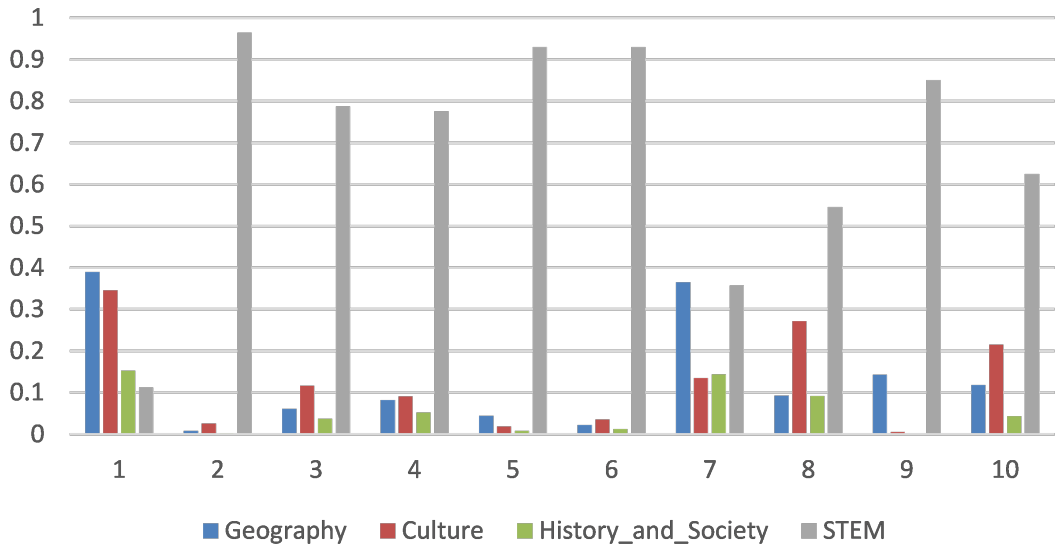


Figure 4.5: Distribution of ORES topics in the 10 largest clusters of the bibliographic coupling network of Wikipedia articles.

Largest clusters Next, we zoom-in the largest clusters of this network, showing their ORES topic distribution in Figure 4.5, and their distribution over WikiProjects in Figure 4.6. The combination of topics and projects is helpful in uncovering what a cluster is about. For example, the largest cluster is devoted to biographies and has a relatively balanced topic distribution across geography, history and society, culture. Importantly though, its top WikiProjects include military history, United States and football, suggesting a certain skew in this respect. Most clusters are, instead, specialized in specific STEM fields. For example, cluster 2 is focused on molecular and cell biology, and cluster 9 on entomology (spiders, in particular). The dominance of STEM in 'journal-article-citing' Wikipedia is further confirmed. Exceptions include history and biographies, and archaeology/physical anthropology (cluster 7).

Network visualization We conclude by a visual comparison of the bibliographic coupling supernetwork using the same layout but different coloring according to its modularity class (Figure A.4), ORES topics (Figure A.5), and WikiProjects (Figure A.6). These visualizations confirm two patterns we have previously

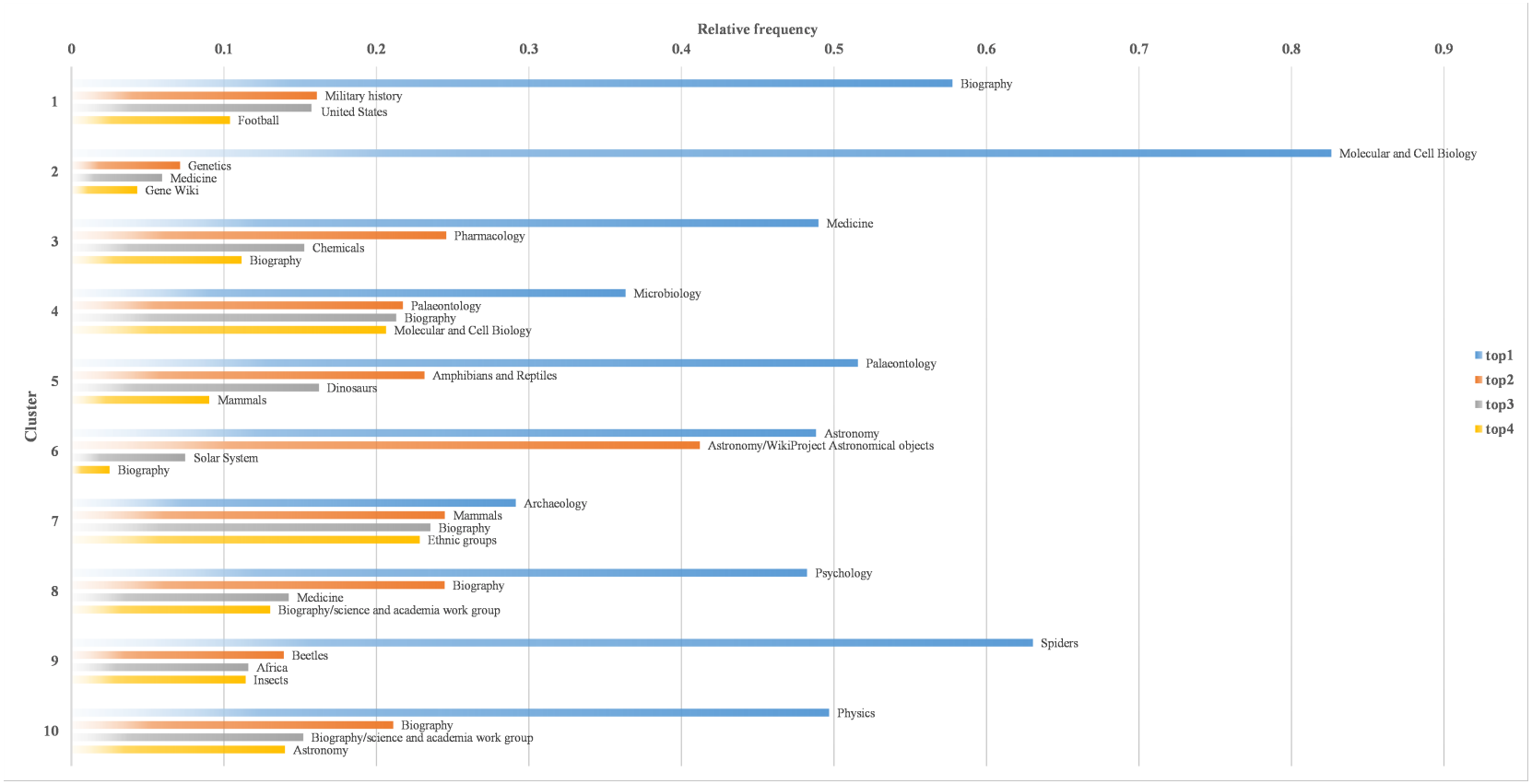


Figure 4.6: Distribution of WikiProjects in the 10 largest clusters of the bibliographic coupling network of Wikipedia articles.

highlighted: a) the important role of biographies in connecting this part of Wikipedia (first visualization); b) the systematic importance of STEM (second visualization). Furthermore, we can also appreciate the apparently effective role played by WikiProjects in helping editors coordinate for the curation of specialized knowledge in a coherent whole (third visualization).

4.4.3 Science from a Wikipedia Perspective

Co-citation network The co-citation network of journal articles cited from Wikipedia has 1,050,686 nodes and 17,916,861 edges, with a density of around 0.00003. The resulting supernetwork has 71,983 nodes (clusters) and 23,668 edges, and a density of 0.00001. The distribution of small to large cluster sizes and the fraction of isolated clusters is, consequently, even more pronounced than for the bibliographic coupling network, as it can be seen from Figure A.3.

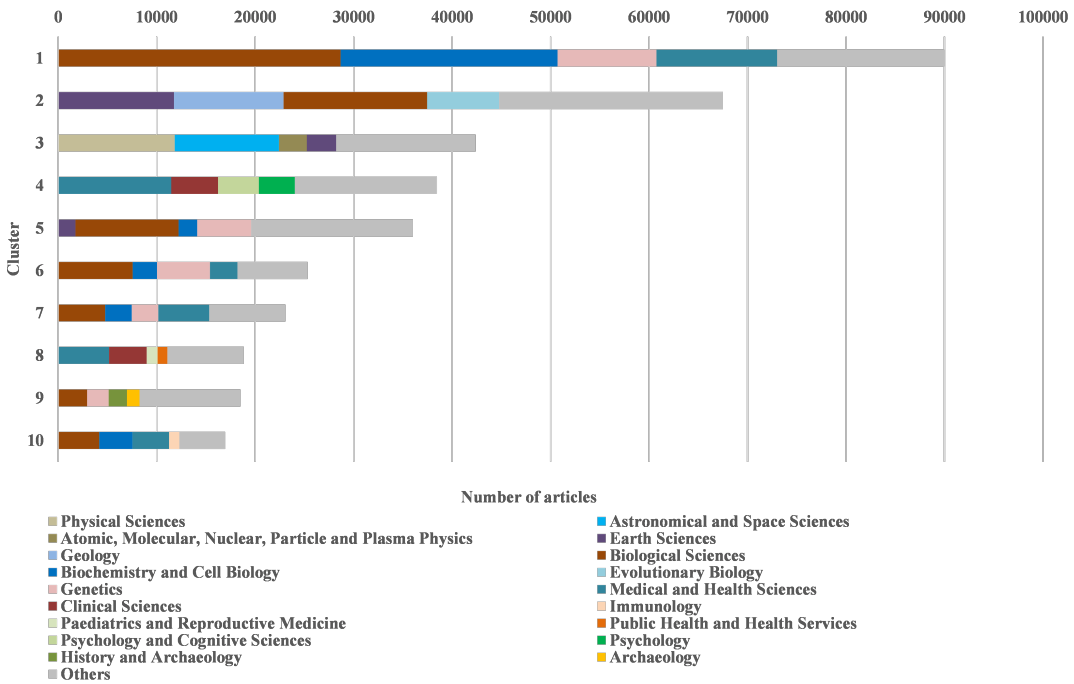


Figure 4.7: Distribution of major Fields of Research in the 10 largest clusters of the co-citation network of journal articles. The categories in each bar are ordered by color (in the legend), from left to right, top to bottom. For example, cluster one top major fields of research include (in order of appearance in the legend): Biological Sciences, Biochemistry and Cell Biology, Genetics, Medical and Health Sciences, Other.

Largest clusters In Figure 4.7 we show the distribution of the top major Fields of Research per top-10 cluster in the co-citation network. This result

strongly confirms the dominance of biology and medicine as the two top fields cited from Wikipedia. Virtually all top clusters include a dominant or at least sizable fraction of citations to these two fields, with the exception of the second cluster which contains a sizeable share of contributions in earth sciences, geology, and evolutionary biology, and the third one, focused on physics and astronomy.

Network visualization Also for the co-citation supernetwork we visualize it using the same layout and different coloring, respectively by modularity class (Figure A.7) and top major Field of Research (Figure A.8). The results are consistent with what we previously discussed. Furthermore, the two top clusters, which from Figure 4.7 we know being roughly focused on cell biology and genetics (cluster 1) and physical anthropology (cluster 2), group differently by modularity, with medicine, psychology and clinical sciences showing a stronger connection to the former than the latter.

4.5 Discussion

We provided a map of science in Wikipedia highlighting clear trends, while rising some further questions. Firstly, our first research question asked which scientific sources are cited from Wikipedia and what are their characteristics. The dominant role of STEM literature clearly emerges, with biology and medicine as the top fields, followed by earth sciences, physics and astronomy. Furthermore, a sizeable fraction of this literature is well cited and published in notorious venues. These results confirm previous studies conducted on smaller datasets (Colavizza, 2020) or at the journal level of analysis (Arroyo-Machado et al., 2020).

We were able to delve deeper as well, in view of our second research question asking which areas of Wikipedia rely on scientific sources and how do they relate. Firstly, the seemingly marginal role of journal articles from non-STEM fields is attenuated by the connecting role of biography articles in Wikipedia, which effectively bridge history, geography and culture with STEM topics. Secondly, the bibliographic coupling network of Wikipedia articles is not only smaller, but also better connected than the co-citation network of journal articles cited from Wikipedia. This might showcase the consolidating encyclopedic role of Wikipedia, as well as the positive impact of WikiProjects as a means to coordinate editorial efforts.

A set of new questions emerge from our work. To begin with, on the characteristics of journal articles cited from Wikipedia: we have found that on average these are well cited outside of Wikipedia too, yet many remain poorly cited or even not cited at all – for example, 70% of journal articles cited from Wikipedia received fewer than 100 citations from other journal articles at the time of the study. Furthermore, the fraction of Open Access articles cited from Wikipedia appears relatively high even though not dominant. Lastly, the age of such jour-

nal articles is mostly distributed within the past 20 years, possibly hinting at a chronological debt in Wikipedia: several citations might have been current when most of Wikipedia articles were created, but might no longer be up to date. Wikipedia is in this respect a 'slow altmetric indicator', which takes time to accumulate (Fang and Costas, 2020). Secondly, our network analysis finds interesting clusters of articles which warrant further study. How the central role of the largest fields of research (biology, medicine) articulates with the networks' 'periphery', in particular, remains an open question.

Some of the limitations of our study constitute possible directions for future work as well. The most important one is that we only considered citations to journal articles. Adding other cited sources, such as books and Web contents, would complement the map which we provided here. Secondly, the dataset we used constitutes a snapshot of Wikipedia at a certain point in time: a study of citations including time information would provide for a clearer picture of the dynamics of negotiation and consolidation of knowledge in Wikipedia. Lastly, Wikipedia's internal structure can also be mapped using information such as the internal link structure or the textual similarity of the articles. A comparison of the citations networks we studies here with these would further enrich the map of Wikipedia.

4.6 Conclusion

In this chapter, we mapped the organization of Wikipedia according to its use of scientific journal articles via citations. We made use of a recent dataset of citations from Wikipedia, and relied on network analysis techniques, in particular network clustering. We were able to show that Wikipedia heavily relies on scientific contents from biology, medicine and a handful of other STEM fields, including physics and earth sciences. Journal articles cited from Wikipedia are, on average, well-cited, published in notorious journals such as Nature or Science, and have been published over the past 20 years. While non-STEM fields are only marginally represented in journal articles cited from Wikipedia, they play an important connecting role via Wikipedia's biographies. This is but an example of how Wikipedia is able to interconnect knowledge across scientific fields and also with other non-scientific topics. In this respect, the most interesting future work which awaits us is the extension of this map of science in Wikipedia to include books, Web contents and all other sources cited in Wikipedia.

These findings show that Wikipedia privileges certain domains of science, while also connecting them to broader areas of human knowledge. This map of Wikipedia's scientific coverage highlights both the strengths and imbalances of the platform. At the same time, Wikipedia's knowledge infrastructure relies not only on scientific literature but also on other types of sources. Among these, news media outlets are particularly influential, as they provide coverage of current

events and often serve as key references in politically or socially sensitive topics.

Having examined the role of academic publications in shaping Wikipedia's scientific content, the next chapter turns to news media sources. There, I analyze how their political polarization and factual reliability may affect Wikipedia's commitment to neutrality and knowledge integrity.

Chapter 5

Polarization and Reliability of News Sources in Wikipedia

5.1 Introduction

Building on Chapter 4, which examined the role of scientific publications in structuring Wikipedia's knowledge base, this chapter shifts attention to news media, another cornerstone of Wikipedia's sourcing practices. Unlike journal articles, news outlets are more immediately responsive to events and central to covering current affairs. At the same time, they carry risks of political polarization and bias, which makes them a crucial site for investigating Wikipedia's knowledge integrity.

Wikipedia is one of the most extensive encyclopedias worldwide, providing an open go-to reference for reliable online content and a key hub to the Web (Piccardi et al., 2020). Wikipedia's articles are contributed by volunteers, following the policies of taking a neutral point of view (NPOV), verifiability of facts and sources, and contributing no original research. In principle, all Wikipedia articles should be "based on reliable, independent, published sources with a reputation for fact-checking and accuracy".[1] Sources are usually cited in footnotes and references. News media outlets provide a sizeable share of Wikipedia's cited sources, yet they often contain both factual contents and opinions or viewpoints around them (Fetahu et al., 2015). News reporting from well-established outlets is generally considered reliable for statements of fact. However, a potential for viewpoint bias remains and may affect the integrity of knowledge in Wikipedia, or at least Wikipedia's neutral point of view. While millions of volunteer contributors create and maintain free knowledge in Wikipedia (Aragón and Saez-Trumper, 2021), new challenges emerge in terms of information quality and reliability (Saez-Trumper, 2019; Morgan, 2019). What is more, news media often polarize around

[1] `https://en.wikipedia.org/wiki/Wikipedia:Verifiability#What_counts_as_a_reliable_source`.

opposite political viewpoints (Patterson, 2011; Sutter, 2000). While previous work has focused on assessing the reliability of Wikipedia contents (Przybyla et al., 2022) and editors' possible biases (Rogers and Sendijarevic, 2012; Yasseri et al., 2013), researchers still have to systematically investigate Wikipedia's knowledge integrity (Wikipedia, 2022; Sugandhika and Ahangama, 2022). We contribute to this line of work by exploring the political polarization of news media sources in Wikipedia. We further assess their reliability and determine whether there is a relationship between the two effects. To this end, we ask the following research questions:

1. RQ1: Is there political polarization in the news media sources cited in Wikipedia?

2. RQ2: What factors influence news media polarization in Wikipedia? Specifically, is there a relationship between news media political polarization and factual reliability?

In order to answer these questions, we rely on the large-scale dataset *Wikipedia Citations* (Singh et al., 2021); we use third-party sources to estimate the political polarization and reliability of news media outlets: the Media Bias Monitor (MBM) and the Media Bias Fact Check (MBCF). Following the approach taken by MBM, we consider political polarization across a mono-dimensional spectrum between the liberal left and conservative right, acknowledging this as a limitation. Reliability conveys an estimate of the factual correctness of an outlet, in terms of contents and framing. We speculate that reliability and polarization might be related, for example with media outlets closer to a given political leaning being considered more reliable on average. Firstly, we provide a quantitative overview of news media sources' political polarization in Wikipedia (RQ1); secondly, we make use of regression analysis to clarify the relationship between news media source polarization, on the one hand, and an article's topic, WikiProject and a source factual reliability, on the other hand (RQ2). Our aim is to inform Wikipedia's knowledge integrity agenda (Taraborelli, 2019) by rising awareness on possible biasing effects in Wikipedia's sources.

5.2 Previous Work

5.2.1 Wikipedia's Core Policies

Wikipedia strives to take a neutral viewpoint and provide reliable contents (Mesgari et al., 2015). To this end, Wikipedia abides to three core content policies:

1. Neutral Point of View (NPOV): "representing fairly, proportionately, and, as far as possible, without editorial bias, all of the significant views that have

been published by reliable sources on a topic."[2]

2. Verifiability: "other people using the encyclopedia can check that the information comes from a reliable source."[3]

3. No original research: "Wikipedia articles must not contain original research."[4]

These three policies could help to improve Wikipedia's article quality (Pavalanathan et al., 2018) and could enable us to collectively address many of the practical issues stemming from collaboratively curating encyclopedic content (Arazy et al., 2006). However, from an epistemic perspective, they lay the responsibility for assessing content quality and reliability to third parties via reliable sources (Saez-Trumper, 2019). On the one hand, reliable sources are essential to Wikipedia's status of a neutral encyclopedia, yet on the other hand the selection of sources invariably leads to controversies (Borra et al., 2014) and even edit wars (Sumi et al., 2011). To be sure, this might largely be a feature as some researchers believe that the existence of such controversies ultimately leads to better quality articles (Shi et al., 2019).

5.2.2 Knowledge Integrity in Wikipedia

As one of the main repositories of free knowledge available today, Wikipedia plays a central role on the Web (Arazy et al., 2006; Smith, 2020). Its very widespread usage and radical openness to readers and contributors make Wikipedia vulnerable to malicious information attacks and disinformation (Saez-Trumper, 2019), which in turn could compromise Wikipedia's knowledge integrity (Aragón and Saez-Trumper, 2021).

Knowledge Integrity is one of the research priorities individuated by Wikimedia Research, whose aim is to identify and address threats to contents in Wikipedia, to increase the capabilities of patrollers, and to provide mechanisms for assessing the reliability of sources (Taraborelli, 2019). As of August 2022, Wikipedia is active in 318 language versions and each version is maintained by a dedicated (or language-specific) community [5], thus, knowledge integrity risks can arise in many different forms. If we compare the size of the active editor communities with the scale of the Wikipedia project, it is clear that resources for patrolling and verifying contents remain on high demand (Morgan, 2019; Saez-Trumper, 2019). Besides, a lack of geographical diversity might favor nationalistic biases (Sato, 2021). From the perspective of contents, disputes between community members due to disagreements about the content of articles (Rogers and Sendijarevic, 2012; Yasseri et al., 2013), content verifiability (Lewoniewski et al., 2019; Redi et al.,

[2] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view.
[3] https://en.wikipedia.org/wiki/Wikipedia:Verifiability.
[4] https://en.wikipedia.org/wiki/Wikipedia:No_original_research.
[5] https://en.wikipedia.org/wiki/List_of_Wikipedias#cite_note-3.

2019) and quality (Lewoniewski et al., 2017; Rogers and Sendijarevic, 2012) are significant and enduring aspects of Wikipedia.

5.2.3 Wikipedia's Sources

The 'verifiability' policy guarantees the existence of an important aspect of Wikipedia: citations (Kaffee and Elsahar, 2021). Citations serve several important roles: "they uphold intellectual honesty and reduce the risk of plagiarism, they attribute prior work and ideas to their authors, they allow the reader to independently determine whether the referenced material supports the statements made by an editor in Wikipedia, and thus they help the reader gauge the strength and validity of the material an editor has relied on."[6] However, evidence shows that references in Wikipedia are not too actively used by readers (Piccardi et al., 2020). In Wikipedia, scientific or scholarly literature takes up a large proportion of citations to sources (Nielsen et al., 2017; Singh et al., 2021), and Wikipedia's citation rates are often aligned with those in the scholarly literature (Shuai et al., 2013; Mesgari et al., 2015; Yang and Colavizza, 2022a). Although it has been found that Wikipedia can have an influence on scientific research (Thompson and Hanley, 2018), and some professional journalists have also begun to use Wikipedia in their work (Messner and South, 2011), the debate on using Wikipedia as a credible academic information resource is still active (Tomaszewski and MacDonald, 2016).

Despite the efforts of Wikipedia's contributors, many or even most articles in Wikipedia may still contain unsubstantiated or outdated claims, especially those flagged as being of lower quality (Lewoniewski et al., 2017). Sometimes editors' might not use citations systematically (Chen and Roth, 2012; Forte et al., 2018) or engage in polarized edit conflicts (Umarova and Mustafaraj, 2019). Research suggests that some editors' violations might be caused by biases (Hube, 2017), such as cultural (Callahan and Herring, 2011), political (Greenstein and Zhu, 2012a) or gender bias (Wagner et al., 2015).

Das and Lavoie (Das and Lavoie, 2014) determine the topics an editor is interested in and the editor's stance by editors' behaviour and interactions, finding that bias exists especially when a single point of view dominates controversial topics. Hube (Hube, 2017) provides a method to detect both explicit and implicit bias in Wikipedia articles and observe its evolution by analyzing language, editing and citation styles. Greenstein and Zhu (Greenstein and Zhu, 2012a) analyse political bias in Wikipedia by measuring the degree of political leaning of an article. They rely on a content-based method (Gentzkow and Shapiro, 2010) which calculates the frequency of particular phrases to measure the degree of political bias. They find that Wikipedia had a liberal bias in the early years, but that bias declines over time, supporting "a narrow interpretation of Linus' Law, namely, [that] a greater number of contributors to an article makes an article more

[6] `https://en.wikipedia.org/wiki/Citation`.

neutral". Besides, since Wikipedia has many language versions, different language versions can also contain specific biases. A study on the Wikipedia pages of UK politicians surfaced a substantial polarization of editors across political lines, in turn reflected in their choice of news media sources (Agarwal et al., 2020). Ewa and Susan (Callahan and Herring, 2011) find systematic biases in the focus on a particular topic or person in Wikipedia versions in different languages. Zhou et al. (Zhou et al., 2015) find that people's attention to war-related topics affects the number of words and the number of subjective concepts, which in turn affects the bias of emotional expression. Last but not least, several studies have analyzed gender differences in Wikipedia. Wagner et al. (Wagner et al., 2015) find that while women on Wikipedia are covered and featured well in many Wikipedia language editions, the way women are portrayed might differ from the way men are portrayed. Reagle and Rhue (Reagle and Rhue, 2011) study gender bias by comparing Wikipedia and Encyclopedia Britannica. They illustrate that while the number of articles related to women is increasing, compared to the articles on men, the articles on women are more likely to be missing on Wikipedia. Researchers have also made efforts to improve the verifiability of Wikipedia's contents for example by flagging unsupported contents in view of adding citations to reliable sources (Fetahu et al., 2016; Redi et al., 2019).

5.2.4 News Media Sources in Wikipedia

Wikipedia supports the use of sources from news media outlets: "news reporting from well-established news outlets is generally considered to be reliable for statements of fact."[7] News media sources are indeed among the most-used in Wikipedia. Fetahu et al. (Fetahu et al., 2015) find that almost 20% of the external references in the English version of Wikipedia are to news articles. In the dataset we use for this contribution such proportion is closer to 30%. Previous work also found that Wikipedia's news sources are overall factually reliable, yet not uniformly so (Yang and Colavizza, 2022b). Nielsen uses a Wikipedia dump from 2008 to find that the BBC, the New York Times, and the Washington Post were the most cited news media outlets at the time, with the BBC far ahead of the other outlets. Among the top 20 most-cited news outlets, most are American and four each being Australian and British (Nielsen, 2010). However, it is difficult to ignore the potential for polarization in news media (Patterson, 2011; Wolton, 2019), as well as their uneven reliability across the spectrum of outlets (Lazer et al., 2018). Previous work has focused on providing methods to assess the reliability of Wikipedia's content. Hube and Fetahu (Hube and Fetahu, 2018) proposed a supervised classification approach based on a self-build bias word lexicon, which could be able to detect biased statements with an accuracy of 74%. Przybyła et al. (Przybyla et al., 2022) collect a corpus of over 50 million citations to 24 million identified sources from

[7] `https://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources`.

Wikipedia Complete Citation Corpus (WCCC) and build a search index using multiple meaning representations, using NLP (Natural Language Processing) and ML (Machine Learning), enabling the automatic retrieval of sources to support or disprove a claim. While a considerable amount of work has been done to assess the polarization and reliability of Wikipedia's contents separately, their systematic and combined assessment for news media sources remains an open challenge.

5.3 Data

Our work is primarily based on *Wikipedia Citations*, a public dataset of citations from English Wikipedia to all its sources, including news media (Singh et al., 2021). We enrich *Wikipedia Citations* with data from the Media Bias Monitor (MBM) and the Media Bias Fact Check (MBFC): two authoritative indices of news media outlets providing an estimate of their political leaning and factual reliability. The combination of these sources allows us to quantify the political polarization and factual reliability of news sources cited from Wikipedia.

5.3.1 Wikipedia Citations

Wikipedia Citations includes more than 29M citations extracted from the over 6M articles English Wikipedia in May 2020. In *Wikipedia Citations*, each Wikipedia page contains several citations pointing to external sources. Of these, 25M (85.2%) are equipped with external links (URLs). For these URLs we extract the domain name using the *tldextract* package.[8] Domain names are critical in our approach as they allow linking to the Media Bias Monitor (MBM) and Media Bias Fact Check (MBFC) indices, which are based on Web domains and sub-domains. For example, given the external link *https://www.nbcnews.com/politics/politics-news/al-gore-compares-climate-deniers-uvalde-law-enforcement-officers-nobod-rcna39707*, after domain extraction we end up with *www.nbcnews.com*. With this method, we extract 1,554,632 unique domains. Then, we link domain names to MBM and MBFC by querying their APIs. The main limitation of this approach is that it works at the news media outlet level, which corresponds to the domain name, and not at the specific source level (the actual cited news article).

We further enrich *Wikipedia Citations* with information about an article's topics and WikiProjects. With the data from ORES Web service[9] and public data (Johnson and Halfaker, 2020), we equip Wikipedia articles with topic (coverage of 99.3%) and WikiProjects (coverage of 17.5%). In this paper, we also use fractional counting to account for an article belonging to multiple topics or projects at the same time.

[8] `https://pypi.org/project/tldextract`.

[9] `https://wiki-topic.toolforge.org/#lang-agnostic-model`.

5.3.2 Media Bias Monitor (MBM)

To estimate the political polarization of Wikipedia citations, we use the Media Bias Monitor (Ribeiro et al., 2018). This system collects demographic data about the Facebook followers of 20,448 distinct news media outlets via Facebook Graph API [10] and Facebook Marketing API [11]. These data include political leanings, gender, age, income, ethnicity and national identity. For political leanings, the Facebook Audience API [12] provides five levels: Very Conservative, Conservative, Moderate, Liberal, Very Liberal. To measure the political leaning of an outlet, MBM firstly finds the fraction of readers having different political leanings, and then multiply the fraction for each category with the following values: very liberal (–2), liberal (–1), moderate (0), conservative (1), and very conservative (2). The sum of such scores provides a single polarization score for the outlet, ranging between –2 and 2, where a negative score indicates that a media outlet is read more by a liberal leaning audience, while a positive score indicates a conservative leaning audience.

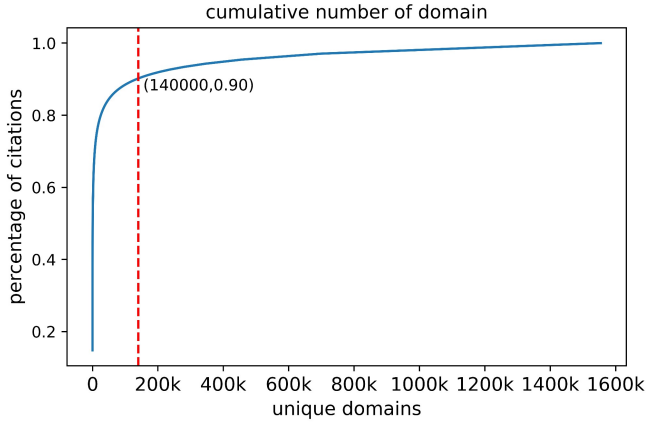


Figure 5.1: Cumulative distribution of Wikipedia citations to distinct domain names.

In the original paper, MBM is compared to alternative approaches used to infer the political leanings of news media outlets, finding that this method highly correlates with most alternatives. What is more, MBM covers 20,448 news media outlets and provides multi-dimension data. On these grounds, we use MBM in our study. As we have 1,554,632 unique domains from *Wikipedia Citations*, which

[10] `developers.facebook.com/docs/graph\protect\discretionary{\char\hyphenchar\font}{}{}api`.

[11] `developers.facebook.com/docs/marketing\protect\discretionary{\char\hyphenchar\font}{}{}apis`.

[12] `developers.facebook.com/docs/marketing\protect\discretionary{\char\hyphenchar\font}{}{}api/audiences\protect\discretionary{\char\hyphenchar\font}{}{}api`.

mostly are not news media outlets, we focus the matching of unique domains to MBM on the most frequently cited domains from Wikipedia. In Figure 5.1, we show that the top 140,000 unique domains cover up to 90% of all *Wikipedia Citations*. We thus decided to only keep these top 140,000 domain names for our study, and match them in MBM (Baigutanova et al., 2023b,a; Zheng et al., 2023; Valentim et al., 2021).

When looking a domain name up via the MBM API [13], we can get four different query results:

1. No match; example: *trove.nla.gov.au*.
2. One exact match; example: *www.breitbart.com*.
3. More than one result, including the exact match; example: *www.abc.com*.
4. More than one result, *not* including an exact match; example: *www.nytimes.com*.

In each case, we proceed as follows. We label domains without a match (result 1) as *NaN*, while we use the polarization score of the exact match for domains with result 2 or 3. For domains with result 4, we use the average polarization score of all the matches, under the assumption that this approximates the polarization score of the exact match. To test our assumption, we use the 1113 unique domains that have multiple results including an exact match (result 3), and compare the distribution of exact polarization scores and average polarization scores in Figure 5.2. We can see that the two distributions are overall comparable, which supports our assumption. Following this procedure we are able to equip 4,866,377 citations (16.6% out of a total of 29.3M) with polarization scores. These 4.9M citations are all the citations in Figure 5.3b, while 29.3M are 100% of citations in Figure 5.3a. We note again that 29.3M is the total number of citations in the dataset, while citations to news media sources are estimated at 8.9M (Yang and Colavizza, 2022b).

5.3.3 Media Bias Fact Check (MBFC)

To answer our research questions, we not only need the political polarization of a news media outlet, but also an estimate of its factual reliability. In order to get the reliability data we use Media Bias Fact Check, which offers the largest set of labels of any news source rating service (Bozarth et al., 2020). For each news media outlet, a minimum of 10 headlines are reviewed and a minimum of 5 news stories are reviewed to get a reasonable factual rating [14]. MBFC classifies reliability in 6 levels: VERY HIGH (a score of 0), which means that the source is considered to be always factual; HIGH (a score of 1 to 2), which means that the

[13] `https://twitter-app.mpi-sws.org/media-bias-monitor/index.php`.
[14] `https://mediabiasfactcheck.com/methodology`.

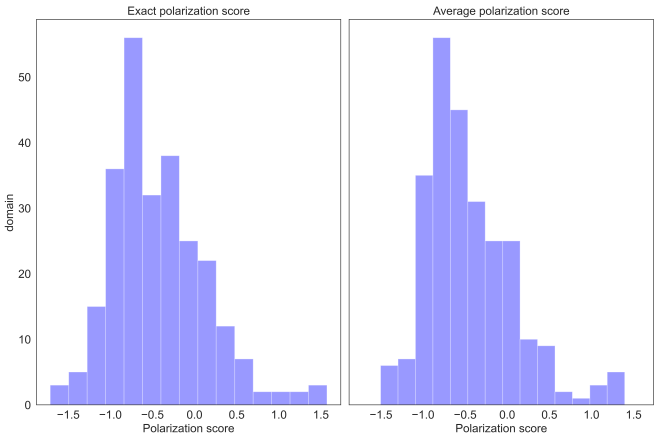
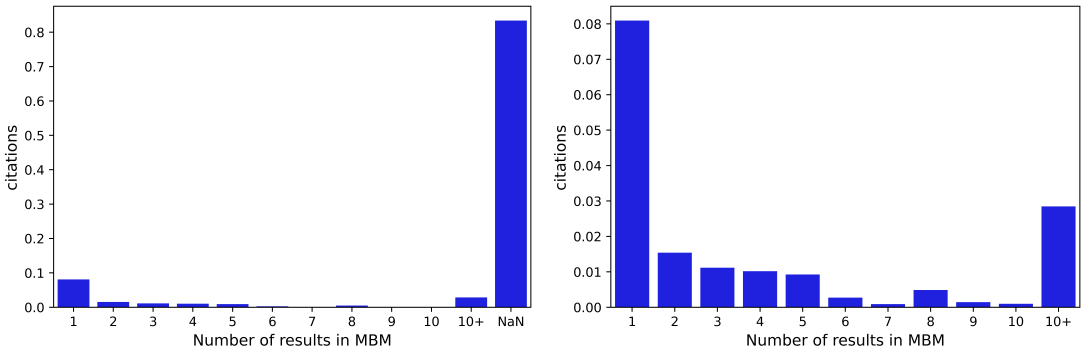


Figure 5.2: Distribution of polarization score for multiple results.



(a) With *NaN*. (b) Without *NaN*.

Figure 5.3: Fraction of matches in MBM for a given domain name, and their citation coverage.

source is considered to be almost always factual; MOSTLY FACTUAL (a score of 3 to 4), which means that the source is considered to be usually factual but may have failed a fact check or two that was not promptly corrected; MIXED (a score of 5 to 6), which means the source does not always use proper sourcing or sources to other mixed factual sources; LOW (a score of 7 to 9), which means the source rarely uses credible sources and is not trustworthy for reliable information; VERY LOW (a score of 10), which means the source rarely uses credible sources and is not trustworthy for reliable information. For example, in MBFC the New York Times is rated HIGH as they are mostly reliable except for some Op-Eds and Fox News is rated MIXED because they may publish misleading reports.

We crawled all the news media outlet data on MBFC and got a dataset including the ratings for 3,586 outlets. Since we already have the domain names of each URLs in *Wikipedia Citations*, we use the same method to extract the domain names from the MBFC dataset as well. We then match the two datasets

via domain names. 689 (19.2% out of 3,586) domains are matched resulting in 3,041,283 Wikipedia citations with both a factual rating and political polarization score, or 10.4% of all citations. In Figure 5.4 we show a bar plot of the number of Wikipedia citations by reliability scores, noting that, while there are only 1467 citations rated as "VERY LOW", there remains a sizable fraction of citations to low or mixed reliability outlets.

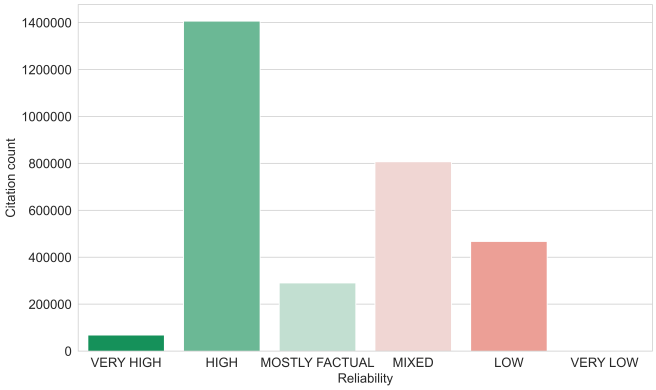


Figure 5.4: Distribution of Wikipedia's news media citation reliability scores.

5.4 Results

We start by providing the overall distribution of Wikipedia's citation political polarization score in Figure 5.5. We remind that the polarization score (x-axis) ranges between -2 (very liberal) and 2 (very conservative). The average Wikipedia citation polarization score (red line) is -0.51 (median -0.52), therefore leaning towards liberal. The bulk of citations also falls between the range -1 and 0.

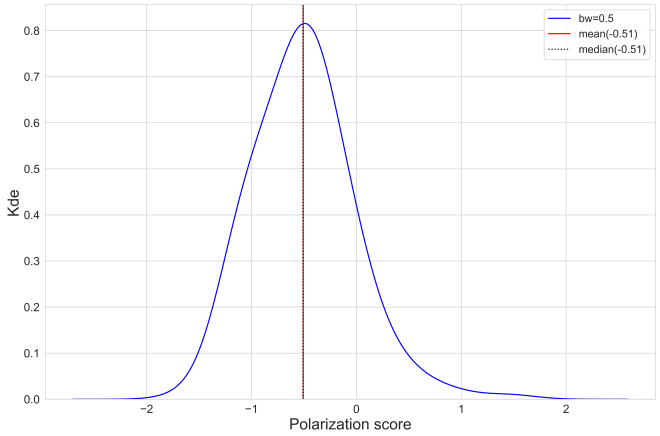


Figure 5.5: Distribution of Wikipedia's news media citation political polarization scores using Kernel Density Estimates (KDE). Negative: liberal; positive: conservative.

We attempt to break down these results using the information on Wikipedia's articles, namely their topics and WikiProject. Topics are organised hierarchically, with four macro topics: Culture, Geography, History and Society, STEM. The overview of citation political polarization per macro topic is given in Figure 5.6. On the left side, we use a violin plot to show the distribution of polarization scores for each macro topic. From this plot, we cannot see differences among macro topics. The bar plot on the right side provides the relative size of a topic in Wikipedia, showing how articles in Culture takes up nearly 50% of all citations, while STEM covers 6.5% of them.

Similarly, we show the distribution for the top 10 topics in Figure 5.7 and for the top 10 WikiProjects in Figure 5.8. We again confirm the general trend discussed above, while also finding minor shifts from it. For example, the topic sports has a higher conservative-leaning fraction of citations, all the while maintaining a liberal-leaning skew. The WikiProjects Politics and India are more liberal-leaning than the average, instead. Taken together, these results confirm that the overall trend towards liberal political polarization is not specific to some areas of Wikipedia, but seems to be widespread across topics and WikiProjects.

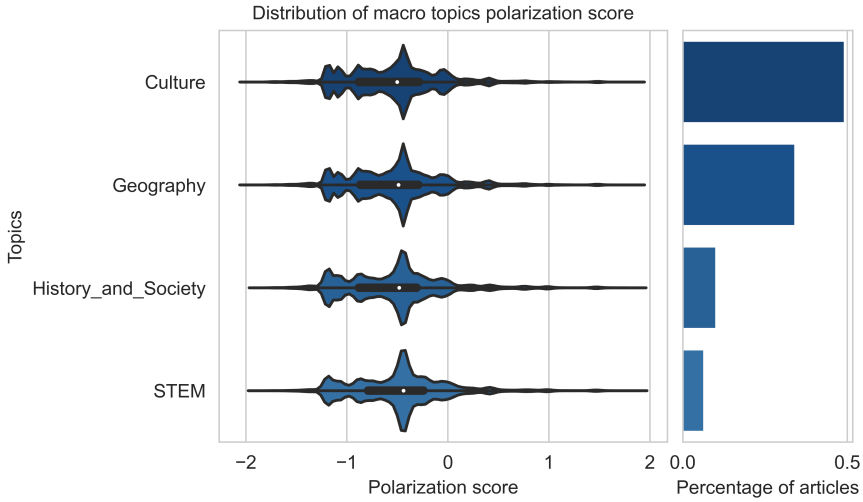


Figure 5.6: Distribution of Wikipedia citation political polarization scores per macro topic. Negative: liberal; positive: conservative.

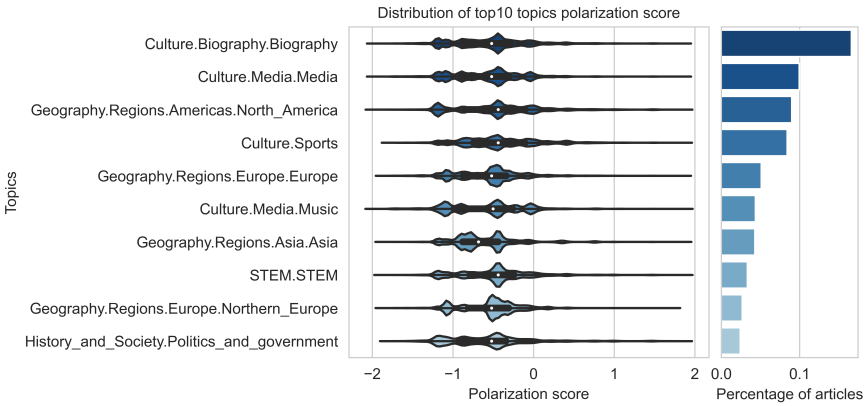


Figure 5.7: Distribution of Wikipedia citation political polarization scores for the top 10 topics. Negative: liberal; positive: conservative.

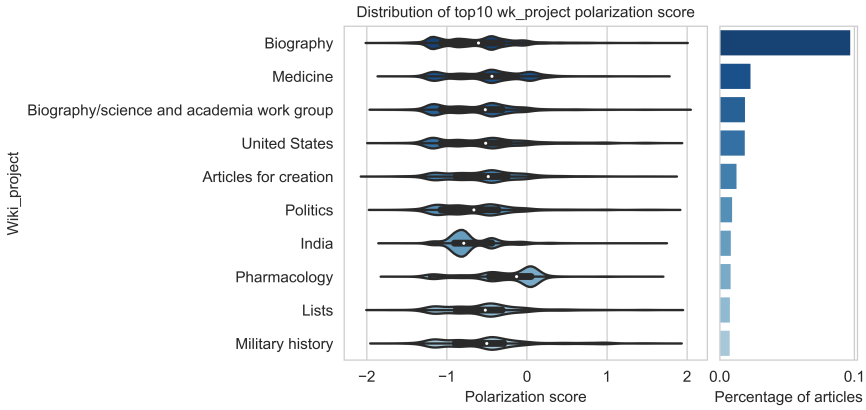


Figure 5.8: Distribution of Wikipedia citation political polarization scores for the top 10 WikiProjects. Negative: liberal; positive: conservative.

In principle, in Wikipedia the neutrality and reliability of contents are tied together. Nevertheless, in practice, we speculate that editors may introduce political polarization in their sources in order to prioritise reliable ones. We have shown before, in Figure 5.4, that most cited news outlets are labelled as highly reliable or mostly factual, even if a significant share of mixed or low reliability sources remains. More details are given in Figure 5.9, where we plot the top 5 news outlets per polarization score group, and show their reliability class as well. In this plot, we divide the news outlets into four groups according to polarization scores: Very Liberal [-2 to -1], Liberal (-1 to 0], Conservative (0 to 1], Very Conservative (1 to 2]. For each group, the x-axis shows the percentage of a news outlet by citations within its group. We can see that, for example, The Guardian is labelled as mixed reliability and takes more than 50% of citations in the Very Liberal group, while the NYT is the second Very Liberal source and is considered highly reliable. Fox News is the top Very Conservative news outlet, with mixed reliability score. We note that in the group Liberal we also have YouTube, with a low reliability score. YouTube is not an outlet with an editorial policy per se, but a repository of contents of any kind. We therefore test whether our results hold when removing citations to YouTube from the dataset, finding that after removal the political polarization distribution moves slightly further liberal overall, while the effect of a conservative polarization in low reliability sources fades substantially (see below). Nevertheless, these changes do not alter out main findings.

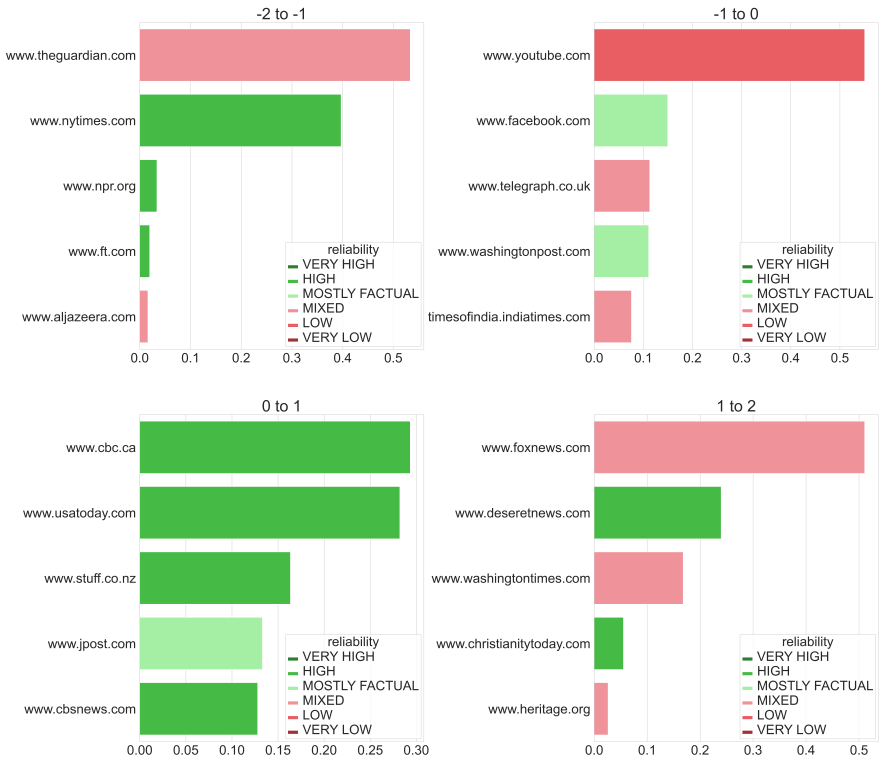


Figure 5.9: Top 5 news outlets reliability class for different political polarization groups.

Next, we make use of multiple linear regression to address RQ2 and explore whether there is a relationship between political polarization and reliability of news media sources in Wikipedia. In our model, we take the political polarization score as the dependent variable, using Wikipedia article topics and WikiProjects as independent variables and set reliability as a control variable. To simplify the model, we proceed as follows. For topics, we use the macro topic: Geography, History and Society, Culture and STEM. For WikiProjects, we focus on the top 10 WikiProjects and set the rest of WikiProjects as "Other". For reliability, we take MOSTLY FACTUAL as the reference class, which is the one in between high and low reliability classes. Thus, our final model is based on the following formula:

$$\begin{aligned} \textit{Polarization score} = {} & \textit{Reliability} + \textit{Geography} + \textit{Culture} + \textit{History and Society} + \textit{STEM} + \\ & \textit{Biography} + \textit{Medicine} + \textit{Biography science and academia work group} + \\ & \textit{United States} + \textit{Articles for creation} + \textit{Politics} + \textit{India} + \\ & \textit{Pharmacology} + \textit{Lists} + \textit{Military history} + \textit{Other} \end{aligned}$$

We provide the results of our regression analysis in Table 5.1. Firstly, while most effects appear to be significant, their coefficient magnitude is always small. We first notice how small the effect of topics is, confirming our previous intuition. Some WikiProjects show slightly larger effects, for example India (more liberal)

and Pharmacology (less liberal). Considering reliability in turn, we cannot see a clear pattern emerge. While high reliability shows a liberal skew, very high reliability shows a conservative skew in turn. Mixed sources tend to be more liberal, while low and very low reliability ones tend to be more conservative.

To test our results, we develop several different models. First of all, we test only for polarization score and Wikipedia topics. In this model, all macro topics have a significant effect on the polarization score with small coefficients, that is, Geography and STEM will bring a less liberal skew while Culture History and Society will have an ever stronger liberal skew. When using a model with reliability and topics, our results converge and become very similar to the model discussed above which also includes WikiProjects. As mentioned previously, we also test our final model without citations to YouTube. After removing them, the most important change is that the low reliability coefficient becomes non significant and goes close to zero, thus making the case for a possible association between low reliability and conservative news outlets disappear.

Table 5.1: Regression results for the effect of news media reliability on political leaning, controlling for Wikipedia topics and projects.

Variables	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.56	0.005	-118.360	0.000	-0.569	-0.550
C(factual, Treatment(reference='MOSTLY FACTUAL'))[T.HIGH]	-0.09	0.002	-37.653	0.000	-0.090	-0.081
C(factual, Treatment(reference='MOSTLY FACTUAL'))[T.LOW]	0.09	0.003	30.143	0.000	0.083	0.095
C(factual, Treatment(reference='MOSTLY FACTUAL'))[T.MIXED]	-0.20	0.002	-82.082	0.000	-0.201	-0.192
C(factual, Treatment(reference='MOSTLY FACTUAL'))[T.VERY HIGH]	0.16	0.004	42.224	0.000	0.149	0.163
C(factual, Treatment(reference='MOSTLY FACTUAL'))[T.VERY LOW]	0.07	0.023	2.833	0.005	0.020	0.112
Geography	0.03	0.002	16.010	0.000	0.023	0.030
Culture	-0.00	0.002	-0.758	0.448	-0.004	0.002
History and Society	-0.02	0.001	-16.579	0.000	-0.027	-0.021
STEM	0.01	0.002	7.098	0.000	0.009	0.016
Biography	-0.04	0.002	-23.958	0.000	-0.047	-0.039
Medicine	0.01	0.004	2.037	0.042	0.000	0.015
Biography_science_and_academic_work_group	-0.04	0.004	-10.502	0.000	-0.047	-0.032
United_States	0.06	0.002	36.096	0.000	0.060	0.067
Articles_for_creation	0.01	0.005	2.874	0.004	0.004	0.022
Politics	-0.01	0.002	-6.233	0.000	-0.017	-0.009
India	-0.09	0.004	-24.027	0.000	-0.101	-0.086
Pharmacology	0.11	0.007	14.672	0.000	0.092	0.120
Lists	0.04	0.003	15.955	0.000	0.037	0.047
Military_history	0.03	0.002	11.020	0.000	0.022	0.032
Other	0.02	0.004	3.986	0.000	0.008	0.023
No. Observations: 604459			R-squared: 0.047			

5.5 Discussion

Wikipedia editors follow core policies when editing articles (Pavalanathan et al., 2018), in an attempt to provide a neutral point of view and reliable contents (Mesgari et al., 2015). Nevertheless, biases might still be found in Wikipedia in a variety of forms, and as such they require a never-ending effort on the part of the community. We find a moderate yet systematic liberal polarization in Wikipedia's news media sources. The average polarization score of Wikipedia sources is -0.5,

with the distribution of polarization scores concentrated around -1 to 0, on a scale between -2 (very liberal) and 2 (very conservative). Our results partially confirm and extend previous ones (Agarwal et al., 2020), while also showing that Wikipedia remains polarized towards liberal news media (Gentzkow and Shapiro, 2010). This finding is relevant as it signals a possible systematic biasing effect whose causes and effects will have to be further studied. News media sources not only select and provide specific information, but also convey it with a certain framing which might influence how a topic is discussed in Wikipedia.

We initially speculated that the presence of political polarization might be partially explained by the editors' need to balance a source factual reliability with its political leaning. Interestingly, we find no clear relationship between reliability and polarization. The relationship between reliability and political polarization is complex, with more conservative sources being associated with both high and low reliability, while liberal sources tend to more often be of mixed reliability. This finding leaves the question of the motivations for political polarization open. We speculate that a multiplicity of factors might play a role, from pre-existing leanings in the composition of the editors' community, to an increasingly polarized media landscape making it difficult to find neutral news media sources to use. Our results may also help the case for changing Wikipedia's sourcing policies, which might prevent information lacking accepted reliable secondary sources from being considered, and at the same time prevent broad and important areas of knowledge from entering the Wikipedia project. This is especially the case for cultures that rely on the non-written transmission and expression of knowledge (such as oral sources) (Taraborelli, 2019). Relying on a more diverse set of sources could be an approach to reduce possible bias in Wikipedia.

We acknowledge several limitations of our study, some of which constitute possible directions for future work. First of all, we rely also on external sources to measure political polarization and reliability. While such sources are considered authoritative, their coverage is only partial and they use specific approaches to score news media outlets which could be complemented in the future. We also mainly focus on the binary political polarization distinction between liberal and conservative sources. Several other political dimensions exist which could be considered in the future. Exploring the relationship between source polarization and reliability remains an open challenge, one that should focus on more dimensions than what we considered here, including time. In this respect, our work scores news media sources at the domain level, while a more granular analysis should be done at the level of the individual source contents. Similarly, we did not consider how a source is used in Wikipedia, and whether its use reflects such polarization or not. Our study remains focused on English Wikipedia, while its extension to more languages would provide for a broader picture.

5.6 Conclusion

In this chapter, we analyzed a potential source of bias in Wikipedia, by considering citations to news media sources and their political polarization. We used a large-scale dataset of citations from Wikipedia, enriching it with metrics of political media polarization from the Media Bias Monitor, and of factual reliability from the Media Bias Fact Check. We found a moderate yet systematic liberal polarization in Wikipedia's news media sources. We also showed that there is no clear relationship between a news media source's reliability and its political leanings.

These results offer a foundation to inform Wikimedia's research agenda about possible sources of disinformation and bias, in view of upholding its neutral point of view policy. Specifically, to keep Wikipedia as a neutral source of information, a better understanding of possible sources of bias is needed not only considering contents or editors, but also Wikipedia's external sources. Here we provided a preliminary analysis of political polarization in Wikipedia from the perspective of citations to news media sources, while much remains to be done. On the one hand, the measurement of political polarization, reliability and other potential signals of bias could be considered more comprehensively, although we have relied here on authoritative sources. On the other hand, a more granular study on the level of the contents of sources and how they are used in Wikipedia would significantly enrich our preliminary findings. The relationship between polarization and reliability would benefit from a higher dimensional analysis. Therefore, we see our work as fostering further attention to the always-open challenge of preserving and improving Wikipedia's knowledge integrity.

By highlighting the role of political polarization in shaping Wikipedia's external references, this chapter complements the analysis in Chapter 4. Taken together, the first two chapters show that both scientific and media sources underpin Wikipedia's epistemic foundations, but in different ways: scientific publications through disciplinary concentrations, and news media through ideological alignments. These insights naturally lead into Chapter 7, which examines a third dimension of sourcing practices: the accessibility of scholarly publications through open access models.

Part Two: The Role of Open Access in the Dissemination of Science

Chapter 6

Overview

This part investigates the role of open access (OA) in shaping how scientific knowledge is disseminated, accessed, and contested on Wikipedia. While OA publishing is often promoted as a means of improving access to scholarly work, its downstream effects on platforms like Wikipedia remain insufficiently understood. In this part, I examine how OA status influences both the visibility of research in Wikipedia and its involvement in citation-related editorial disputes. Chapters 7 and 8 offer two complementary perspectives: the first on the dissemination of OA publications via Wikipedia, and the second on their mobilization in areas of epistemic conflict. Together, these studies shed light on how openness intersects with credibility, access, and contestation in one of the most influential public knowledge infrastructures. I address the following research questions:

- **To what extent does open access publishing increase the visibility and uptake of scientific work on Wikipedia?**
  In Chapter 7, I revisit prior claims about the OA advantage in Wikipedia by focusing on individual articles rather than journals, and by leveraging a large-scale citation dataset linked with OpenAlex and Scimago metadata. Using regression modeling and descriptive analysis, I find that OA articles are more likely to be cited from Wikipedia, and that this effect is partially mediated by citation count. The chapter also explores how this relationship has evolved over time, contributing new insight into the mechanisms through which OA supports public engagement with science.

- **Do open access publications play a distinctive role in scientific disputes on Wikipedia?**
  Chapter 8 turns to the contested side of science communication by analyzing citation disputes within Wikipedia. Drawing on a newly constructed dataset of over 3,500 publications involved in more than 2,200 dispute cases, I evaluate whether OA articles are disproportionately cited in contentious editing contexts. By comparing dispute intensity and employing logistic

regression models, I assess how publication characteristics—including OA status, field, and citation impact—influence the likelihood of being used in epistemic conflicts. The findings suggest that OA publications are more accessible not only to the general public but also to editors engaged in debate, and may thus be more readily mobilized in citation-driven disputes.

These chapters together highlight that open access is not only a policy goal but also a practical factor shaping the real-world use and contestation of science in public knowledge platforms. They offer empirical insight into how accessibility, authority, and controversy are intertwined in the digital dissemination of science.

Chapter 7

Open Access Improves the Dissemination of Science: Insights from Wikipedia

7.1 Introduction

Having established in Chapters 4 and 5 that Wikipedia relies heavily on certain types of sources and that these can reflect disciplinary and political biases, we now turn to the question of access. Beyond the issue of which sources are cited, it is equally important to consider which sources are available for citation in the first place. The open access (OA) movement directly influences which scientific knowledge is accessible to both editors and readers, and therefore has important implications for Wikipedia's role as a public knowledge infrastructure.

Open access (OA) publishing has emerged as a popular alternative to traditional subscription-based models, with the goal of making research more widely accessible to the public. This movement has gained momentum over the years, with many scholars recognizing the benefits of open access in promoting the dissemination of scientific knowledge and funding bodies adopting OA mandates (Piwowar et al., 2018; Holmberg et al., 2020). In this context, OA not only expands readership within academia but also provides a crucial channel through which reliable scientific knowledge can enter public-facing platforms such as Wikipedia.

Citations play a crucial role in supporting Wikipedia's mission to provide reliable and verifiable information[1]. Among various sources, academic and peer-reviewed publications are widely regarded as the most reliable[2]. The OA movement provides Wikipedia with a valuable opportunity to access a vast repository of reliable and verifiable scientific knowledge. By incorporating OA citations, Wikipedia can enhance its role in scientific communication. (Tattersall et al., 2022). As a dynamic platform for sharing and disseminating knowledge across the globe, Wikipedia is relied upon by millions of users every day to satisfy a wide range

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

[2] `https://en.wikipedia.org/wiki/Wikipedia:Verifiability#Reliable_sources`.

of information needs (Singer et al., 2017). It has become a critical source of information for both the general public and academic researchers, and its impact is extending beyond the realm of general knowledge and into the academic sphere (Park, 2011; Kousha and Thelwall, 2017; Tohidinasab and Jamali, 2013).

Wikipedia's extensive use of citations makes it possible to analyze its reliance on academic publications, which is a central aspect of our investigation. Previous research utilized the Scopus database and an English Wikipedia database dump extracted from 2014, culminating in the identification of 32,361 unique articles for analysis. They found that articles from OA journals exhibit 47% higher odds of being cited in Wikipedia compared to those from paywalled journals (Teplitskiy et al., 2017). Notably, their adoption of journals as the unit of analysis, rather than individual articles, has the possible drawback of wrongly estimating the influence of OA on scientific knowledge dissemination through Wikipedia. This limitation also arises from overlooking articles accessible via green or hybrid routes (ElSabry, 2017). Moreover, their manual matching approach imposed constraints on the scale of their research. Therefore, an exploration conducted at the granularity of individual articles not only promises a more nuanced understanding of the relationship between OA and the dissemination of scientific knowledge through Wikipedia but also unveils the role of citation count in this process.

In light of this, this chapter seeks to fill this gap by examining how OA publications affect Wikipedia at the article-level granularity. Specifically, we aim to answer the following research questions:

1. RQ1: To what extent does Wikipedia rely on open access publications? How has this been changing over time?

2. RQ2: To what extent does the open access status of an article influence its likelihood of being cited in Wikipedia?

To address these questions, we will use descriptive statistics and regression analysis based on the *Wikipedia Citations* dataset (Singh et al., 2021). To identify the information in article-level granularity such as the OA status of publications, we will use the OpenAlex and Scimago data. Our research contributes to understanding the role of OA in the dissemination of scientific knowledge and the impact of Wikipedia in this process, as well as informing policy and practice in the realm of open scholarly communication.

The remainder of this chapter is structured as follows. Section 2 provides an overview of existing research in the field. Section 3 describes our dataset and methodology. Section 4 presents descriptive statistics of OA publications in Wikipedia (RQ1), and then uses regression analysis to model the influence of OA status on the likelihood of a paper being cited in Wikipedia (RQ2). Finally, Sections 5 and 6 offer a discussion and conclusion of our findings.

7.2 Previous Work

7.2.1 Open Access in Science

The key idea behind OA is to provide unrestricted and free access to scientific outcomes, thus enhancing their visibility and reach regardless of financial or geographical constraints (Tennant et al., 2016; Redalyc et al., 2003). The increasing popularity of OA in academic publications has generated extensive discussions among scholars in recent years. Empirical studies have shown that OA has had a significantly positive impact on the accessibility of scientific journal articles (Björk et al., 2010). However, the distribution of OA publications varies depending on the data source. A comprehensive analysis of OA publications based on Crossref data shows that at least 27.9% of the total 19 million scientific articles are OA (Piwowar et al., 2018). In contrast, studies report that around 55% of articles in the Google Scholar database from 2009 to 2014 are OA, and more than 50% of scientific papers published since 2007 can be accessed freely (Martín-Martín et al., 2018; Archambault et al., 2014). Among the various OA policies, Bronze OA is the most common type (Piwowar et al., 2018). Although the distribution of OA varies across different fields, General Science, Technology, and Biomedical research have relatively higher OA rates, while Engineering and Arts&Humanities have lower rates (Archambault et al., 2014; Martín-Martín et al., 2018).

An “open access citation advantage" (OACA) has also been a topic of ongoing debate. Some researchers have observed that a citation advantage linked to OA exists, although the effect magnitude varies based on the dataset and methods used. For example, OA articles have been found to receive 18% more citations than average based on Web of Science, while Scopus reports an even higher, positive 40% effect (Piwowar et al., 2018; Archambault et al., 2014). Kristin found that in four disciplines—philosophy, political science, electrical and electronic engineering, and mathematics, OA articles exhibit a greater research impact (Antelman, 2004). Distinct advantages are found for green OA articles hosted in institutional repositories, receiving 106% more citations than gold OA or non-OA articles, and OA articles receive up to 36% more diverse, interdisciplinary citations than non-OA articles (Young and Brandes, 2020). Despite these findings, a recent systematic review of OACA suggests that the debate continues, revealing diverse outcomes across different studies (Langham-Putrow et al., 2021). Out of 134 included studies, 47.8% confirm the existence of OACA, 27.6% deny it, 23.9% find OACA only in subsets, and 0.8% are inconclusive, with a notable association between the focus on multiple disciplines and the identification of OACA in subsets (Langham-Putrow et al., 2021). Therefore, the effects of OA on citation patterns remain a topic of interest and active investigation.

7.2.2 Science and Wikipedia

With the rapid development of the internet, traditional peer review processes need to adapt to keep pace with the rapid knowledge creation in the 21st century (Black, 2008). As the largest encyclopedia worldwide, Wikipedia aims to effectively and globally distribute information based on scientific findings[3], thereby making it a valuable altmetric source (Sugimoto et al., 2017; Mesgari et al., 2015). Evans and Krauthammer observed higher citation counts for articles linked in Wikipedia, suggesting its potential for impact assessment (Evans and Krauthammer, 2011). Altmetric.com integrated Wikipedia mentions into its tracking in 2015[4], but doubts have arisen about Wikipedia's reliability for impact assessment. Lin and Fenner found that only 4% of PLOS articles were cited in Wikipedia (Lin and Fenner, 2014), and Kousha and Thelwall concluded that Wikipedia citations are insufficient for impact assessment in most fields (Kousha and Thelwall, 2017).

Previous research indicates that Wikipedia's topical coverage is similar to that of scientific disciplines. With 13.44% of its citations coming from OA journals (Arroyo-Machado et al., 2020) and 31.2% of Wikipedia citations associated with OA sources, this percentage has exhibited an upward trend over the years (Pooladian and Borrego, 2017). Additionally, STEM fields, particularly biology and medicine, comprise the most prominently featured scientific topics on Wikipedia (Yang and Colavizza, 2022a). Fields such as medicine and psychology have a comparatively high number of citations to research papers on Wikipedia and are sometimes used as a gateway to further academic research (Maggio et al., 2017; Schweitzer, 2008). Furthermore, journal articles cited in Wikipedia tend to be published in high-impact journals (e.g., with higher impact factors) and are more frequently OA than the average article (Nielsen, 2007; Teplitskiy et al., 2017).

Science significantly contributes to Wikipedia, but the influence is reciprocal. Previous studies have established that Wikipedia can enhance the citation impact of the articles it cites (Thompson and Hanley, 2018). Furthermore, Wikipedia has demonstrated its ability to rapidly and reliably incorporate novel scientific findings in response to ongoing public events or crises (Colavizza, 2020).

7.2.3 Citation Analyses of Wikipedia

The open release of citation datasets from Wikipedia has led to a surge in studies examining citation analysis on Wikipedia (Singh et al., 2021; Zagorova et al., 2021). Among the articles on Wikipedia, 6.7% cite at least one journal article with an associated digital object identifier (DOI) (Singh et al., 2021), and the majority of these cited journal articles were published in the past two decades (Yang and Colavizza, 2022a). Benjakob et al. (2022) conducted a study on the quality of

[3] https://wikimediafoundation.org/about/mission/

[4] https://www.altmetric.com/blog/new-source-alert-wikipedia/

citations in Wikipedia during COVID-19 and found that Wikipedia mostly cites reliable sources and prefers OA articles. Some researchers have focused on user behavior related to reference usage on Wikipedia. Piccardi et al. (2020) found that engagement with citations on Wikipedia is generally low, but references are more frequently looked up when the information is not included.

Despite the growing number of citation studies on Wikipedia, the relationship between OA and Wikipedia still requires further exploration. Previous research has examined the effect of OA on Wikipedia, and found that OA articles were 47% more likely to be cited than paywalled articles when controlling for journal and research fields (Teplitskiy et al., 2017). However, their focus on analyzing journals rather than individual articles leads to in an underestimation of OA impact on disseminating scientific knowledge through Wikipedia. This limitation stems from the oversight of articles accessible through green or hybrid routes (ElSabry, 2017). Additionally, their manual matching approach constrained the scope of the research. Consequently, this study aims to build on previous findings by employing a more rigorous and comprehensive methodology, examining individual articles, and accounting for additional confounding factors to better understand the relationship between OA and Wikipedia.

7.3 Data and Methods

The data collection process adhered to the workflow outlined below. First, we obtained all citations from English Wikipedia to any source using the open dataset called *Wikipedia Citations* (Kokash and Colavizza, 2024). Next, to identify journal articles, we used the classification and DOI information provided by *Wikipedia Citations*. Then, to enrich the journal articles with article-level data such as citation counts, OA status, and OA policy, we used the OpenAlex API to retrieve relevant information through DOIs for each journal article. Finally, we used data from Scimago to obtain relevant information for each journal. The following sections provide a detailed description of the main datasets used in the study.

7.3.1 Wikipedia Citations

The primary dataset used in this research is *Wikipedia Citations*, a comprehensive dataset of over 45 million citations extracted from the February 2024 dump of English Wikipedia (Kokash and Colavizza, 2024). This is an updated version of the 2020 dataset (Singh et al., 2021). Of these, approximately 2.2 million citations are classified as journal articles, with 2,197,461 of them containing a DOI.

7.3.2 OpenAlex and Scimago

To examine the impact of OA articles, we used OpenAlex, a free and open platform providing data on academic papers and researchers (Priem et al., 2022). OpenAlex draws data from sources such as Microsoft Academic Service (MAG) and Crossref, containing more than 240 million academic works. These works are useful for research in bibliometrics, science and technology studies, and science of science policy (Bredahl, 2022; Hao et al., 2022). To obtain the necessary data for journal articles in *Wikipedia Citations*, we utilized the OpenAlex API[5] to retrieve relevant article details such as OA status, OA policy, publication date, publisher, and concepts, among others, for each DOI. After matching, we retrieved article information from OpenAlex for 2,154,524 journal articles.

In our paper, we used OA policy following the classification scheme proposed by OpenAlex, which includes the following categories:

1. Gold: Published in a fully OA journal.
2. Green: Toll-access on the publisher landing page, but there is a free copy in an OA repository.
3. Hybrid: Free under an open license in a toll-access journal.
4. Bronze: Free to read on the publisher landing page, but without any identifiable license.
5. Closed: All other articles.

OA status is treated as a binary variable in our analysis, defined as either True or False. According to OpenAlex, and as supported by previous literature (Piwowar et al., 2018), an article is considered OA if it has a URL where the full text can be read without payment or login.

Additionally, we collected journal information to conduct a regression analysis on the influence of OA. We obtained this information using data downloaded from Scimago[6]. Scimago is an OA resource that provides an internationally recognized journal rank indicator for analysis in the fields of scientometrics and informetrics (Falagas et al., 2008; Yuen, 2018; González-Pereira et al., 2010). We equipped each journal with the SCImago Journal Rank indicator (SJR) and other relevant information.

Model Specification

Dependent variable To assess the potential advantage of OA articles in Wikipedia, we defined a binary dependent variable, ***is_wiki***, which indicates

[5] `https://docs.openalex.org/how-to-use-the-api/get-single-entities`.
[6] `https://www.scimagojr.com/aboutus.php`.

whether an article has been cited in Wikipedia or not. Since our primary dataset consists solely of articles cited in Wikipedia, we use OpenAlex to obtain negative samples of articles not cited in Wikipedia, via stratified sampling.

Independent variable To assess the impact of OA articles on their citation rates in Wikipedia, we analyze two types of variables: article-level and journal-level. At the article level, we consider the number of citations (***times_ cited***), whether the article is OA (***is_ oa***), the time of publication (***article_ age***), and the field of research (***concept***). These features have been shown to influence citation impact in previous studies (Colavizza et al., 2020; Gargouri et al., 2010; Yegros-Yegros et al., 2015; Struck et al., 2018; Teplitskiy et al., 2017; Nielsen, 2007). At the journal level, we primarily consider the Scimago Journal Rank (***SJR***). To accurately represent the SJR for each article, we use the rank assigned to journals for the same year in which the article was published. Since a year-by-year breakdown of journal ranks is only available from 1999 to 2020, we assign the rank for 1999 to articles published before 1999, as it is the earliest available representation. This range (i.e., published before 1999) accounts for 29% of the citations in our curated set from Wikipedia.

Although these variables have been widely used to model citation impact in previous studies, few analyses have directly linked these indicators to whether an article is cited in Wikipedia, particularly concerning different OA policies.

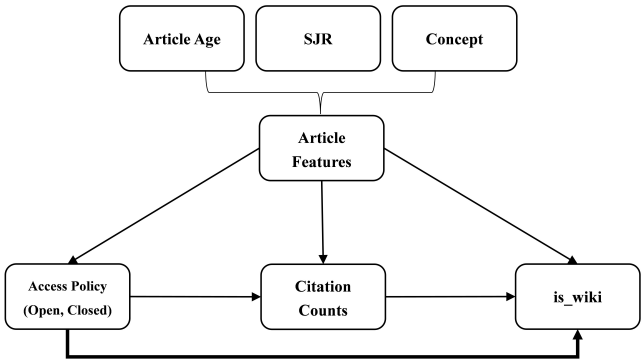


Figure 7.1: Assumed causal structure of Wikipedia's OA citation adoption effect, with a black line representing an assumed causal relationship between two variables.

In this study, we use logistic regression to analyze the relationship between a binary dependent variable and one or more independent variables. The logistic regression coefficients represent the size of each predictor variable's contribution to the target variable. Figure 7.1 illustrates the assumed causal structure of Wikipedia's OA citation adoption effect, with a black line depicting an assumed causal relationship between two variables. Specifically, we assume that the likelihood of a journal article being cited in Wikipedia is directly influenced by its

features, citation counts, and OA status. At the same time, OA status can also influence citation counts, leading to a mediated effect on the article's adoption in Wikipedia. Our models measure both the direct effect and the total effect of OA status on being cited in Wikipedia. The direct effect is shown as a thick black line in Figure 7.1, while the total effect includes both direct and mediated (via citation counts) effects.

Dataset construction

We aim to create a balanced dataset of journal articles suitable for regression analysis. The initial dataset, sourced from Wikipedia, contained 1,499,021 unique scientific articles. To initiate our regression analysis, we constructed a dataset by adding Journal, Year of Publication, and Concept as stratifying variables. This dataset will be used to account for the influence of ***concept***. To restrict our focus to root-level concepts and avoid ambiguity, we filtered the citations to include only those with a single associated concept, resulting in a set of 410,573 articles. Subsequently, we assembled corresponding sets of articles for these two datasets from OpenAlex based on the stratifying variables, excluding those already cited in Wikipedia. To reduce noise in the sampling strategy, we removed journals with no corresponding name in Scimago and those with fewer than 20 citations. We also removed all articles published before 1900 to eliminate sparsely mentioned dates and accept a slight recency bias.

After pre-processing, we group the articles within each stratum and proceed as follows:

1. Filter the entire set of OpenAlex articles to include only those matching the fields in the strata.

2. If the number of articles in this filtered set is fewer than in the curated Wikipedia dataset, discard the strata and remove the corresponding articles from the curated dataset.

3. Otherwise, randomly sample an equal number of articles from the filtered set and add them to the set of negative samples.

After iterating through all strata (90,019 in total), we derived a final negative set comprising 261,230 entries. When combined with the corresponding sets of Wikipedia-cited articles, this results in a comprehensive dataset totalling 522,460 entries.

To ensure the robustness of our sampling methodology, we repeated the process five times, resulting in five different datasets that were used in the analyses. Although our method of matching strata to construct a set of negative samples approximates the more rigorous method of propensity score matching (PSM), the discrete nature of our strata and the large population size contribute to the

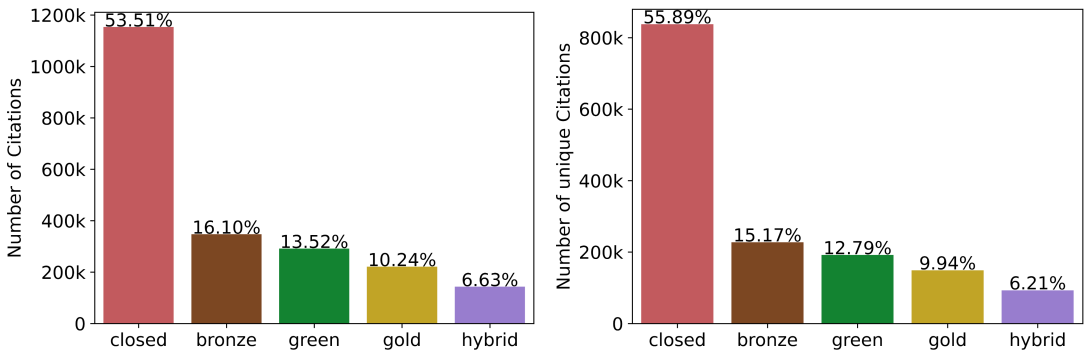
robustness of our analysis. A descriptive overview of this curated dataset is provided in Tables B.1, B.2 and B.3 in the appendix.

7.4 Results

We augmented our analysis by incorporating additional metadata from OpenAlex and Scimago, enabling us to obtain information for 98.0% (2,154,524) of the 2,197,461 citations with a valid DOI. From these, we extracted 1,499,021 unique publications (DOIs) and their associated OA status. Our findings show that 46.5% (1,021,820 out of 2,197,461) of the citations and 44.1% (661,068 out of 1,499,021) of the publications were OA.

Characterizing open access Articles within Wikipedia

We present our findings on the distribution of OA policy in Wikipedia citations in figure 7.2. Our results show that the most commonly observed OA policy in Wikipedia citations is the bronze policy, which is aligns with trends in scholarly literature (Piwowar et al., 2018). The second most common OA policy observed in Wikipedia citations is green, which is significantly more prevalent than the gold policy.



(a) Over citations. (b) Over unique journal articles.

Figure 7.2: Distribution of open access policy in Wikipedia.

Figure 7.3 illustrates the distribution of OA articles based on the publication year. The grey bars represent the total number of articles cited by Wikipedia each year, encompassing both OA and non-OA articles. The green bars indicate the number of OA articles within each grey bar. The red dotted line indicates the annual proportion of OA articles from the Web of Science database, the blue dashed line represents the annual OA article ratio from the OpenAlex database, and the black line shows the proportion of OA articles cited in Wikipedia by

publication year. All three lines are plotted against the right y-axis, representing the fraction of OA articles, while the left y-axis shows the total article count.

Overall, the data reveal a consistent increase in the proportion of OA articles cited by Wikipedia over the past four decades. Compared to the overall scientific literature, as measured by the OpenAlex and Web of Science databases, Wikipedia's proportion of OA articles is notably higher. This trend suggests a growing reliance on OA articles within Wikipedia, indicating their significant influence on scientific representation on the platform. Specifically, since 2015, the percentage of OA article citations in Wikipedia has consistently exceeded 50%.

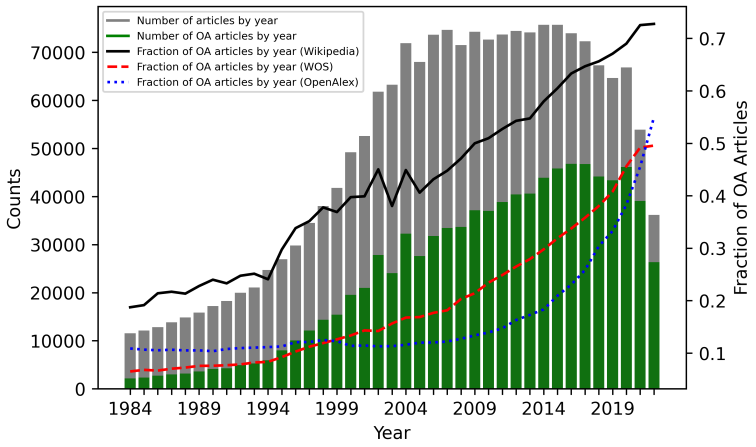


Figure 7.3: Fraction of OA citations by publication date of citation.

We examined the breakdown of OA status and OA policies across the 40,806 journals in our dataset. To effectively visualize this information, we calculated the number of citations for each journal and selected the top 20 for further analysis. Figure B.2 displays the total number of citations for the top 20 journals, where blue represents OA articles and orange represents non-OA articles. Consistent with previous studies, high-impact journals such as *Nature*, *PNAS*, and *Science* appear frequently on Wikipedia (Nielsen, 2007), accounting for 5.7% of all citations. However, inferring the OA status of articles based on whether journals are classified as "Open Access" or "Closed Access" can be misleading (Teplitskiy et al., 2017), due to the high variance in OA status among articles within the same journal. For instance, although some articles in *Nature* and *Science* are OA, others are non-OA.. Therefore, studying the relationship between OA and Wikipedia at the journal level is inappropriate.

To further explore the distribution of OA policies among the top 20 journals, we visualized the data in Figure B.3. Our analysis shows a growing trend towards bronze OA policies among journals. However, some journals that classify themselves as OA, such as "*Journal of Biological Chemistry*"[7], contain a significant

[7] `https://www.elsevier.com/journals/journal-of-biological-chemistry/`

proportion of articles classified as Hybrid or Gold OA. Despite potential limitations in OpenAlex's classification of OA articles, we adhered to their classifications in our study.

Additionally, we analyzed the distribution of OA status across OpenAlex concepts, as shown in Figure 7.4 and Figure B.1. The analysis utilized the OpenAlex dataset, which comprises 65,000 concepts, including 19 root-level concepts. We employed fractional counting to assess the number of citations for each root-level concept. In Figure 7.4, the left side displays the percentage of cited publications with OA status for each concept, while the right side shows the total number of citations per concept, ordered from the largest to the smallest. The blue bars represent the fraction of OA citations within each OpenAlex concept, while the red bars represent the fraction of paywalled articles. Given that 46.5% of citations on Wikipedia are OA, we used this percentage as a baseline for OA proportionality, represented by the black dotted line in Figure 7.4. Additionally, the black star in the same figure denotes the percentage of OA articles for each concept across the entire OpenAlex dataset, serving as a benchmark for the broader scientific landscape. Our analysis revealed significant variance in OA proportions across different fields.

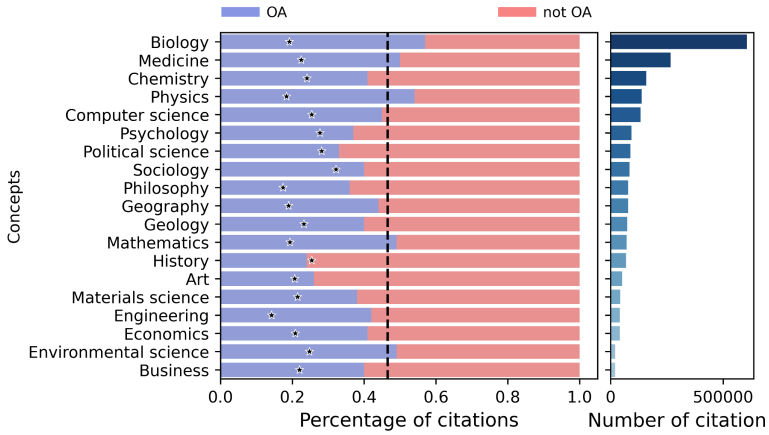


Figure 7.4: Distribution of OA status and count of citations by OpenAlex concept.

Notably, the OA proportions for all concepts in Wikipedia significantly exceed those observed in OpenAlex, underscoring the critical role of OA articles in shaping Wikipedia's information sources. Although the overall proportion of OA on Wikipedia is 46.5%, certain concepts have a relatively higher OA citation rate. Specifically, Biology (57%), Physics (53%), Medicine (50%), Environmental Science (49%), and Mathematics (49%) demonstrate a greater reliance on OA publications for shaping their scientific content. Conversely, Political Science (33%), Art (27%), and History (24%) show the lowest proportions of OA articles among Wikipedia's

`0021-9258/open-access-journal`.

cited sources. Generally, Wikipedia exhibits a stronger dependence on OA articles in STEM-related fields than in the humanities, where citations of scientific articles are less prevalent.

OA Citation Advantage

To comprehensively assess the impact of an article's OA status on its likelihood of being cited by Wikipedia, we developed a series of statistical models utilizing the datasets outlined in the data and methods section. The objective of this analysis is to elucidate the role of OA articles within the scientific discourse of Wikipedia, and identify any potential advantages associated with their citation patterns.

Model results

We use logistic regression for its interpretability and expressiveness, and we apply log transformations to continuous variables. To thoroughly evaluate the overall impact of OA status on citation adoption, our primary logistic regression model, designed to examine the influence of *is_ oa*, is formulated as follows:

$$\begin{aligned} is\_\,wiki = {} & is\_\,oa + \ln(article\_\,age) + \ln(SJR) + concept \\ & + is\_\,oa \times \ln(article\_\,age) \end{aligned} \tag{7.1}$$

To assess the direct influence of OA status on the likelihood of being cited in Wikipedia, while accounting for the interplay between citation count and OA status, we introduce the second formula:

$$\begin{aligned} is\_\,wiki = {} & is\_\,oa + \ln(article\_\,age) + \ln(SJR) + concept \\ & + \ln(times\_\,cited + 1) + is\_\,oa \times \ln(article\_\,age) \end{aligned} \tag{7.2}$$

To validate the robustness of our model, we evaluated the statistical significance of each coefficient across all five samples. A coefficient is considered statistically insignificant if it lacks significance in at least one of the samples. We then present the effects in terms of odds ratios, calculated from the mean odds ratios across all five samples. Additionally, we conducted a multicollinearity check on the variables in the model and found that all Variance Inflation Factors (VIF) were below 10, indicating the absence of significant multicollinearity issues.

Our analysis of the results is summarized in Table 7.1, which lists the regression results of model 1 and model 2. From Table 7.1, it is evident that OA articles exhibit substantially higher odds of being cited in Wikipedia compared to paywalled articles. Specifically, in model 1, OA articles have 80% higher odds of being cited, and this becomes 64.7% in model 2 when considering citation counts. These findings highlight that the OA status of an article plays a crucial role in its

likelihood of being used as a reference on Wikipedia, suggesting that Wikipedia is more inclined to cite OA articles over paywalled ones.

Incorporating citation counts into the model enhances the interpretability and reveals additional insights. Similar to OA status, citation counts play a significantly positive role in the odds of scientific articles being cited in Wikipedia. This suggests that articles with higher citation counts are more likely to be referenced on Wikipedia, reflecting their impact and visibility within the scientific community. In terms of conceptual classification analysis, we use biology, which has the highest citation count on Wikipedia, as our reference point. By incorporating citation counts into our analysis, we identified 14 concepts that significantly influence the likelihood of articles categorized as OA being cited on Wikipedia. Notably, several concepts from the humanities and social sciences, such as Art, History, Philosophy, and Political Science, exhibited notably positive coefficients. This finding reflects the unique characteristics of these domains, known for being low-citation fields (Patience et al., 2017), despite their substantial importance within the Wikipedia ecosystem. Additionally, Environmental Science also demonstrated a high coefficient, likely due to its interdisciplinary nature, which incorporates knowledge from both natural and social sciences.

Furthermore, the age of the article demonstrates a modest yet significantly negative effect on the likelihood of OA articles being cited by Wikipedia. This finding implies that newer publications are more likely to be cited by Wikipedia compared to older articles, indicating a preference for recent and up-to-date scientific content on the platform.

Despite the negative effect of the SJR, insights can be gleaned from the distribution of SJR among Wikipedia citations, as shown in Figure B.4. In Figure B.4, the x-axis represents the SJR value obtained from Scimago, while the y-axis represents the proportion of Wikipedia citations. It is evident that nearly 90% of the cited journals on Wikipedia have an SJR value of less than 10. Additionally, the mean SJR in our dataset is 3.68, with a third quartile of 4.38. This finding further supports our regression results, indicating that in Wikipedia, most citations come from journals with small SJR values. Thus, as the SJR decreases, the likelihood of OA articles being cited by Wikipedia increases.

To gain insight into the interaction between OA status and citation counts in Wikipedia, we use Formula 2 to create a graph that plots these two variables along with article features.

The graph, shown in Figure 7.5, displays the dependent variable, ***is_wiki***, on the y-axis and the citation counts (variable ***times_cited***) on the x-axis. Articles are grouped according to their OA status. We plot the average model prediction for each group using the first data sample and provide 95% bootstrapped confidence intervals for each group (faded color). The red line illustrates the trend of OA adoption by citation count under the condition that the OA status is closed, while the blue line shows the trend under the condition that the OA status is open. This graph reveals several insights. Firstly, when the citation counts are

Table 7.1: Regression results for the first sample with models 1 and 2.

Regression Model	Model 1 ($R^2 = 0.00034$)			Model 2 ($R^2 = 0.07182$)		
Feature	Coef	Odds Ratios	P>z	Coef	Odds Ratios	P>z
Intercept	-0.363	0.695	0	-0.979	0.376	0
ln1p_times_cited				0.442	1.557	0
ln(article_age)	0.064	1.066	0	-0.067	0.935	0
ln(SJR)	-0.002	0.998	0.487	-0.246	0.782	0
is_oa	0.588	1.800	0	0.499	1.647	0
is_oa:ln_article_age	-0.091	0.900	0	-0.103	0.902	0
Art	-0.001	0.999	0.980	0.669	1.952	0
Business	0.001	1.001	0.992	0.38	1.462	0
Chemistry	-0.002	0.998	0.845	0.014	1.014	0.237
Computer science	-0.004	0.996	0.846	0.251	1.285	0
Economics	-0.006	0.994	0.885	0.042	1.043	0.387
Engineering	0.000	1.000	0.999	0.523	1.688	0
Environmental science	0.002	1.002	0.987	0.75	2.118	0
Geography	-0.007	0.993	0.877	0.583	1.791	0
Geology	0.008	1.008	0.626	0.09	1.095	0
History	-0.004	0.996	0.914	0.646	1.908	0
Materials science	0.009	1.009	0.822	-0.069	0.934	0.109
Mathematics	0.002	1.002	0.931	0.415	1.514	0
Medicine	0.003	1.003	0.675	0.03	1.03	0
Philosophy	-0.003	0.997	0.915	0.703	2.02	0
Physics	-0.012	0.988	0.386	0.227	1.255	0
Political science	-0.006	0.994	0.851	0.681	1.977	0
Psychology	-0.002	0.998	0.905	-0.069	0.934	0
Sociology	0.002	1.002	0.981	0.466	1.594	0

very low (near 0), there is a significant initial citation advantage for OA articles compared to the paywalled articles. As the citation count increases (up to 200), this advantage gradually expands. However, as the citation count continues to grow, this advantage becomes less distinguishable. Our previous work (Yang and Colavizza, 2022a) shows that articles cited fewer than 100 times account for 70% of all cited articles, while only about 3% are cited 1,000 times or more. Therefore, most citations in Wikipedia benefit from this OA effect. We speculate that the OA adoption effect arises because Wikipedia editors may find it easier to discover and access open research results earlier in the publication timeline, before these articles accumulate citations and gain broader peer recognition.

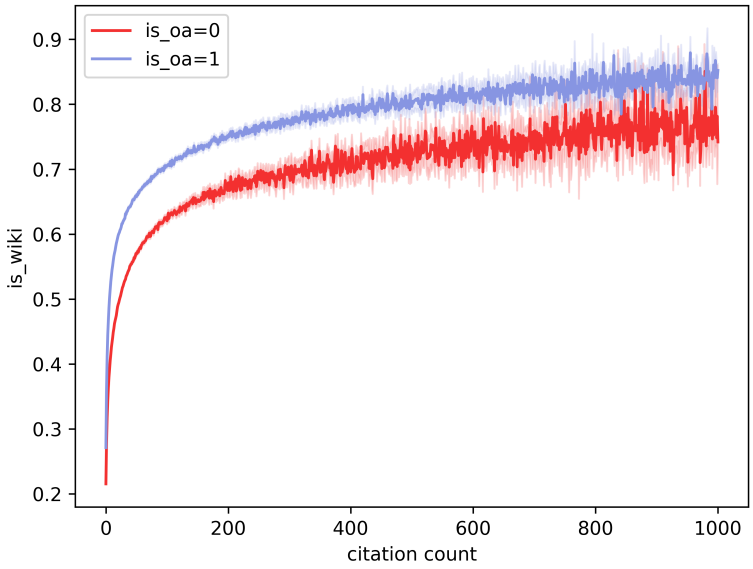


Figure 7.5: OA adoption effect at varying citation counts, based on model 2.

In addition, we examined the interaction between OA status and article age on Wikipedia using model 2, as illustrated in Figure 7.6. The figure reveals a significant advantage for younger articles, especially those aged less than 48 months (4 years). OA articles in this age group have a 10% higher likelihood of being cited by Wikipedia compared to a paywalled article. However, as the age of the articles increases to around 240 months (20 years), the likelihood of adoption begins to decline. This trend highlights Wikipedia's preference for newer articles, particularly those published within the last four years, over older publications.

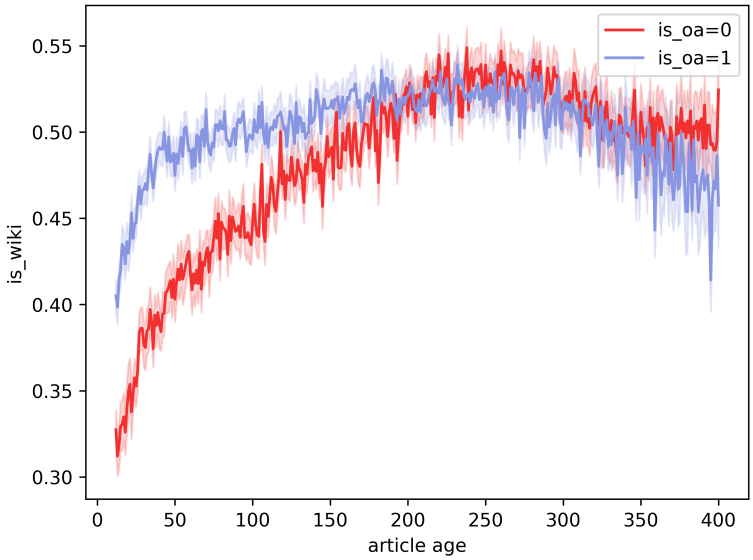


Figure 7.6: OA adoption effect at varying article age, based on model 2.

Moreover, we employed two regression models to investigate the impact of OA policy on citation adoption, using "closed" as the baseline. The results, presented in Table 7.2, demonstrate that all OA policies significantly enhance the overall adoption rate for OA articles. Additionally, the interaction between OA policies and article age shows a significant negative effect on OA adoption. Table 7.3 presents the results from the second model, which explores the indirect effect of OA policy and reveals a similar trend. However, the bronze policy exhibits a slightly significant negative impact. To validate the robustness of our findings, we conducted regressions across all five samples, with the results reported in Tables B.6 and B.7.

Table 7.2: Coefficients for OA adoption by policy. Results for the first sample, model 1, $R^2 = 0.0013$.

Feature	coef	odds_ratios	P>z
bronze	0.101	1.106	0.000
gold	0.032	1.032	0.000
green	0.162	1.176	0.000
hybrid	0.190	1.210	0.000
ln_article_age	0.019	1.019	0.000
ln(SJR)	-0.036	0.965	0.000
bronze:ln_article_age	-0.0398	0.961	0.000
gold:ln_article_age	-0.1392	0.870	0.000
green:ln_article_age	-0.1868	0.830	0.000
hybrid:ln_article_age	-0.2230	0.800	0.000

Table 7.3: Coefficients for OA adoption by policy. Results for the first sample, model 2, $R^2 = 0.073$.

Feature	coef	odds_ratios	P>z
bronze	-0.111	0.895	0.054
gold	1.090	2.974	0.000
green	0.834	2.302	0.000
hybrid	1.156	3.177	0.000
ln_article_age	-0.070	0.933	0.000
ln(SJR)	-0.244	0.783	0.000
bronze:ln(article_age)	0.029	1.030	0.007
gold:ln_article_age	-0.245	0.783	0.000
green:ln_article_age	-0.171	0.843	0.000
hybrid:ln_article_age	-0.274	0.760	0.000
ln(1 + *times_cited*)	0.447	1.564	0.000

7.5 Discussion

The surge in popularity and growth of OA has significantly contributed to the dissemination of scientific knowledge. Our research highlights Wikipedia's increasing reliance on OA articles, constituting 46.5% of all scientific citations on the platform, a notable rise from the 31.2% reported in the prior study (Pooladian and Borrego, 2017). This trend has shown continuous growth, particularly evident in scientific articles cited by Wikipedia that were published after 2011, where at least 50% are OA. In comparison, only 30% of articles in the Web of Science database and 20% in the OpenAlex database were OA during the same period These findings align with the broader scientific community's trend, as evidenced by the percentage of OA articles steadily increasing to 28% in 2018, with OpenAlex reporting 47% (Piwowar et al., 2018). Despite high-impact journals remaining a preferred source for Wikipedia (Nielsen, 2007), variations in the distribution of OA articles within journals emphasize the necessity for a nuanced, article-level approach.

Our examination of OA policies in scientific articles and Wikipedia unveils a parallel trend (Piwowar et al., 2018). Bronze policy (16.10%) and green policy (13.52%) dominate as the most common OA policies in Wikipedia. The higher prevalence of green policy in Wikipedia compared to scientific articles suggests differences in reference acquisition methods between Wikipedia editors and researchers. This trend further reinforces the importance of not overlooking articles accessible through green routes, thereby avoiding underestimating the impact OA can have on disseminating scientific knowledge through Wikipedia (ElSabry, 2017).

Our study further reveals disparities in OA Wikipedia citations across disci-

plines, with biology, physics, and mathematics exhibiting higher OA citation rates, while social sciences and humanities show comparatively lower rates. Nevertheless, Wikipedia's robust reliance on OA articles persists across all OpenAlex root concepts.

In addition, our analysis reveals an "OA citation advantage" in Wikipedia, meaning that OA publications are more likely to be cited as references in Wikipedia compared to paywalled publications. Specifically, under similar conditions, OA articles have a 64.7% higher likelihood of being cited in Wikipedia than their paywalled counterparts. Despite the significantly negative effect of the SJR on citation likelihood, we found that over 90% of articles cited in Wikipedia have an SJR below 10, with nearly 80% below 5. This distribution indicates that Wikipedia editors prioritize the accessibility of reliable sources over the prestige of the journals in which they are published. Furthermore, the likelihood of an OA article being cited increases with its citation count but decreases as the article ages. Wikipedia's editors demonstrate strong responsiveness to new scientific developments, frequently updating content referencing OA articles published within the past four years, reflecting a clear preference for recent and easily accessible scientific knowledge.

We acknowledge certain limitations in our study. First, by focusing exclusively on articles with DOIs, we excluded conference papers and earlier literature. Future research could benefit from including these additional sources. While our regression model accounted for significant factors, such as OA status, OA policy, and citation counts, other causal variables like article length may influence article citations on Wikipedia. Furthermore, our study did not consider time as an analytical dimension, prompting future research to delve into Wikipedia's edit history for specific data at the time of article citation, facilitating a deeper understanding of the causal mechanisms underpinning the interplay between OA and Wikipedia. Additionally, Wikipedia supports dual citations, allowing both paywalled and OA versions of a source to be cited together. This functionality, supported by tools like Wikipedia's OABOT[8], facilitates the addition of OA links to paywalled citations, improving access without violating copyrights. As a result, some paywalled publications can even be accessed through these OA links. Future studies could more comprehensively explore the impact of OA on Wikipedia by considering this broader context.

7.6 Conclusion

This chapter assessed the impact of open access (OA) on Wikipedia by analyzing article-level features using a comprehensive dataset of Wikipedia citations, OA metrics from OpenAlex, and journal data from Scimago. Our findings reveal that

[8] https://en.wikipedia.org/wiki/Wikipedia:OABOT#:%7E:text=Wikipedia%20links%20to%20hundreds%20of,does%20not%20violate%20any%20copyrights

OA articles are increasingly cited over time, with their proportion on Wikipedia significantly exceeding that in the broader scientific literature. Moreover, OA articles enjoy a citation advantage on Wikipedia, with a greater likelihood of being referenced compared to similar paywalled articles. This advantage is particularly pronounced for highly cited articles and those published within the past four years. These results underscore the importance of OA in broadening the dissemination of scientific knowledge, especially on influential platforms like Wikipedia, where newer and more impactful articles are more likely to reach a wider audience.

Our study lays the groundwork for further research on Wikipedia and open science. Future studies should consider a broader range of sources and variables to more fully understand the OA effect on Wikipedia. Additionally, exploring other aspects of open science, such as open research data and software, through similar methodologies could provide further insights. In conclusion, our study highlights the significance of OA in Wikipedia and its potential broader impact, offering a foundation for future research and contributing to the understanding of OA's role in the dissemination of scientific knowledge.

By demonstrating the citation advantage of OA publications, this chapter shows how accessibility reshapes which science is visible on Wikipedia. At the same time, the increased prominence of OA sources also raises questions about how they are used in contested settings, where openness may amplify both visibility and conflict. These issues are taken up in Chapter 8, which investigates how OA publications are implicated in editorial disputes and what this reveals about the negotiation of scientific authority on Wikipedia.

Chapter 8

Weaponized Citations: The Role of Open Access Publications in Wikipedia's Scientific Disputes

8.1 Introduction

Chapters 4, 5 and 7 examined which sources enter Wikipedia, how they may be biased, and how accessibility affects their visibility. In this final empirical chapter, we shift from coverage and access to conflict. Disputes are a central aspect of Wikipedia's collaborative process and provide a window into how scientific knowledge is contested, defended, and negotiated in public. This chapter therefore examines the role of OA publications within editorial disputes.

Wikipedia has emerged as one of the largest and most accessible online encyclopedias in the digital age, serving millions of users worldwide across a wide range of fields. Its open editing model allows virtually anyone to contribute or modify content, which has facilitated rapid growth and continuous updates. However, this openness also introduces challenges, including vandalism, the spread of misinformation, and edit conflicts (Priedhorsky et al., 2007; Kittur et al., 2007b; Sumi et al., 2011). These conflicts often arise when multiple editors repeatedly undo each other's contributions due to disagreements (Kittur et al., 2007b; Yasseri et al., 2012). A core principle guiding editorial practice on Wikipedia is the Neutral Point of View (NPOV), which requires that articles be written fairly, accurately, and supported by verifiable sources. This principle is especially important for articles addressing contentious issues or biographies of living persons[1]. Within this framework, not only Wikipedia articles content but also the citations used to support claims can become central to editorial disputes.

In addition to its encyclopedic function, Wikipedia increasingly serves as a platform for public engagement with scientific knowledge (Jemielniak and

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

Aibar, 2016; Segev and Sharon, 2017; Shafee et al., 2017). Its commitment to transparency, evident in open revision histories, publicly accessible discussion pages, and community-enforced guidelines, has contributed to its credibility among both general readers and professionals. Unlike academic journals, which often rely on closed peer review processes and involve long publication timelines, Wikipedia allows for immediate updates. Editors can incorporate newly published scientific findings in near real-time. This responsiveness makes Wikipedia a dynamic platform where scientific knowledge is not only disseminated but also interpreted, negotiated, and at times contested in public view (Benjakob and Aviram, 2018; Black, 2008).

Recent research has recognized Wikipedia as a space where scientific disputes become visible (Weltevrede and Borra, 2016; Wilson and Likens, 2015). Editors frequently disagree on how to deal with contested scientific claims, and such disputes are often mediated through community policies concerning verifiability, reliable sourcing, and neutrality (Hara and Doney, 2015; Wyatt et al., 2016; Steiert, 2025). These editorial interactions offer valuable insights into how scientific authority is constructed, reinforced, or challenged within a collaborative and public knowledge environment.

While previous research on Wikipedia conflicts has primarily focused on editorial behavior and content disputes, often using metrics such as edit frequency, revert patterns, and editor activity (Yasseri et al., 2012), the role of scientific citations within these disputes has received comparatively less attention. Scientific publications, typically regarded as authoritative sources, are frequently cited to support or refute competing interpretations. Examining how these sources are mobilized in editorial conflicts can offer important insights into the relationship between scientific knowledge production and its contestation in digital public settings.

One particularly relevant dimension of this issue concerns the role of open access (OA) publications. OA publications, which are freely available to all readers, have become increasingly common and are often associated with greater visibility and citation impact (Björk and Solomon, 2012; Piwowar et al., 2018). Their accessibility may make them more likely to be cited in contentious discussions, as they are easier to retrieve, read, and assess by a broader range of contributors. This raises the possibility that OA publications are more frequently involved in editorial disputes on Wikipedia. Despite this plausible connection, empirical research assessing whether OA publications are disproportionately cited in contentious editing contexts remains limited.

In light of these gaps, this chapter aims to investigate the role of OA publications in disputes identified within Wikipedia articles. Specifically, we address the following research questions:

- **RQ1**: What are the most important characteristics of scientific publications that are involved in scientific disputes?

- **RQ2**: Are OA publications more likely to be used in scientific disputes compared to paywalled publications?

To answer these questions, we construct a comprehensive dataset that integrates Wikipedia edit histories with metadata from Crossref and OpenAlex. This dataset includes 3,514 scientific publications cited across 2,221 identified dispute cases on Wikipedia. We analyze the characteristics of these publications and compare OA and paywalled articles in terms of their involvement in disputes, using multiple indicators of *dispute intensity*. Furthermore, we employ logistic regression models to estimate the likelihood that a scientific publication is cited in a dispute, controlling for factors such as article age, citation count, field of research, and journal type.

By examining the relationship between OA status and the likelihood of involvement in citation-related conflicts, this study offers new insights into how scientific knowledge is mobilized, contested, and negotiated within the digital public sphere. Our findings contribute to the literature on science communication, OA, and the governance of user-generated knowledge platforms such as Wikipedia.

8.2 Previous Work

8.2.1 Wikipedia and its Controversies

Wikipedia is widely recognized as a pioneering open knowledge platform, characterized by its collaborative and open editing model (Kittur et al., 2007b). While this model fosters inclusiveness and facilitates the rapid dissemination of knowledge, it also creates potential for editorial disagreements. One prominent form of conflict is known as an *edit war*, where contributors repeatedly override each other's edits due to differing perspectives (Yasseri et al., 2012). These conflicts raise important questions about the reliability and neutrality of collaboratively produced content (Arazy et al., 2011).

In most cases, Wikipedia articles evolve through a predominantly constructive and cooperative process. Editors tend to improve each other's contributions by expanding content, correcting factual or grammatical errors, and gradually converging toward a stable, consensus-based version (Wilkinson and Huberman, 2007). According to Wilkinson and Huberman, nearly 99% of articles on the English Wikipedia develop through this relatively smooth and incremental editorial dynamic. Representative examples of such evolution include entries like `Benjamin Franklin`, `Pumpkin`, and `Helium`, which have seen steady refinement without major conflict.

However, this collaborative ideal does not always hold. In articles dealing with controversial, politicized, or high-profile topics, editorial collaboration can give way

to conflict. Disputes often take the form of so-called *edit wars*[2], where opposing groups repeatedly override each other's changes. Schneider et al. (Schneider et al., 2010) estimate that for frequently edited or heavily viewed pages—a pair of characteristics that tend to co-occur (Ratkiewicz et al., 2010)—around 12% of talk page discussions focus on reversions or vandalism[3]. These findings suggest that article popularity and controversy are strongly linked, with heightened attention often coinciding with editorial instability.

To address such challenges, the Wikipedia community has established a comprehensive set of governance mechanisms aimed at mitigating conflict and preserving content quality. These include the well-known "three-revert rule," temporary or indefinite page protection, editorial tagging to flag controversial material, and administrative measures such as user warnings, temporary suspensions, or permanent bans for persistently disruptive behavior.[4]

Research in this area has largely focused on editorial behaviors and article-level indicators of controversy. Various approaches have been developed to quantify editorial conflict on Wikipedia. For instance, one of the earliest and most widely used metrics is the count of reverts and total number of edits, where a high number of reverts suggests a contentious article (Kittur et al., 2007b). Another commonly used metric is the number of words deleted by one user from another's contributions. This approach defines disputes between user pairs based on the amount of content removed by each party from the other's edits (Vuong et al., 2008). A third approach involves analyzing mutual reverts between two editors. When two contributors repeatedly undo each other's changes, it often signals the presence of a sustained disagreement (Suh et al., 2007). Temporal features have also been employed in measuring controversy. One such measure is the inverse of the time interval between consecutive edits made by different users. Shorter intervals between conflicting edits may reflect intense editorial disputes (Brandes and Lerner, 2008). Similarly, West et al. (West et al., 2010) and Adler et al. (Adler et al., 2011) utilized temporal patterns of editing behavior to detect instances of vandalism on Wikipedia. Their central argument is that edits deemed problematic tend to be reverted more rapidly than non-problematic ones, suggesting that the time interval between an edit and its subsequent reversion can serve as a reliable indicator of disruptive activity. Yasseri et al. (Yasseri et al., 2013) proposed a composite metric of controversy, denoted as M, which aggregates the weights of mutually reverting editor pairs and multiplies this sum by the total number of editors engaged with the article. This method captures both the depth and breadth of conflict within a Wikipedia page.

Although these methods provide valuable insights into the dynamics of editorial behavior and have been instrumental in identifying controversial content, they

[2] `https://en.wikipedia.org/wiki/Wikipedia:Edit_warring`

[3] Vandalism on wikis includes the insertion of hoaxes, offensive language, or nonsense text. Reversions are edits used to restore content to its previous, stable version.

[4] See `https://en.wikipedia.org/wiki/Wikipedia:Edit_warring` for policy details.

primarily focus on surface-level interactions such as editing frequency, reversion patterns, and word-level deletions. The role of references and citations, particularly scientific ones, has been relatively understudied in this body of work. Given Wikipedia's emphasis on verifiability and reliable sourcing—especially in articles on contentious topics—citations themselves can become a site of disagreement. Understanding how scientific publications are cited, disputed, or removed during Wikipedia conflicts could offer new perspectives on the intersection between scientific knowledge and public discourse.

8.2.2 Scientific Citations and Disputes in Wikipedia

Scientific citations serve not only to acknowledge prior research but also function as instruments of persuasion, authority, and boundary-setting within both academic and public discourse (Nigel Gilbert, 1977; Cozzens, 1989). In contested contexts, citations are often used strategically to support specific claims, challenge opposing arguments, or align with particular epistemic communities (Gilbert and Mulkay, 1984; Hyland, 1999).

On Wikipedia, where editorial guidelines such as "verifiability" and "no original research" place strong emphasis on reliable sourcing, scientific references play a particularly central role (Benjakob et al., 2022; Lewoniewski et al., 2023; Konieczny, 2016). Editors rely on scientific publications not only to substantiate statements, but also to resolve disputes, justify inclusion or exclusion of content, and negotiate editorial consensus. In articles addressing topics often surrounded by controversy, such as climate change, COVID-19, or gender identity, the selection and interpretation of specific scientific sources frequently becomes a focal point of disagreement (Esteves Gonçalves da Costa and Cukierman, 2019; Benjakob et al., 2022; Currie, 2012). Despite the importance of citations in these disputes, there remains a lack of research on the characteristics of the cited literature and how these characteristics may influence the likelihood of citation-related disputes.

One potentially important factor that has received limited attention is the role of accessibility to scientific publications. OA publications, which are freely available to the public, are generally associated with increased visibility, broader dissemination, and greater uptake in social media and public discourse compared to subscription-based articles (Piwowar et al., 2018; Björk and Solomon, 2012). This broader exposure may contribute to a higher likelihood of being cited on platforms like Wikipedia (Yang et al., 2024). Because OA publications are more readily accessible to editors and readers alike, they may be more frequently included in articles, more actively scrutinized by the community, or more easily drawn into points of contention. However, whether OA articles are more likely to be involved in citation-related disputes remains an open empirical question.

Previous research has provided important insights into how scientific knowledge circulates in digital environments. Yet, few studies have directly examined how specific attributes of scientific publications—such as OA status, citation impact,

disciplinary field, or publication venue—are associated with their presence in Wikipedia disputes. Moreover, most existing work has examined either editorial behavior or citation patterns in isolation, without integrating these perspectives into a unified empirical framework.

This study seeks to address these gaps by investigating how features of scientific publications, particularly OA status, are related to their involvement in citation disputes on Wikipedia. By combining large-scale Wikipedia editing histories with detailed bibliometric metadata from sources such as Crossref and OpenAlex, this research provides new insights into how scientific literature is not only cited but also challenged and contested within one of the world's most widely used open knowledge platforms.

8.3 Methodology

8.3.1 Data Collection and Sources

This study draws on data from three primary sources. First, we used Crossref Event Data curated by CWTS, which records citation-related events on Wikipedia. These events capture the addition and removal of references to scientific publications identified by Digital Object Identifiers (DOIs). Second, we utilized the Wikipedia API to retrieve article revision histories, editor identifiers, and page-level metadata such as namespace classification. Third, we integrated metadata from the OpenAlex dataset to obtain attributes of the cited scientific publications, including OA status, publication year, citation count, disciplinary classification, and retraction status.

8.3.2 Temporal Scope, Dispute Definition, and Filtering

Crossref Event Data has recorded citation events from April 11, 2017 onward. To ensure the completeness of revision histories from article inception, we restricted our sample to English-language Wikipedia articles whose earliest revision date occurred on or after January 1, 2018. This yielded 609,256 Wikipedia articles with a combined total of over 71 million revisions.

Previous studies on disputes in Wikipedia have primarily focused on edit wars, and adopt the definition provided directly by Wikipedia[5]. These studies have centred on detecting such conflicts, typically using one of two approaches. The first involves analyzing the associated talk pages[6] for indicators such as cleanup tags, strong language, or hostile interactions. In some cases, the length of the talk page may exceed that of the article itself, requiring the archiving of earlier discussions. The second approach examines the revision history of an article,

[5] `https://en.wikipedia.org/wiki/Wikipedia:Edit_warring`

[6] `https://en.wikipedia.org/wiki/Help:Talk_pages`

where edit war behavior can often be identified by patterns of repeated reverts among contributors (Sumi et al., 2011).

In a broader sense, scientific disputes have been characterized as sustained and public disagreements involving conflicting knowledge claims. According to McMullin, a scientific controversy exists when both sides assert the validity of their own views while challenging the other's, and when these views are presented as being grounded in scientific reasoning (McMullin, 1987). The disagreement must also be visible to a wider audience through written or spoken communication, rather than remaining private, allowing others to evaluate the competing claims.

Building on these prior perspectives and considering the specific context of scientific publications, we operationalise a scientific dispute as a sequence of citation changes involving repeated addition and removal of the same DOI by at least two distinct editors within a short time frame. Specifically, a dispute is identified when the same DOI is inserted or removed at least three times within a seven-day period.

We evaluated alternative temporal windows of one, three, seven, and thirty days. The seven-day window was selected as a compromise: shorter intervals risked excluding slower-developing disputes, while longer periods tended to capture unrelated content updates, obsolecense or page evolution. Applying this definition, and restricting the scope to article pages (namespace = 0)[7], we identified 193,416 candidate scientific disputes across 2,807 Wikipedia pages and 9,602 unique DOIs.

8.3.3 Dispute Validation and Consolidation

An initial manual inspection of 100 randomly sampled events revealed that only 33 percent represented genuine disputes. The remaining events primarily stemmed from incomplete or incorrect records in the Crossref Event Data. To improve reliability, we implemented a two-step verification process.

First, for each event, we used the Wikimedia API to retrieve the article text before and after the corresponding revision. We then compared different versions of the same article and excluded events where the detected text differences did not include the corresponding DOI. Second, we randomly sampled 100 events from the filtered dataset for manual validation. Among these, 88 percent were confirmed as genuine scientific disputes. The remaining cases typically involved legitimate DOI changes, but lacked clear evidence of disagreement, such as edit summaries indicating a dispute. The random sampling was designed to ensure representativeness of the filtered dataset.

After validation, we retained 10,884 confirmed dispute events involving 1,681 Wikipedia articles and 3,548 distinct scientific publications. To account for prolonged or fragmented disputes, we consolidated overlapping events on the same

[7]A Wikipedia namespace is a categorization system that groups pages by their function or content type, helping distinguish between articles, user pages, discussions, drafts, and project-related content. `https://en.wikipedia.org/wiki/Wikipedia:Namespace`

article into unified conflict episodes. Specifically, two events were merged if they occurred within overlapping time windows and involved at least one common revision, the process can be found in 8.1. This conservative merging criterion minimized the risk of erroneously combining unrelated events, helping ensure that each episode reflected a coherent editorial disagreement over a specific scientific source. The final consolidation process yielded 2,221 distinct scientific dispute episodes.

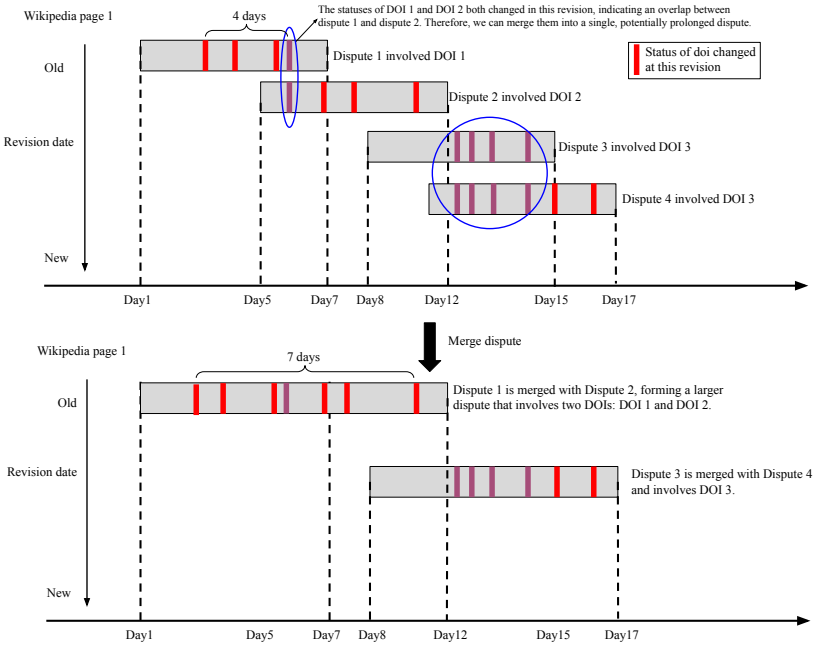


Figure 8.1: Merging procedure for scientific disputes.

8.3.4 Metadata Enrichment and Final Dataset

To contextualize the dispute episodes, we enriched the dataset with structured metadata from both OpenAlex and Wikipedia API. For each cited scientific publication, we collected information on OA status, publication year, citation volume, primary disciplinary concept, and retraction status. Additionally, we extracted topic classifications for each Wikipedia article using the ORES ArticleTopic[8] model and retrieved structural characteristics of the articles.

After removing entries with incomplete metadata, the final dataset consisted of 2,221 consolidated scientific dispute episodes. These episodes involved 1,669 Wikipedia articles and 3,514 unique scientific publications, all with complete metadata coverage.

[8] `https://www.mediawiki.org/wiki/ORES/Articletopic`

8.4 Results

8.4.1 Characteristics of Publications Involved in Scientific Disputes

We begin by examining the distribution of scientific disputes across academic disciplines. Figure 8.2 presents the frequency and duration of disputes for each primary concept, based on scientific articles. Concepts are ordered by the total number of dispute episodes (shown on the right), while the left panel displays the distribution of dispute durations in seconds. To aid interpretation, three vertical reference lines are included: a red dotted line marking one day, a light blue line marking one week, and a dark blue line indicating two weeks.

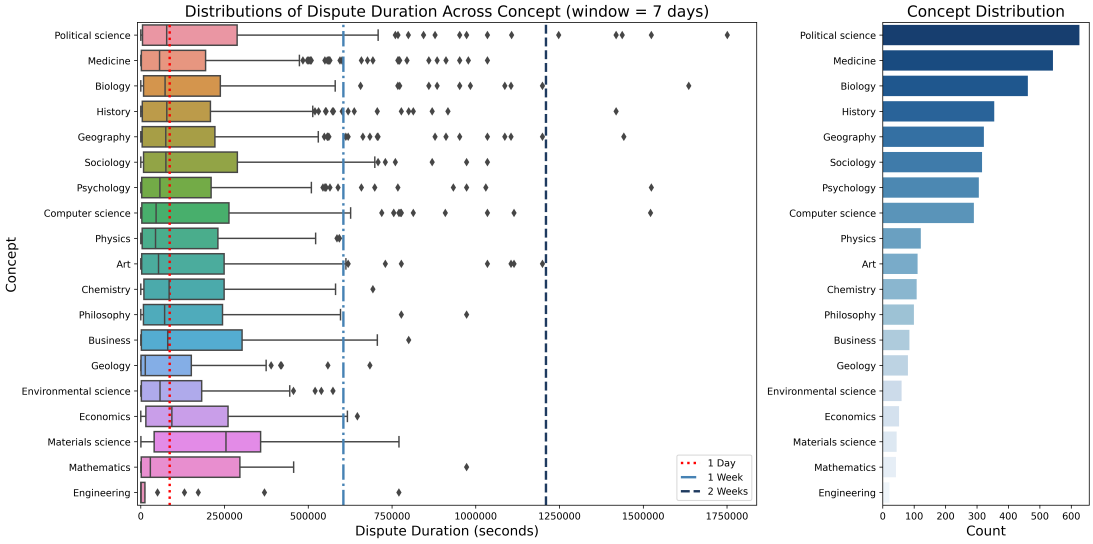


Figure 8.2: Distribution of scientific dispute durations across disciplines.

The most frequently contested fields include Political Science, Medicine, and humanities-related disciplines such as History, Geography, and Sociology. In contrast, foundational sciences like Physics, Chemistry, Mathematics, and Engineering are markedly less represented. Mathematics and Engineering show the lowest dispute frequencies.

In terms of duration, most disputes are short-lived, typically resolving within three days. The average dispute lasts less than one day, suggesting that many editorial disagreements are addressed or lose momentum quickly. Among the exceptions, materials science exhibits the longest average dispute durations, whereas geology shows the shortest. The relatively brief duration in Engineering may reflect the small number of disputes observed (only 21 articles). Overall, we find limited variation in dispute duration across disciplines, indicating that Wikipedia editors tend to resolve disputes at similar speeds regardless of topic. Alternatively,

it is possible that unresolved disputes migrate to talk pages rather than continuing on the main article pages.

To further understand the temporal dynamics of scientific disputes, Figure 8.3 visualizes when disputes occurred relative to publication year. Each point represents a scientific article involved in a dispute, with the x-axis showing publication year and the left y-axis representing dispute count. The point shape indicates access type (circles for OA, crosses for paywalled), while the point size corresponds to the number of dispute episodes per article. The red trend line, plotted against the right y-axis, represents the total number of disputes per year with a shaded 95% confidence interval.

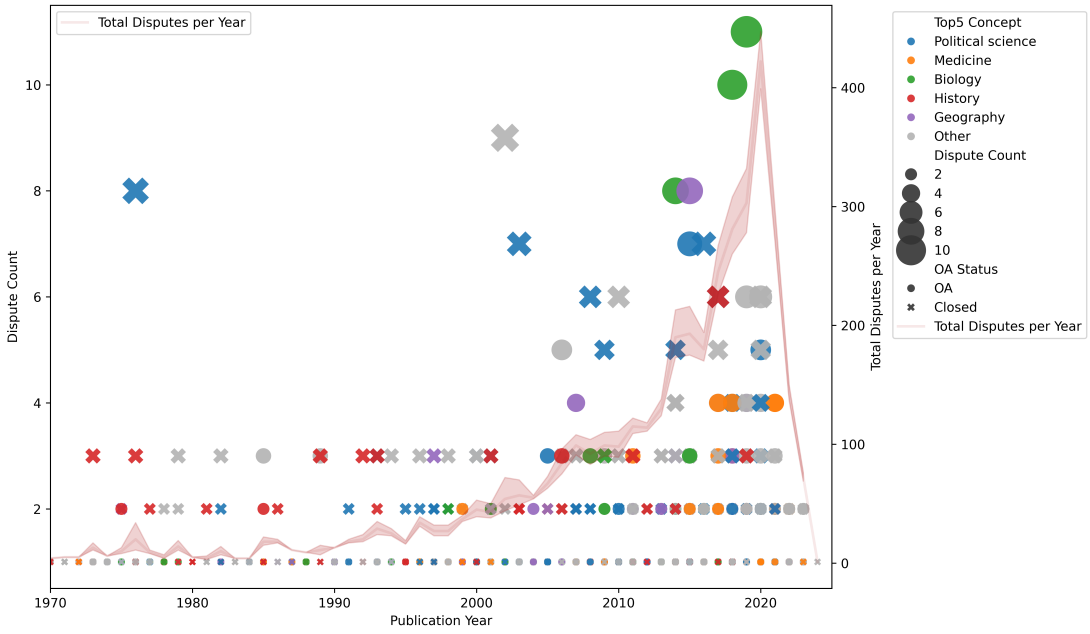


Figure 8.3: Temporal trends and disciplinary distribution of scientific disputes.

To enhance visual clarity, we color only the top 5 most disputed concepts; all others are grouped under "Other" in light gray. Most articles appear in only a single dispute episode. However, since the early 2000s, an increasing number of articles have been involved in multiple disputes. The article with the highest dispute frequency is a 2019 publication in Frontiers in Genetics titled "Assortative Mating on Ancestry-Variant Traits in Admixed Latin American Populations", which was cited in 11 separate dispute episodes. These disputes revolved around contested interpretations of genetic admixture and its implications for demographic classification and racial identity. Editors contested the reliability of the study, juxtaposing it with sources such as the CIA World Factbook and the Latinobarómetro survey. Disagreements involved content addition/removal, accusations of bias or vandalism, and broader tensions over the role of scientific sources in politically sensitive topics.

Overall, articles in Political Science and Biology are more frequently involved in high-frequency disputes. In contrast, articles in History, Sociology, and Psychology tend to be associated with more moderate or low-intensity dispute activity. Furthermore, we observe a growing number of disputes over time, especially concerning recently published articles, which are more likely to be cited and edited.

Next, we assess whether citation impact differs between disputed and non-disputed publications. Using log-transformed citation counts, we find that disputed articles exhibit slightly higher citation levels (median = 3.66, mean = 3.69, SD = 1.81) than non-disputed ones (median = 3.56, mean = 3.49, SD = 1.74). A Mann–Whitney U test indicates that this difference is statistically significant ($p = 0.0002$), though modest in effect size. These findings suggest that articles involved in disputes tend to be marginally more cited, potentially reflecting their broader visibility or relevance to contested topics.

We also examine the role of journal prestige by comparing the top 10 journals most frequently associated with disputed publications. Figure 8.4 shows each journal's total number of disputes and corresponding average SNIP (source normalized impact per paper) scores[9]. Highly prestigious journals such as Nature, Science, and the New England Journal of Medicine exhibit both high dispute frequencies and high SNIP scores. This suggests that articles from high-impact journals may be more likely to attract editorial scrutiny on Wikipedia.

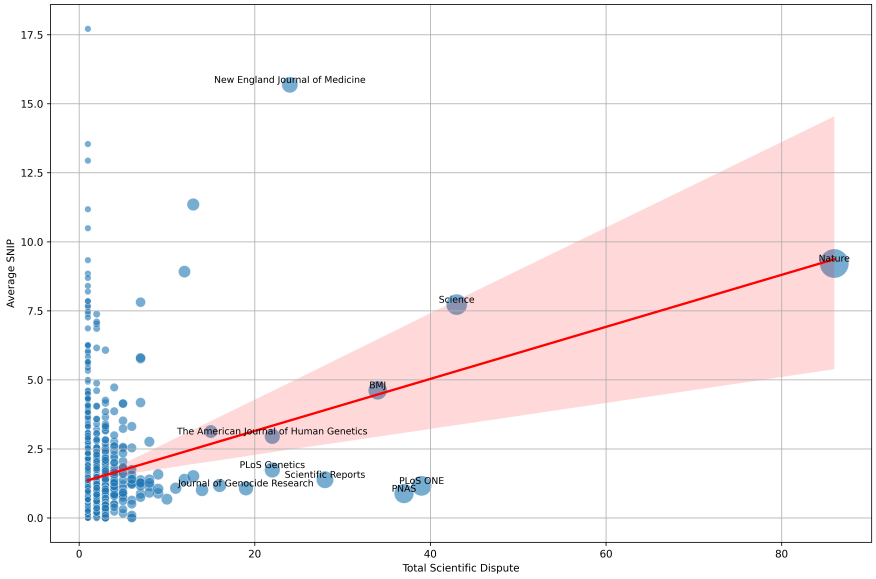


Figure 8.4: Dispute frequency and journal Prestige (SNIP).

Notably, some journals, such as PLoS ONE and PNAS, also rank high in dispute frequency despite relatively modest SNIP scores. These cases may reflect

[9]SNIP is a key metric from CWTS that reflects the average citation impact of a journal's publications.`https://www.journalindicators.com/`

their multidisciplinary scope, high publication volumes, or frequent citation across diverse Wikipedia articles. This pattern suggests that both journal prestige and publication exposure may influence the likelihood of scientific disputes.

In addition, we distinguished between registered and anonymous editors by identifying whether an editor's username was an IP address. Based on this classification, we analyzed the duration of scientific disputes across different primary concepts and editor types, as shown in Figure 8.5.

Each dispute was categorized into one of three editor-type groups: disputes exclusively involving registered editors, those involving only anonymous editors, and those jointly edited by both types. The y-axis in the figure represents different scientific concepts, while each cell in the heatmap indicates the average duration (in seconds) of disputes for the corresponding editor type and concept combination.

The results reveal several noteworthy patterns. First, the majority of scientific disputes occur either among registered editors or between registered and anonymous editors. In contrast, disputes occurring solely between anonymous editors are relatively rare and tend to be short-lived. An exception is observed in the field of Geography, where anonymous-only disputes lasted around four days on average.

Moreover, disputes involving both registered and anonymous editors generally lasted longer than those limited to registered editors, particularly in domains such as Materials Science, Political Science, Art, and Sociology. This suggests that dispute dynamics are not only influenced by the scientific topic at hand, but also by the type of contributors involved in the editing process. The presence of both editor types may indicate more persistent disagreements or challenges in reaching consensus.

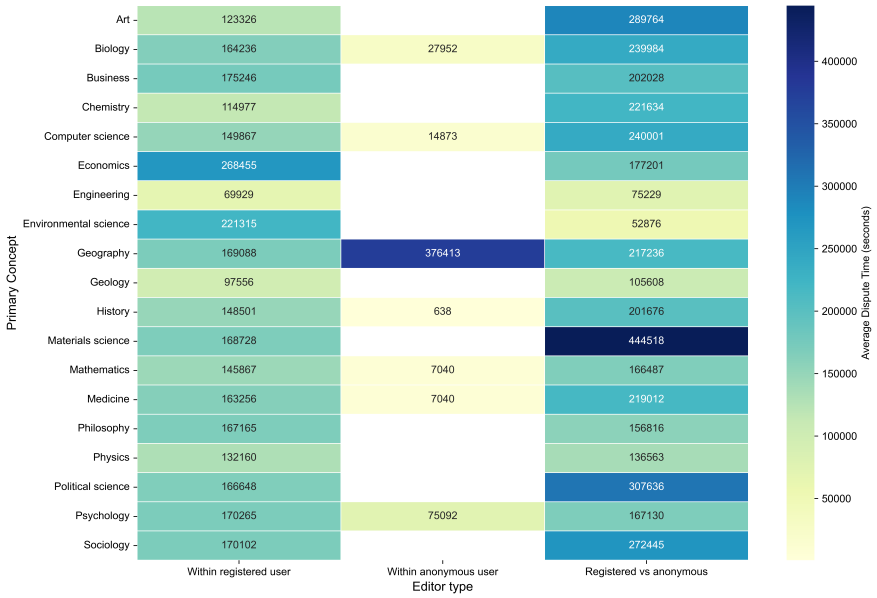


Figure 8.5: Average dispute duration by editor type and concept.

8.4.2 Descriptive Analysis of Open Access Articles and Disputes

To address the second research question regarding whether OA publications are more likely to be involved in scientific disputes, we begin with a descriptive overview.

First, we examined the distribution of OA types among the disputed articles in our dataset. Paywalled publications account for the largest share, comprising 54.3% of all disputed articles. This pattern is consistent with earlier findings (Yang et al., 2024) that indicate approximately 55% of Wikipedia-cited scientific articles are not openly accessible. Among OA categories, the most frequent are Bronze (14%), followed by Green (11.5%), Gold (10.5%), Hybrid (7.5%), and Diamond (2.2%).

Next, we examine how the proportion of OA articles involved in disputes has evolved over time. Figure 8.6 shows a stacked bar chart based on publication year, with the left y-axis indicating the count of disputed articles and the right y-axis showing the proportion of OA articles. Closed articles are shown in red, and OA articles are in blue. The grey dotted line represents the annual proportion of OA articles among disputed publications, while the solid green line depicts the overall OA publication rate in the OpenAlex dataset.

The figure illustrates two key trends. First, more recently published articles tend to be involved in a higher number of disputes. Second, the proportion of OA articles among disputed publications has increased steadily over time. This proportion often exceeds the overall OA prevalence observed in the broader scientific literature, particularly after the year 2000. These patterns suggest that OA articles may play an increasingly visible role in the formation of disputed knowledge on Wikipedia.

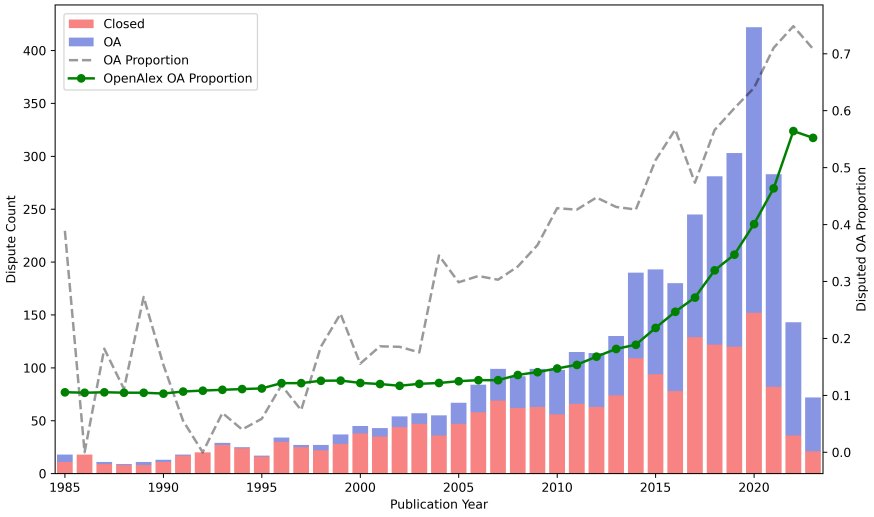


Figure 8.6: Temporal trends in OA proportion among disputed articles.

To explore disciplinary variation in OA involvement, we constructed a heatmap that illustrates the number of disputes involving each OA type across different scientific concepts (Figure 8.7). Each cell reflects the frequency of disputes, with darker colors indicating higher counts.

The heatmap highlights clear differences across domains. Political Science exhibits the highest dispute frequency, with approximately 72% of these disputes involving paywalled articles. In comparison, Biology and Medicine show a more balanced distribution. Although paywalled still dominates, the proportions of Green, Bronze, and Gold OA are relatively higher in these fields. Diamond OA appears least involved across all domains, which may reflect either lower publication volumes or different editorial practices.

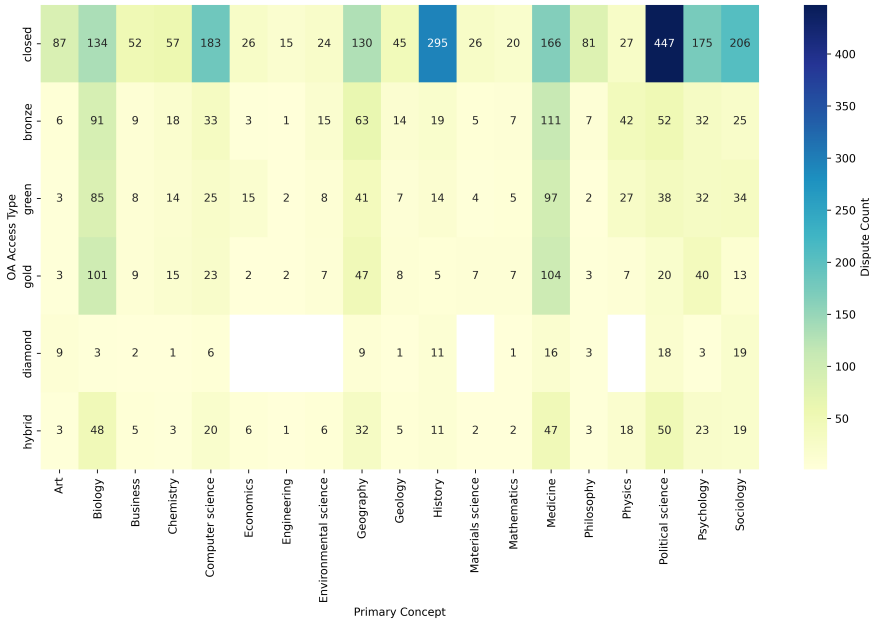


Figure 8.7: Heatmap of dispute frequency by OA type and concept.

These descriptive findings provide initial evidence that both OA status and disciplinary domain are associated with patterns of Wikipedia disputes involving scientific literature. Subsequent regression and survival analyses further test the statistical significance and interaction effects of these variables.

8.4.3 The Role of Open Access Articles in Scientific Disputes

To examine whether OA articles are more likely to become subjects of scientific disputes, we first classified our dataset into two groups: 3,514 articles identified as involved in scientific disputes, and all remaining articles without dispute records (N = 1,140,242). A binary variable `has_dispute` was created to indicate dispute involvement (1 = involved, 0 = not involved).

We constructed a logistic regression model with `has_dispute` as the dependent variable. Key continuous predictors include:

- **Article Age** (`ln_article_age`): The age of the article since publication, measured in months and log-transformed to reduce skewness.
- **Citation Count** (`ln1p_cited_by_count`): The number of citations received, log-transformed after adding 1 to handle zero values.
- **OA Status** (`is_oa`): A binary indicator where 1 represents OA articles and 0 represents paywalled articles.
- **Concept**: A categorical variable representing the subject area of the article. Dummy coding was applied, with `Engineering` used as the reference category.
- **Retraction Status** (`is_retracted`): A binary indicator of whether the article has been retracted.
- **Number of References** (`n_refs`): The total number of references cited by the article.
- **Topic**: A categorical variable representing the topic of the article, extracted from OpenAlex. Dummy coding was applied, with `Modeling the Dynamics of COVID-19 Pandemic` used as the reference category.

All predictors exhibited variance inflation factor (VIF) values below 10, suggesting no multicollinearity issues. The model formula is as follows:

```
has_dispute = is_oa + ln_article_age*is_oa + ln1p_cited_by_count +
   C(concept, Treatment('Engineering')) + n_refs + is_retracted
```

The logistic regression results are presented in Table 8.1.

Results indicate that OA articles are significantly more likely to be involved in disputes ($\beta = 0.891$, $p < 0.001$), suggesting that publicly accessible research is more frequently cited and scrutinized. Articles with higher citation counts also show a positive association with dispute likelihood ($\beta = 0.197$, $p < 0.001$), consistent with the idea that high-impact or widely discussed works attract more editorial attention. Conversely, older articles are less likely to be disputed ($\beta =$ -0.521, $p < 0.001$), especially among OA publications, as indicated by a significant interaction term ($p < 0.001$).

Disciplinary differences are also evident. Compared to Engineering articles, those in Political Science ($\beta = 1.09$), Sociology ($\beta = 0.99$), History ($\beta = 0.79$), and Philosophy ($\beta = 0.64$) are significantly more likely to be disputed ($p < 0.01$ for all). This aligns with the intuition that fields dealing with social, historical, or

Table 8.1: Logistic regression predicting likelihood of Wikipedia dispute involvement.

Variable	**Coefficient**	***p*-value**	**Odds Ratio**
Intercept	-4.069	<0.001	0.017
OA (is_oa)	0.891	<0.001	2.438
Log Article Age (ln_article_age)	-0.521	<0.001	0.594
OA × Log Article Age	-0.212	<0.001	0.809
Log Citation Count (ln1p_cited_by_count)	0.197	<0.001	1.218
# References (n_refs)	-0.0017	<0.001	0.998
Retracted (is_retracted)	0.562	0.429	1.754
Art	0.427	0.062	1.532
Biology	-0.465	0.038	0.628
Business	0.247	0.307	1.280
Chemistry	-0.310	0.165	0.734
Computer Science	0.145	0.511	1.156
Economics	0.284	0.226	1.328
Environmental Science	0.543	0.076	1.721
Geography	0.345	0.119	1.412
Geology	-0.609	0.044	0.544
History	0.793	0.001	2.210
Materials Science	-0.312	0.210	0.732
Mathematics	-0.234	0.406	0.792
Medicine	0.271	0.218	1.311
Philosophy	0.645	0.008	1.906
Physics	-0.054	0.821	0.947
Political Science	1.090	<0.001	2.974
Psychology	0.464	0.056	1.591
Sociology	0.986	<0.001	2.680

Note: Reference category for disciplines is Engineering. Pseudo $R^2 = 0.0357$.

ideological content are more prone to editorial conflict. In contrast, articles from Biology and Geology are significantly less likely to be involved in disputes.

We conducted a similar analysis using article topic instead of concept. The model specification was:

```
has_dispute = is_oa + ln_article_age*is_oa + ln1p_cited_by_count +
    C(topic_top10, Treatment('Modeling the Dynamics of COVID-19
                Pandemic')) + n_refs + is_retracted
```

We limited topics to the top 10 most frequently occurring among disputed articles, grouped all others under an "Other" category, and used the smallest of the top 10—*Modeling the Dynamics of COVID-19 Pandemic*—as the reference. Results are presented in Table 8.2.

Table 8.2: Logistic regression predicting likelihood of Wikipedia dispute by topic.

Variable	**Coefficient**	***p*-value**	**Odds Ratio**
Intercept	-2.092	<0.001	0.124
OA (is_oa)	0.603	0.005	1.828
Log Article Age (ln_article_age)	-0.482	<0.001	0.618
OA × Log Article Age	-0.176	<0.001	0.839
Log Citation Count	0.143	<0.001	1.153
# References	-0.0021	<0.001	0.998
Retracted	0.346	0.628	1.413
American Political Thought and History	-0.681	0.021	0.506
Coronavirus Disease 2019	0.214	0.459	1.239
COVID-19 Research	-0.008	0.974	0.992
Genomic Analysis of Ancient DNA	1.237	<0.001	3.445
Intersectionality in LGBTQ+ Mental Health	-0.008	0.977	0.992
Other	-1.776	<0.001	0.169
Islamic Reform in Middle East	-0.115	0.672	0.892
Populism in Contemporary Politics	0.744	0.012	2.104
Stellar Astrophysics	-1.598	<0.001	0.202
Politics in Turkey	0.622	0.030	1.863

Note: Reference topic is "Modeling the Dynamics of COVID-19 Pandemic". Pseudo $R^2 = 0.04195$.

Again, OA status is positively associated with dispute involvement (OR = 1.83, $p = 0.005$). Newer articles are more likely to be disputed, particularly when they are OA, as indicated by the significant interaction effect between OA status and article age. Citation count is also positively associated with dispute likelihood, whereas the number of references has a weak negative association. Retraction status does not significantly predict dispute involvement.

At the topic level, articles categorized under *Genomic Analysis of Ancient DNA* exhibit significantly higher odds of dispute (OR = 3.44, $p < 0.001$), potentially reflecting ethical or interpretive controversies. Topics related to contemporary politics—such as *Populism in Contemporary Politics* and *Politics in Turkey*—are also associated with higher dispute likelihoods. Conversely, technical fields like *Stellar Astrophysics* show significantly lower odds of dispute, suggesting that such topics are less contentious in the public or editorial sphere. Articles grouped under "Other" topics are the least likely to be disputed, further confirming that less socially relevant or debated topics are less prone to editorial contention.

While logistic regression provides insight into the likelihood of dispute involvement, we were also interested in understanding how quickly OA articles become subject to disputes compared to paywalled articles. We defined survival time as the number of days between an article's publication and its first involvement in a dispute. We then applied Kaplan-Meier survival analysis, stratified by OA status, to visualize the time-to-dispute distributions, as shown in Figure 8.8.

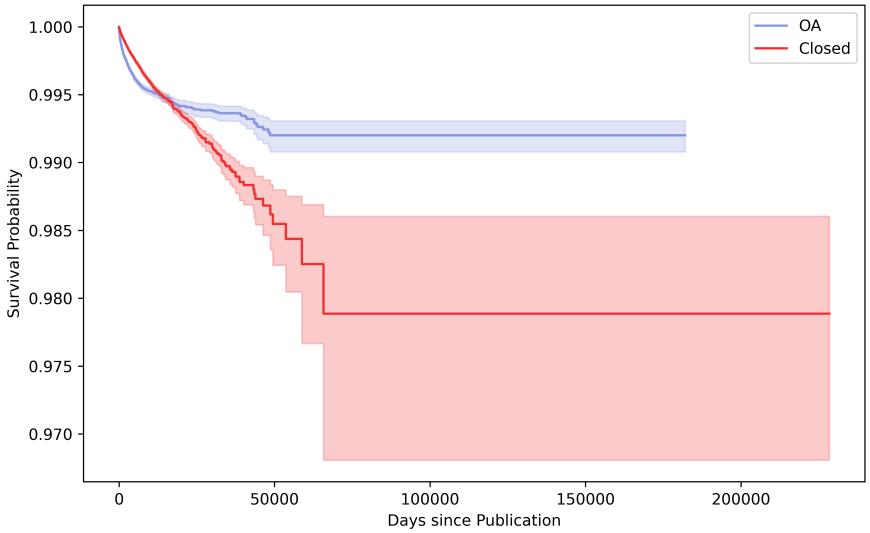


Figure 8.8: Kaplan-Meier survival curves for time to first dispute by OA status.

In the figure, the x-axis represents days since publication, and the y-axis indicates survival probability (i.e., the probability of not having been involved in a dispute). Blue and red lines correspond to OA and paywalled articles, respectively, with shaded areas indicating 95% confidence intervals.

The survival curves reveal notable differences between the two groups. In the early stage after publication, OA articles tend to become involved in disputes more quickly than paywalled articles, as indicated by the steeper, near-exponential decline of the OA curve compared to the more gradual, linear-like decrease of the paywalled curve. At later time points, however, this pattern reverses: the OA curve stabilizes, whereas the paywalled curve continues to decline, reflecting a

more persistent accumulation of disputes over time. A log-rank test confirms that the difference is statistically significant ($p < 7.6 \times 10^{-15}$).

It should also be noted that the majority of articles in the dataset (approximately one million) never become involved in disputes, while only about 3,000 do. This substantial imbalance explains the narrow confidence intervals observed across most of the curves, alongside the much wider intervals in the long-tail region, where data are sparse.

To further explore the timing of dispute involvement, we fitted a Cox proportional hazards model using OA status, citation count, and concept as predictors. Results are shown in Table 8.3.

Table 8.3: Cox Proportional Hazards Regression Results

Variable	**Coef**	**HR (exp(coef))**	**p-value**	**Significance**
OA (is_oa)	0.37	1.45	<0.005	***
Citation count (log, ln1p_cited_by_count)	0.11	1.11	<0.005	***
Art	-0.31	0.73	0.20	
Biology	0.10	1.11	0.65	
Business	0.56	1.75	0.02	*
Chemistry	-1.24	0.29	<0.005	***
Computer science	-0.03	0.97	0.90	
Economics	-0.43	0.65	0.09	
Environmental science	1.67	5.29	<0.005	***
Geography	0.21	1.23	0.35	
Geology	0.61	1.84	0.01	**
History	1.92	6.79	<0.005	***
Materials science	-0.60	0.55	0.02	*
Mathematics	-0.39	0.68	0.15	
Medicine	0.77	2.15	<0.005	***
Philosophy	0.62	1.86	0.01	**
Physics	0.03	1.03	0.89	
Political science	2.37	10.72	<0.005	***
Psychology	1.80	6.07	<0.005	***
Sociology	0.27	1.32	0.22	

Note: Reference category for discipline is **Engineering**. Concordance: 0.60; Significance levels: *** $p < 0.005$, ** $p < 0.01$, * $p < 0.05$.

The analysis reveals that OA articles are significantly more likely to be disputed sooner (HR = 1.45, $p < 0.005$). Citation count is slightly positively associated with dispute timing (HR = 1.11, $p < 0.005$), suggesting that highly cited papers may attract disputes more quickly, possibly because they draw greater attention and scrutiny. Relative to Engineering, articles in Chemistry and Materials Science are significantly less likely to be disputed promptly. In contrast, articles in Environmental Science, Geology, History, Medicine, Philosophy, Political Science, and Psychology exhibit substantially higher hazards of dispute. This disciplinary variation may reflect differences in public visibility, societal relevance, and the extent to which research in these areas intersects with contentious or value-laden

issues.

Lastly, we also investigated whether scientific disputes involving OA publications tend to reach consensus more quickly, as reflected by shorter dispute durations, compared to disputes involving paywalled publications. Using the Mann–Whitney U test, we found no statistically significant difference in dispute duration between OA and paywalled articles ($p = 0.274$). Additionally, a multivariate regression analysis controlling for primary concept, citation count, and article age showed that OA status was not a significant predictor of dispute duration ($\beta = 0.1215$, $p = 0.213$).

8.5 Discussion

This study systematically analyzed the characteristics of scientific disputes on Wikipedia and examined the role of OA publications in such disputes. The discussion is organized around the two research questions (RQs) and grounded in the empirical findings presented in the Results section.

Regarding **RQ1**: *What are the characteristics of scientific publications that are involved in scientific disputes?* Our findings show that scientific disputes occur more frequently in the social sciences and humanities, particularly in fields such as Political Science, History, and Sociology. These disciplines often engage with issues related to social values, ideological differences, and political viewpoints, which are associated with higher epistemic uncertainty and interpretative flexibility. In contrast, fields such as the natural sciences and engineering, which tend to rely on more empirically stable knowledge, are associated with fewer citation-related disputes.

Although the focus of our study differs, this disciplinary pattern aligns with prior research by Yasseri et al. (Yasseri et al., 2013), who found that the most controversial articles on English Wikipedia often concern topics such as individuals, politics, religion, nations, and global warming. A similar pattern was observed in their cross-linguistic study of the 1,000 most controversial Wikipedia articles across ten language editions (Sumi et al., 2011). While these articles are not necessarily based on scientific publications, they reveal a thematic overlap with some fields in our dispute dataset.

To explore this further, we compared our list of scientific disputes with the M-score of article-level controversy as defined by Yasseri et al. (Sumi et al., 2011). The results indicate little overlap between highly controversial Wikipedia articles and those containing numerous scientific disputes. For example, the English Wikipedia article for "Jesus," which ranks among the top 10 most controversial with an M-score of 8,421,728, contains only five scientific disputes. This divergence suggests that public attention to controversial topics does not necessarily translate into contention over scientific sources. Scientific disputes differ in that they require a different type of engagement—one rooted in interpretive legitimacy and epistemic

trust.

Furthermore, the resolution time for scientific disputes is relatively short. Across nearly all Wikipedia concepts in our dataset (with the exception of materials science), the median time to resolution is less than one day. This may reflect the perception of scientific publications as authoritative sources, which facilitates faster consensus-building. Although prior work has not provided exact durations, agent-based modeling studies suggest that reaching consensus typically requires more than 20 rounds of edits (Kalyanasundaram et al., 2015). Additionally, the presence of heterogeneous editor types (e.g., registered and anonymous users) has been associated with longer conflict duration (Kalyanasundaram et al., 2015). Our findings support this observation: disputes involving both registered and anonymous editors generally last longer than those involving a single user type. One possible explanation is that anonymous contributors are often viewed as less credible, and previous studies have shown that edits by anonymous users are more likely to be reverted and contain disruptive content (Tran et al., 2020).

We also found that disputed articles tend to have higher citation counts and are frequently published in prestigious journals such as *Nature*, *Science*, and *PNAS*. These journals are also among the most cited sources on Wikipedia overall (Arroyo-Machado et al., 2020; Yang and Colavizza, 2022a). This aligns with previous findings that scientific publications from high-status journals tend to be more visible to both editors and readers (Teplitskiy et al., 2017). Their broad reach and authoritative status increase the likelihood of being cited but also raise the potential for contention when used as evidence in public knowledge spaces like Wikipedia.

Regarding **RQ2**: *Are OA publications more likely to be used in scientific disputes compared to paywalled publications?* Our analysis shows that OA publications are significantly more likely to be involved in disputes, and that disputes concerning OA articles tend to emerge sooner after publication. This pattern can be explained by the increased accessibility of OA articles (Yang et al., 2024), which facilitates rapid dissemination and critical evaluation by a broader and more diverse editor base. These results extend existing literature on OA's role in enhancing visibility, readership, and scholarly engagement (Yang et al., 2024). Furthermore, OA articles may disproportionately address emerging, interdisciplinary, or controversial topics, which naturally attract greater contestation and discussion. This is consistent with prior findings that disputes around contested content on Wikipedia often mirror broader societal controversies (Borra et al., 2015).

The steady increase in the proportion of disputed OA articles, outpacing the general growth of OA publishing, suggests that openness not only increases accessibility but also accelerates the dynamics of knowledge contestation. These patterns support the idea that OA fosters a more participatory and transparent knowledge ecosystem, where scientific claims are more rapidly scrutinized and negotiated. However, the relationship between OA and disputes varies across

disciplines. For example, in Political Science, disputed articles more frequently involve paywalled publications, possibly reflecting disciplinary norms regarding information gatekeeping or the prominence of certain traditional sources (Severin et al., 2020). In biological and medical sciences, the distribution of OA and paywalled articles in disputes is more balanced, indicating different editorial cultures and accessibility patterns. These disciplinary differences suggest that openness interacts with field-specific epistemic cultures and norms, reinforcing the idea that the impact of OA on knowledge contestation is context-dependent.

Although previous research suggests that non-experts may endorse scientific norms more strongly in highly contested fields (Schug et al., 2025), potentially facilitating consensus-building, our findings do not support the notion that open access significantly shortens the duration of disputes. This suggests that while openness may enhance transparency and participation, dispute resolution is likely shaped more by the nature of contested content and the dynamics of editorial deliberation than by access alone.

Taken together, these findings point to several implications. The higher involvement of OA publications in disputes indicates that OA plays a critical role in enabling public scrutiny and facilitating dynamic knowledge negotiation in collaborative platforms. This reinforces the broader vision of OA as not only a means for democratizing access to scientific information but also a catalyst for enhanced critical engagement and contestation. Concurrently, the prevalence of disputes in social sciences and humanities underscores the ongoing challenges of reconciling diverse perspectives in fields where social, political, and ideological values are deeply embedded.

Despite these contributions, this study has limitations. First, we relied on the Crossref Event Data, which, while valuable for efficiently capturing citation events on Wikipedia, has limitations in temporal coverage and completeness due to discrepancies with Wikipedia's own data records. Although strict filtering ensured data quality, some relevant dispute events may have been missed, potentially biasing results toward more recent or well-documented cases. Second, our operationalization of scientific disputes, defined by a seven-day time window, while justified through sensitivity analyses, may benefit from further refinement and validation in future work. Third, as our analysis focused solely on Wikipedia's main article pages and did not account for activity on associated talk pages (Arroyo-Machado and Torres-Salinas, 2024), we may have underestimated the duration or complexity of some disputes. Future research should incorporate talk page discussions to more comprehensively understand how disputes involving scientific articles emerge, evolve, and are eventually resolved. Finally, since disputes may manifest differently across language editions of Wikipedia (Yasseri et al., 2013), and our study only examined the English version, further work is needed to explore cross-lingual variations in the dynamics and representation of scientific disputes.

In summary, this chapter answers the research questions by demonstrating that scientific disputes on Wikipedia are shaped by the interplay of disciplinary

context, publication impact, editor identity, and openness. OA publications emerge as prominent actors in these disputes, facilitating earlier and more frequent contestation. These insights contribute to a deeper understanding of how scientific knowledge is constructed and contested in an open, collaborative digital environment and highlight the complex role of openness in shaping knowledge dynamics.

8.6 Conclusion

This chapter investigated the characteristics of scientific publications involved in disputes on Wikipedia and examined the role of OA publications in these conflicts. Our findings reveal that scientific disputes are more prevalent in social sciences and humanities, where epistemic uncertainty and value-laden topics are common, compared to the natural sciences and engineering. Publications that attract greater attention, such as those with high citation counts and those published in prestigious journals, are more likely to be contested. Registered editors play a central role in sustaining these disputes, indicating the importance of experienced contributors in the editorial process.

Furthermore, OA publications are significantly more likely to be used in scientific disputes and tend to become subjects of dispute earlier than paywalled publications. This suggests that increased accessibility and visibility foster more rapid and frequent contestation of scientific knowledge. Nevertheless, the relationship between OA and disputes varies across disciplines, highlighting the influence of field-specific norms on the dynamics of knowledge negotiation.

These findings highlight a dual role of OA: while it democratizes access to scientific knowledge, it also intensifies public scrutiny and epistemic contestation. Wikipedia serves as a key arena where scientific authority is continuously negotiated, not only through content but through the strategic deployment of citations. Our results offer implications for open science, scholarly communication, and digital platform governance, suggesting that the architecture of openness affects not only access, but also the trajectories of scientific controversy in public discourse.

By situating OA within the dynamics of conflict, this chapter complements earlier findings on coverage, bias, and access. Taken together, the four chapters show that Wikipedia is simultaneously a site of dissemination, bias, accessibility, and contestation. The final chapter of the thesis will integrate these insights and reflect on what they reveal about Wikipedia's role as a socio-technical infrastructure for the public communication of science.

Chapter 9

Conclusion

Scientific knowledge plays a crucial role in shaping public understanding, policy, and civic life. In the digital era, platforms like Wikipedia have become key infrastructures for mediating access to science, integrating diverse sources, and enabling public participation in the making of knowledge. In this thesis, I have examined how scientific knowledge is disseminated, structured, and contested on Wikipedia, with special attention to citation practices, open access publishing, and editorial dynamics. This investigation has been carried out in two main parts. In Part One, I analyzed how Wikipedia structures scientific knowledge and reflects the broader information ecosystem through its citation behavior. In Part Two, I explored how open access (OA) affects the visibility and contestation of scientific knowledge on the platform. I summarize the findings of these parts in the next section.

9.1 Findings

Part One This part focused on how Wikipedia integrates and filters scientific information. In Chapter 4, I examined large-scale citation patterns from Wikipedia to scientific journals and found that the platform disproportionately draws from STEM fields, particularly biology and medicine, followed by earth sciences, physics, and astronomy. A sizeable fraction of this literature is well cited and published in prestigious venues such as Nature or Science. Biographical articles emerged as important structural bridges, linking scientific content to broader social and cultural domains. The seemingly marginal role of non-STEM journal articles is thus attenuated by their connective function: biographies in particular serve to bridge history, geography, and culture with STEM topics.

Network analysis further showed that Wikipedia's bibliographic coupling network is not only smaller but also more tightly connected than the co-citation network of the underlying academic literature. This reflects both the consolidating encyclopedic role of Wikipedia and the coordinating influence of WikiProjects.

More broadly, the results indicate that Wikipedia's knowledge structure is not a simple mirror of the academic citation landscape. Instead, editorial practices actively reshape the organization of knowledge by consolidating disparate fields into a more integrated whole, thereby serving the platform's encyclopedic mission.

In Chapter 5, I examined the epistemic and political dimensions of citation by analyzing the selection of news sources. The results show that most citations are drawn from reliable outlets, though a non-trivial fraction still comes from sources with low or mixed reliability. With respect to political orientation, I found a moderate yet systematic liberal polarization: the average polarization score of cited news outlets is -0.5, with most sources clustered between -1 and 0 on a scale from -2 (very liberal) to 2 (very conservative). Notably, there is no clear relationship between the reliability of a news outlet and its political leanings.

These findings illustrate how Wikipedia's neutrality principle is operationalized in practice. Although the platform enforces strict sourcing guidelines, editorial choices are shaped by the availability, prominence, and perceived legitimacy of external media. In this way, Wikipedia simultaneously reflects broader patterns of media polarization and refracts them through its own governance mechanisms.

Taken together, the findings of Part One underscore that Wikipedia does not merely aggregate external references but actively reconfigures them. Scientific journal articles and news media are integrated in ways that both reinforce existing hierarchies, through reliance on prestigious STEM journals, and open new pathways for interdisciplinarity and public engagement, through biographical and cross-domain linkages. Similarly, while news media introduce potential biases, Wikipedia's editorial norms mediate their influence, revealing the platform's role as both a gatekeeper and a negotiator of scientific credibility in the public sphere.

Part Two This part explored the role of open access in mediating science on Wikipedia.

In Chapter 7, I first examined the distribution of OA articles cited on Wikipedia across disciplines. Biology, physics, and mathematics exhibit higher citation rates of OA publications, whereas social sciences and humanities show comparatively lower OA uptake. Over time, the proportion of OA articles cited from Wikipedia has steadily increased, reflecting broader trends in scholarly publishing and suggesting that editors actively leverage freely accessible sources to enhance verifiability and reach. Controlling for citation count, publication age, and disciplinary field, OA articles are significantly more likely to be cited than paywalled articles. This "open access advantage" is particularly pronounced for recent and highly cited publications, highlighting OA's role in amplifying scientific visibility on this public platform.

In Chapter 8, I examined how scientific sources are involved in editorial disputes, with a particular focus on OA publications. Scientific disputes occur most frequently in fields with higher epistemic uncertainty and interpretive flexibility, such as social sciences and humanities, while natural sciences and engineering experience fewer disputes. Publications that attract greater attention—such as

those with high citation counts or published in prestigious journals—are more likely to be contested. OA articles are disproportionately involved in disputes and tend to become subjects of contention sooner after publication, suggesting that accessibility accelerates both engagement and scrutiny. This pattern is particularly evident in socially sensitive domains, indicating that OA not only facilitates broader readership but also exposes scientific claims to rapid evaluation and negotiation by a diverse editor base.

Taken together, these findings indicate a complex dual role of OA on Wikipedia. On one hand, OA democratizes access to scientific knowledge, enabling editors and readers to interact with research that might otherwise be behind paywalls. On the other hand, this increased accessibility amplifies exposure to scrutiny and contestation, revealing a dynamic where visibility, authority, and dispute intersect. The influence of OA varies across disciplines, reflecting how field-specific norms, editorial cultures, and the inherent interpretive flexibility of certain topics shape the likelihood and timing of disputes. Additionally, these results highlight that Wikipedia is not simply a passive repository; editorial practices, access models, and community governance collectively mediate how scientific knowledge is presented, contested, and legitimized in the public sphere.

Overall, Part Two complements the findings from Part One by showing that not only the selection of sources but also their accessibility critically shapes the public visibility and contestation of science. OA functions as both an enabler of knowledge dissemination and a catalyst for public epistemic engagement, reinforcing Wikipedia's role as a participatory platform where authority and credibility are continually negotiated. These insights contribute to a deeper understanding of how digital platforms and open science intersect to influence the dynamics of knowledge production, dissemination, and contestation.

9.2 Limitations and Future Work

Part One One limitation of this part lies in the reliance on citation data and metadata, which do not fully capture the interpretive, discursive, and contextual dimensions of knowledge integration on Wikipedia. While citation patterns provide a powerful tool for mapping the flow of scientific influence, they cannot reveal how sources are actually employed in the construction of narrative, argumentation, or explanation within articles. Our analysis focused exclusively on journal articles, excluding other types of sources such as books, reports, conference papers, and web-based content, all of which contribute substantially to Wikipedia's knowledge ecosystem. Including these sources in future studies would allow for a more comprehensive mapping of knowledge integration across both scientific and non-scientific domains. Additionally, future research could combine citation analysis with close reading, natural language processing, or topic modeling to examine the rhetorical and narrative roles of citations, including how editors frame, interpret,

and integrate evidence in shaping content.

In Chapter 5, our analysis of news media citations also has several limitations. We relied on external sources to measure political polarization and factual reliability. Although these sources are considered authoritative, their coverage is partial and the scoring methodologies they use are constrained. Future research could adopt multiple, complementary approaches to assess media bias and reliability, including cross-validation with alternative datasets or more dynamic, longitudinal measures. Our study focused on a binary distinction between liberal and conservative media, but political orientations exist along multiple dimensions, including regional, economic, and cultural axes. Future work could examine these additional dimensions and explore how they interact with reliability and editorial selection. Furthermore, our analysis was conducted at the domain level, while a more granular investigation at the level of individual articles or specific news items could clarify how political orientation and reliability influence citation practice. We also did not examine how news sources are used in context within Wikipedia articles, such as whether they are cited to support factual claims, provide background, or illustrate controversy. Expanding the study to include the functional role of sources would deepen understanding of Wikipedia's epistemic practices. Finally, our work focused solely on English Wikipedia; examining multiple language editions could reveal whether different editorial cultures or local information environments produce different patterns of source selection and polarization.

Another important limitation concerns temporal dynamics. Our analysis is based on a snapshot of citations at a single point in time. Wikipedia is a constantly evolving platform, with articles and citations being continually updated, added, or removed as new knowledge emerges or consensus shifts. Longitudinal studies that track changes in citation patterns over time would provide insights into the processes of knowledge consolidation, revision, and obsolescence of references. In addition, we did not integrate structural information such as internal article links, WikiProject participation, or textual similarity. Combining citation networks with these structural features could illuminate how editorial coordination, collaborative efforts, and networked connections shape the organization and integration of knowledge across different scientific and non-scientific domains. Future research could also examine the interplay between article edit histories, contributor experience, and citation patterns to better understand the social and procedural mechanisms through which knowledge is curated, contested, and maintained on Wikipedia.

Part Two In Chapter 7, several limitations of this study suggest directions for future research. The focus on articles with DOIs means that other scholarly outputs, including conference papers, earlier literature, and non-journal publications, were excluded. Incorporating these additional sources could provide a more comprehensive picture of the references that shape Wikipedia content. Although regression models accounted for OA status, OA policy, and citation counts, other

factors such as article length, editorial prominence, or topic relevance may also influence the likelihood of citation, and their inclusion could help clarify the mechanisms driving reference patterns. Temporal dynamics were also not explicitly considered in the current analysis. Wikipedia articles evolve continuously, and the timing of edits, citations, and disputes may influence the visibility, integration, and contestation of scientific sources. Examining the revision history could capture these changes, revealing how OA and paywalled publications are incorporated, updated, and challenged over time. This approach would also provide insight into the speed at which OA articles are cited after publication and how their prominence shifts as new evidence emerges.

In Chapter 8, limitations also arise from the reliance on Crossref Event Data to identify citation-related disputes. While this source efficiently captures many events, it may be incomplete, particularly for earlier disputes or less prominent articles, potentially biasing the results toward more recent or highly documented cases. Defining disputes using a seven-day observation window captures short-term conflicts but may miss longer or more complex contestations. Refining the operationalization of disputes and exploring alternative measures could strengthen the analysis of conflict dynamics. The study focused solely on edits to main article pages, excluding activity on associated talk pages where extended deliberation and negotiation often occur. Integrating talk page discussions in future research would provide a more complete understanding of how disputes arise, unfold, and are resolved. Finally, the restriction to English Wikipedia limits the generalizability of the findings. Citation practices, OA adoption, and dispute dynamics may differ across language editions due to regional access to literature, local editorial cultures, and distinct norms regarding verification. Comparative research across multiple languages could illuminate these variations and provide a broader view of Wikipedia as a global knowledge infrastructure.

Overall, these limitations indicate that future studies should adopt a broader, temporally informed, and multilingual approach, integrating diverse source types, structural features, and editorial contexts to more fully understand how OA shapes the accessibility, contestation, and consolidation of scientific knowledge on Wikipedia.

9.3 Final Remarks

In this thesis, I have explored how Wikipedia functions as a socio-technical infrastructure for the public communication of science. The platform's open and collaborative nature allows a wide range of actors, from professional scientists to volunteer editors, to participate in knowledge production. At the same time, this openness raises challenges regarding authority, credibility, transparency, and neutrality, highlighting the complex dynamics inherent in public knowledge-making. The aim of this work has been to investigate how scientific knowledge is dissemi-

nated, structured, and contested on Wikipedia, with a particular focus on citation practices, open access publishing, and editorial dynamics.

The findings reveal that Wikipedia is not merely a repository of references but an active arena where knowledge is filtered, legitimized, and negotiated. Part One showed that citation patterns favor STEM fields, particularly biology, medicine, and earth sciences, while biographical articles and cross-domain linkages bridge these with non-STEM topics, enabling interdisciplinarity. Editorial norms mediate the use of news media sources, balancing reliability with political bias. Part Two demonstrated that open access publications enjoy a citation advantage, are integrated earlier, and are more frequently involved in editorial disputes. This dual effect of openness underscores Wikipedia's role in both amplifying scientific visibility and facilitating contestation, especially in socially sensitive domains. Together, these results illustrate how citation practices, access models, and community governance shape which scientific knowledge becomes visible and how it is interpreted in the public sphere.

This thesis also offers avenues for future research. Longitudinal studies could track how citation patterns and editorial practices evolve, revealing the temporal processes of knowledge integration and contestation. Comparative analyses across different language editions of Wikipedia would illuminate the influence of regional editorial cultures and access to resources on scientific knowledge dissemination. Expanding the scope of sources beyond journal articles to include books, preprints, institutional repositories, and web content could provide a more comprehensive understanding of Wikipedia's knowledge network. Finally, integrating textual analysis, network structures, and talk page discussions could deepen insights into how editorial decisions, argumentation, and dispute resolution shape the representation and legitimacy of scientific knowledge.

Overall, this work emphasizes that Wikipedia is a key site where science meets society. By bridging empirical analysis with questions of epistemic authority, openness, and editorial governance, the thesis highlights both the opportunities and challenges of public knowledge infrastructures. The findings aim to inform future studies on digital platforms, open science, and the circulation of scientific knowledge in the twenty-first century.

Appendix A

Appendix to Chapter 4

A.1 Top-10 Most Cited Journal Articles

1. Laemmli U. K. (1970) "Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4". *Nature*. **Open access: closed**. **Research organization country: United Kingdom**. **Number of citations: 214,886**. **Number of recent citations: 6,111**.

2. Bradford M. M. (1976) "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding". *Analytical Biochemistry*. **Open access: closed**. **Research organization country: United States**. **Number of citations: 193,330**. **Number of recent citations: 16,578**.

3. Perdew J. P., Burke K., Ernzerhof M. (1996) "Generalized Gradient Approximation Made Simple". *Physical Review Letters*. **Open access: closed**. **Research organization country: United States**. **Number of citations: 99,164**. **Number of recent citations: 27,949**.

4. G.M. Sheldrick. (2007) "A short history of SHELX". *Acta Crystallographica Section A: Foundations and advances*. **Open access: open**. **Research organization country: Germany**. **Number of citations: 72,560**. **Number of recent citations: 7,355**.

5. Axel D. Becke. (1993) "Density-functional thermochemistry. III. The role of exact exchange". *The Journal of Chemical Physics*. **Open access: closed**. **Research organization country: Canada**. **Number of citations: 69,187**. **Number of recent citations: 8,871**.

6. Chengteh Lee, Weitao Yang, Robert G. Parr. (1988) "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron

density". *Physical Review B*. **Open access: closed**. **Research organization country: United States**. **Number of citations: 66,421**. **Number of recent citations: 8,872**.

7. Marshal F. Folstein, Susan E. Folstein, Paul R. McHugh. (1975) ""Mini-mental state" A practical method for grading the cognitive state of patients for the clinician". *Journal of Psychiatric Research*. **Open access: closed**. **Research organization country: United States**. **Number of citations: 64,625**. **Number of recent citations: 8,428**.

8. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman. (1990) "Basic local alignment search tool". *Journal of Molecular Biology*. **Open access: closed**. **Research organization country: United States**. **Number of citations: 63,340**. **Number of recent citations: 10,717**.

9. F. Sanger, S. Nicklen, A. R. Coulson. (1977) "DNA sequencing with chain-terminating inhibitors". *PNAS*. **Open access: open**. **Research organization country: United Kingdom**. **Number of citations: 58,637**. **Number of recent citations: 1,046**.

10. Piotr Chomczynski, Nicoletta Sacchi. (1987) "Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction". *Analytical Biochemistry*. **Open access: closed**. **Research organization country: United States**. **Number of citations: 56,286**. **Number of recent citations: 993**.

A.2 Figures

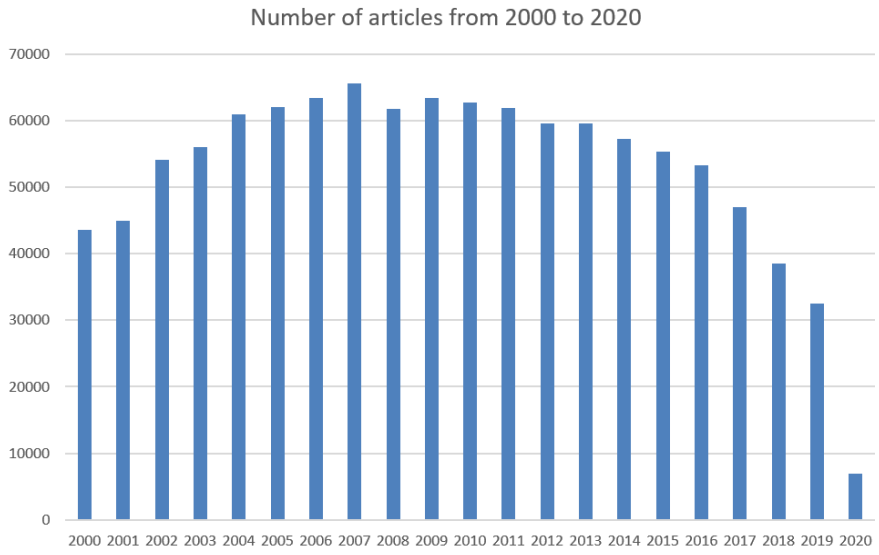


Figure A.1: Publication year of journal articles cited from Wikipedia, from 2000 to 2020.

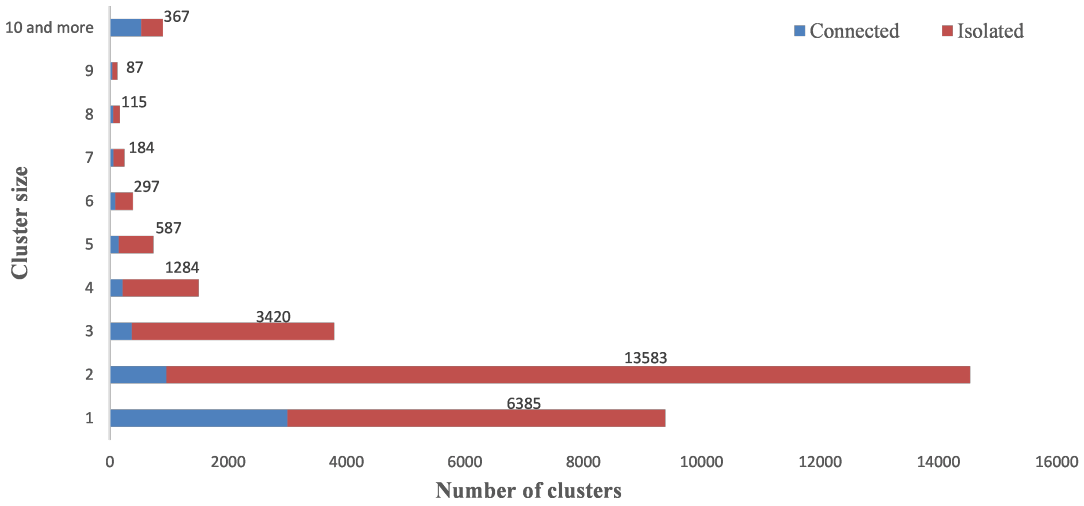


Figure A.2: Distribution of clusters by size in the bibliographic coupling network of Wikipedia articles.

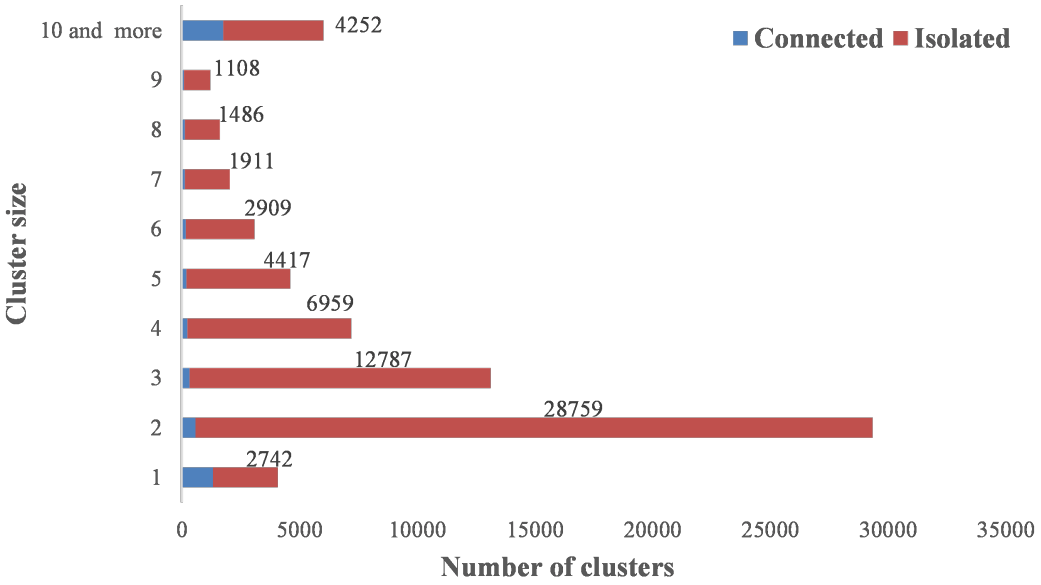


Figure A.3: Distribution of clusters by size in the co-citation network of journal articles.

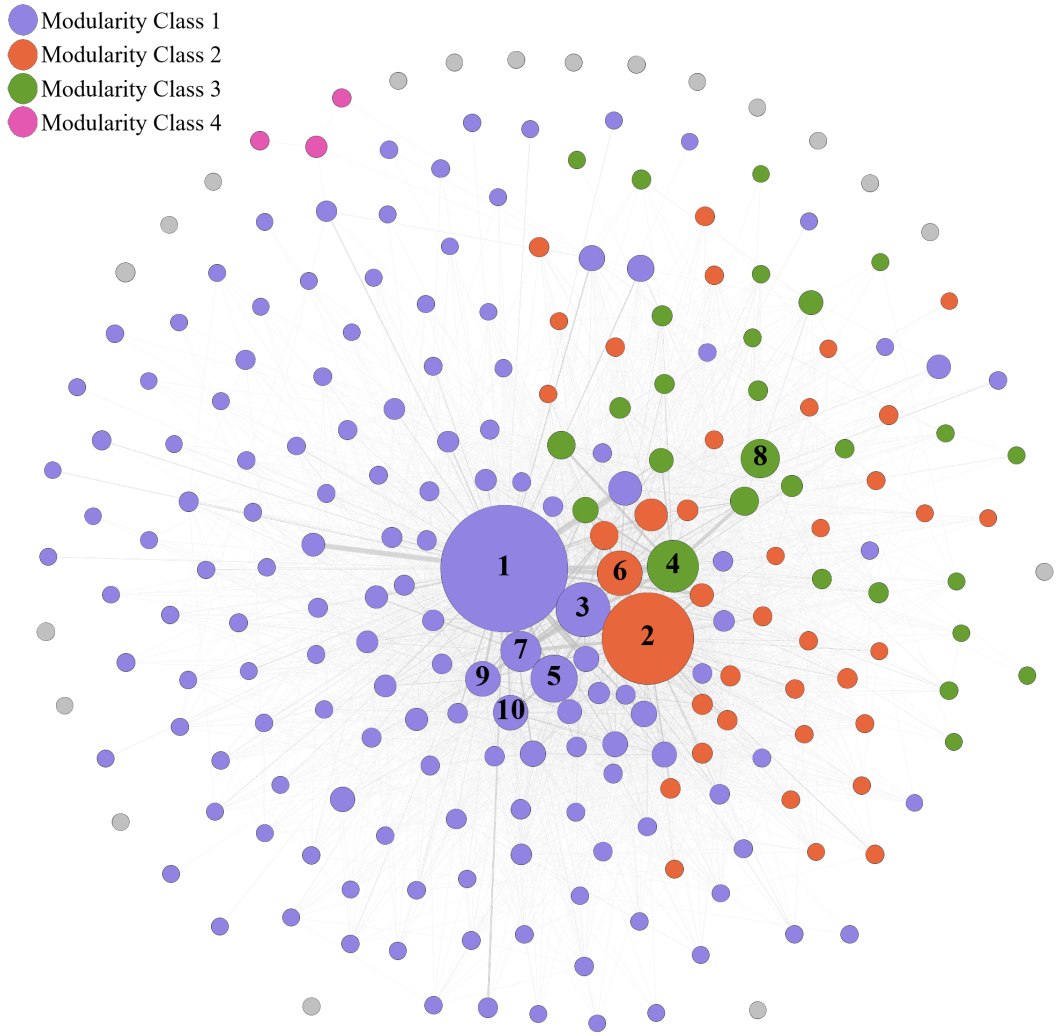


Figure A.4: Bibliographic coupling supernetwork coloured by modularity class.

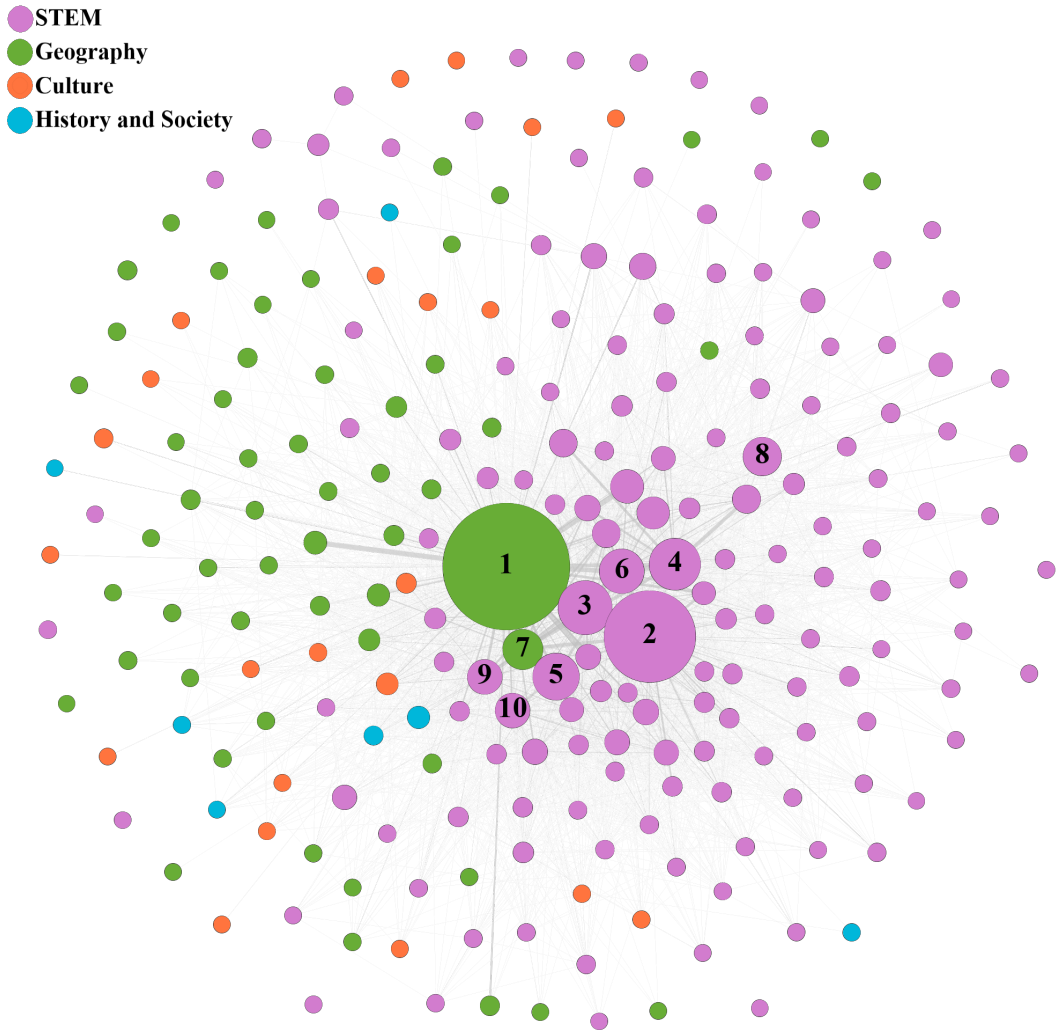


Figure A.5: Bibliographic coupling supernetwork coloured by top ORES topic within a node/cluster. Compare with Figure 4.5.

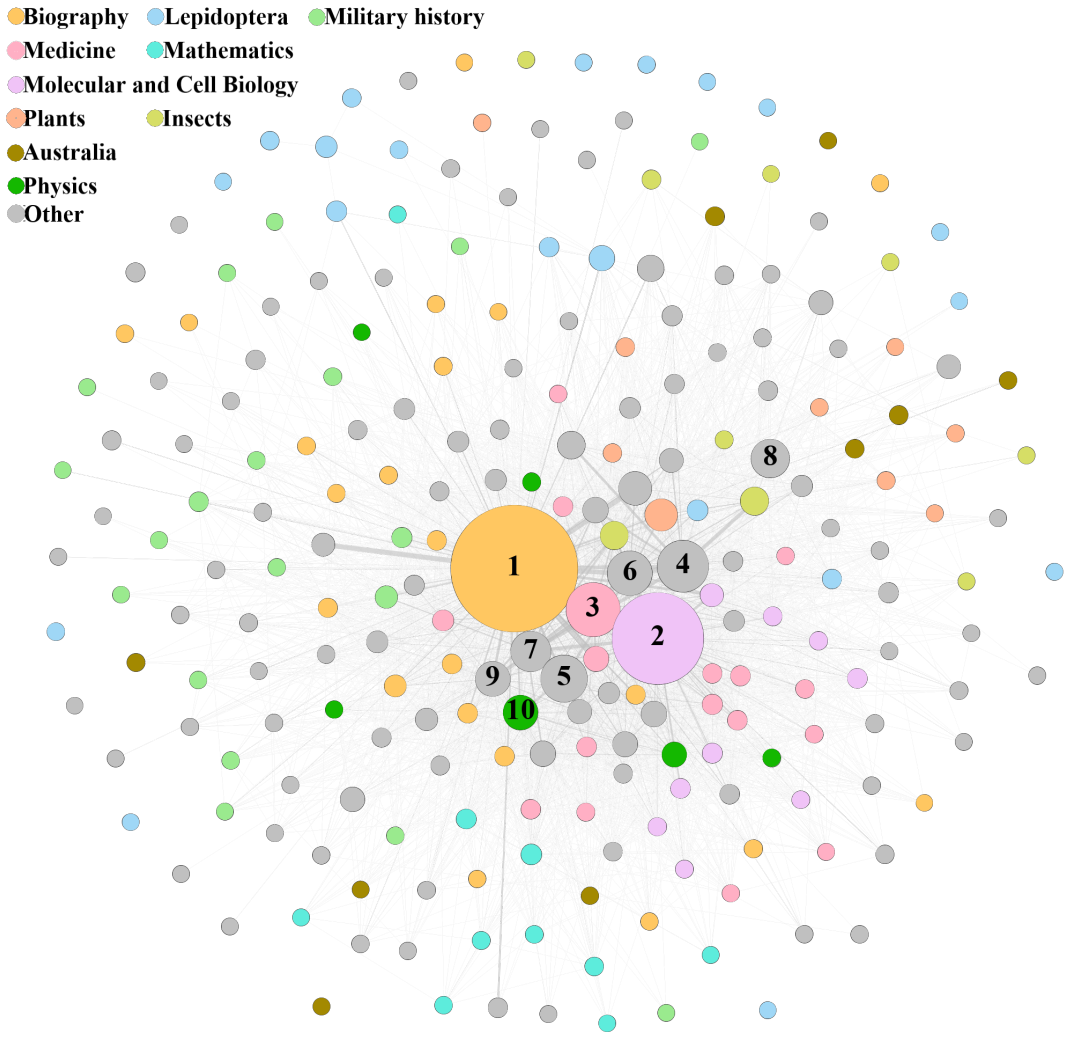


Figure A.6: Bibliographic coupling supernetwork coloured by top WikiProject within a node/cluster. Compare with figure 4.6.

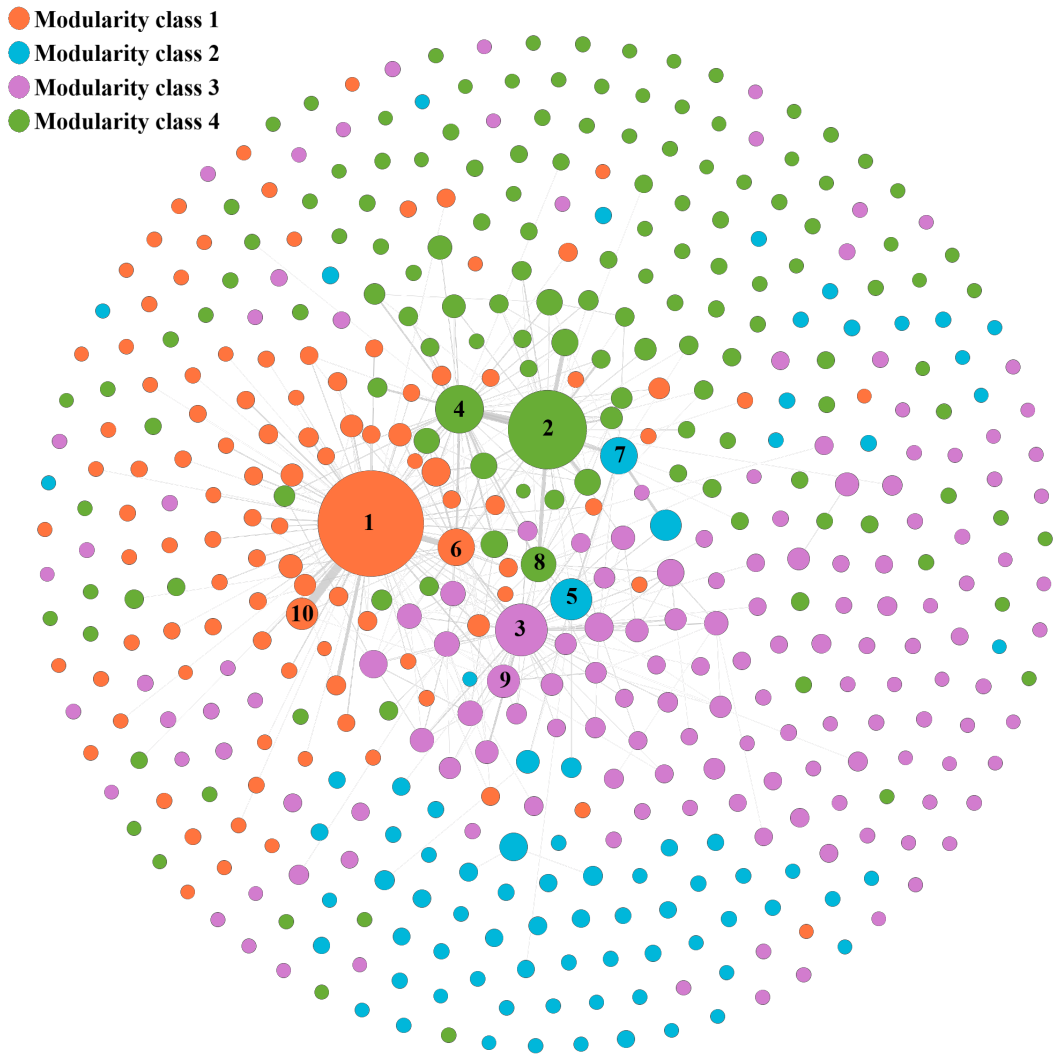


Figure A.7: Co-citation supernetwork coloured by modularity class.

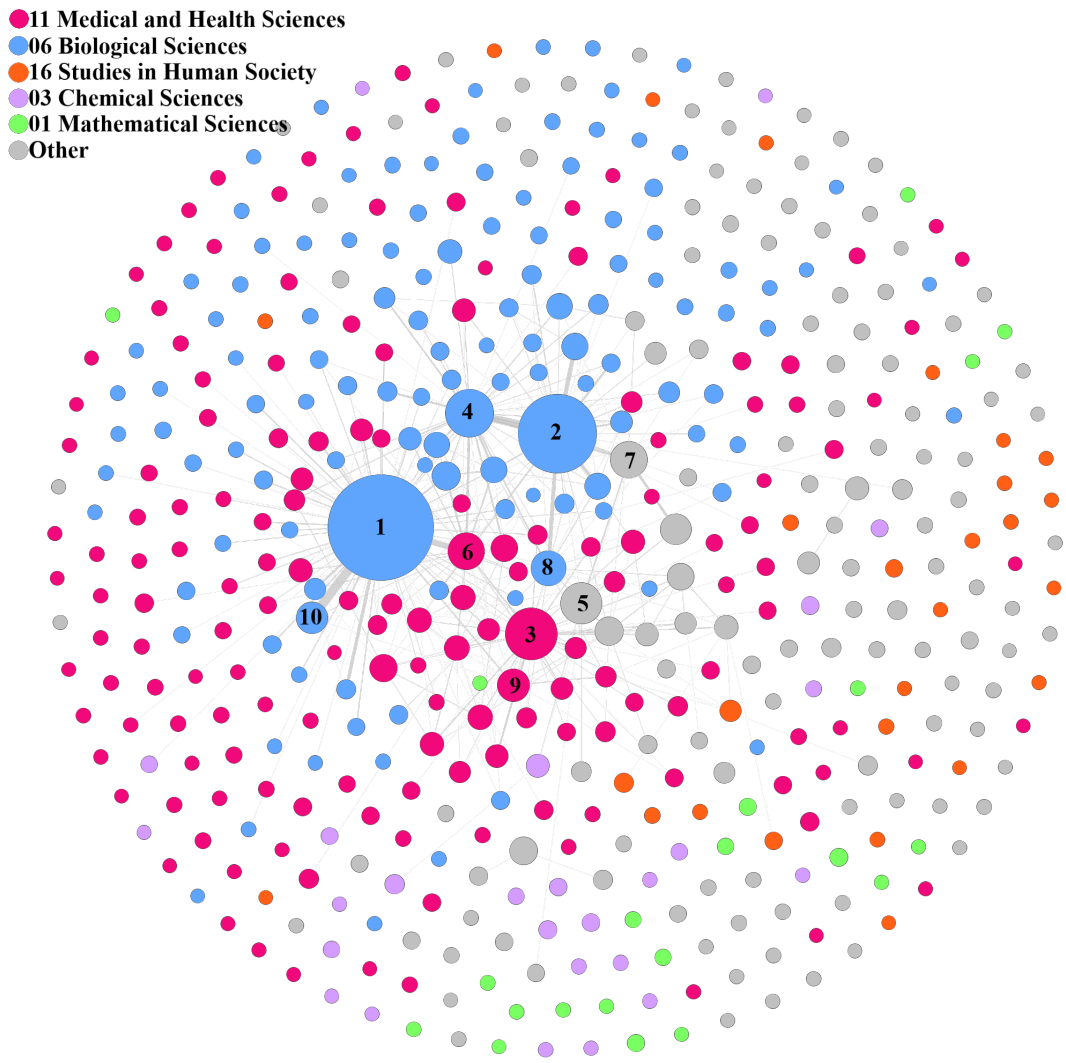


Figure A.8: Co-citation supernetwork coloured by top major Field of Research. Compare with Figure 4.7.

Appendix B

Appendix to Chapter 7

B.1 Figures

Presented below are two figures depicting the distribution of OA status and policies among the top 20 journals. We have discussed it in the results part. This observation underscores the significance of conducting an article-level analysis for a more comprehensive understanding of the subject matter.

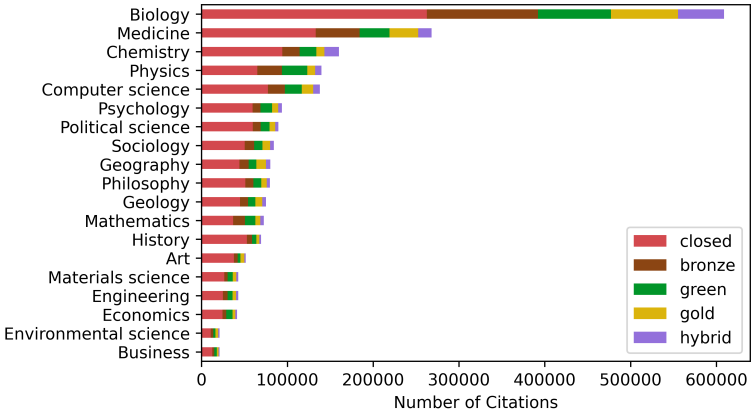


Figure B.1: Distribution of OA policies by OpenAlex concept.

In Figure B.1, we illustrate the distribution of OA policies across various concepts. Our analysis reveals that bronze and green policies predominantly characterize most concepts in OA articles, except for Art, where the gold policy assumes significance.

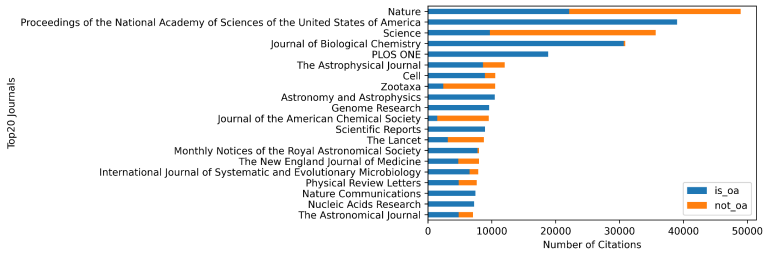


Figure B.2: Distribution of OA status by top 20 journals.

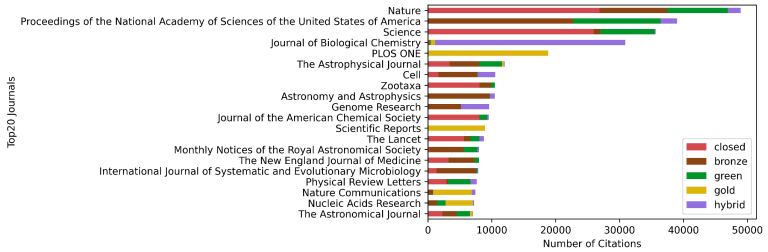


Figure B.3: Distribution of OA policies by top 20 journals.

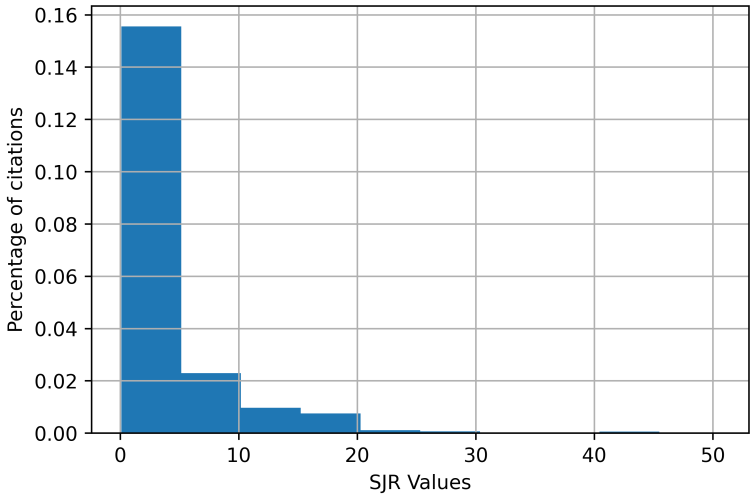


Figure B.4: Distribution of SJR among Wikipedia citations.

B.2 Tables

The quality of our stratified samples is demonstrated through the descriptive statistics provided in Table B.1, B.2. Additionally, Table B.3 presents a count of articles by concepts within our dataset. The regression results for formulas 1 and 2 for the entire sample are displayed in Table B.4, B.5, B.6 and B.7.

Table B.1: Descriptive statistics for the articles cited in Wikipedia.

	cited_by_count	num_references	article_age	H index	is_oa	ln(SJR)
count	261,230	261,230	261,230	261,230	261,230	261,230
mean	146.133	37.797	252.171	249.348	0.504	0.632
std	834.450	49.357	201.939	259.576	0.500	1.166
min	0	0	12	0	0	-2.303
25%	14	10	122	87	0	-0.140
50%	45	28	208	164	1	0.554
75%	124	48	310	305	1	1.460
max	304,415	1,976	1,487	1,331	1	3.922

Table B.2: Descriptive statistics for the articles not cited in Wikipedia. Average over all samples.

	cited_by_count	num_references	article_age	H index	is_oa	ln(SJR)
count	261,230	261,230	261,230	261,230	261,230	261,230
mean	59.008	28.316	252.137	249.348	0.497	0.632
std	228.084	38.377	202.060	259.576	0.500	1.166
min	0	0	12	0	0	-2.303
25%	3	2	122	87	0	-0.140
50%	18	21	208	164	0	0.554
75%	54	41	310.2	305	1	1.460
max	44,132.4	2,891.2	1,487	1,331	1	3.922

B.3 Supplementary Regression Results

This section provides an in-depth analysis of OA citation advantage for each OpenAlex concept. To achieve this, we developed 19 distinct regression models, each dedicated to analyzing the adoption of OA citation for a single concept. We use the second formulation for each model, with data pertaining solely to the corresponding concept being considered in each case.

To gain insight into the effect of OA adoption on each concept, we present the coefficients for the *is_oa* variable in Table B.8 and the coefficients for the $\ln(1 + times\_cited)$ variable in Table B.9.

Table B.8 indicates that OA articles across most concepts exhibit a positive OA Wikipedia citation advantage, with five concepts showing statistically significant advantages. The top five concepts with the highest OA adoption advantage are Chemistry, Economics, Psychology, Business, and Physics, suggesting that STEM-related subjects attract more attention on Wikipedia.

Regarding $\ln(1 + times\_cited)$ in each concept, citation counts demonstrate a significantly positive effect in nearly all concepts, although Environment Science and Engineering do not show significance. OA articles in several OpenAlex concepts, including Biology, Computer Science, Chemistry, Medicine, Psychology,

Table B.3: Count of articles by concepts in the final combined dataset.

Num.	Concept	Counts	Num.	Concept	Counts
1	Biology	144,307	11	History	1,618
2	Medicine	48,286	12	Art	1,589
3	Chemistry	18,135	13	Materials science	1,256
4	Physics	10,835	14	Economics	1,005
5	Psychology	8,491	15	Geography	911
6	Geology	8,429	16	Business	480
7	Mathematics	5,620	17	Sociology	466
8	Computer science	5,274	18	Engineering	241
9	Philosophy	2,145	19	Environmental science	211
10	Political science	1,931			

Table B.4: Coefficients for overall OA adoption. Average results across all 5 samples, model 1, $R^2 = 0.00032$.

index	coef	odds_ratios	P>z
ln_article_age	0.064	1.066	0.000
ln(SJR)	-0.002	0.998	0.482
is_oa	0.583	1.791	0.000
is_oa:ln_article_age	-0.104	0.901	0.000

Table B.5: Coefficients for overall OA adoption. Average results across all 5 samples, model 2, $R^2 = 0.072$.

index	coef	odds_ratios	P>z
ln1p_times_cited	0.442	1.556	0.000
ln_article_age	-0.068	0.934	0.000
ln(SJR)	-0.245	0.782	0.000
is_oa	0.494	1.639	0.000
is_oa:ln_article_age	-0.103	0.902	0.000

Mathematics, Economics, Geology, Materials Science, and Physics, exhibit, on average, over a 30% higher likelihood of being cited in Wikipedia compared to paywalled articles. Citation counts remain important factors in these concepts.

Table B.6: Coefficients for OA adoption by the policy. Average results across all 5 samples, model 1, $R^2 = 0.0013$.

	coef	odds_ratios	P>z
Bronze	0.194	1.215	0.002
Gold	0.672	1.959	0.000
Green	1.147	3.149	0.000
Hybrid	1.092	2.982	0.000
ln_article_age	0.064	1.066	0.000
Bronze:ln_article_age	-0.031	0.970	0.010
Gold:ln_article_age	-0.138	0.871	0.000
Green:ln_article_age	-0.189	0.828	0.000
Hybrid:ln_article_age	-0.232	0.793	0.000
ln(SJR)	-0.001	0.999	0.654

Table B.7: Coefficients for OA adoption by the policy. Average results across all 5 samples, model 2, $R^2 = 0.073$.

	coef	odds_ratios	P>z
Bronze	-0.158	0.855	0.015
Gold	1.071	2.920	0.000
Green	0.834	2.303	0.000
Hybrid	1.212	3.361	0.000
ln1p_times_cited	0.447	1.564	0.000
ln_article_age	-0.071	0.932	0.000
Bronze:ln_article_age	0.038	1.039	0.002
Gold:ln_article_age	-0.241	0.786	0.000
Green:ln_article_age	-0.173	0.841	0.000
Hybrid:ln_article_age	-0.284	0.753	0.000
ln(SJR)	-0.244	0.784	0.000

Table B.8: Coefficients for OA adoption by concept for all samples (*is_oa*).

concept	min OR	max OR	OR mean	Highest P-value	Mean R^2
Biology	1.589	1.684	1.621	0.000	0.067
Computer science	1.314	1.737	1.471	0.264	0.052
Chemistry	3.828	4.068	3.918	0.000	0.060
Medicine	2.177	2.390	2.280	0.000	0.134
Psychology	2.524	3.554	3.150	0.001	0.090
Mathematics	1.536	1.748	1.672	0.227	0.073
Economics	2.277	3.635	3.250	0.276	0.104
Geology	1.347	1.531	1.428	0.136	0.053
Sociology	0.937	4.174	2.095	0.967	0.014
History	1.407	2.168	1.768	0.483	0.018
Geography	1.490	2.689	2.066	0.505	0.009
Philosophy	0.400	0.649	0.487	0.301	0.005
Materials science	1.511	3.379	2.253	0.497	0.098
Art	0.783	1.175	0.961	0.902	0.008
Environmental science	0.132	0.551	0.418	0.647	0.008
Physics	2.163	2.559	2.318	0.000	0.064
Engineering	0.863	1.965	1.313	0.911	0.006
Business	2.241	3.556	3.101	0.380	0.032
Political science	0.602	0.805	0.732	0.617	0.017

Table B.9: Coefficients for OA adoption by concept for all samples (ln(1 + *times_cited*)).

concept	min OR	max OR	OR mean	Highest P-value	Mean R^2
Biology	1.594	1.602	1.599	0.000	0.067
Computer science	1.303	1.314	1.308	0.000	0.052
Chemistry	1.545	1.566	1.554	0.000	0.060
Medicine	1.820	1.835	1.824	0.000	0.134
Psychology	1.473	1.490	1.482	0.000	0.090
Mathematics	1.431	1.454	1.442	0.000	0.073
Economics	1.461	1.519	1.490	0.000	0.104
Geology	1.476	1.487	1.481	0.000	0.053
Sociology	1.113	1.211	1.153	0.003	0.014
History	1.252	1.297	1.274	0.000	0.018
Geography	1.124	1.149	1.137	0.000	0.009
Philosophy	1.109	1.126	1.116	0.000	0.005
Materials science	1.551	1.570	1.562	0.000	0.098
Art	1.191	1.267	1.213	0.000	0.008
Environmental science	1.011	1.149	1.069	0.824	0.008
Physics	1.409	1.418	1.413	0.000	0.064
Engineering	1.015	1.144	1.062	0.779	0.006
Business	1.205	1.240	1.221	0.000	0.032
Political science	1.203	1.240	1.227	0.000	0.017

Bibliography

Adams, C. E., Montgomery, A. A., Aburrow, T., Bloomfield, S., Briley, P. M., Carew, E., Chatterjee-Woolman, S., Feddah, G., Friedel, J., Gibbard, J., et al. (2020). Adding evidence of the effects of treatments into relevant Wikipedia pages: A randomised trial. *BMJ open*, 10(2):e033655.

Adler, B. T., De Alfaro, L., Mola-Velasco, S. M., Rosso, P., and West, A. G. (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 277–288. Springer.

Agarwal, P., Redi, M., Sastry, N., Wood, E., and Blick, A. (2020). Wikipedia and westminster: Quality and dynamics of Wikipedia pages about uk politicians. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 161–166, New York, NY, USA. Association for Computing Machinery.

Almeida, R. B., Mozafari, B., and Cho, J. (2007). On the evolution of Wikipedia. *International Conference on Weblogs and Social Media (ICWSM), Colorado, USA*.

Antelman, K. (2004). Do open-access articles have a greater research impact? *College & research libraries*, 65(5):372–382.

Aragón, P. and Saez-Trumper, D. (2021). A preliminary approach to knowledge integrity risk assessment in Wikipedia projects.

Arazy, O., Morgan, W., and Patterson, R. (2006). Wisdom of the crowds: Decentralized knowledge construction in Wikipedia. In *16th Annual Workshop on Information Technologies & Systems (WITS) Paper*.

Arazy, O., Nov, O., Patterson, R., and Yeo, L. (2011). Information quality in Wikipedia: The effects of group composition and task conflict. *Journal of management information systems*, 27(4):71–98.

Archambault, É., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L., and Roberge, G. (2014). Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013. *Copyright, Fair Use, Scholarly Communication, etc.*

Arroyo-Machado, W. and Torres-Salinas, D. (2024). Citation practices in Wikipedia talk pages: First insights from an unexplored discussion channel.

Arroyo-Machado, W., Torres-Salinas, D., Herrera-Viedma, E., and Romero-Frías, E. (2020). Science through Wikipedia: A novel representation of open knowledge through co-citation networks. *PloS one*, 15(2):e0228713.

Avieson, B. (2022). Editors, sources and the'go back'button: Wikipedia's framework for beating misinformation. *First Monday*.

Baigutanova, A., Myung, J., Saez-Trumper, D., Chou, A.-J., Redi, M., Jung, C., and Cha, M. (2023a). Longitudinal Assessment of Reference Quality on Wikipedia. In *Proceedings of the ACM Web Conference 2023*, pages 2831–2839. arXiv:2303.05227 [cs].

Baigutanova, A., Saez-Trumper, D., Redi, M., Cha, M., and Aragón, P. (2023b). A Comparative Study of Reference Reliability in Multiple Language Editions of Wikipedia. arXiv:2309.00196 [cs].

Benjakob, O. and Aviram, R. (2018). A clockwork Wikipedia: From a broad perspective to a case study. *Journal of Biological Rhythms*, 33(3):233–244.

Benjakob, O., Aviram, R., and Sobel, J. A. (2022). Citation needed? Wikipedia bibliometrics during the first wave of the COVID-19 pandemic. *GigaScience*, 11:giab095.

Björk, B.-C. and Solomon, D. (2012). Open access versus subscription journals: a comparison of scientific impact. *BMC medicine*, 10(1):73.

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., and Guðnason, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009. *PLOS ONE*, 5(6):e11273. Publisher: Public Library of Science.

Black, E. W. (2008). Wikipedia and academic peer review: Wikipedia as a recognised medium for scholarly publication? *Online Information Review*, 32(1):73–88.

Borra, E., Weltevrede, E., Ciuccarelli, P., Kaltenbrunner, A., Laniado, D., Magni, G., Mauri, M., Rogers, R., and Venturini, T. (2015). Societal controversies in Wikipedia articles. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 193–196.

Borra, E., Weltevrede, E., Ciuccarelli, P., Kaltenbrunner, A., Laniado, D., Magni, G., Mauri, M., Rogers, R., Venturini, T., et al. (2014). Contropedia-the analysis and visualization of controversies in Wikipedia articles. In *OpenSym*, pages 34–1.

Bozarth, L., Saraf, A., and Budak, C. (2020). Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 US presidential nominees. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 48–59.

Brandes, U. and Lerner, J. (2008). Visual analysis of controversy in user-generated encyclopedias. *Information visualization*, 7(1):34–48.

Bredahl, L. (2022). Chapter 1. Introduction to Bibliometrics and Current Data Sources. *Library Technology Reports*, 58(8):5–11. Number: 8.

Brezar, A. and Heilman, J. (2019). Readability of English Wikipedia's health information over time. *WikiJournal of Medicine*, 6(1):1–6.

Brossard, D. (2013). New media landscapes and the science information consumer. *Proceedings of the National Academy of Sciences*, 110(supplement_3):14096–14101.

Callahan, E. S. and Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915.

Chen, C.-C. and Roth, C. (2012). {{Citation needed}} the dynamics of referencing in Wikipedia. In *Proceedings of the eighth annual international symposium on wikis and open collaboration*, pages 1–4.

Colavizza, G. (2020). COVID-19 research in Wikipedia. *Quantitative Science Studies*, 1(4):1349–1380.

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., and McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*, 15(4):e0230416. Publisher: Public Library of Science.

Costas, R., Rijcke, S., and Marres, N. (2021). "Heterogeneous couplings": Operationalizing network perspectives to study science-society interactions through social media metrics. *Journal of the Association for Information Science and Technology*, 72(5):595–610.

Cozzens, S. (1989). What do citations count? the rhetoric-first model. *Scientometrics*, 15(5-6):437–447.

Currie, M. (2012). The feminist critique: Mapping controversy in Wikipedia. In *Understanding Digital Humanities*, pages 224–248. Springer.

Das, S. and Lavoie, A. (2014). Automated inference of point of view from user interactions in collective intelligence venues. In *International Conference on Machine Learning*, pages 82–90. PMLR.

ElSabry, E. (2017). Who needs access to research? exploring the societal impact of open access. *Revue française des sciences de l'information et de la communication*.

Esteves Gonçalves da Costa, B. and Cukierman, H. L. (2019). How anthropogenic climate change prevailed: A case study of controversies around global warming on portuguese Wikipedia. *New Media & Society*, 21(10):2261–2282.

Evans, P. and Krauthammer, M. (2011). Exploring the use of social media to measure journal article impact. In *AMIA Annual Symposium Proceedings*, volume 2011, page 374. American Medical Informatics Association.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., and Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. *The FASEB Journal*, 22(8):2623–2628. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1096/fj.08-107938.

Fallis, D. (2008). Toward an epistemology of Wikipedia. *Journal of the American Society for Information science and Technology*, 59(10):1662–1674.

Fang, Z. and Costas, R. (2020). Studying the accumulation velocity of altmetric data tracked by Altmetric.com. *Scientometrics*, 123(2):1077–1101.

Fetahu, B., Anand, A., and Anand, A. (2015). How much is Wikipedia lagging behind news? In *Proceedings of the ACM Web Science Conference*, pages 1–9.

Fetahu, B., Markert, K., Nejdl, W., and Anand, A. (2016). Finding news citations for Wikipedia. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 337–346.

Forte, A., Andalibi, N., Gorichanaz, T., Kim, M. C., Park, T., and Halfaker, A. (2018). Information fortification: An online citation behavior. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork*, pages 83–92.

Forte, A., Larco, V., and Bruckman, A. (2009). Decentralization in Wikipedia governance. *Journal of Management Information Systems*, 26(1):49–72.

Fraser, N., Momeni, F., Mayr, P., and Peters, I. (2019). The effect of biorxiv preprints on citations and altmetrics. *BioRxiv*, page 673665.

Fuchs, C. (2007). *Internet and society: Social theory in the information age*. Routledge.

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., and Harnad, S. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLOS ONE*, 5(10):e13636. Publisher: Public Library of Science.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from US daily newspapers. *Econometrica*, 78(1):35–71.

Gilbert, G. N. and Mulkay, M. (1984). *Opening Pandora's box: A sociological analysis of scientists' discourse*. CUP Archive.

Glott, R., Schmidt, P., and Ghosh, R. (2010). Wikipedia survey–overview of results. *United Nations University: Collaborative Creativity Group*, 8:1158–1178.

González-Pereira, B., Guerrero-Bote, V. P., and Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, 4(3):379–391.

Greenstein, S. and Zhu, F. (2012a). Collective intelligence and neutral point of view: the case of Wikipedia. Technical report, National Bureau of Economic Research.

Greenstein, S. and Zhu, F. (2012b). Is Wikipedia biased? *American Economic Review*, 102(3):343–348.

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.

Hao, H., Cui, Y., Wang, Z., and Kim, Y.-S. (2022). Thirty-Two Years of IEEE VIS: Authors, Fields of Study and Citations. arXiv:2208.03772 [cs].

Hara, N. and Doney, J. (2015). Social construction of knowledge in Wikipedia. *First Monday*.

Heilman, J. M., Kemmann, E., Bonert, M., Chatterjee, A., Ragar, B., Beards, G. M., Iberri, D. J., Harvey, M., Thomas, B., Stomp, W., et al. (2011). Wikipedia: a key tool for global public health promotion. *Journal of medical Internet research*, 13(1):e1589.

Heilman, J. M. and West, A. G. (2015). Wikipedia and medicine: quantifying readership, editors, and the significance of natural language. *Journal of medical Internet research*, 17(3):e62.

Herzog, C., Hook, D., and Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1):387–395.

Holmberg, K., Hedman, J., Bowman, T. D., Didegah, F., and Laakso, M. (2020). Do articles in open access journals have more frequent altmetric activity than articles in subscription-based journals? An investigation of the research output of Finnish universities. *Scientometrics*, 122(1):645–659.

Hu, B. (2024). Negotiation, power and ethics in online collaborative translation: translation of "covid-19" by Wikipedia translator-editors. *The Translator*, 30(1):78–95.

Hube, C. (2017). Bias in Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 717–721.

Hube, C. and Fetahu, B. (2018). Detecting biased statements in Wikipedia. In *Companion proceedings of the the web conference 2018*, pages 1779–1786.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied linguistics*, 20(3):341–367.

Jemielniak, D. (2020). *Common knowledge? An ethnography of Wikipedia*. Stanford University Press.

Jemielniak, D. and Aibar, E. (2016). Bridging the gap between Wikipedia and academia. *Journal of the Association for Information Science and Technology*, 67(7):1773–1776.

Jemielniak, D., Masukume, G., and Wilamowski, M. (2019). The most influential medical journals according to Wikipedia: quantitative analysis. *Journal of medical Internet research*, 21(1):e11429.

Johnson, I., Gerlach, M., and Sáez-Trumper, D. (2021). Language-agnostic Topic Classification for Wikipedia. In *Companion Proceedings of the Web Conference 2021*, pages 594–601.

Johnson, I. and Halfaker, A. (2020). Wikipedia Articles and Associated WikiProject Templates. *Figshare*.

Kaffee, L.-A. and Elsahar, H. (2021). References in Wikipedia: The editors' perspective. In *Companion Proceedings of the Web Conference 2021*, pages 535–538.

Kalyanasundaram, A., Wei, W., Carley, K. M., and Herbsleb, J. D. (2015). An agent-based model of edit wars in Wikipedia: How and when is consensus reached. In *2015 Winter Simulation Conference (WSC)*, pages 276–287. IEEE.

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, J., Kim, S., and Lee, C. (2019). Anticipating technological convergence: Link prediction using Wikipedia hyperlinks. *Technovation*, 79:25–34.

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., and Mytkowicz, T. (2007a). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World wide web*, 1(2):19.

Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. (2007b). He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462.

Kokash, N. and Colavizza, G. (2024). A Comprehensive Dataset of Classified Citations with Identifiers from English Wikipedia (2024).

Konieczny, P. (2016). Teaching with w ikipedia in a 21st-century classroom: Perceptions of w ikipedia and its educational benefits. *Journal of the Association for Information Science and Technology*, 67(7):1523–1534.

Kousha, K. and Thelwall, M. (2017). Are Wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*, 68(3):762–779.

Langham-Putrow, A., Bakker, C., and Riegelman, A. (2021). Is the open access citation advantage real? a systematic review of the citation of open access and subscription-based articles. *PloS one*, 16(6):e0253129.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lemmerich, F., Sáez-Trumper, D., West, R., and Zia, L. (2019). Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 618–626, New York, NY, USA. Association for Computing Machinery.

Lewoniewski, W., Węcel, K., and Abramowicz, W. (2017). Relative quality and popularity evaluation of multilingual Wikipedia articles. In *Informatics*, volume 4, page 43. MDPI.

Lewoniewski, W., Węcel, K., and Abramowicz, W. (2019). Multilingual ranking of Wikipedia articles with quality and popularity assessment in different topics. *Computers*, 8(3):60.

Lewoniewski, W., Węcel, K., and Abramowicz, W. (2020). Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. *Information*, 11(5):263.

Lewoniewski, W., Węcel, K., and Abramowicz, W. (2023). Understanding the use of scientific references in multilingual Wikipedia across various topics. *Procedia Computer Science*, 225:3977–3986.

Lin, J. and Fenner, M. (2014). An analysis of Wikipedia references across plos publications. In *Altmetrics14: Expanding impacts and metrics an ACM web science conference 2014 workshop*, pages 23–26.

Lin, J., Yu, Y., Zhou, Y., Zhou, Z., and Shi, X. (2020). How many preprints have actually been printed and why: a case study of computer science preprints on arxiv. *Scientometrics*, 124(1):555–574.

Maggio, L. A., Steinberg, R. M., Piccardi, T., and Willinsky, J. M. (2020). Meta-Research: Reader engagement with medical content on Wikipedia. *Elife*, 9:e52426.

Maggio, L. A., Willinsky, J. M., Steinberg, R. M., Mietchen, D., Wass, J. L., and Dong, T. (2017). Wikipedia as a gateway to biomedical research: The relative distribution and use of citations in the English Wikipedia. *PLOS ONE*, 12(12):e0190046. Publisher: Public Library of Science.

Martín-Martín, A., Costas, R., van Leeuwen, T., and Delgado López-Cózar, E. (2018). Evidence of open access of scientific publications in Google Scholar: A large-scale analysis. *Journal of Informetrics*, 12(3):819–841.

McMullin, E. (1987). Scientific controversy and its termination. *Scientific controversies: Case studies in the resolution and closure of disputes in science and technology*, pages 49–92.

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., and Lanamäki, A. (2015). "the sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245.

Messner, M. and South, J. (2011). Legitimizing Wikipedia: How US national newspapers frame and use the online encyclopedia in their coverage. *Journalism Practice*, 5(2):145–160.

Morgan, J. (2019). Research:patrolling on Wikipedia.

Murray, D., Siler, K., Lariviére, V., Chan, W. M., Collings, A. M., Raymond, J., and Sugimoto, C. R. (2018). Gender and international diversity improves equity in peer review. *BioRxiv*, 10:400515.

Nicholson, J. M., Uppala, A., Sieber, M., Grabitz, P., Mordaunt, M., and Rife, S. C. (2021). Measuring the quality of scientific references in Wikipedia: an analysis of more than 115M citations to over 800,000 scientific articles. *The FEBS Journal*, 288(14):4242–4248.

Nielsen (2010). Top news cites referenced from Wikipedia.

Nielsen, F. A. (2007). Scientific citations in Wikipedia. arXiv:0705.2106 [cs].

Nielsen, F. Å., Mietchen, D., and Willighagen, E. (2017). Scholia, scientometrics and wikidata. In *European Semantic Web Conference*, pages 237–259. Springer.

Nigel Gilbert, G. (1977). Referencing as persuasion. *Social studies of science*, 7(1):113–122.

Nisbet, M. C. and Scheufele, D. A. (2009). What's next for science communication? promising directions and lingering distractions. *American journal of botany*, 96(10):1767–1778.

Park, T. K. (2011). The visibility of Wikipedia in scholarly publications.

Patience, G. S., Patience, C. A., Blais, B., and Bertrand, F. (2017). Citation analysis of scientific categories. *Heliyon*, 3(5):e00300.

Patterson, T. E. (2011). *Out of Order: An incisive and boldly original critique of the news media's domination of Ameri*. Vintage.

Pavalanathan, U., Han, X., and Eisenstein, J. (2018). Mind your pov: Convergence of articles and editors towards Wikipedia's neutrality norm. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23.

Perianes-Rodriguez, A., Waltman, L., and van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, 10(4):1178–1195.

Piccardi, T., Gerlach, M., Arora, A., and West, R. (2023). A large-scale characterization of how readers browse Wikipedia. *ACM Transactions on the Web*, 17(2):1–22.

Piccardi, T., Redi, M., Colavizza, G., and West, R. (2020). Quantifying engagement with citations on Wikipedia. In *Proceedings of The Web Conference 2020*, WWW '20, pages 2365–2376, New York, NY, USA. Association for Computing Machinery.

Piccardi, T., Redi, M., Colavizza, G., and West, R. (2021). On the Value of Wikipedia as a Gateway to the Web. In *Proceedings of the Web Conference 2021*, WWW '21, page 249–260, New York, NY, USA. Association for Computing Machinery.

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., and Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, 6:e4375. Publisher: PeerJ Inc.

Pooladian, A. and Borrego, Á. (2017). Methodological issues in measuring citations in Wikipedia: a case study in library and information science. *Scientometrics*, 113(1):455–464.

Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., and Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007 ACM international conference on supporting group work*, pages 259–268.

Priem, J., Piwowar, H., and Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833 [cs].

Przybyla, P., Borkowski, P., and Kaczyński, K. (2022). Countering disinformation by finding reliable sources: a citation-based approach. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE. ISSN: 2161-4407.

Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F., and Vespignani, A. (2010). Characterizing and modeling the dynamics of online popularity. *Physical review letters*, 105(15):158701.

Reagle, J. and Rhue, L. (2011). Gender bias in Wikipedia and britannica. *International Journal of Communication*, 5:21.

Redalyc, L., Clase, R., and IN-COM UAB, S. (2003). Berlin declaration on open access to knowledge in the sciences and humanities.

Redi, M., Fetahu, B., Morgan, J., and Taraborelli, D. (2019). Citation needed: A taxonomy and algorithmic assessment of Wikipedia's verifiability. In *The World Wide Web Conference*, pages 1567–1578.

Ribeiro, F. N., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., and Gummadi, K. P. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Twelfth international AAAI conference on web and social media*.

Rogers, R. and Sendijarevic, E. (2012). Neutral or national point of view? a comparison of srebrenica articles across Wikipedia's language versions. *Wikepedia Academy*.

Roy, D., Bhatia, S., and Jain, P. (2021). Information asymmetry in Wikipedia across different languages: A statistical analysis. *Journal of the Association for Information Science and Technology*.

Saez-Trumper, D. (2019). Online disinformation and the role of Wikipedia.

Salah, A. A., Gao, C., Suchecki, K., and Scharnhorst, A. (2012). Need to categorize: A comparative look at the categories of universal decimal classification system and Wikipedia. *Leonardo*, 45(1):84–85.

Sato, Y. (2021). Non-english editions of Wikipedia have a misinformation problem.

Schneider, J., Passant, A., and Breslin, J. G. (2010). A content analysis: How Wikipedia talk pages are used. In *Proceedings of the 2nd International Conference of Web Science*, pages 1–7.

Schug, M., Bilandzic, H., and Kinnebrock, S. (2025). Endorsement of scientific norms among non-scientists: The role of science news consumption, political ideology, and science field. *Public Understanding of Science*, page 09636625251315882.

Schweitzer, N. J. (2008). Wikipedia and Psychology: Coverage of Concepts and Its Use by Undergraduate Students. *Teaching of Psychology*, 35(2):81–85. Publisher: SAGE Publications Inc.

Segev, E. and Sharon, A. J. (2017). Temporal patterns of scientific information-seeking on google and Wikipedia. *Public understanding of science*, 26(8):969–985.

Severin, A., Egger, M., Eve, M. P., and Hürlimann, D. (2020). Discipline-specific open access publishing practices and barriers to change: an evidence-based review. *F1000Research*, 7:1925.

Shafee, T., Mietchen, D., and Su, A. I. (2017). Academics can help shape Wikipedia. *Science*, 357(6351):557–558.

Shi, F., Teplitskiy, M., Duede, E., and Evans, J. A. (2019). The wisdom of polarized crowds. *Nature human behaviour*, 3(4):329–336.

Shuai, X., Jiang, Z., Liu, X., and Bollen, J. (2013). A comparative study of academic and Wikipedia ranking. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 25–28.

Silva, F., Viana, M., Travençolo, B., and da F. Costa, L. (2011). Investigating relationships within and between category networks in Wikipedia. *Journal of Informetrics*, 5(3):431–438.

Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M., and Leskovec, J. (2017). Why We Read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1591–1600, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Singh, H., West, R., and Colavizza, G. (2021). Wikipedia citations: A comprehensive data set of citations with identifiers extracted from english Wikipedia. *Quantitative Science Studies*, 2(1):1–19.

Smith, D. A. (2020). Situating Wikipedia as a health information resource in various contexts: A scoping review. *PloS one*, 15(2):e0228786.

Steiert, O. (2025). Declaring crisis? temporal constructions of climate change on Wikipedia. *Public Understanding of Science*, 34(2):188–203.

Struck, D. B., Durning, M., Roberge, G., and Campbell, D. (2018). Modelling the Effects of Open Access, Gender and Collaboration on Citation Outcomes: Replicating, Expanding and Drilling. *STI 2018 Conference Proceedings*, pages 436–447. Publisher: Centre for Science and Technology Studies (CWTS).

Sugandhika, C. and Ahangama, S. (2022). Assessing information quality of Wikipedia articles through google's eat model. *IEEE Access*, 10:52196–52209.

Sugimoto, C. R., Work, S., Larivière, V., and Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and technology*, 68(9):2037–2062.

Suh, B., Chi, E. H., Pendleton, B. A., and Kittur, A. (2007). Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. In *2007 IEEE symposium on visual analytics science and technology*, pages 163–170. IEEE.

Sumi, R., Yasseri, T., et al. (2011). Edit wars in Wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 724–727. IEEE.

Sundin, O. (2011). Janitors of knowledge: constructing knowledge in the everyday life of Wikipedia editors. *Journal of documentation*, 67(5):840–862.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

Sutter, D. (2000). Can the media be so liberal-the economics of media bias. *Cato J;*, 20:431.

Taraborelli, D. (2019). File:knowledge integrity - wikimedia research 2030.

Tattersall, A., Sheppard, N., Blake, T., O'Neill, K., and Carroll, C. (2022). Exploring open access coverage of Wikipedia-cited research across the white rose universities. *Insights*, 35.

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., and Hartgerink, C. H. J. (2016). The academic, economic and societal impacts of Open Access: an evidence-based review. Technical Report 5:632, F1000Research. Type: article.

Teplitskiy, M., Lu, G., and Duede, E. (2017). Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, 68(9):2116–2127.

Thompson, N. and Hanley, D. (2018). Science is shaped by Wikipedia: Evidence from a randomized control trial. *MIT Sloan Research Paper*, 5238-17.

Tohidinasab, F. and Jamali, H. R. (2013). Why and where Wikipedia is cited in journal articles? *Journal of Scientmetric Research*, 2:231–238.

Tomaszewski, R. and MacDonald, K. I. (2016). A study of citations to Wikipedia in scholarly publications. *Science & technology libraries*, 35(3):246–261.

Torres-Salinas, D., Romero-Frías, E., and Arroyo-Machado, W. (2019). Mapping the backbone of the Humanities through the eyes of Wikipedia. *Journal of Informetrics*, 13(3):793–803.

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

Tran, C., Champion, K., Forte, A., Hill, B. M., and Greenstadt, R. (2020). Are anonymity-seekers just like everybody else? an analysis of contributions to Wikipedia from tor. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 186–202. IEEE.

Umarova, K. and Mustafaraj, E. (2019). How partisanship and perceived political bias affect Wikipedia entries of news sources. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1248–1253, New York, NY, USA. Association for Computing Machinery.

Valentim, R., Comarela, G., Park, S., and Saez-Trumper, D. (2021). Tracking Knowledge Propagation Across Wikipedia Languages. arXiv:2103.16613 [cs].

Viégas, F. B., Wattenberg, M., Kriss, J., and Van Ham, F. (2007). Talk before you type: Coordination in Wikipedia. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 78–78. IEEE.

Visser, M., van Eck, N. J., and Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 2(1):20–41.

Voss, J. (2005). Measuring Wikipedia. In *Proceedings of ISSI*, volume 1, pages 221–231.

Vuong, B.-Q., Lim, E.-P., Sun, A., Le, M.-T., Lauw, H. W., and Chang, K. (2008). On ranking controversies in Wikipedia: models and evaluation. In *Proceedings of the 2008 international conference on Web search and data mining*, pages 171–182.

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a man's Wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 454–463.

Weltevrede, E. and Borra, E. (2016). Platform affordances and data practices: The value of dispute on Wikipedia. *Big Data & Society*, 3(1):2053951716653418.

West, A. G., Kannan, S., and Lee, I. (2010). Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata? In *Proceedings of the Third European Workshop on System Security*, pages 22–28.

Wikipedia (2022). Vandalism on Wikipedia.

Wilkinson, D. M. and Huberman, B. A. (2007). Assessing the value of coooperation in Wikipedia. *arXiv preprint cs/0702140*.

Wilson, A. M. and Likens, G. E. (2015). Content volatility of scientific topics in Wikipedia: a cautionary tale. *PloS one*, 10(8):e0134454.

Wolton, S. (2019). Are biased media bad for democracy? *American Journal of Political Science*, 63(3):548–562.

Wyatt, S., Harris, A., and Kelly, S. E. (2016). Controversy goes online: Schizophrenia genetics on Wikipedia. *Science & Technology Studies*, 29(1):13–29.

Yang, P. and Colavizza, G. (2022a). A map of science in Wikipedia. In *Companion Proceedings of the Web Conference 2022*, pages 1289–1300.

Yang, P. and Colavizza, G. (2022b). Science in context: An overview of all Wikipedia's sources. In *26th International Conference on Science, Technology and Innovation Indicators (STI 2022), Granada*. Zenodo.

Yang, P., Shoaib, A., West, R., and Colavizza, G. (2024). Open access improves the dissemination of science: insights from Wikipedia. *Scientometrics*, 129(11):7083–7106.

Yarovoy, A., Nagar, Y., Minkov, E., and Arazy, O. (2020). Assessing the Contribution of Subject-matter Experts to Wikipedia. *ACM Transactions on Social Computing*, 3(4):1–36.

Yasseri, T., Spoerri, A., Graham, M., and Kertesz, J. (2013). The most controversial topics in Wikipedia: A multilingual and geographical analysis. *SSRN Electronic Journal*.

Yasseri, T., Sumi, R., Rung, A., Kornai, A., and Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PloS one*, 7(6):e38869.

Yegros-Yegros, A., Rafols, I., and D'Este, P. (2015). Does Interdisciplinary Research Lead to Higher Citation Impact? The Different Effect of Proximal and Distal Interdisciplinarity. *PLOS ONE*, 10(8):e0135095. Publisher: Public Library of Science.

Young, J. S. and Brandes, P. M. (2020). Green and gold open access citation and interdisciplinary advantage: A bibliometric study of two science journals. *The Journal of Academic Librarianship*, 46(2):102105.

Yuen, J. (2018). Comparison of Impact Factor, Eigenfactor Metrics, and SCImago Journal Rank Indicator and h-index for Neurosurgical and Spinal Surgical Journals. *World Neurosurgery*, 119:e328–e337.

Zagorova, O., Ulloa, R., Weller, K., and Flöck, F. (2021). "I updated the <ref>": The evolution of references in the English Wikipedia and the implications for altmetrics. *Quantitative Science Studies*, pages 1–27.

Zheng, X., Chen, J., Yan, E., and Ni, C. (2023). Gender and country biases in Wikipedia citations to scholarly publications. *Journal of the Association for Information Science and Technology*, 74(2):219–233. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24723.

Zhou, Y., Cristea, A., and Roberts, Z. (2015). Is Wikipedia really neutral? a sentiment perspective study of war-related Wikipedia articles since 1945. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 160–168.

Samenvatting

Wikipedia is uitgegroeid tot een centraal platform voor het publieke gebruik van wetenschappelijke kennis. Als een open, collaboratief bewerkt encyclopedisch project weerspiegelt het niet alleen het bredere informatie-ecosysteem, maar beïnvloedt het ook hoe wetenschap wordt geraadpleegd, geïnterpreteerd en bediscussieerd door miljoenen gebruikers wereldwijd. In dit proefschrift onderzoek ik hoe wetenschappelijke kennis wordt verspreid en betwist op Wikipedia, met speciale aandacht voor de rol van open access-publicaties (OA). Door gebruik te maken van grootschalige datasets en een combinatie van kwantitatieve analyse en netwerkmodellering, analyseer ik de structurele en epistemische dynamieken die ten grondslag liggen aan het gebruik van wetenschappelijke literatuur op het platform.

Het proefschrift bestaat uit twee hoofdonderdelen. In het eerste deel onderzoek ik hoe Wikipedia wetenschappelijke kennis integreert en structureert door grote citatiepatronen van Wikipedia naar wetenschappelijke publicaties te analyseren. Ik laat zien dat Wikipedia sterk leunt op tijdschriftartikelen uit STEM-disciplines, met name biologie en geneeskunde, en dat biografische artikelen fungeren als belangrijke bruggen tussen wetenschappelijke en geesteswetenschappelijke domeinen. Daarnaast onderzoek ik de politieke en epistemische dimensies van citatiepraktijken door de selectie van nieuwsbronnen op Wikipedia te analyseren. Hierbij komt een consistente maar gematigde liberale voorkeur in bronselectie naar voren, zelfs wanneer rekening wordt gehouden met verschillen in feitelijke betrouwbaarheid.

Het tweede deel richt zich op de impact van open access op de verspreiding en betwisting van wetenschappelijke kennis binnen Wikipedia. Op basis van citatiegegevens op artikelniveau constateer ik dat OA-publicaties significant vaker worden geciteerd op Wikipedia dan betaalmuurpublicaties, zelfs wanneer er wordt gecontroleerd voor factoren zoals het aantal citaties en de leeftijd van de publicatie. Dit effect is vooral sterk voor recent gepubliceerde en vaak geciteerde artikelen, wat de versterkende rol van OA in publieke kennisinfrastructuren onderstreept. In het laatste hoofdstuk onderzoek ik hoe wetenschappelijke bronnen worden

ingezet in redactiedisputen op Wikipedia. Ik vind dat OA-publicaties niet alleen vaker worden geciteerd, maar ook vaker onderwerp zijn van geschillen, vooral in sociaal gevoelige vakgebieden. Deze bevindingen suggereren dat toegankelijkheid zowel zichtbaarheid als betrokkenheid verhoogt, maar ook de kans op kritiek en controverse vergroot.

Al met al biedt dit proefschrift een veelzijdig perspectief op hoe Wikipedia fungeert als publiek intermediair voor wetenschappelijke communicatie. Door het analyseren van citatiedynamieken, redactionele praktijken en toegangsvormen draagt het bij aan ons begrip van hoe wetenschap wordt vertaald naar publieke kennis en hoe digitale platforms de grenzen van wetenschappelijke autoriteit vormgeven.

Abstract

Wikipedia has become a central venue for the public consumption of scientific knowledge. As an open, collaboratively edited encyclopedia, it not only reflects the broader information ecosystem but also shapes how science is accessed, interpreted, and debated by millions of users worldwide. In this thesis, I investigate how scientific knowledge is disseminated and contested on Wikipedia, with a particular focus on the role of open access (OA) publications. Drawing on large-scale datasets and combining quantitative analysis with network modeling, I examine the structural and epistemic dynamics that underpin the use of scientific literature on the platform.

The thesis consists of two main parts. In the first part, I examine how Wikipedia integrates and structures scientific knowledge by analyzing large-scale citation patterns from Wikipedia to scientific publications. I show that Wikipedia heavily relies on journal articles from STEM fields, especially biology and medicine, and that biographical articles serve as important bridges between scientific and humanistic domains. I also explore the political and epistemic dimensions of citation practices by examining the selection of news media sources on Wikipedia, revealing a consistent but moderate liberal bias in source selection, even after accounting for differences in factual reliability.

The second part focuses on the impact of open access publishing on the dissemination and contestation of science within Wikipedia. Using article-level citation data, I find that OA publications are significantly more likely to be cited in Wikipedia than paywalled ones, even after accounting for factors such as citation count and publication age. This effect is especially pronounced for highly cited and more recent articles, highlighting the amplifying role of OA in public knowledge infrastructures. In the final chapter, I study how scientific sources are mobilized in Wikipedia's editorial disputes. I find that OA publications are not only more likely to be cited, but also more likely to be contested in disputes, particularly in socially sensitive fields. These findings suggest that accessibility increases both visibility and engagement, but may also heighten exposure to scrutiny and disagreement.

Taken together, this thesis offers a multifaceted view of how Wikipedia serves as a public interface for scientific communication. By analyzing citation dynamics, editorial practices, and access models, it contributes to our understanding of how science is translated into public knowledge and how digital platforms shape the boundaries of scientific authority.

Titles in the ILLC Dissertation Series:

ILLC DS-2021-03: **Seyyed Hadi Hashemi**
Modeling Users Interacting with Smart Devices

ILLC DS-2021-04: **Sophie Arnoult**
Adjunction in Hierarchical Phrase-Based Translation

ILLC DS-2021-05: **Cian Guilfoyle Chartier**
A Pragmatic Defense of Logical Pluralism

ILLC DS-2021-06: **Zoi Terzopoulou**
Collective Decisions with Incomplete Individual Opinions

ILLC DS-2021-07: **Anthia Solaki**
Logical Models for Bounded Reasoners

ILLC DS-2021-08: **Michael Sejr Schlichtkrull**
Incorporating Structure into Neural Models for Language Processing

ILLC DS-2021-09: **Taichi Uemura**
Abstract and Concrete Type Theories

ILLC DS-2021-10: **Levin Hornischer**
Dynamical Systems via Domains: Toward a Unified Foundation of Symbolic and Non-symbolic Computation

ILLC DS-2021-11: **Sirin Botan**
Strategyproof Social Choice for Restricted Domains

ILLC DS-2021-12: **Michael Cohen**
Dynamic Introspection

ILLC DS-2021-13: **Dazhu Li**
Formal Threads in the Social Fabric: Studies in the Logical Dynamics of Multi-Agent Interaction

ILLC DS-2021-14: **Álvaro Piedrafita**
On Span Programs and Quantum Algorithms

ILLC DS-2022-01: **Anna Bellomo**
Sums, Numbers and Infinity: Collections in Bolzano's Mathematics and Philosophy

ILLC DS-2022-02: **Jan Czajkowski**
Post-Quantum Security of Hash Functions

ILLC DS-2022-03: **Sonia Ramotowska**
Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences

ILLC DS-2022-04: **Ruben Brokkelkamp**
How Close Does It Get?: From Near-Optimal Network Algorithms to Suboptimal Equilibrium Outcomes

ILLC DS-2022-05: **Lwenn Bussière-Carae**
No means No! Speech Acts in Conflict

ILLC DS-2022-06: **Emma Mojet**
Observing Disciplines: Data Practices In and Between Disciplines in the 19th and Early 20th Centuries

ILLC DS-2022-07: **Freek Gerrit Witteveen**
Quantum information theory and many-body physics

ILLC DS-2023-01: **Subhasree Patro**
Quantum Fine-Grained Complexity

ILLC DS-2023-02: **Arjan Cornelissen**
Quantum multivariate estimation and span program algorithms

ILLC DS-2023-03: **Robert Paßmann**
Logical Structure of Constructive Set Theories

ILLC DS-2023-04: **Samira Abnar**
Inductive Biases for Learning Natural Language

ILLC DS-2023-05: **Dean McHugh**
Causation and Modality: Models and Meanings

ILLC DS-2023-06: **Jialiang Yan**
Monotonicity in Intensional Contexts: Weakening and: Pragmatic Effects under Modals and Attitudes

ILLC DS-2023-07: **Yiyan Wang**
Collective Agency: From Philosophical and Logical Perspectives

ILLC DS-2023-08: **Lei Li**
Games, Boards and Play: A Logical Perspective

ILLC DS-2023-09: **Simon Rey**
Variations on Participatory Budgeting

ILLC DS-2023-10: **Mario Giulianelli**
Neural Models of Language Use: Studies of Language Comprehension and Production in Context

ILLC DS-2023-11: **Guillermo Menéndez Turata**
Cyclic Proof Systems for Modal Fixpoint Logics

ILLC DS-2023-12: **Ned J.H. Wontner**
Views From a Peak: Generalisations and Descriptive Set Theory

ILLC DS-2024-01: **Jan Rooduijn**
Fragments and Frame Classes: Towards a Uniform Proof Theory for Modal Fixed Point Logics

ILLC DS-2024-02: **Bas Cornelissen**
Measuring musics: Notes on modes, motifs, and melodies

ILLC DS-2024-03: **Nicola De Cao**
Entity Centric Neural Models for Natural Language Processing

ILLC DS-2024-04: **Ece Takmaz**
Visual and Linguistic Processes in Deep Neural Networks: A Cognitive Perspective

ILLC DS-2024-05: **Fatemeh Seifan**
Coalgebraic fixpoint logic Expressivity and completeness result

ILLC DS-2024-06: **Jana Sotáková**
Isogenies and Cryptography

ILLC DS-2024-07: **Marco Degano**
Indefinites and their values

ILLC DS-2024-08: **Philip Verduyn Lunel**
Quantum Position Verification: Loss-tolerant Protocols and Fundamental Limits

ILLC DS-2024-09: **Rene Allerstorfer**
Position-based Quantum Cryptography: From Theory towards Practice

ILLC DS-2024-10: **Willem Feijen**
Fast, Right, or Best? Algorithms for Practical Optimization Problems

ILLC DS-2024-11: **Daira Pinto Prieto**
Combining Uncertain Evidence: Logic and Complexity

ILLC DS-2024-12: **Yanlin Chen**
On Quantum Algorithms and Limitations for Convex Optimization and Lattice Problems

ILLC DS-2024-13: **Jaap Jumelet**
Finding Structure in Language Models

ILLC DS-2025-01: **Julian Chingoma**
On Proportionality in Complex Domains

ILLC DS-2025-02: **Dmitry Grinko**
Mixed Schur-Weyl duality in quantum information

ILLC DS-2025-03: **Rochelle Choenni**
Multilinguality and Multiculturalism: Towards more Effective and Inclusive Neural Language Models

ILLC DS-2025-04: **Aleksi Anttila**
Not Nothing: Nonemptiness in Team Semantics

ILLC DS-2025-05: **Niels M. P. Neumann**
Adaptive Quantum Computers: decoding and state preparation

ILLC DS-2025-06: **Alina Leidinger**
Towards Language Models that benefit us all: Studies on stereotypes, robustness, and values

ILLC DS-2025-07: **Zhi Zhang**
Advancing Vision and Language Models through Commonsense Knowledge, Efficient Adaptation and Transparency

ILLC DS-2025-08: **Sophie Klumper**
The Gap and the Gain: Improving the Approximate Mechanism Design Frontier in Constrained Environments

ILLC DS-2026-01: **Bryan Eikema**
A Sampling-Based Exploration of Neural Text Generation Models

ILLC DS-2026-02: **Marten Folkertsma**
Empowering Quantum Computation with: Measurements, Catalysts, and Guiding States

ILLC DS-2026-03: **Valentin Richard**
Presuppositional and Dynamic Aspects of Questions