A Tale of Two Communities: Exploring Academic References on Stack Overflow

Run Huang
University of Southern California
Los Angeles, USA

Souti Chattopadhyay
University of Southern California
Los Angeles, USA

ABSTRACT

Stack Overflow is widely recognized by software practitioners as the go-to resource for addressing technical issues and sharing practical solutions. While it is not typically seen as a forum for scholarly discourse, users on Stack Overflow often refer to academic sources in their discussions. Yet, little is known about these referenced works from the academic community and how they intersect the needs and interests of the Stack Overflow community. To bridge this gap, we conducted a large-scale study on academic references in Stack Overflow. Our findings reveal that Stack Overflow communities with different domains of interest engage with academic literature at varying frequencies and speeds. The contradicting patterns suggest that some disciplines may have diverged in their interests and development trajectories from the corresponding practitioner community. Finally, we discuss the potential of Stack Overflow in gauging the real-world relevance of academic research.

CCS CONCEPTS

• **Information systems** → **World Wide Web**.

KEYWORDS

Stack Overflow, citation analysis, industry impact

ACM Reference Format:
Run Huang and Souti Chattopadhyay. 2024. A Tale of Two Communities: Exploring Academic References on Stack Overflow. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651464

1 INTRODUCTION

Stack Overflow (SO) is a popular Q&A platform among software practitioners and developers [2] for discussing technical issues [1]. These discussions might go beyond just finding immediate technical solutions, as SO users often delve deeper into the problems at hand and share their own knowledge and expertise in the process. Organically, SO users refer to a wide range of sources to support their claims [9], including academic articles published in conferences and journals. Such practices suggest that SO may be serving as a potential bridge between the two communities—developers and academic researchers—for knowledge dissemination.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651464

However, we currently lack an understanding on the contribution of these academic references in SO discussions and how the SO community interacts with them. This limits our insights into SO's role in facilitating knowledge transfer and reflecting the practical relevance of academic research. Investigating this gap could uncover potential discrepancies between cutting-edge research and the current industry norms and interests. It could also guide both communities in identifying how academic advancements can match the needs of the SO community and the broader developer community it serves. To the best of our knowledge, this is the first study to examine the recognition of academic research within SO, and to investigate how scientific knowledge is referenced and consumed in real-world, problem-solving contexts.

Prior works have studied academic references on Twitter [16] and Wikipedia [17], emphasizing the general visibility and popularity of academic research rather than its practical relevance. Similarly, a platform called Altmetric included SO citations as one of its alternative metrics for measuring academic impact. However, it only quantifies how frequently academic works are mentioned but overlooks the context of these references and their connections to the original discourse. As a result, Altmetric suffers from the similar limitation of serving merely as an indicator of visibility.

In this paper, we aim at a deeper understanding of how academic research is recognized and applied in solving real-world software development challenges across different communities of interest on SO, by answering the following research questions (**RQ**s):

RQ1: What academic articles are cited on Stack Overflow?
RQ2: Which parts of Stack Overflow rely on academic sources?
RQ3: How do various Stack Overflow communities interact with academic research?
RQ4: How quickly are academic articles recognized on Stack Overflow?

We sifted through 44 million URLs on SO and identified 15,066 references to 10718 unique academic articles. Leveraging topic modeling and social network analysis techniques, we explored the patterns of interaction between SO communities with diverse technical interests and academic research from various fields. To support future research, we have made our dataset publicly available [5].

2 METHODOLOGY

Unlike Wikipedia or academic settings where citations follow a standard format, users on Stack Overflow (SO) commonly use hyperlinks to reference academic sources. This practice makes it challenging to distinguish academic references from other types of links, as both appear as bare URLs without bibliographic information [2]. In this section, we describe our heuristic method for identifying links leading to academic articles and the data collection process.

i) Filtering for Potential Candidates. Instead of examining every external link on SO, we narrowed our search to those that are most likely to host academic documents. Prior efforts [10] have limited to links containing recognized identifiers such as DOIs and ISBNs. We significantly expanded this scope by incorporating links originating from recognized academic repositories, e.g., the *ACM Digital Library*.

Links containing DOIs were identified using regex. For the remaining, we checked if their web domains appeared on a list of domain names affiliated with academic repositories. In compiling this list, we considered various academic entities, including publishers (e.g., Springer), academic societies (e.g., ACM) and databases (e.g., ResearchGate). However, given the sheer number of these potential sources, it is infeasible to include them all. Instead, we adopted a best-effort approach, focusing on the most significant ones. To better align with Stack Overflow's emphasis on software and computing, we included all 31 publishers indexed by DBLP, which is a major Computer Science bibliography. Moreover, with inputs from experts and authoritative sources [11], we further enrich the selection with 13 academic societies renowned for organizing prestigious conferences (e.g., CVPR) and ten well-known academic databases. One author then iteratively curated all relevant web domains belonging to these entities (e.g., *ieeexplore.ieee.org* from IEEE Xplore).

Although our selection[1] is not exhaustive, it effectively covers a substantial part of the academic publishing landscape. The 31 included publishers issue over 16,000 journals in various fields [12], and the 13 academic societies host more than 500 conferences annually, not to mention the extensive reach of the included databases.

ii) Retrieving and Validating Bibliographic data. For every candidate link identified previously, we assumed its content to be academic and attempted to retrieve possible bibliographic information. For links directing to PDF files, we extracted titles and DOIs using Grobid, a popular tool for parsing scientific documents. If the link leads to a webpage, we assumed it to be the landing page of a research article, and retrieved its potential titles from various HTML tags, such as `<h1>`,`<meta property="og:title">`, `<title>`, `<h2>`, etc. Additionally, we searched for DOIs within the HTML file using regex. For non-functional links, we accessed their archived versions via Internet Archive's Wayback Machine.

In cases where the candidate link does not lead to academic content as we assumed, the data heuristically extracted from the HTML or PDF file would not be valid bibliographic data. With this in mind, we can filter out ineligible candidates by cross-validating their extracted metadata against two major academic databases: Semantic Scholar and OpenAlex [7]. A matching record in the title or DOI would confirm the academic nature of the link. We then gathered detailed metadata for verified academic references from the two databases, including abstract, venue, citation counts, etc.

Data Collection. Based on the official Stack Overflow data dump released on December 8, 2023, we extracted 44 million URLs found in the edit histories of 59 million posts (totaling 160 million revisions). Similar to prior work [8], URLs embedded within code blocks were omitted, as they are often irrelevant to knowledge sharing. After this refinement, we obtained a dataset [5] of 30.9 million links, from which 15,066 references to 10718 academic articles made by 12,963 posts were identified through the previously described process.

[1]The included academic repositories and detailed methodology can be found here.

3 RESULTS

***RQ 1**: What academic articles are cited on Stack Overflow?*

For a fine-grained characterization of the cited articles, we categorize each article into its corresponding Field of Research (FoR) using OpenAlex's concept tagging model [13]. This model analyzes the abstract and title and generates an initial list of relevant research fields, along with a confidence score. The fields are assigned in six levels, from the broadest to highly specific, following Wikidata's taxonomy. For each article, we selected the second level field (e.g., *World Wide Web*) with the highest confidence score as its FoR.

The articles referenced on SO span 218 FoR across 19 disciplines. Table 1 presents key characteristics of the ten largest fields in terms of SO reference counts. Notably, *Artificial Intelligence* (AI) makes up over 30% of all references. These AI articles are impactful and recent, averaging 1331 academic citations and an age of 3.75 years at the time of mention. In contrast, SO references from fields such as *Programming Languages* and *Algorithms* tend to be older on average. Such variations indicate the diverse patterns of interacting with academic research on SO, from embracing cutting-edge developments to relying on more established and foundational works.

Table 1: Top 10 Fields of Research (FoR)

Field of Research	#Ref	Avg. Age ♣	Avg. Citation ♠
Artificial Intelligence	5082	***3.75*** ± 0.07	1331.5
Algorithms (ALG)	1373	9.80 ± 0.29	428.5
Prog. Lang. (PL)	721	8.36 ± 0.38	210.4
Natural Language Processing (NLP)	536	4.59 ± 0.29	1087.1
Statistics (Stat)	481	10.36 ± 0.50	780.2
Combinatorics	470	***15.12*** ± 0.61	***173.2***
Parallel Computing	443	6.07 ± 0.36	219.8
Mathematical Optimization	438	8.73 ± 0.44	***1356.2***
Theoretical Computer Science (TCS)	398	7.51 ± 0.42	592.4
Data Mining	387	5.83 ± 0.31	1104.2

Table 2: Venues

Venue	#Ref	h5i ★	Field
arXiv*	1655	N/A	Various
NeurIPS	597	309	ML
ICLR	530	303	ML
CVPR	517	***422***	CV
ICML	361	254	ML
ICCV	189	228	CV
ACL	174	192	NLP
EMNLP	170	176	NLP
TPAMI	144	179	CV
ECCV	136	238	CV
JMLR	129	106	ML
NAACL	113	133	NLP
ATC	109	***59***	OS
Crypto	106	***59***	Security
VLDB	95	79	DB

♣: the average duration (± SEM) between publication and every reference ♠: the average academic citation counts of articles ★: h5-index retrieved from Google Scholar; all venues ranked within top 20 of their respective fields, except arXiv

Table 2 lists the 15 venues with the most SO references. Consistent with existing studies on Wikipedia citations [17], these venues are highly regarded in their respective fields (★), with an interesting exception of *arXiv**. We manually inspected a random sample of the referenced articles from *arXiv* and found that many of them were later published in peer-reviewed venues with slightly different titles. An explanation can be that Stack Overflow users tend to (i) integrate academic insights at a fast pace, and (ii) prefer open-access content. We observe a similar dominance of AI-related research (e.g., ML, NLP, CV) among the venues, with 14 out of the 20 venues and 51% of the articles on *arXiv* being related to AI.

***RQ 2**: Which parts of Stack Overflow rely on academic sources?*

We discerned which technical domains and user communities were associated with the posts citing academic literature on Stack Overflow by analyzing the overarching themes of these discussions. Although using the user-generated tags attached to each SO post for topic modeling may seem intuitive, prior research suggests that such tags often fall short in reflecting the actual discussions accurately [6]. Moreover, SO has over 65,000 existing tags that are too parochial to capture each post's broader themes and areas of

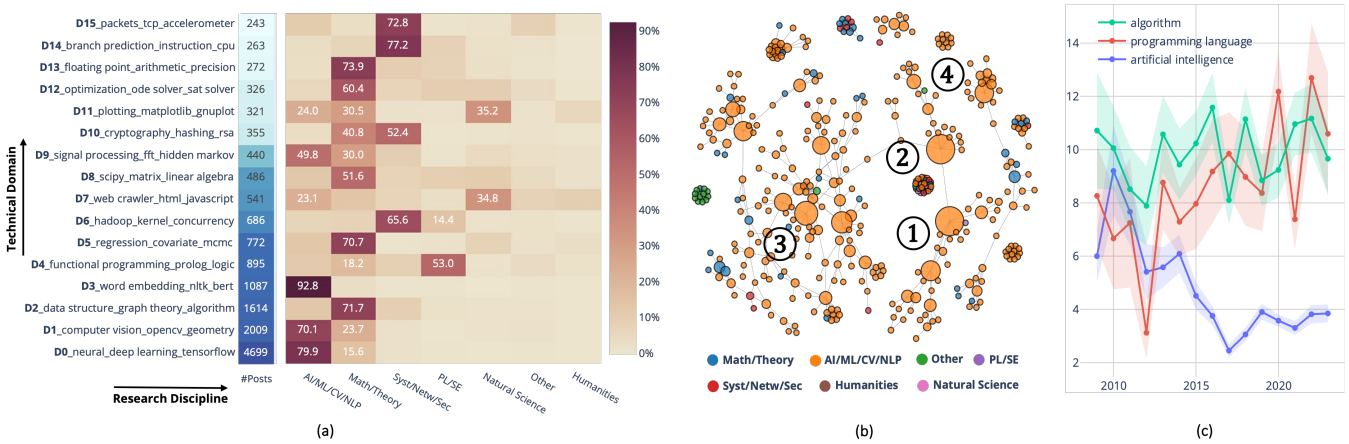

Figure 1: (a) Heatmap depicting the relationships between technical domains (Y-axis) and research disciplines (X-axis). The blue bar along Y-axis represents the number of posts within each technical domain. Cell *(x, y)* denotes the percentage of papers referenced by domain *x* that originate from discipline *y*. (b) Ten largest components in the co-citation network of academic references on SO. ①-④ are key nodes with the highest PageRank score. (c) The average "First-cite Interval" (Y-axis) for articles in three major fields (FoR) cited in SO posts each year (X-axis).

interest. Following existing works [15], we utilized BERTopic [4] to categorize the discussions into distinct technical domains. Initially, the model was fine-tuned for more granularity and coherence, producing 109 preliminary topic clusters. Two authors then discussed and qualitatively merged related topics into broader *domains*. For example, clusters about *Named Entity Recognition* and *Sentiment Analysis* were grouped under the broader domain of *NLP* (D3). Eventually, we compiled all SO topics into 16 technical domains that serve as focal points of interest for distinct communities on SO.

The leftmost column in Figure 1(a) shows the number of posts containing academic references within each technical domain (*Y-axis*). Notably, the domains of D0 (*machine learning*), D1 (*vision/graphics*), and D2 (*algorithms*) were most actively incorporating academic knowledge. Whereas more traditional and application-oriented domains such as D15 (*data communications*) and D14 (*computer architecture*) exhibit less integration of academic research.

***RQ 3**: How do various Stack Overflow communities interact with academic research?*

To examine how different SO communities interact with academic research, we mapped the citation flow between 218 research fields (FoR) and the 16 technical domains. However, the resulting bipartite network is hard to interpret and visualize due to its high dimensionality. We simplified its complexity by aggregating the 218 research fields into seven broader *Disciplines*. The simplification was carefully executed to preserve the underlying citation patterns, guided by two criteria: (i) the similarity in how different FoRs are referenced together across technical domains, signaled by the pairwise Spearman's correlation (where high correlations suggest similar citation patterns), and (ii) their hierarchical relationships within an FoR ontology [14], as to ensure the aggregation also respects established disciplinary structures. For example, *Computer Network*, *Distributed Computing*, and *Computer Security* were grouped into the *Syst/Netw/Sec* category for their high inter-correlations in citation patterns (above 0.6) and shared academic lineage.

Figure 1(a) illustrates the relationships between 16 technical domains and seven research disciplines. Cell (x, y) denotes the percentage of papers referenced by domain x that originate from discipline y. For example, 40.8% of the articles referenced in domain D10 are from the *Math/Theory* discipline. This figure reveals that SO discussions typically rely on one single research discipline, with notable exceptions being D7 (*web scraping*) and D11 (*data visualization*), where the distributions are relatively even. We analyzed sample posts from these two domains and discovered that posts within D7 often seek guidance on downloading research articles or scraping bibliographic data programmatically, while the discussions in D11 typically revolve around replicating data visualizations in scientific articles. These activities (downloading and visualizing) are universal and not confined to any field, leading to the balanced distribution observed. In such cases, academic references primarily serve as illustrative examples or supplementary materials, rather than as integrated sources of knowledge for problem-solving.

We further explored which articles were referenced together on SO and mapped the structure of scientific knowledge through a co-citation analysis. Each node in the co-citation network represents an academic article and is connected to another if they were jointly referenced by the same SO post. The resulting network was highly fragmented and sparse, with 2541 nodes (23.7% of all cited articles on SO) connected by merely 3089 edges. This limited connectivity likely stems from the highly specialized nature of many SO discussions, which often demand niche and domain-specific knowledge, resulting in less overlap among the cited articles. Confirming this hypothesis, however, requires a more detailed contextual analysis of the isolated dyads and fragments within the network, a task we reserved for future studies.

Meanwhile, our current study focused on the ten largest components in the co-citation network, comprising 381 nodes and 1358 edges, as depicted in Figure 1(b). The size of a node reflects the article's PageRank score, denoting its importance and influence within the network. Notably, we observed that pivotal and pioneering works—those that lay new foundations and advance the field—are associated with highest PageRank scores. Noteworthy examples include papers that introduced the ① *Transformer Architecture*, ② *BERT language model*, ③ *Deep Residual Network*, and ④ *Generative Adversarial Networks*. Our qualitative analysis showed that

users often cite these articles alongside others in a post to provide essential background knowledge for understanding its content. A practice that makes academic findings more accessible to a wider audience, highlighting SO's role in bridging the knowledge gap between forefront academic researchers and software practitioners.

***RQ 4**: How quickly does Stack Overflow integrate academic research?*

We investigated the pace at which academic articles were integrated into Stack Overflow communities by tracking the "First-cite Interval" — the time lag between a paper's first mention on SO and its publication date. This backward tracing approach circumvents the potential right censoring bias [3]. On average, articles took 6.6 years to be recognized on SO, although this interval varied significantly across fields. For instance, AI-related papers were typically referenced within 3.7 years, whilst PL papers have a longer latency of 9.3 years. Figure 1(c) further illustrates the trend of diffusion rates of three major research fields (see Table 1) over time. Prior to 2010, these fields exhibited similar first-cite intervals of around 6-10 years (Note that SO was launched in late 2008). However, from 2010 to 2017, the rate at which AI articles were referenced accelerated rapidly. This timeframe coincides with significant breakthroughs in AI, such as the *Adam optimizer*, *Batch normalization*, *Attention mechanism*, etc. Conversely, the First-cite Interval for Algorithms papers saw little changes, while that for PL papers even noticeably increased, implying that earlier foundational works in these areas might hold more relevance in Stack Overflow discussions.

4 DISCUSSION AND CONCLUSION

This study presents the first large-scale analysis of academic references on Stack Overflow, aiming to understand how scholarly knowledge diffuses into this practitioner-centric community. In this section, we discuss the implications of our results.

Divergent trajectories. Our analysis suggested that the research trajectories in certain fields do not always align with the practical discussions on Stack Overflow. For instance, discussions in D15 (*data communications*) referred mostly to *computer network* articles (as expected), but the number of references is low, and there's a noticeable preference for older publications. This could imply a potential disconnect from the latest developments in this field. Additionally, a significant number of documents referred in domains such as D13 (*floating point*), D14 (*computer architecture*), and D15 were protocols like IEEE Standard Specifications and IETF Request for Comments (RFCs). Although these documents have DOIs, they are typically not classified as academic literature. This could indicate that communities with greater focus on application and engineering problems find more value in industry standards and experiential knowledge over academic research, which further points to a possible divergence between cutting-edge academic findings and the practical, day-to-day challenges faced by professionals in certain technical domains.

Feasibility as an Altmetric. Academic references on Stack Overflow offer a novel lens to observe how scientific knowledge diffuses into the developer communities, suggesting their potential as an alternative metric (altmetric) for gauging the industry impact and practicality of academic research.

We observed a tangible link between the volume of SO references and an article's influence in practical settings. For example, articles that emerged as the central nodes in the SO co-citation network typically represent groundbreaking research with extensive application in the industry. Furthermore, the time interval between an article's publication and its acknowledgment on SO is comparatively shorter than that observed with other altmetrics, such as Patent [3] and Wikipedia [17] citations, suggesting that SO references may offer a more immediate measure of the influence.

However, there are challenges to consider. For instance, not all references signify an effective transfer of scholarly knowledge, as seen in D11 (*data visualization*) and D7 (*web scraping*), where academic references serve merely as contextual support. As such, to make SO a meaningful indicator of research impact in the industry, future research needs to develop rigorous methods to evaluate the intention and contribution of these academic references. Specifically, to contextually assess whether a SO reference genuinely facilitates the diffusion of knowledge and provides tangible solutions to the challenges faced by software practitioners and developers.

REFERENCES

[1] Rabe Abdalkareem, Emad Shihab, and Juergen Rilling. 2017. On code reuse from stackoverflow: An exploratory study on android apps. *Information and Software Technology* 88 (2017), 148–158.

[2] Sebastian Baltes, Christoph Treude, and Martin P. Robillard. 2022. Contextual Documentation Referencing on Stack Overflow. *IEEE Trans. Software Eng.* (2022). https://doi.org/10.1109/TSE.2020.2981898

[3] Hancheng Cao, Yujie Lu, Yuting Deng, Daniel A. McFarland, and Michael S. Bernstein. 2023. Breaking Out of the Ivory Tower: A Large-scale Analysis of Patent Citations to HCI Research. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2023). https://doi.org/10.1145/3544548.3581108

[4] Maarten R. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv* (2022). https://doi.org/10.48550/arXiv.2203.05794

[5] Run Huang and Souti Chattopadhyay. 2024. The SORef Dataset. *figshare* (2024). https://doi.org/10.6084/m9.figshare.25195805

[6] Von Ithipathachai and Maral Azizi. 2022. Are tags 'it?': analysis of the impact of tags on StackOverflow questions. *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing* (2022). https://doi.org/10.1145/3477314.3506985

[7] Heather Piwowar Jason Priem and Richard Orr. 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833

[8] Jiakun Liu, Xin Xia, David Lo, Haoxiang Zhang, Ying Zou, Ahmed E. Hassan, and Shanping Li. 2022. Broken External Links on Stack Overflow. *IEEE Transactions on Software Engineering* (2022). https://doi.org/10.1109/TSE.2021.3086494

[9] Jiakun Liu, Haoxiang Zhang, Xin Xia, David Lo, Ying Zou, Ahmed E. Hassan, and Shanping Li. 2022. An exploratory study on the repeatedly shared external links on Stack Overflow. *Empirical Software Engineering* (2022). https://doi.org/10.1007/S10664-021-10028-Y

[10] Altmetric LLC. 2024. *How does Altmetric work.* Retrieved Mar 11, 2024 from https://www.altmetric.com/about-us/our-data/how-does-it-work/

[11] Google LLC. 2024. *Google Scholar Metrics.* Retrieved Mar 11, 2024 from https://scholar.google.com/citations?view_op=metrics_intro

[12] Andreas Nishikawa-Pacher. 2022. Who are the 100 largest scientific publishers by journal count? A webscraping approach. *Journal of Documentation* (2022). https://doi.org/10.1108/JD-04-2022-0083

[13] OpenAlex.org. 2024. *Automated concept tagging for OpenAlex, an open index of scholarly articles.* Retrieved Feb 3, 2024 from https://docs.google.com/document/d/1OgXSLriHO3Ekz0OYoaoP_h0sPcuvV4EqX7VgLLblKe4

[14] Angelo Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. 2018. The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas. In *International Workshop on the Semantic Web.* https://doi.org/10.1007/978-3-030-00668-6_12

[15] Abhinandan Udupa, K N Adarsh, Anvitha Aravinda, Neelam H Godihal, and N Kayarvizhy. 2022. An Exploratory Analysis of GSDMM and BERTopic on Short Text Topic Modelling. *International Conference on Cognitive Computing and Information Processing* (2022). https://doi.org/10.1109/CCIP57447.2022.10058687

[16] Iain Xie Weissburg, Mehir Arora, Liangming Pan, and William Yang Wang. 2024. Tweets to Citations: Unveiling the Impact of Social Media Influencers on AI Research Visibility. arXiv:2401.13782

[17] Puyu Yang and Giovanni Colavizza. 2022. A Map of Science in Wikipedia. In *Companion Proceedings of the Web Conference 2022 (WWW '22).* https://doi.org/10.1145/3487553.3524925