Assessing the Contribution of Subject-matter Experts to Wikipedia

ALEX YAROVOY, YIFTACH NAGAR, EINAT MINKOV, and OFER ARAZY, University of Haifa

Attempts to explain the success of knowledge co-production communities have focused on organizational design, including structure, motivation, roles, and coordination mechanisms. Meantime, the role that subject-matter-experts play in these knowledge production settings has largely been left in a theoretical and empirical void; its existence has been assumed, but we know little about its nature and scope, as it is difficult to observe. In this article, we start filling that void, using Wikipedia as the setting for our empirical investigation. First, we carefully crossed information from individual Wikipedia editor pages with external sources such as Google Scholar to reliably identify editors who are credentialed experts. Matching these credentialed experts with their Wikipedia editing patterns, we used this dataset to train a machine learning classifier that we then employed to identify additional expert editors and assess the nature and the scope of their work across Wikipedia. Our results suggest that the scope of expert involvement is substantial, albeit with considerable differences across topics. We estimate that approximately 10%–30% of Wikipedia's contributors have substantial subject-matter expertise in the topics that they edit. We discuss implications for theory and practice of peer-production.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; Collaborative content creation;

Additional Key Words and Phrases: Wikipedia, subject-domain experts, classification

ACM Reference format:
Alex Yarovoy, Yiftach Nagar, Einat Minkov, and Ofer Arazy. 2020. Assessing the Contribution of Subject-matter Experts to Wikipedia. *ACM Trans. Soc. Comput.* 3, 4, Article 21 (October 2020), 36 pages.
https://doi.org/10.1145/3416853

1 INTRODUCTION

What is the role that subject-matter experts play in collaborative knowledge co-production systems? Extant literature that attempts to explain the success of these systems and communities has focused on organizational design, including structure, motivation, roles, and coordination mechanisms [18, 81, 96, 100, 101]. Yet, the role that subject-matter experts play in knowledge production in those settings has largely been left in a theoretical and an empirical void. In this article, we aim to address this gap in the context of Wikipedia, the free online encyclopedia.

Authors' addresses: A. Yarovoy, Y. Nagar, E. Minkov, and O. Arazy, Dep. of Information Systems, University of Haifa, 199 Abba Hushi Blvd., Mount Carmel, Haifa 3498838, Israel; emails: ayarovoy@campus.haifa.ac.il, ynagar@alum.mit.edu, {einatm, oarazy}@is.haifa.ac.il.

2469-7818/2020/10-ART21 $15.00
https://doi.org/10.1145/3416853

Since its launch in 2001, Wikipedia has become a prodigious one-stop-shop for knowledge, covering a remarkable breadth of topics. It is one of the world's most visited websites, and its articles frequently appear at the top of search results. While still occasionally critiqued as a less-than-reliable source (e.g., Reference [19]), many Wikipedia readers clearly consider it a valuable repository of knowledge, and studies suggest that the quality of articles is generally good [34, 60, 62] and that Wikipedia articles are shaping scientific discourse [106]. Wikipedia employs a unique collaborative-authoring process, whereby anyone—even those not registered as members—can contribute contents, and any revision made to an article automatically "goes live."

Given that Wikipedia's reliance on a large crowd of lay authors for peer-production stands in stark contrast to traditional models for creating encyclopedias (notably, *Britannica*) that rely on credentialed experts for producing articles [83], Wikipedia's success in becoming a credible source of knowledge has surprised many people. Even after it became clear that Wikipedia's model was working well, claims were still raised that experts do not have much incentive to contribute [61], and some critics—including one of Wikipedia's cofounders—continued to critique its "anti-expert tendencies" [94]. Naturally, the success of Wikipedia's approach, which few thought would work, drew the attention of the academic community. Numerous studies have tried to explain how quality of content is achieved in this co-production model despite the inclusive approach that does not screen for relevant expertise.

A common line of explanation argues that in Wikipedia, expertise is distilled from the "wisdom of the crowd" [103] through a sophisticated aggregation process. By this argument, "the expert is re-constructed" from a large number of contributors with relevant "local" (or "lay") knowledge [74]. Studies adopting this rationale have attributed importance to Wikipedia's socio-technical system and its mechanisms for aggregating and synthesizing a mass of divergent positions to achieve a coherent knowledge product [81, 100, 101]. Scholars have described the robust social organization that has developed alongside Wikipedia's automated processes [18, 96]. In particular, the literature explored the intricacies of Wikipedia's governance mechanisms and provided rich accounts of its self-organizing mechanisms [7, 11] and elaborate bureaucracy, emphasizing the importance of policies [1, 25, 29, 54, 68, 80], roles [13, 15, 30], and the affordances of the underlying wiki platform [29, 80, 109].

That being said, beyond effective processes, the production of any good also requires the input of high-quality ingredients. In the creation of knowledge-based products these inputs reflect knowledge and expertise. Remarkably, despite the importance of expertise for organizations involved in innovation and the development of knowledge-based products [77], particularly for self-organized teams [73], thus far the discussion regarding the role of experts in citizens-based peer-production (and in particular, Wikipedia's collaborative authoring process) has taken place in an empirical void. To date, only few have attempted to assess the nature of experts' contributions in Wikipedia, and these attempts provide only limited information. A survey of Wikipedia editors that was conducted by the Wikimedia Foundation in 2011[1] pertained to demographic aspects such as education level, but did not establish whether the editors considered themselves as subject-matter experts or whether experts contribute to their actual domains of expertise. We suspect that one of the primary reasons for the absence of empirical evidence regarding the expertise of those contributing to peer production projects has been the difficulty associated with identifying experts and their work on a large scale. As Wikipedia's editors are not required to use their true identities, or even to register, many of them do not disclose their credentials nor provide details regarding their expertise. This makes any attempt to estimate the scope of expert contributions challenging. Thus,

[1] https://en.wikipedia.org/wiki/Wikipedia:Wikipedians#Demographics.

somewhat similarly to *dark matter* of the universe, we have good reasons to believe it is there, but it is difficult to measure, and our understanding of its nature and distribution is still incomplete.

The key objective of this work, therefore, was to advance the study of expertise in peer-production by empirically studying the nature and scope of experts' work in Wikipedia's co-production activity. Beyond addressing a gap in the literature, we believe that such work may also provide insights that can inform the people who participate in, operate, and steer such platforms. For instance, we propose that the tools developed here may be helpful for locating people with relevant expertise among the crowd, allowing to direct their efforts to relevant tasks.

Attaining this objective required that we develop a reliable process for identifying subject-domain experts among Wikipedia's contributors. To that end, we developed a lightly supervised classification approach for identifying domain experts at scale. Using a sample of 1K Wikipedia articles, we began by carefully annotating positive examples of Wikipedia editors with well-defined and provable notion of expertise with respect to the topic of an article that they have edited. Namely, for the purpose of identifying the positive examples for our training set, we focused on *scholarly* expertise.[2] The pool of identified scholarly experts consisted of accounts of registered users, whose identity was known, having expertise validated by correlating information from their Wikipedia user pages and external sources. Matching these positive examples of expert editors with a set of counter examples, the goal of our learning tool was to determine whether an editor contributing to a particular article is a subject-matter expert on the article's topic.

Crucially, the learning process was agnostic to any details about the editor's identity; instead, we aimed at detecting domain experts solely based on their editorial history. We hand-crafted a diverse set of features for learning purposes, which describe the users' editorial activity in terms of intensity, relevance, and focus within the focal article, as well as across other Wikipedia articles. Consequently, the model of expertise prediction is applicable to the global population of Wikipedia editors, including those who are non-registered. We applied the model to assess and rank-order more than 300K contributors in terms of their subject matter expertise for a sample of articles distributed across Wikipedia's topical categories. Our qualitative analysis shows that the patterns that characterize scholarly experts apply to other sources of expertise, detected by our model; thus, although our model was trained on examples of scholarly experts it also detects other non-scholar subject-matter experts.

Altogether, based on our study results, we believe it is a conservative estimate that subject-matter experts make up roughly 10%–30% of Wikipedia's editor population, and that, in turn, they are responsible for a similar portion of the edits.

To the best of our knowledge, this is the first work to empirically characterize and quantify the scope of experts' involvement in Wikipedia's co-production activity. In particular, our study sheds light on several key questions, including: (a) the patterns of activity that characterize subject-matter experts; (b) the level of experts' involvement within the online community; and (c) the varying impact of experts' work across different Wikipedia articles.

2 EXPERTS AND EXPERTISE IN KNOWLEDGE PRODUCTION

There are myriad definitions of experts and expertise in the literature [49, 76]. While, naturally, not all definitions include all aspects, it is generally agreed that an *expert* is someone with extensive knowledge and honed skills who reliably demonstrates superior abilities and performance in a certain domain and who is recognized as such by her peers and/or the public [49, 76]. For the purpose of this article, where not explicitly stated otherwise, we shall accordingly use the term

[2]The focus of our empirical investigation on scholarly experts is in line with the approach taken in the survey of experts by the Wikimedia Foundation.

"expertise" to denote that individual capacity,[3] and the term "subject matter expert" (or SME) to denote an individual who has expertise with respect to a subject domain.

Subject-matter expertise is normally attained through a prolonged, extensive, and deliberate learning process. This continuum is often portrayed as having multiple levels that people can gradually ascend. A three-level division (*Novice* or *Non-expert*; *Semi-expert*; and *Expert*) is most common (e.g., References [51, 63, 87]), though other models have been proposed [47]. People with various degrees of expertise in a certain domain may differ in many respects, including their mental models, assessments of situations, repertoires of routines and solutions, and—central to our context of knowledge production and curation—the breadth and depth of their knowledge [43, 92]. In this article, we shall primarily be discussing subject-matter experts, focusing almost exclusively on the *depth and breadth of their knowledge of specific topics* as the main characteristic that sets them apart from less-knowledgeable people. It is worth noting that since expertise is domain-specific, one might be a novice in a certain domain, a subject-matter expert in another domain, and perhaps a semi-expert in a neighboring domain—all at the same time.

In some domains, formal recognition of a person's expertise is customary, in the form of diplomas (as in academia), titles, professional licenses and accreditations, and so on. Yet in other domains a person's expertise may be socially acknowledged without any official or formal recognition [85]. Thus, while many subject-matter experts are credentialed academics, many other subject-matter experts are people who have acquired their expertise through practice, hobby, or otherwise.

2.1 Experts and Expertise in Online Co-production Communities

In traditional organizations, people are usually appointed to roles that carry decision-making authority and titles that reflect recognition and set expectations of their expertise. This is true not only in hierarchical organizations, but also in other organizational forms, such as matrix organizations [57]. From an information processing perspective, there is an advantage to assigning tasks to members of the organization who are skilled to perform them, and a great deal of the discussion of organizational design considers people's levels and domains of expertise (e.g., References [57, 59]). Traditional organizations therefore dedicate substantial efforts and resources to find, recruit, and retain experts.

However, throughout thepast decades, rapid advances in computation (notably, the introduction of the personal computer and later the world-wide-web) led to dramatic declines in the cost of information production and communication, thereby enabling unprecedentedly large numbers of people the means to collaborate in knowledge production, curation, and dissemination. In the early 21st century, new patterns of organizing emerged, such as commons-based peer-production, crowdsourcing, and social computing, affording alternative avenues for knowledge production and innovation, and challenging existing organizational models [22, 75, 84, 104]. For instance, whereas before corporate innovation was almost exclusively done within companies and laboratories, now they are aided by crowds who participate in innovation contests; and whereas before a large encyclopedia such as the *Encyclopedia Britannica* was compiled by paid subject domain experts through organized, scripted, hierarchical processes, now Wikipedia is run by a "ragtag band of volunteers" [110]. Books and articles (e.g., References [73, 97, 102–104]) celebrated these new organizational forms, highlighting the way in which they eschew the reliance on expert employees in favor of the new king: the crowd. Prominent among the examples celebrated in many of these essays was Wikipedia.

[3]Expertise can also emerge from interaction in groups and organizations [24, 107], but our focus here is on the contribution of individual experts.

2.2 What Do We Currently Know about Expert Involvement in Wikipedia?

As of today, Wikipedia is one of the most prominent examples of communal co-production of knowledge. It is the world's largest and most-viewed encyclopedia, offering unparalleled breadth. As of January 2019, there were instances of Wikipedia in nearly 300 languages, offering a total of 49.4M articles.[4] The English language Wikipedia is the largest project, with more than 5.7M articles.[5] Although it is generally not accepted as a primary, conclusive source of credible information in formal institutions (e.g., academia or in court), it is widely used as a reference, and several studies have found its articles to be of good quality.[6]

Since its inception as a side project that was aimed to feed its faltering predecessor—the expert-written, expert-reviewed Nupedia—Wikipedia has adopted an approach that has been described as "*anti-credentialist.*" Wikipedia's articles are open for all to edit, regardless of their credentials or level of expertise, and the voices or opinions of credentialed subject-matter experts do not get any extra weight in decision-making processes. As cofounder Jimmy Wales told *The Guardian*: "*To me, the key thing is getting it right, and if a person's really smart, and they're doing fantastic work, I don't care if they're a high-school kid or a Harvard professor*" [28]. Not everyone agreed with that approach, though. Notably, the other cofounder, Larry Sanger (who left the project in 2002) expressed concerns about Wikipedia's perceived credibility and a "*crying need to get more experts on board and a publicly credible review process in place*" [93]. Within the Wikipedia community itself, the discussion about expert involvement has often focused on credentialed experts. For instance, under the section "What is an expert editor?" the Wikipedia Essay Wikipedia:Expert retention[7] suggests that *"an expert editor is a user with an advanced degree, such as a PhD, a professional degree, such as a Juris Doctor or equivalent professional expertise (e.g., a widely published novelist) who is contributing to Wikipedia in his or her field of expertise. Some editors may consider graduate students who are working on doctoral degrees to be functioning at a high level of expertise, though lower than a Professor with a PhD."* This debate about the way Wikipedia should treat expert credentials has continued ever since, within the Wikipedia community, as well as among scholars and critics [32, 97].

Though, as we explained, expertise and credentials are not one and the same, it appears that to Sanger (and others), they were. One reason for that might be that although there are many true experts who do not have diplomas or similar kinds of formal credentials, credentials—so long as they are genuine—are a helpful and sometimes critical means of identifying true experts. While in some social settings such as long-standing organizations and small communities-of-practice where people know each other for a long time, people are often able to recognize one's expertise, this naturally becomes a challenge in a large online community such as Wikipedia, with loose membership, where most people do not know each other, and where a lot of interactions are transactional. In such a setting, verifiable credentials can become an important signal of expertise.

The anti-credentialist stance of Wikipedia does not stem from a lack of appreciation of the value that experts can contribute to Wikipedia. It is almost self-explanatory why contributions by experts can be very valuable. But as the failures of Nupedia and Citizendium have demonstrated, relying *solely* on experts inevitably leads to a bottleneck and to a slow publishing process. However, Wikipedia's anti-credentialism also led to issues and challenges, including article quality and public perception. Importantly, dealing with a large crowd of non-experts, at times including trolls, proved challenging to many expert editors [93], and some of them have vocally left. This has

[4]Source: Wikipedia, https://en.wikipedia.org/wiki/List of Wikipedias.
[5]Source: Wikimedia Foundation, https://stats.wikimedia.org.
[6]For a comprehensive review of relevant studies, see https://en.wikipedia.org/wiki/Reliability of Wikipedia.
[7]https://en.wikipedia.org/wiki/Wikipedia:Expert_retention.

resulted in a backlash within the community, as reflected in several highly viewed essays that called for reflection and changes in Wikipedia's policies.[8] For example, one of the essays raised the question, "*How can we attract and retain experts in Wikipedia? If Albert Einstein were alive today, would he want to contribute?*."[9] Some attempts to change were made—specifically, Wales himself tried to push for a "credential policy" that may give more consideration to the opinions of credentialed experts, but it was not accepted, and for now, Wikipedia's anti-credentialist stance remains.

Although Wikipedia decided not to grant any special authority to editors based on any external credentials, recognition of the potential value of experts' involvement and other strategic concerns have led the community, as well as Wikimedia foundation, to seek ways to better the relationship between Wikipedia and various expert populations [39]. Two notable initiatives are *GLAM* and the *Education Program*. The GLAM (galleries, libraries, archives, and museums) outreach[10] assigns experienced Wikipedia editors to work with local representatives at cultural institutions in an attempt to bring certified knowledge to Wikipedia articles.[11] The program started in 2010 and established cooperation with institutions such as the *British Museum in London*, the *Smithsonian Archives of American Art*, *Museu Picasso in Barcelona*, and more. In the first couple of years, the GLAM program contributed to over 2K articles in over 50 languages. The Wikipedia Education Program invites "*educators and students around the world [to] contribute to Wikipedia and other Wikimedia projects in an academic setting*." Over four years, more than 10K students have participated in the program, contributing to more than 10K Wikipedia articles in multiple languages.[12] Interestingly, calls to involve experts also originated outside of Wikipedia, and papers published in scientific journals discuss the potential merits that scientists can gain from using and contributing to Wikipedia [52, 58, 71]. While initiatives such as GLAM tell us something about experts' involvement in editing Wikipedia (e.g., when it is performed through these initiatives), there are many expert contributions that are done independently.

To date, only few have attempted to estimate the extent of experts' contributions to Wikipedia. Some rough aggregate survey data appears in Wikipedia.[13] Specifically, a survey of Wikipedia editors that was conducted by the Wikimedia Foundation in 2011 and focused on scholars in academe, estimated that 61% of participants were graduates of degree-level education: 35% with undergraduate degrees, 18% with a master's degree, and another 8% with doctorates [55]. While these figures suggest that Wikipedians are by-and-large educated, undergraduates (or even graduate) students are not usually considered as subject-matter experts [24, 36, 49–51]. Around the same period, the Wikimedia Foundation conducted a study of experts' involvement in Wikipedia, focusing on scholars in academe [105]. Beyond serving as indication of Wikipedia involving experts, this survey has revealed some interesting insights. Notably, it found that scholars contributing to Wikipedia hold various positions: PhD candidates, to postdocs, non-tenured faculty, tenured faculty, researchers in industry, and other professional experts (the former and latter contribute the most, possibly because they have more time on their hands). Further, findings from the survey indicate that many experts contribute primarily in their area of expertise (48%), yet many contribute outside their domain (35%); and some contribute both in and out of their area of expertise (17%) (interestingly, the majority of survey respondents recognized that Wikipedia policies discourage writing about their own research, i.e., it is considered self-promotion). Notwithstanding the value of such surveys, much is still unknown about the actual impact of experts on Wikipedia. These data do not

[8]See https://en.wikipedia.org/wiki/Category:Wikipedia_essays_on_experts_and_expertise.
[9]https://en.wikipedia.org/wiki/Wikipedia:Expert_retention.
[10]https://en.wikipedia.org/wiki/Wikipedia:GLAM.
[11]https://en.wikipedia.org/wiki/Wikipedia:GLAM.
[12]https://outreach.wikimedia.org/wiki/Education/About.
[13]https://en.wikipedia.org/wiki/Wikipedia:Wikipedians#Demographics.

tell us what are the levels of involvement of those participants, nor do they tell us whether these participants' contributions were done in their respective areas of expertise.

How, then, can insights about experts' involvement in Wikipedia be gained? As of today, Wikipedia does not require editors to identify themselves or even to register, and its editors are often anonymous. Growing to recognize the value of experts' contributions, Wikipedia now provides a standard linked-data mechanism for experts to manually identify themselves and establish their credentials (e.g., link to an academic institution) using the ORCID identifier.[14] However, the adoption of this mechanism has been rather slow, hampering any investigation of experts' involvement and the scope of their contribution.

In summary, we know that Wikipedia is interested in experts' contributions, although it is not willing to privilege any formal mark of expertise nor to grant recognized experts any special roles within the community. We also know that many editors edit anonymously—either without registering at all, without logging in, or using nicknames—which poses a challenge for any attempt of quantitative empirical inquiry into the scope and nature of expert work in Wikipedia. It also makes it challenging for Wikipedians to recognize expert editors whose help may be needed in various tasks. An automated tool that can help identify experts and their work may therefore be desired and useful, for both research and practice.

3 PREVIOUS EFFORTS ON DETECTING EXPERTS' CONTRIBUTION TO PEER-PRODUCTION

Given our working definition of experts as people with extensive knowledge and honed skills, it follows that an automatic procedure for expert detection should seek evidence for knowledgeability in the contributions one makes to the shared repository.

Relevant to our work, the topic of finding experts in online communities has been an active research area. Online communities are often used for advice or help seeking [99], and it is therefore important for members of the community (or guests) to be able to find and identify experts within the community who may possess relevant knowledge for addressing their issues or queries [2]. In some settings, this task has been referred to as "competency management," seeking to identify "who knows what" [116].

Most works on expert-finding describe the members of the online community using some content-based model, mapping users' past contributions onto lexical or topic-based space; the relevance of users to a focal question is then assessed in terms of their content similarity in that space [27, 31, 42, 69, 115, 116]. Several studies further relied on interaction patterns and social network information to gauge the level of knowledgeability or authority of users [3, 17, 26, 31, 79, 115]. For example, these works assume that community members who tend to answer rather than ask questions have a higher level of knowledge than their peers, or consider available feedback scores as proxy to the quality of past answers, e.g., Reference [26]. Pal et al. [86] further claimed that when questions are marked as "poorly" or "partly" answered, this attracts the attention of expert users. Such explicit indications available in certain online question-answering communities are not used by the Wikipedia co-production community, however.

To the best of our knowledge, there has been little work on identifying subject-matter experts in online co-production communities such as Wikipedia. Demartini et al. [46] proposed to create a profile for each Wikipedia user based on the articles they edited in the past, assuming that these pages reflect the user's knowledge areas. They further assumed that editors of highly linked articles in Wikipedia were experts in the respective knowledge domain. No empirical evaluation was included in their work, due to the lack of community judgements or other indicators of expertise

[14]See https://orcid.org/, https://en.wikipedia.org/wiki/Wikipedia:ORCID.

within the Wikipedia community. We believe our study to be first to provide empirical evidence regarding experts' involvement and its characteristics in Wikipedia.

Few works have attempted to estimate the quality (or trustworthiness) of particular revisions of Wikipedia articles, for example the Wikimedia Foundation's ORES (Objective Revision Evaluation Service) project,[15] which uses automatic machine learning methods and is trained on Wikipedia's community-based article scoring procedure [45, 95].[16] Analyzing the quality difference of consecutive revisions allows deducing the quality added/lost in a particular edit, and aggregating these scores per editor may provide a proxy for the overall quality of their contributions. However, improved quality may result from added clarity and better articulation, which do not necessarily require domain-specific knowledge. Some other works studied an edit's "persistence" (or longevity, i.e., how long the content remains in the article until removed or overwritten through the community's ongoing refactoring process) and derived an editor's aggregate "reputation" score [4, 5, 64]. Whereas such approaches measure indicators that may possibly correlate with expertise, they do not gauge domain expertise directly. Since these elements of content relevance and contribution persistence could in theory be positively correlated with domain expertise, we have included them in our modeling, among many other types of evidence.

We are not aware of any work that studied the activities of experts in Wikipedia, nor in similar citizen-based co-production initiatives. There is an active area of research studying contributors' trajectories of participation [7, 10, 70, 114], often linking activity profiles to personal characteristics, (e.g., motivation [20]), or to functional roles within the Wikipedia community [15]. Particularly relevant to our investigation are prior studies that distinguished between content-oriented and community-oriented contributors [14, 16], suggesting that those focusing on content co-production are on average less active and tend to focus their activity in few pages. These works suggest that experts are likely to concentrate on content-related activities. In this work, we include and put to test similar information in modeling and characterizing the activity of subject-matter experts.

4 RESEARCH QUESTIONS

To date, there is a dearth of empirical research on expertise in Wikipedia (and more broadly, in knowledge co-production). Our review has highlighted some noticeable gaps in the literature. The goal of this study is to fill in these gaps and characterize the experts that contribute to Wikipedia, as well as to assess the scope of their contribution.

Our study aims to address the following research questions:

- **RQ1** *What are the characteristics of SMEs' work that distinguish it from other editorial work in Wikipedia?* In other words, *what activities do SMEs do? And how do these differ from the activities of non-experts?*
- **RQ2** *What portion of Wikipedia's editors are SMEs (with respect to the articles edited)? What is the scope of SME work in Wikipedia?*
- **RQ3** *To what extent does SMEs' involvement vary across Wikipedia articles?*

To address these research questions, which involve the assessment of SMEs' contributions to Wikipedia at scale, we devised an automatic approach that reliably identifies SMEs in Wikipedia. Formally, our goal was to automatically determine, with a good degree of confidence, whether a given editor e, who has edited Wikipedia article a, was an expert with respect to that article's subject domain. This task is challenging, since little is known about the editors: The majority do

[15] The ORES tool; see https://meta.wikimedia.org/wiki/Objective_Revision_Evaluation_Service.
[16] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wikipedia/Assessment.

not register within the community (following previous work, we shall identify the non-registered users by their device's IP address [14]); among those who are registered, only a small portion disclose personal details, such as their name or self-description. And, rarely do editors explicitly mention subject-matter expertise—possibly a reflection of the community's stance on the issue. Even if there exists an indication of subject-matter expertise (e.g., on one's personal *user page* on Wikipedia),[17] ideally this self-reported evidence should be verified through external sources.[18]

Consequently, a comprehensive approach for identifying subject-matter expertise must rely on *implicit* markers of editorial activity, as opposed to *explicit* information about the editor. We address this aim using a lightly supervised learning framework, based on the following key components:

i. We observe that a subset of editor-article pairs {<*e,a*>} may in fact be attributed with consensual and provable notion of subject-matter expertise, namely, *scholarly expertise*. To that end, we identified and manually annotated examples of editor-article pairs, for which there exists strong evidence of editor *e* being an SME with respect to an article *a* that she has edited in the form of relevant academic credentials. Only a small minority of editor-article pairs could be verified as scholar SMEs, however; to overcome this problem of searching for a needle in a haystack, we devised a semi-automatic annotation procedure that first scans a large volume of editor-article pairs, tracking possible indicators of scholarly expertise within the editors' Wikipedia pages, as well as outside of Wikipedia. We then focused our annotation effort accordingly on candidate editor-article pairs that were likely to include relevant evidence of scholarly expertise.
ii. Given the labeled SME and counter examples, we apply learning to predict subject-matter expertise for unlabeled editor-article pairs. While explicit personal information must be used for manually determining and validating expertise, the automatic expertise prediction module only considers *implicit* statistical descriptors of editorial activity as available evidence. This makes the learning model applicable to all kinds of editors, including a large population of anonymous non-registered editors, for whom no personal information is available. To the extent that SMEs share some characteristics of editorial activity, regardless of their credentials, we expect, and indeed show, that a model of expertise prediction trained with respect to scholars applies also to domain experts who are not credentialed.

5 ANNOTATION OF SUBJECT-MATTER EXPERTS

Our first challenge was the detection and annotation of SME editor-article pairs. In this section, we describe in detail the data sampling and annotation procedures designed for this study and the resulting labeled dataset.

5.1 A Wikipedia Article Sample

We refer in this study to a corpus of 1K Wikipedia articles composed of a balanced mixture of articles across different topical categories and article maturity levels, having been constructed using a double-stratified sampling procedure [7]. Importantly for our purposes, the articles in this collection have been analyzed with respect to the types of *production* activities (i.e., co-authoring of encyclopedic entries on Wikipedia's Main namespace) performed by editors throughout the articles' history. Concretely, each *revision* (defined as the difference between two consecutive versions of an article) of every article in the corpus, starting from the article's creation date, was tagged with the relevant activities, where a single revision may contain multiple types of edits.

[17] https://en.wikipedia.org/wiki/Wikipedia:User_pages.
[18] For a controversy regarding a Wikipedian's falsified credentials, please see https://en.wikipedia.org/wiki/Essjay.

The underlying taxonomy of wiki-work includes well-defined categories, such as: *add substantive new content*, *rephrase existing text*, *add references to external sources*, *remove vandalism*, *fix typos*, and more. The tagging procedure involved a combination of manual annotation and a machine learning approach. We refer the reader to Reference [7] for additional details regarding this sample of articles and how it was processed.[19]

5.2 Annotation Guidelines

While there exist various types and levels of expertise, we focused our labeling efforts on a well-defined notion of expertise, namely, *scholarly expertise*. Having identified editor-article pairs for which we detected evidence of in-domain scholarly expertise, such pairs were annotated as positive examples of SMEs. Concretely, we labeled as SMEs editors with PhD-level education in a knowledge area relevant to the focal article. To determine whether this condition holds, the annotation process relied on all evidence available, both within and outside Wikipedia, including: indications of expertise on the editor's personal page; being able to discern an editor's first and last name so her identity could be tracked and expertise validated in external Web resources; and, verification that the editor's area of expertise matched the topical domain of the particular article in question.[20]

Table 1 illustrates the annotation guidelines through several examples. The table includes the descriptions provided by editors on their personal Wikipedia pages. As shown, the editor in (a) self-attests to be a biological oceanographer. As the full name of the editor was disclosed in this case, we were able to find that editor's researcher profile on *Google Scholar*[21] and validate this information. Considering the relevance of the editor's area of expertise to the focal article ("Algae"), this instance is to be labeled as SME. Example (b) belongs to a Professor of Computer Science. Also in this case, the editor's real name was disclosed, and we were able to track their public researcher profile. Since the subject domain of the editor's expertise and the focal article edited ("king's graph") match, this example was labeled as SME. As mentioned on the personal page of editor (b), that editor also edited articles on other topics, such as *California Geography*, and others; notice that if the focal page concerned a geography concept, this particular editor would not have been labeled as SME, due to mismatch between the focal topic and formal expertise. Example (c) is yet another example of editor-article pair labeled as SME; in this case, the editor added structured infoboxes to his Wikipedia personal page indicating that they had a PhD degree, as well as included a link to their researcher profile.

Examples (d)–(f) in Table 1 include indications of some level of expertise mentioned on the user personal page, where the editors describe themselves as an academic staff, a software engineer, and a hobbyist of Genealogy, respectively. Since relevant formal education could not be verified based on the information on the editors' personal pages, nor based on external sources, neither of these editor-article pairs were labeled as SMEs for our purposes (despite some topical relevance between the editor's self-description and the focal article in example (d)). Finally, examples (i)–(j) lack any indication of domain expertise on the editors' pages or elsewhere and were not labeled as SMEs.

5.3 Example Sampling: Overcoming the Needle in a Haystack Problem

The annotation of scholarly expertise according to the meticulous guidelines outlined above involved manual search and verification of relevant evidence within Wikipedia and the Web, and was therefore costly and limited in scope. Crucially, the proportion of editor-article pairs for which

[19]The original corpus extracted and annotated dates to the January 2012 dump of the English Wikipedia; see https://dumps.wikimedia.org/. As part of our prediction experiments reported in Section 7.2, we extended this analysis until July 1, 2018.
[20]We emphasize that the annotators have not examined the editor's activity on the focal page or elsewhere on Wikipedia for identifying expertise.
[21]https://scholar.google.co.il/.

Table 1. Annotation Guidelines, Illustrated by Example Editor-article Pairs

Strong evidence of scholarly expertise:	(a)	*"Algae"*–"I am a biological oceanographer working the Station Biologique in Roscoff, France, a marine biology laboratory located off the coast of northern Brittany. The Station Biologique is affiliated with CNRS and the Universite Pierre et Marie Curie in Paris. My research focuses on picoplankton, the smallest creatures of marine plankton with size below 2 μm. ... Publications: Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga, Bolidomonas: a new genus with two species belonging to a new algal class, the Bolidophyceae (Heterokonta), and more."
	(b)	*"King's graph"*–"I'm a computer science professor at NA. Much of my Wikipedia editing is on mathematics articles, but I've also edited articles on computer science, academic biography, the arts, and California geography. I've also contributed many diagrams and photographs to Wikipedia and Wikimedia Commons. As an employee of a public university I believe that public outreach is part of my job description, and in that sense that my edits here to subjects within my professional expertise are paid edits. However, the topics and content of my editing here are wide-ranging and entirely self-directed. I neither participate in, nor condone, paid edits for specific articles or content."
	(c)	*"Arthur Eddington"*–"XXX is known in the real world as XXX, who lives in XXX. I've been making small edits to Wikipedia since November 2007." User box: "MPhys This user has a Master of Physics degree."; "PhD This user has a Doctor of Philosophy degree." http://www.astro.cardiff.ac.uk/contactsandpeople/XXX Image: Dr. XXX https://scholar.google.co.il/XXX Image: scholar, postdoc
Insufficient evidence of scholarly expertise:	(d)	*"Ayn Rand"*–"I worked for Indiana University at Indiana University Purdue University Indianapolis for several years, where I earned degrees in philosophy and political science."
	(e)	*"Same-sex marriage in Iowa"*–"I am a software engineer in the Seattle area. I have a master's degree in computer science from the University of Wisconsin-Madison and a bachelor's degree from Iowa State University in computer engineering. My personal interests include computers, curling, traveling, bridge, board games, science fiction, hiking, politics, investing, and music."
	(f)	*"Contraflow lane reversal"*–"I've spent hundreds of hours doing online genealogical research. As much as I've learned about my ancestors (over 500 of them and almost 2000 relatives!), I've also learned a tremendous amount of US and European history and may make an occasional edit on a subject relating to my research."
No indication of expertise:	(g)	*"Contraflow lane reversal"*–"I am a Wikipedia:administrator since July 5, 2004. I first contributed to Wikipedia in April of 2001 (oldest edits lost in early software upgrades). My first edit with my username appears to be on April 26. I have been active on a number of online projects—active and passive (SETIHome, Distributed Proofreaders, Mars crater labeling, Star- dustHome, FoldingHome on the Wikipedia team: Team page, etc.)"
	(h)	*"Gastrointestinal tract"*–"Hi, I'm XXX, I'm a Geek and Wikipedia administrator living in Cambridge."

The table quotes the editors' Wikipedia pages: some pages (examples a–c) include evidence of relevant scholarly expertise as defined in this work, namely, formal PhD-level qualification; some other editors may be domain experts, but do not qualify as scholarly SMEs according to the descriptions on their pages (examples d–f); and some pages do not indicate expertise (examples g–h). The process of annotating editors as scholarly SMEs relied on their page descriptions and complementary Web searches. In addition, it required a match between the editor's area of expertise and the topic of the focal page.

sufficient evidence may be found is small; annotating a random sample of editor-article pairs would therefore yield a poor number of provable positive examples. To overcome this problem of "finding a needle in a haystack," we devised a heuristic rule-based scoring strategy, screening a large volume of editor-article pairs for their potential as provable SMEs *before* proceeding to the manual analysis.

In brief, the automatic scoring procedure aims to identify relevant evidence that might be useful for the human annotators. It scans two sources of potential evidence: the personal Wikipedia page of the editor in question (if exists), and indications of the editor's publication history on the Web. We devised a set of scoring rules, searching for the presence of concrete terms on the editor's personal Wikipedia page that we believed were highly indicative of scholarly expertise, e.g., "PhD" and "doctorate," as well as more general and possibly ambiguous terms, such as "academic" and "expert." We assigned a weight to each term to reflect their presumed predictive power, respectively. The scoring heuristic computes the total weight of the terms found on the editor's personal page, as pages that contain multiple expertise-related terms are more likely to indeed include self-attested indication of scholarly expertise. To track relevant scholarly publication history, we devised an automatic procedure that aims to match an editor with her researcher profile on Google Scholar and assess the topical relevance of her publications with respect to the Wikipedia article that she has edited.

We applied the automatic scoring procedure to all of the distinct editor-article pairs that comprise our reference corpus (Section 5.1). Based on the outcomes of this procedure, we were able to identify candidate SME editor-article pairs and prioritize them for further manual investigation and verification. Indeed, high ratios (as high as 69%) of verifiable SME were found among editor-article pairs that have been highly scored by our heuristic. To complement and diversify the set of annotated examples, editor-article pairs for which partial or weaker indications of expertise have been tracked by the heuristic were sampled and annotated.

Complete details about the scoring heuristic, its results, and the sampling of candidate pairs for manual annotation, are provided in the Appendix.

5.4 The Labeled Dataset

The manual annotation of candidate editor-article pairs was performed by a member of our research team and two paid undergraduate students. Following a training phase, which included a discussion of points of disagreement per a set of co-annotated examples, the remaining examples were qualified by each rater independently. Inter-annotator agreement among the three annotators, measured on a set of 100 editor-article pairs, was estimated at Fleiss agreement rate [53] of 0.77, which is considered to be "substantial" [53] or "excellent" [37]. Once we ensured inter-rater agreement was good, the annotators continued to manually assess the remaining candidate SME examples.

Overall, 506 distinct editor-article pairs were manually validated as SMEs. Table 2 details the distribution of the positively labeled pairs across articles, and across Wikipedia's topical categories. As shown, the positive examples were drawn from the histories of 211 distinct articles in the reference corpus.

To distinguish SME from non-SME editor-article pairs, we aimed to complement our dataset with *counter* examples. Crucially, while the annotation of positive examples involved the identification of concrete and definite evidence of domain expertise, it was impossible to ascertain that an editor was *not* an SME with respect to an article that they have edited. Mainly, an editor may choose not to publicly disclose, or conceal, relevant information about their qualifications or background—on their personal page or elsewhere. Manual annotation of non-SME examples would have therefore been costly and ineffective, and would have likely yielded low agreement rates.

Instead, we collected counter examples in the following fashion: For each article that contained positively labeled examples, we sampled a similar number of editors, who have not been labeled as SMEs, considering these editor-article pairs as counter examples. To diversify our sample, we distributed the selection of counter examples according to their assessment by our automatic scor-

Table 2. Dataset Statistics: The Number of Distinct Editor-article Pairs Labeled as SMEs and Counter Examples, and the Number of the Respective Source Articles, per Topical Category

Topical category	Articles	SME examples	Counter examples
Agriculture	10	19	31
Arts	4	5	7
Business	4	5	10
Chronology	2	5	8
Concepts	11	31	35
Culture	6	12	14
Education	10	20	16
Environment	10	17	27
Geography	8	10	13
Health	15	56	43
History	5	12	14
Humanities	6	9	7
Humans	7	21	19
Language	9	26	29
Law	7	9	12
Life	7	15	15
Mathematics	24	80	46
Medicine	21	42	34
Nature	14	42	47
People	3	5	5
Politics	5	12	17
Science	14	43	53
Society	6	7	10
Technology	3	3	2
Total	211	506	514

ing procedure. Thus, our samples include Wikipedia editors who own personal user pages, with no formal expertise words mentioned in them (as in examples (e)–(g), Table 1); editors for whom low topical similarity was found between their publication profile and the focal article; and editors with insufficient evidence altogether of their background or interests. Concrete details on the sampling of the counter examples are provided in the Appendix.

As shown in Table 2, the resulting dataset is balanced, containing a similar number of positive and counter examples across articles and topical categories. (It is possible that this dataset over-represents the SME examples compared with their true prevalence in Wikipedia, or vice versa—indeed, the true proportions of SMEs are unknown.) We follow the common practice of constructing a balanced dataset, where this setup has been shown to support effective learning (e.g., References [89, 91]).

We emphasize that compared with the meticulously annotated SME examples, the counter examples reflect lower level of certainty. Mainly, it is possible that the counter examples include SME editor-article pairs, which merely lack indication of relevant editor expertise. Yet, while mixed and noisy in this sense,[22] the overall distribution of the counter examples is expected to exhibit

[22]Because of possible label noise, and to avoid confusion, we use the term "counter examples," rather than "negative" examples.

substantially lower levels of SME editor-article pairs compared with the distribution of the positively labeled examples.[23]

Indeed, we later show that the two populations—the SME-labeled and the counter examples—differ significantly on multiple dimensions of editorial activity, thus qualifying as valuable data for learning purposes.

6 AUTOMATIC IDENTIFICATION OF SUBJECT-MATTER EXPERTS

We now turn to examine the learning task that is central to this research. Given an editor *e* who has made edits to Wikipedia article *a*, we wish an automatic agent to determine whether *e* is subject-matter expert with respect to *a*. Importantly, the automatic expertise identification process is denied access to explicit information about the editor, either within or outside Wikipedia. While we used such information for detecting and manually verifying SME, a large mass of editors do not maintain a personal page on Wikipedia, nor disclose their true identity. Therefore, a generalizable learning model must rely on implicit indicators of editorial behavior. In what follows, we describe a set of features designed to capture the editorial activity patterns of editor *e* within the focal article *a*, as well as across Wikipedia. We then assess learning performance using the labeled dataset, measuring the extent to which the labeled SME editor-article pairs are identified as such by several learning algorithms. In addition, we assess the classifier's performance on a separate, unlabeled dataset, observing its capability to identify experts "in the wild."

6.1 Features Describing Editorial Activity

The feature types that are proposed and examined in this work are summarized in Table 3. Hereby, we describe and motivate these features in detail.

6.1.1 Focal Page Activity. We were first interested in encoding relevant aspects of editorial activity performed by editor *e* on the focal article *a* throughout the article's history. A set of features were designed to capture the *intensity* of that activity; possibly, large and repeated contributions indicate on expertise in the subject-domain. We further considered the *nature of co-authorship activity* on the focal article. Previous work by Arazy et al. [7] employed a taxonomy of wiki-work, including the categories of: *adding references to external sources*, e.g., scientific citations; *adding substantive new content*; *adding or modifying Wiki Markups*; *moving or creating a new article*; *deleting substantive content*; *fixing typos and grammatical errors*; *reorganizing existing text*; *rephrasing existing text*; *inserting vandalism*; *deleting vandalism*; *adding of hyperlinks* (to other Wikipedia pages); or, *miscellaneous* activities. We wished to explore whether the distribution of the different activity types performed by editors are indicative of domain expertise. Having the revisions in the reference corpus assigned with these wiki-work categories (see Section 5.1), we encoded the proportion of each of the activity types (12 in total) for each editor-article pair. In addition, for the same purpose, we measured the count and ratio of *content words* (as opposed to functional and punctuation words) contributed by the editor as part of their co-authorship activity on the focal article.[24]

The proposed features further assess the *topical relevance* of the editor's contributions to the article—how central are the editor's contributions to the article's main theme? Topical relevance was modeled based on the explicit semantic analysis (ESA) semantic similarity measure [48, 56],[25] gauging and summing over the semantic overlap score between each of the text fragments added by the editor throughout the article's history and the focal article's theme; specifically, we

[23] Learning in the presence of labeling noise is common practice in some settings, e.g., References [44, 90, 108].

[24] We used Scikit-Learn to distinguish between content and non-content words [82].

[25] The ESA measure maps each term to Wikipedia's articles in which it is pronounced; similarity between texts is then computed using cosine similarity in this so-called explicit representation space.

Table 3. Description of the Feature Types Used to Model the Editor's Activity on a Focal Article, as well as across Wikipedia

Category	Feature type	Reg.
	Activity on the focal article:	
Intensity of activity in page	*total page edits:* total number of edits made by editor e in the focal article a throughout its history.	
	page edits ratio: the ratio of edits by e out of all edits made on article a.	
	mean edit interval: the frequency in which e performed edits to article a.	
	mean edit size: the mean length of the modified text in a single edit, measured in tokens and in bytes.	
	edits period: the ratio of edits by editor e in each quarter of the life span of article a.	
Nature of co-authorship activity	*edit types:* the distribution (proportions) of editor activities mapped to wikiwork categories defined by Reference [7].	
	content token count: the total number of content words contributed by editor e to article a throughout its history.	
	content token ratio: the ratio between content and non-content words ("stop words," punctuations) contributed by editor e to article a throughout its history.	
Topical relevance	*contributions' relevance:* measured as the summation over the ESA similarity scores between each of the textual fragments added by e to article a throughout its history and the (current) article's summary.	
Longevity of contributions	*persistence:* average persistence period per token added by editor e to article a, i.e., the length of period until being modified by other editors.	
Coordination activity	*edits on "talk page":* total number of edits made by e on the talk page of article a.	√
	Cross-Wikipedia activity:	
Topical focus	*structural similarity:* topical coherence among other articles edited by e and the focal article a measured in terms of common hyperlinks and common Wikipedia categories.	
	content similarity: topical coherence among other articles edited by e and focal article a measured in terms of ESA similarity, considering the articles' titles or summaries.	
Production activity (main namespace)	*distribution:* the entropy of the edits by e across all articles that they contributed to.	
Activity in Community	*tenure:* total length of activity of editor e in the Wikipedia community since their registration date, and the total number of pages they edited in all of Wikipedia's namespaces.	√
	activity focus: the ratio between the total number of edits by e in Wikipedia's main namespace and the total number of edits across all namespaces.	
	distribution of activity: the proportion of the number of edits by e in each of the non-main namespaces of Wikipedia and the total number of edits by e across all namespaces	√

As indicated, some of the proposed feature types are limited to registered editors (Reg).

considered the summary paragraph of the article, which typically describes the focal concept at high level.[26]

Presumably, the content contributed by subject-matter experts would be valued by the community. Following previous work [64], we assessed the *persistence* of individual contributions in terms of the time duration of being included in the article before being removed or modified in subsequent revisions. Having all of the tokens modified by the editor annotated with their persistence in seconds,[27] we modeled the average token-level persistence of the editor's contributions to the focal page.

Finally, editors who are registered Wikipedia users may contribute to the "talk page" associated with the article, which is used for inter-editor discussions.[28] For those editors, we modeled their involvement in this *coordination activity* as the total number of edits made by them on the talk page; possibly, expert editors differ from non-expert contributors on this dimension.

6.1.2 Other Activity in Wikipedia. In addition to the editor's history on the focal page, we considered her past edits across all of Wikipedia's articles, aiming to capture a broader view of the editor's interests and activity patterns. Presumably, SMEs would tend to edit articles related to their area of expertise. To capture such *topical focus*, we identified up to five articles besides the focal article, which the editor contributed to the most, and evaluated the similarity between this set of articles and the focal article. As detailed in Table 3, both structural and content-based measures were employed in assessing inter-article similarity for this purpose, including the average number of common hyperlinks and Wikipedia categories between each of the reference articles and focal article, as well as the average of the respective ESA semantic similarity scores.

The proposed features also measure the editor's general tendency to focus her attention on a limited number of articles. We modeled the distribution of the editor's past contributions across articles in terms of entropy, where entropy would be generally higher for editors who have made small changes across a large number of articles, and low for editors who have focused their editorial efforts on a limited set of articles.

Finally, we were interested in representing aspects related to the editor's *activity in the community*. For those editors who are registered, one may accurately measure their overall period of activity in the Wikipedia community; we encoded this "tenure" period as a feature. Additional features were designed to capture the editor's activity across all of Wikipedia's namespaces.[29] Unlike the main namespace, which accommodates co-authorship activities, the other namespaces of Wikipedia are used for coordination and policy management. The proposed feature types measure the editor's *concentration on co-production activity* as the ratio of their past edits on the main namespace as opposed to the other namespaces of Wikipedia, as well as the proportion of her edits in each namespace of Wikipedia; presumably, SMEs' focus of activity across Wikipedia's namespaces might differ from that of non-SMEs. Notably, the other namespaces of Wikipedia, besides the main namespace that is dedicated to encyclopedic entries, are typically edited by committed registered members of the community, whom we refer to as *Wikipedians*. Given our desire to learn a model of wide applicability and the desire not to bias prediction towards sub-class of the population, we excluded these and other feature types that may be characteristic of Wikipedians (see Table 3) in learning, and consequently, in predicting expertise among the general population of editors.

[26] https://en.wikipedia.org/wiki/Wikipedia:Summary_style.
[27] We acknowledge Aaron Halfaker for the provided dataset with a pre-calculated persistency of every edit in English Wikipedia.
[28] https://en.wikipedia.org/wiki/Wikipedia:Talk_page_guidelines.
[29] https://en.wikipedia.org/wiki/Wikipedia:Namespace.

Table 4. 10-fold Cross-validation Classification Results Using Our Labeled Dataset and Different Learners

Algorithm	Precision	Recall	F1	ROC (AUC)
KNN	0.601	0.617	0.608	0.641
SVM	0.602	0.598	0.598	0.663
Random forest	0.736	0.769	0.750	0.818
XGBoost	0.737	0.772	0.752	0.832

In favor of generalization to all types of users, only those features that pertain to activity in public main namespace of Wikipedia are used.

6.2 Classification Setup

In learning, we describe each editor-article pair {<*e,a*>} in terms of the proposed feature functions, ignoring any other information that may be available, such as the editor's self-description on their personal page. That is, learning must track subject-matter expertise from information about the editor's past activity on the focal article and across Wikipedia as represented by the feature descriptions. The learning model is therefore applicable also to those editors who do not disclose any personal information.

We used the labeled dataset for training classifiers to identify SME editor-article pairs, having learning performance evaluated by the extent to which the classifiers successfully distinguished the labeled SME examples from the counter examples. We experimented with several learning algorithms, including the example-based KNN, max-margin linear SVM classifier,[30] and the tree-based algorithms of random forest and XGBoost [67] as implemented in the "Scikit-Learn" learning suite [88]. In our experiments, we computed ESA similarity using a publicly available implementation.[31]

Table 4 details the classification results, measuring the extent to which the learned models identified the SME-labeled pairs in terms of precision, recall, F_1, and area under the ROC curve (ROC AUC). Since the dataset is limited in size, learning performance is evaluated using a 10-fold cross-validation (CV) procedure. As shown, the classification results are well above an uninformative random guess baseline. The tree-based algorithms yielded the best performances, measuring up to 0.74 and 0.77 in precision and recall, respectively, and 0.83 in ROC AUC. Henceforth, we adopt the best-performing XGBoost [35] as the classifier of choice. As discussed below, we are generally interested in high-precision classification and further improve precision by considering the classifier's confidence in its predictions.

6.2.1 Impact of Feature Configuration. Our classification model should apply to both registered and non-registered Wikipedia editors. That is, while our labeled SME examples are composed of registered editors, we expect the learned model to *generalize* to all editors regardless of their registration status. We therefore excluded those features that describe activity that is typical of registered "Wikipedians" (Table 3) from the example representations. Nevertheless, in another experiment, we examined learning performance on our labeled dataset, which consists of registered Wikipedia editors, using the full set of features. Table 5 shows 10-fold cross-validation results using XGBoost for the two feature setups, i.e., *generalizable* vs. *full.* As expected, the full model yields better classification results compared with the partial model: Precision and recall are higher by 0.042 and 0.068 absolute points, yielding performance scores of 0.78 and 0.84, respectively. Indeed,

[30] We also experimented with RBF-kernel SVM, which yielded similar results to the linear SVM.
[31] We used an ESA implementation based on the English Wikipedia, available at https://github.com/pvoosten/explicit-semantic-analysis. See http://www.cs.technion.ac.il/~gabr/resources/code/esa/esa.html.

Table 5. 10-fold Cross-validation Classification Results Using Our Labeled Dataset and the Best-performing XGBoost Classifier

Features	Precision	Recall	F1	ROC (AUC)
Generalizable	0.737	0.772	0.752	0.832
Full	0.779	0.840	0.806	0.885

While we limit our final model to those features that pertain to activity in the public name space of Wikipedia in which both registered and non-registered users are active, and are thus applicable to all user types ("generalizable"), better performance may be obtained should this limitation be removed ("Full").

we found that registered SMEs differ from the counter examples with respect to the distribution of their editorial activity across the namespaces of Wikipedia (with SMEs being more active on *talk pages* that discuss co-authorship activities, but less active on *user* and *template talk pages*, which discuss personal pages and article templates). While these features are informative, they are not globally available—we therefore use the *generalizble* feature set for expertise prediction henceforth.

6.2.2 Confidence-based SME Prediction. So far, we evaluated classification performance on our labeled dataset and selected the best-performing XGBoost as our classifier of choice. Our main aim, however, is to apply the classifier (once trained on all of the labeled examples) towards the identification of SMEs within a large pool of unlabeled editor-article pairs. Notably, this target distribution moves away from our labeled examples, as it includes both registered and non-registered editors, and is further expected to exhibit various degrees and flavors of SME. We address these caveats by further tuning the classifier for this purpose; concretely, we improved the precision of our classification model.

In general, there exists a tradeoff between classification precision and recall that may be tuned according to the task preferences [72, 78]. Here, we follow the common practice of setting a threshold over the classifier's confidence in its predictions [33, 98]. Concretely, the XGBoost classifier outputs prediction scores in the range [0,1], which correspond to the probability of the example label to be positive [35]. By default, those examples that are assigned prediction scores above 0.5 are considered positive, and vice versa. Setting a higher threshold over the classifier's prediction scores would result in improved precision, albeit at the cost of lower recall.

Figure 1 depicts 10-fold CV precision and recall results using our labeled dataset while varying the threshold over the prediction confidence scores. As shown, precision rises steadily for higher thresholds, whereas recall plummets. The figure further shows multiple curves of F-measure performances, varying the relative importance of precision and recall: the F_1 measure assumes equal importance to precision and recall, while $F_{0.25}$ attributes four times greater importance to precision over recall.

We shall favor narrowing the scope of editor-article pairs identified as SMEs to those with high classification confidence. Based on our results in Figure 1, we chose to set the threshold over the classifier's confidence scores at 0.8, optimizing $F_{0,25}$ performance on the labeled dataset. Thus, in applying the classifier to unlabeled Wikipedia editor-article pairs, only when the classifier's confidence equals or exceeds 0.8, we shall identify the editor-article pair as SME. At this threshold level, classification precision is estimated at high 0.833, whereas recall is estimated at low 0.364. Thus, our classifier is conservative in that it avoids the classification as SMEs of many true editor-article pairs for which relevant available evidence may be limited.

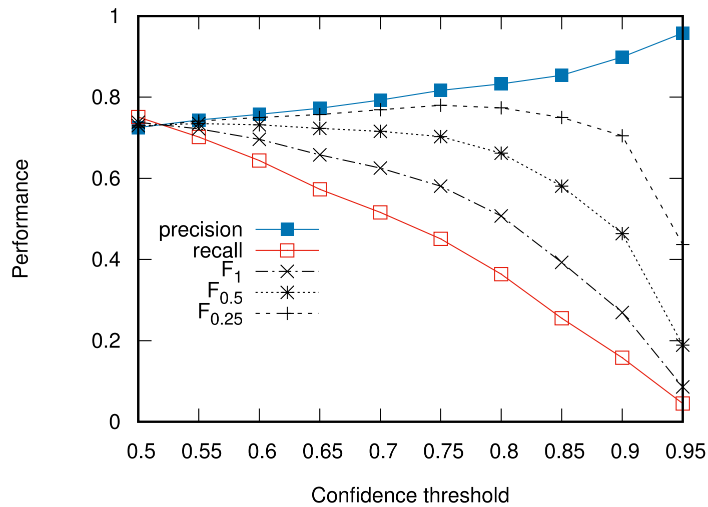

Fig. 1. Cross-validation classification performance using varying thresholds over the prediction scores generated by XGBoost. The F-measure curves reflect cumulative performance for different precision-recall preferences.

6.3 Expertise Identification in the Wild

Let us recall that for practical reasons, our learning model was trained to detect domain expertise based on identifiable example records of scholarly SMEs. Nevertheless, we assumed that the editorial patterns of scholar editors that are learned by the classifier would apply to domain experts of various types. To confirm that this was indeed the case and gauge expert identification "in the wild," we applied our trained model of expertise prediction to a broad sample of unseen examples and manually assessed the classification outcomes.

Having tuned and trained our classifier using all of the labeled examples available, we applied the classifier to our reference corpus of Wikipedia articles (Section 5.1). For the purpose of manual inspection, we focused our attention on a set of 100 randomly selected articles from our reference corpus, similarly distributed across topical categories.[32] For every article, we extracted the set of distinct editors who have edited them throughout the article's history and processed each of the unlabeled editor-article pairs using our classification model. We then considered a sample of 500 top-scoring predicted SMEs, composed of five editors per article, who were non-anonymous (registered) users of Wikipedia. Only 186 of the sampled editors had an active user page that included their self-description. We focused our evaluation effort on a subset of 120 such examples.

Quantitative analysis of the classification outcomes. Ideally, we would have reported classification precision and recall using the unlabeled examples. Unfortunately, for the reasons discussed thus far, we cannot provide accurate estimates of these measures. One may rarely ascertain domain expertise for a given editor-article pair based on the information provided by the editor on their personal page. And it is practically impossible to refute the existence of domain expertise, since the lack of relevant evidence does not imply that the editor must not be a domain expert. Since expertise cannot be manually determined, it is further impossible to gauge recall—the proportion of true SMEs that are captured by our approach. Our cross-validation study simulates the application of learning to held-out examples, where we provide performance estimates of these measures.

[32]To avoid bias, articles that included labeled editor-article pairs, which we used for training the classification model, were not sampled.

Nevertheless, we have made a genuine effort to assess the extent to which the unlabeled examples predicted as SMEs are indeed domain experts. The sample of predicted SME editor-article pairs was manually inspected by the first two authors of this article in an attempt to establish, or refute, the expertise of each editor in respect to the article they edited. We read the text available on the user pages and further attended possible evidence on the web (e.g., when the user page included links to other web pages outside of Wikipedia). Following a stage of independent work, the two annotators had agreed on 80% of the cases and further discussed and reached agreement about the remaining cases. (Disputable cases were resolved by consulting with the other authors of the paper.)

As expected, in the vast majority of cases, there was not enough information available to reach a verdict. For example, this identified SME who has edited the "Anxiety" concept merely described themselves on their page as follows: *"My interest and focus is in trying to help wiki have excellent and leading edge content."* In the lack of concrete details about the editor's background, we could not determine her level of expertise with respect to "Anxiety." Another editor of an article titled "Nontreponemal tests for syphilis" wrote on their page: *"I am the Communications Coordinator at AMMI Canada. Interests in Microbiology and Infectious Diseases."* We could not establish that the coordinator for such an organization was indeed an expert in the topics of medical microbiology and infectious diseases, as one might come to a coordinator role from a background in journalism, communications, media, and so on. In this case, and in some other similar cases, we took a conservative approach and considered the available evidence to be inconclusive.

Overall, we have reached a definite conclusion regarding the level of expertise in 22 cases in our sample (18.5%). Out of those, we established that the editor was indeed an expert in relation to the article in 18 cases (81.2%). Beyond scholars, the identified experts included, for example, a marine biologist who edited the article "Argochampsa" (an extinct genus of eusuchian crocodylomorph); a Native-American who edits articles related to the Indigenous peoples of the Americas (e.g., the "Manuelito Complex" article); and a Pakistani editor who has been working in the food industry for many years, writes blogs about ethnic food, and also edits Wikipedia articles on this topic. The remaining 4 cases (18.8% of the flagged examples) were identified as classification mistakes. In three of these, the editor was a bot (i.e., the same bot, having performed edits in three articles in the sample), which was not eliminated *a priori* using our list of bot accounts. In the other case, an editor of the "Biconditional elimination" article, which is related to the rather complex topic of propositional calculus, declared on his page on special interest in the creation of article outlines[33] and has been involved in editing outlines on many diverse topics; we thus concluded that he was not a domain expert with respect to the topic edited. (In general, outlining may be a type of editorial work that experts are more involved with than others, where that signal might have been captured by the classifier; the literature does suggest that experts are more equipped for this task [111].) Otherwise, we conjecture that the classifier might be misled in cases where the editorial history of the editor or the article is very limited; e.g., an editor who starts a new article and imports relevant text from some existing resource might be mistaken to be an expert.

Sources of expertise. The sampled editors were similarly inspected with respect to the types of predicted expertise. In addition to *scholars*—editors with record of relevant scholarly expertise, as defined in our annotation guidelines (Section 5.2)—we distinguished between *practitioners*, with indication found of current or past personal practice in the focal article's area; or, knowledgeable laypersons whose expertise has been acquired informally and outside work. Consider for example the self-description on a personal page of an editor who has contributed to the "Headache" article

[33] https://en.wikipedia.org/wiki/Wikipedia:Outlines.

and was identified as an SME with respect to that article: *"I am an emergency physician by profession and have been working on Wikipedia for more than ten years."* Based on this description, we flagged this editor as a *practitioner* expert. Next, consider an editor identified as an SME with respect to the article "Tomb Raider: Underworld" (pertaining to a computer game), described on his page as: *"philosopher of video games ... I went to * University to study philosophy and later to University * to study history ... In my spare time I like to play video games. I own a PlayStation 3, PlayStation 4 and PS Vita, and with PlayStation Plus I got a whole bunch of games. With my studies I also look for religious, philosophical and ethical aspects in video games..."* This description suggests that this editor is an avid and knowledgeable, yet non-credentialed, expert in the focal topic. Whenever relevant information was not been accessible, however, we could not determine the editor's source of expertise.

Overall, among the examined sample of identified SMEs, we found roughly equal proportions of scholars, practitioners, and other non-credentialed experts. Based on this assessment of the automatically identified SMEs, we conclude that the characteristics of editorial work of *scholarly experts*, on which our model was trained, carry to non-credentialed SMEs. Our evaluation, albeit limited, further suggests that the experts who edit Wikipedia gain their expertise through various avenues, including academia, practice, unique life experiences, or prolonged interest.

7 RESULTS: CHARACTERIZING SUBJECT-MATTER EXPERTS AND ESTIMATING THE SCOPE OF THEIR WORK

Having outlined our empirical framework for identifying SMEs, we next employ this framework to address our research questions *RQ1–RQ3* regarding the scope of SMEs' work in Wikipedia and the characteristics of their work.

7.1 What are the Activity Profiles that Characterize Subject-matter Experts?

First, we address *RQ1*: *What are the activity profiles that characterize subject-matter experts?* To determine differences between the observed behaviors of SMEs and other editors, we attended our labeled dataset and performed a series of two-tailed unpaired t-tests, comparing the means of the observed feature values for the populations of SMEs and the counter examples. In various applications, it is common to apply a statistical method for counteracting the problem of multiple comparisons (in our case, tests for multiple features). The Bonferroni correction method [112], which hypothesizes that all experiments are independent, is considered the most stringent approach. In our study, we investigated both Bonferroni and False Discovert Rate (FDR) correction methods [21], which are commonly used in areas of data mining of large datasets and in machine learning applications, e.g., Reference [113], and found that the results differ only slightly. Hence, to be more conservative, we opted to report here on our results after applying the Bonferroni correction.

Table 6 lists the features where the differences in means between the two groups were statistically significant ($p < 0.05$). We observe that in comparison with the counter examples, SMEs are characterized by: (a) overall lower cumulative activity levels across all of Wikipedia's namespaces; (b) concentrating on co-authoring of encyclopedic entries, while doing less administrative tasks such as combating vandalism; (c) meaningful edits, i.e., adding content and re-structuring of articles, as opposed to minor edits and clean-up such as fixing typos; (d) persistent edits that survive the continuous refactoring process over longer periods; (e) substantial work on the article on which they were labeled as SMEs (in terms of both edit count and the percentage of overall edits to that page); and (f) focused activity: centered activity in fewer articles (lower entropy of activity) and topical cohesion of the articles that they contribute to (measured either in terms of category or hyperlinks' similarity).

Table 6. Feature Means That Were Found to Be Statistically Significant (p-val<0.05) between SMEs and the Counter Examples in Our Manually Annotated Dataset Using Two-tailed Unpaired t-tests and Bonferroni Correction

Feature category	Feature	SMEs	Counter
Focal page, intensity of activity	total edits (count)*	12.1	3.2
	page edit ratio (×1000)*	10.4	2.3
Focal page, nature of activity	edit types: "fix vandalism" (%)*	0.07	0.30
	edit types: "add content" (%)*	0.28	0.16
	edit types: "reorganize existing content" (%)	0.03	0.02
	edit types: change/add "Wiki MarkUp" (%)*	0.5	0.4
Focal page, longevity of edits	persistence (days)*	208.4	116.0
Focal page, coordination activity	edits on "talk page"*	3.3	0.4
Wikipedia, topical focus	structural similarity, hyperlink overlap (×1000)*	15.7	6.1
	structural similarity, category overlap	0.033	0.028
Wikipedia, distribution	entropy of edit count*	5.8	6.7
Wikipedia, tenure	total edits in all namespaces*	1,584.2	2,642.9

(Features for which differences were found to be statistically different with p-val<0.01 are further marked with asterisk.) The table lists features that apply to both registered and non-registered editors.

Overall, we find the observed differences to be intuitive, profiling SMEs' activity as being focused on making meaningful content contributions to a small number of articles within their area of expertise. Interestingly, our findings echo related work, by which community members of question-answering websites who are perceived as knowledgeable tend to focus their activity on a small number of topical categories [3]. As we have shown, our classifier trained to distinguish SME based on the editorial activity patterns learned from scholars further identified practitioners and other types of non-credentialed experts. Accordingly, we believe that the editorial features listed are generally characteristic of SMEs in Wikipedia.

7.2 What Is the Scope of Subject-matter Experts' Contribution to Wikipedia?

For estimating *the scope of SMEs' work in Wikipedia (RQ2)*, we processed the editorial history of all of the articles within the reference corpus. Let us recall that this corpus comprises a representative sample of 1K Wikipedia articles uniformly distributed over various topical categories and article maturity levels (see Section 5.1). For each of these articles, we extracted and classified the full set of distinct editors who edited them throughout the article's history.[34] These histories comprised 457,840 unlabeled distinct editor-article pairs and 305,574 distinct editors, in total. We note that similarly to the general population of Wikipedia editors, the majority of the editors in this reference sample of articles were not registered members of the Wikipedia community, i.e., anonymous users. In line with prior studies on Wikipedia [14], we considered the IP address as a proxy for an unregistered member's identity.

Each of the unlabeled editor-article pairs was processed by our classification model, trained using all of the labeled examples available. Overall, 20.2% of the evaluated distinct editor-article pairs were classified as SMEs with high confidence (>0.8). Those editors classified as SMEs were responsible for a somewhat larger portion of the co-production work: 22.5% of the total edits made to the respective articles.

[34]Non-human editors, i.e., software robots, were excluded based on a list of bot users publicly available at https://dumps.wikimedia.org/other/pagecounts-ez/wikistats/csv_wp_main.zip.

We performed several sensitivity analyses of these estimates; mainly, we explored alternative thresholds on classifier confidence. We found that the proportions of predicted SME editor-article pairs were 31.8% and 11.6% when employing the thresholds of 0.75 and 0.85, respectively; the overall share of SMEs' activity among the co-production work was 32.2% and 13.5% when employing the same thresholds, respectively.

While we acknowledge that the classifier's predictions are imperfect, let us recall that classification precision was high for this range of confidence thresholds in our controlled experiments, evaluated at 0.82–0.85. However, the respective recall rates were low, in the range of 0.45–0.26, indicating that many true SMEs were also missed by our precision-oriented classification approach. Specifically, we expect editors not to be classified as SME in case of insufficient evidence, e.g., due to limited editorial history.

Altogether, based on these results, we believe it is a conservative estimate that SMEs make up roughly 10%–30% of Wikipedia's editor population, and that, in turn, they are responsible for a similar portion of the edits.

To the best of our knowledge, these estimates provide the first approximation of the amount of domain experts' contributions to Wikipedia. We stress that, in reality, subject-matter expertise, as well as the editorial preferences of experts, are not dichotomous; possibly, there are contributors to Wikipedia with varying levels of domain-specific expertise who follow editorial patterns that possibly differ from those learned by our model. Thus, our estimates could serve as a lower bound for the true scope of experts' contributions to Wikipedia.

7.3 How Does Experts' Involvement Vary across Wikipedia Articles?

Finally, we explore variability in SMEs' involvement across Wikipedia's articles (*RQ3*). It is sensible that the involvement of SMEs vary by the focal concept, as some knowledge domains are more specialized than others. Hence, we now focus our attention on gauging SMEs' involvement on a per-article basis.

Figure 2 illustrates the ratio of editors identified as SMEs for the articles included in our reference corpus, for which the editorial history involved more than 250 distinct editors at the time of the analysis. As shown, there exists high variance with respect to the ratio of identified SMEs across articles. The article with the lowest ratio of identified SMEs among the displayed examples is "pancake": This page has been edited by roughly 2,800 distinct editors at the time of the analysis, out of which only 2.9% were identified as SMEs. Very low levels of SMEs' involvement were observed also for the articles "Ambergris" (4.8%) and, "Baritone" (19.5%), "Social liberalism" (17.2%), and "W. Edwards Deming" (a statistician; 16.8%). Interestingly, other articles with high ratios of identified SMEs are the fictional TV character "Peggy Mitchell" (42.5%), the movie "Crossroads" (27.3%), and the video game "Tomb Raider: Underworld" (16.6%). Based on our analysis of the sources of detected expertise, we conjecture that the identified SMEs in these articles in the entertainment domain correspond to avid hobbyists. (This is most likely the case for articles on video games, and possibly the case for articles about movies and TV shows.) Also of note, health-related and medical topics were not associated with particularly high ratios of identified SMEs: the concepts of "Gastrointestinal tract" (6.0%), "Headache" (7.2%), and "Lymphatic system" (7.5%) reside below the sample median of expert involvement at 10.7%. This is somewhat counter-intuitive, as medical knowledge is presumably owned by expert physicians. A possible explanation is that these articles draw much editorial attention, which is mostly administrative. Or, perhaps physicians edit a wide variety of health-related topics, which are perceived as diverse rather than cohesive, and therefore are not recognized as SMEs by our model (some of the key features of our expertise prediction model pertain to focused activity on a closely related and limited number of articles). Correspondingly, the ratio of identified SME editors per the topic "Anxiety," which may concern

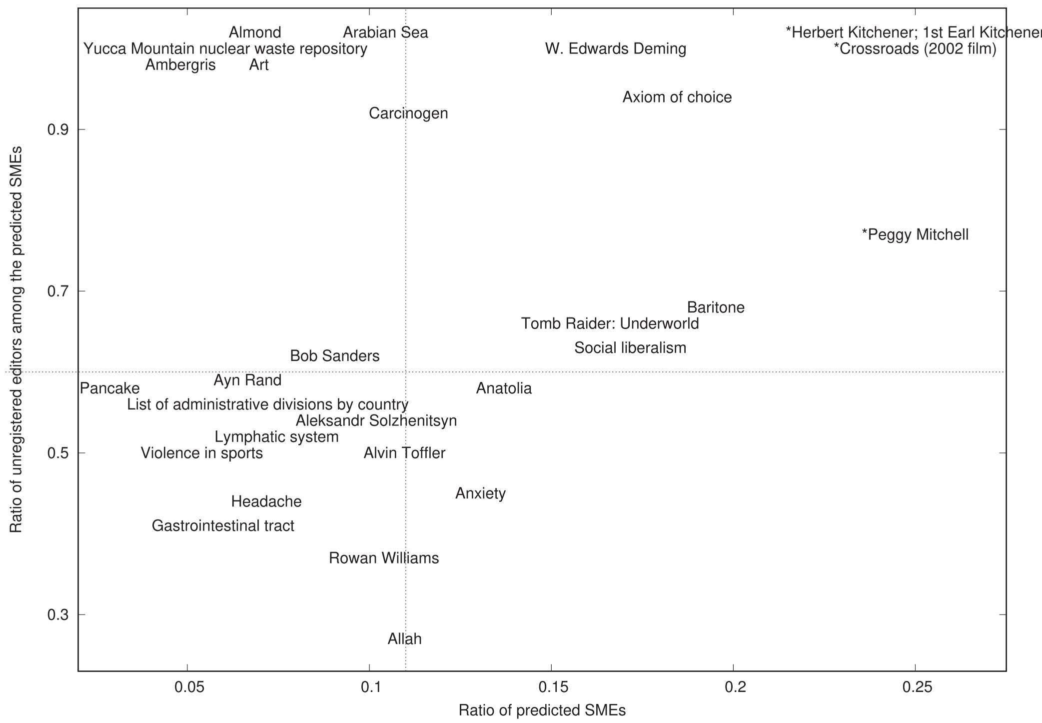

Fig. 2. Distribution of sampled articles, each edited by at least 250 distinct users, in terms of (a) ratio of users predicted as SMEs with high confidence (>0.8), and (b) the proportion of predicted SMEs who are unregistered to Wikipedia. The median value along each dimension is denoted with a dotted line. The articles denoted by asterisk had particularly high ratios of predicted SMEs, beyond the range shown; specifically, the predicted ratios for "Crossroads," "Peggy Mitchell," and "Herbert Kitchener" are 0.27, 0.43, and 0.50, respectively.

a narrower specialization, namely, Psychiatry, was estimated at a higher 13.1%. These conjectures may be examined more closely in future work.

Figure 2 further displays the proportion of SME editors who are not registered to Wikipedia for each of the articles. Again, large variance is observed, having the proportion of non-registered identified SMEs distributed over the range of 0.27–1.00. Interestingly, there exists some correlation between the two measured dimensions. Concretely, for those articles with high ratio of identified SME editors (above 0.15), the proportion of SME editors who are unregistered is also high (measuring 0.63 or more)—these articles populate the top-right part of the figure. In contrast, articles with low ratios of identified SME editors are often associated with low ratios of non-registered SMEs (between 0.25–0.6); consider, for example, the articles of "Ayn Rand" or "violence in sports," shown at the bottom-left part of the figure. We conjecture that for those articles that correspond to popular knowledge domains, relevant SME editors are prevalent within the registered pool of Wikipedia editors. Otherwise, in specialized subject domains, SMEs are more rare, where this gap is filled by expert editors who are not registered. Notably, some articles were associated with low ratios of identified SMEs; yet, almost *all* of the contributors to these articles are non-registered, e.g., "Yucca mountain nuclear waste repository" or "Arabian sea." Possibly, relevant knowledge, or acquaintance, with these topics is biased geographically, such that there are fewer contributors to these articles (including SMEs) amongst Wikipedia's registered editors. This conjecture also deserves further attention in future research.

In summary of this section, we observed that the ratio of identified SME editors indeed varies considerably across articles, conceivably due to varying characteristics of the respective knowledge domain. We further inspected the ratio of non-registered SMEs and found some correlation with the overall ratio of SMEs per article. Future research following these findings may shed more light on the topical areas that are characterized (or call for) expertise involvement in Wikipedia.

8 DISCUSSION

Like other goods, producing knowledge goods at high quality requires both effective processes and high-quality ingredients. Though high-quality outputs can sometimes be distilled or assembled from many low-quality signals (as done in some crowdsourcing settings), the construction of complex intellectual artifacts, such as encyclopedic articles, is likely to be more effective with higher-quality inputs. Yet, while extant literature has paid substantial attention to peer-production methods (including studies on governance mechanisms [18, 96], quality assurance processes [100], norms and policies [1, 25, 29, 54, 68, 80], role system [13, 15, 30], and underlying technology infrastructure [29, 80, 109]), research on the "ingredients," namely, contributors' qualifications and expertise and the nature and scope of their contributions, is scant.

As mentioned above, we suspect that one of the primary reasons for this empirical void, especially regarding Wikipedia, has been the difficulty to trace the work of subject-matter experts, since most of the editors remain anonymous or do not disclose their expertise. Our first task, therefore, was to overcome this obstacle by finding a way to detect expert editors (even when they are anonymous) and their work in Wikipedia. Importantly, for us to be able to tackle the rest of our research questions, we had to develop automated methods, which will allow us to work at scale and with a high degree of confidence.

This was not a trivial task. To achieve it, we referred to a diverse set of 1K Wikipedia articles, with over a million edits, and carefully scanned these data for indications of editorial expertise through an algorithmic process and manual work of human annotators. We were able to verify relevant scholarly expertise for editors who were registered Wikipedia members, who disclosed their identity, and for whom evidence on relevant PhD-level academic education, such as scholarly papers published in the focal article's domain, could be found. Having constructed a dataset composed of expertise-labeled editor-article pairs and counter examples, learning was applied to automatically identify the expert editor-article pairs based on *implicit* patterns and markers of the editors' past contributions to the focal page, as well as across Wikipedia. Despite the inherent bias in our training set towards SME editors with certain traits (registered and scholars), we found that the implicit markers associated with their editorial activity carry to experts who are not characterized by these traits. In particular, qualitative analysis has shown that in addition to scholars, the learned model also identified as experts knowledgeable practitioners and other non-credentialed editors. Thus, the development of the expertise-annotated resource and the automatic expertise detection model serves to remove a major barrier that held back empirical, quantitative, large-scale probing of the nature of expert work in Wikipedia.

8.1 Main Contributions

Upon validating the algorithm, we turned to explore the work of experts. To the best of our knowledge, our analyses provide the first large-scale empirical assessments of expert work in Wikipedia. Our findings offer several contributions to research in the area. First, our results show that subject-matter experts are responsible for a substantial amount of Wikipedia's activity: We estimate that approximately 10%–30% of Wikipedia's contributors have substantial subject-matter expertise in the topics that they edit. We further estimate that these subject-matter experts are responsible for a similar portion of the co-produced content, performing roughly 10%–30% of total edits. Since our classification model emphasized precision over recall, it is certainly possible that it overlooked some domain experts. Thus, the approximations reported here are potentially an underestimate for the magnitude of experts' contributions.

Second, results from this study suggest that experts are characterized by distinctive activity patterns. Specifically, we found that they: (a) concentrate their activity within a highly inter-related set of Wikipedia articles; (b) focus their attention on specific types of editing activities (e.g., adding content, referencing external sources) and avoid other types (e.g., participation in vandalism or in the combat against vandalism). Whereas, we are not aware of prior studies that sought to characterize the profiles of subject-matter experts, our findings correspond to prior works that profiled Wikipedia contributors. For example, prior studies have demonstrated that "content-oriented" contributors (as opposed to those with an orientation towards the community and administrative work) tend to have topic-concentrated activity patterns and are often non-registered [14]. We further found that edits done by experts tend to persist in the face of the ongoing refactoring process; this finding echoes a previous study, which suggested that the contributions by non-registered editors tend to survive longer periods [6] and include essential content [12]. In light of our study's findings, we suspect that those content-oriented non-registered contributors whose edits persist longer are likely subject-matter experts.

Finally, we find that experts' involvement varies greatly across articles. As may be expected, articles about presumably narrow and high-proficiency concepts, e.g., "Axiom of Choice" or "1st Earl Kitchener," involve high ratios of contributions by domain experts. We further examined some articles about video games, movies, and TV, where subject-matter experts make up a considerable fraction of the editors, even up to 40%–50%. Following qualitative analysis, we believe that most of the these experts are non-credentialed experts who have acquired substantial relevant knowledge about their favorite games, movies, and so on. However, for some articles, e.g., "Headache" or "Gastrointestinal tract," expert involvement was low (under 8% of the editors). Prior research has demonstrated that coverage of various topical domains in Wikipedia is uneven, where some domains attract more activity than others [14, 65, 66]. We conjecture that the variability in the level of expert involvement across articles has to do with characteristics of the differences across knowledge domains, but our inquiry is not conclusive, and this area warrants further exploration.

8.2 Limitations

Let us recall that our goal was to identify domain expertise among Wikipedia editors, who do not typically disclose relevant details about their background and qualifications. Furthermore, many of Wikipedia's editors are non-registered and are therefore strictly anonymous. Our approach for tackling this challenge was based on learning editorial patterns, which are implicit and applicable to all editor-article pairs, from detectable examples of *scholar* SMEs; as this type of expertise is well-defined, we were able to devise a rule-based semi-automatic procedure that scaled up the annotation process and consequently achieved high inter-annotator agreement rates. We have assumed, and shown, that the editorial patterns associated with our annotated subset of scholar SMEs apply to other types of non-credentialed experts. Thus, having applied our model of expertise identification at scale, we provided the first estimate regarding the scope of SMEs' involvement in Wikipedia. Hereby, we discuss in detail the various limitations and potential biases that may apply to our work and findings.

Discretization of expertise. One may argue that any contributor who makes a meaningful addition to Wikipedia has a certain degree of expertise with respect to the topics that they edit. In accordance with the literature, we view expertise as a continuum; any editor in Wikipedia is somewhere on that continuum in relation to a topic edited. Due to the lack of explicit and fine-grained information about editorial expertise in Wikipedia, we have identified domain expertise in a binary fashion. Our approach practically identifies editor and article pairs for which the observed editorial patterns resemble the patterns exhibited by verified scholars within a topic of their expertise (maintaining focus on that particular domain, performing certain editorial activities, etc.). While this approach may not distinguish between a credentialed expert and a knowledgeable autodidact who attained their knowledge through years of reading or practice, it relies on clear differences found between the editorial behaviors of verified scholar experts and random editors. Our results further indicate that the notion of expertise may differ across domains. For instance, it is probably easier to distinguish between experts (at the higher end of the continuous spectrum) and non-experts (at the lower end of the spectrum) in more complex topics such as molecular biology, nuclear physics, or neuroscience, than it is on such general topics that many, if not most people, are familiar with, such as "pancake" or "toys." Of course, there are people who are more knowledgeable than others even on those latter topics, but it is likely that knowledge about them is more uniformly dispersed. Accordingly, we observed low rates of identified SMEs for general topics like "pancake."

Annotation and learning. Annotation-wise, a main limitation of this work is the size of our dataset, which was bound due to bottlenecks of the annotation process. The semi-automatic procedure devised in our work, targeted at screening large volumes of past edits with respect to evidence of scholarly expertise, may be further expanded in future work. Mainly, detecting possible evidence of expertise among a very large pool of article-editor pairs, as well as scaling the annotation process, e.g., via crowd-sourcing, should yield more labeled examples for learning purposes. Further, the annotated dataset should ideally represent various types of SMEs. One may adapt our proposed procedure for scanning user personal pages on Wikipedia to support the manual identification and annotation of more diverse SME. We note, however, that the manual assessment of varying flavors of expertise is subjective to a large extent and may result in low inter-annotator agreement rates. Also, non-formal expertise is often hard to verify.

Another inherent limitation of example annotation involved the construction of the counter-SME examples, as the lack of personal information about an editor in general, and the lack of relevant evidence regarding expertise with respect to an article that they have edited in particular, do not imply that they are *not* SMEs in that topic. Consequently, our sampled counter examples might include noisy false-negative labeled examples. While learning in the presence of noise is common practice and generally considered to be effective, such noise may affect learning performance and its assessment to some degree. Unfortunately, we cannot ascertain nor quantify such impact.

Learning-wise, classification is not error-free. In this work, we set a high threshold over the classifier confidence, where this is intended to achieve high precision, but hurts recall. In other words, a large portion of SMEs is to be missed by the classifier due to inconclusive evidence. In our experiments, we observed low classification confidence in the lack of sufficient statistics, e.g., for SME editors with limited history on the target page or in Wikipedia in general. However, precision-wise, the classifier may fail to identify SMEs who demonstrate atypical behaviors, e.g., spread their contributions across diverse topics, some of which they do not specialize in, or take atypical interest in typo correction.

Classification errors can also stem from data distribution shift, in case that our dataset of labeled examples is under-representative of the unlabeled data. Importantly, we have shown through manual validation checks that our algorithm was able to generalize beyond the labeled examples to various types of SME, providing credence to our findings. While we were not able to determine precision and recall for the unlabeled examples (due to the lack of ground truth labels), we expect precision to remain high, where possible lack of generalization might impair recall. Thus, our claim by which our precision-oriented classifier provides with conservative estimates of SME holds.

In future work, we expect that classification can be improved in different ways, e.g., by modeling additional meaningful features; by using more advanced text processing analyses (e.g., embedding-based text similarity measures as alternative to ESA); or by employing neural "feature free" learning schemes, should more data be available for learning.

SME identification bias. In applying our classifier for expertise identification at scale, we imposed intentional bias, by which only those article-editor pairs that are predicted as SME with high classification confidence are predicted to be SMEs. Having tuned the respective confidence threshold value to achieve high precision at the cost of recall, we believe that our classification results under-estimate the true prevalence of SMEs in Wikipedia.

Notably, we assess domain expertise for both registered and non-registered editors. In the lack of ground-truth labels across these sub-populations, we make the assumption that the editorial patterns that characterize SME editors are independent of their registration status. In addition, we make the assumption that non-registered editors may be identified by IP. In practice, anonymous users may use multiple IPs over time. We speculate that being exposed to partial information about the editor's history due to change of IP would lead to low confidence of the classifier, i.e., yield somewhat lower SME identification rates among non-registered users.

8.3 Practical Implications of This Study

Beyond the contributions to the scholarly literature, our study also has important practical implications for custodians of online co-production communities. Experts differ from novices on many aspects; not only are they more knowledgeable, but through training and experience they also often become more skilled and develop different mental models and routines that they employ. Importantly for communities that produce and curate knowledge, research suggests that experts exert better critical-thinking skills, specifically in how they assess sources [111]. This sort of skill is especially relevant for writing a reference work such as an encyclopedia; indeed, Wikipedia's core content policies—Neutral Point of View, No Original Research, and Verifiability—deal with this issue.

The question of experts' role in Wikipedia has been extensively debated. Although the formal stance of the community does not prefer subject-matter experts, there is an active undercurrent that recognizes the potential value of experts' contributions and seeks to increase their involvement in the coauthoring process. In fact, many within the community believe that experts can potentially play important roles in improving the quality of Wikipedia's contents, as acknowledged by Wikipedia's founder, Jimmy Wales, who was quoted saying, "Most entries are edited by enthusiasts, and the addition of a researcher can boost article quality hugely" [83]. A straightforward implication is the need to attract and retain experts and channel their involvement to the articles most needing their contributions. There is an active stream of research on the identification of experts in online communities, seeking to channel their efforts to particular areas that could benefit from their expertise, e.g., Reference [3]. In the context of Wikipedia, prior work has attempted to direct editors to articles of their interest [40, 41], and a similar approach may be applied to routing experts' efforts. Wikipedia has taken some steps to involve more domain experts, such as initiating collaborations with museums, libraries, and so on; here, we suggest that the involvement of expert may be more targeted. Thus, a practical implication of this study is the development of a tool set for identifying experts in Wikipedia, which could serve for directing experts to articles and tasks that could particularly benefit from their contributions. For example, experts' work may be most needed in low-quality articles or in contentious articles where facts are disputed.

A more complex issue—and a highly sensitive one—concerns experts' roles and privileges. To date, the Wikipedia community maintains its original stance of not yielding experts more power in content-related decisions or assigning them special privileges (those special rights are reserved for the "janitors of knowledge" [101] who are experts in Wikipedia's processes and procedures [13, 29], but are experts in articles' topics). In contrast, consider a professional organization such as a hospital, often characterized by a dual management structure: professional (i.e., based on expertise) and administrative. May such a structure be applicable for Wikipedia? Can we envision an administration that—in addition to process specialists—involves domain experts? Are there issues that more naturally lend themselves to the decision of expert committees (e.g., articles' categorical organization, splitting or merging of articles, resolving content-related disputes)? These difficult questions warrant the attention of the Wikipedia community and potentially of other open knowledge-creation and knowledge-curation platforms. Findings from our study could serve to inform this discussion. In the past, Wikipedia's community has demonstrated much agility in adopting its governance mechanisms in an organic manner, including the introduction of new access privileges, development of new policies, creation of bots to streamline some processes, and the application of controls such as page protection. In light of our findings regarding the scope of experts' contribution, an open question for the community is, then, whether to revisit the discussion on the role of credentialed experts.

Beyond online communities, key principles from the community-based peer-production model have begun "spilling over" into traditional organizations [102–104]. Many companies use wikis as a knowledge management tool [3, 104], particularly for developing organizational encyclopedias and knowledge-sharing tools [8, 9], adopting in part the organic processes that typify wiki-based collaboration. Given the different composition of the crowds who author these wikis, where it is likely that most authors will be topical experts, it is an interesting open question how the dynamics of expert work play out in those settings.

9 CONCLUSION

Expertise is a key ingredient in the production of knowledge and innovation. Yet, the literature on the co-creation of knowledge-based products has thus far focused mostly on the processes of co-production, extensively investigating topics such as coordination mechanisms, conflict resolution, quality assurance, roles, and function, and has paid relatively little attention to a key ingredient of these processes, namely: to experts and the expertise they share. Lacking the ability to empirically observe, probe, and measure it, until now the scope and nature of the contribution of experts in citizen-based knowledge peer-production systems, and specifically in Wikipedia, has been like dark matter: assumed by most, debated by some, and mostly unknown.

To the best of our knowledge, this work is first to offer a quantitative empirical investigation and account of the impact of subject-matter experts in Wikipedia, which serves as an exemplar of peer-production [23]. Our results are preliminary, and like any attempt to estimate dark matter, exact numbers should be taken with a grain of salt. However, our findings quite clearly indicate that notwithstanding the importance of process and platform, the inclusion of experts in co-production is an essential component in Wikipedia's success. We expect that our findings will be informative and actionable to scholars of knowledge co-production, as well as to the Wikipedia community.

We further hope that our work will open the door for future research into the role that experts play in peer production of knowledge, as many questions remain open. Future research may be able to provide further data on the experts themselves, on their relationship and interaction with the community, and on the impact of their work on knowledge products. In terms of experts characteristics, open questions include: what their demographics are, as well as the source of their expertise; what motivates them to contribute and what deters them from contributing; when do they produce content, and how. In terms of their relationship with the community, questions include: how they engage and interact with the community, how they contribute to group processes, how experts and expertise are viewed and accepted by the community, and how that affects their motivation and participation trajectories. Finally, and importantly, future research should provide more detail on the impact of expert work, specifically, by tying expert involvement with group-level outcomes such as the quality of co-produced knowledge. Providing comprehensive answers to these questions will require a combination of quantitative and qualitative approaches, and potentially also the integration of both process- and resource-centric perspectives.

APPENDIX

A AUTOMATIC SCREENING OF EVIDENCE ON EXPERTISE

Following is a full description of our procedure for automatically screening editor-article pairs with respect to their potential as verifiable SMEs. We further detail the results of applying this automatic assessment procedure to the editorial history of our reference articles, and the sampling of example editor-article pairs based on these results for further manual investigation and labeling.

A scoring heuristic of editor-articles pairs. We manually constructed the following weighted lexicon of expertise-related terms and had the automatic procedure search for these terms on the editor's personal page on Wikipedia: {"professor" (10); "PhD" (10); "researcher" (10); "Doctor of philosophy" (10); "doctorate" (10); "doctoral" (9); "doctor" (9); "professional" (5); "university" (5); "college" (5); "scientist" (5); "academic" (5); "academy" (5); "diploma" (5); "philosophy" (3); "research" (3); "publish" (3); "mathematician" (3); "statistician" (3); "paper" (3); "expert" (3)}. Considering the total weights of terms found on the editor's page, the possible outcomes of this automatic assessment were (Table 7): *high match*, if the overall weighted term-based evidence exceeded a threshold (manually set to 9); *low match*, if relevant terms were found, but their overall weight was lower than the threshold; *no match*, if none of the pre-specified terms was found; or, *no page*, if the editor in question did not maintain a personal page on Wikipedia.

The automatic assessment procedure further aimed to link the editor with her researcher profile on *Google Scholar*, whenever possible. Having submitted a query with the user's name (if available) to *Google Scholar*'s search engine, the relevance of the top-ranked researcher profiles retrieved was assessed. First, to account for possibly erroneous name mismatch, the similarity between the editor's user name and the retrieved researcher name was measured using the Levenshtein metric, which is often used for personal name matching [38]. In addition, to account for possible name ambiguity and establish the relevance of the researcher profile retrieved, we evaluated the topical similarity between the publications by the candidate researcher profile and the focal article using the explicit semantic analysis (ESA) measure. Concretely, for each candidate researcher profile retrieved, topical similarity was evaluated between each of the (up to 10 most-cited) scholar's publications and the contents of the focal Wikipedia article, having the pairwise ESA similarity scores aggregated. We empirically set a threshold over the total ESA score to 0.03, above which topical relevance was considered to be high.[35] The possible outcomes of this step are either: *no match*; *name mismatch*; *topical mismatch*; *low topical match*; or *high topical match*, as detailed in Table 7.

Example sampling for manual annotation. We applied the above scoring procedure to all of the distinct editor-article pairs that comprise our reference corpus (Section 5.1). The distribution of outcomes is detailed in Table 8. Table 9 specifies the ratio of editor-article pairs that we sampled for manual verification and labeling per each of the cells in Table 8. As shown, our manual annotation

[35] As ESA computes cosine similarity in a high-dimensional space, it outputs small values in absolute terms.

Table 7. The Possible Outcomes of the Automatic Screening of Article-edit Pairs and Their Interpretation

i. Assessed evidence on the personal page:	
No page	The editor does not have an active (non-empty) personal page.
No match	No expertise-related terms found on the editor's personal page.
Low match	Expertise-related terms found on the editor's personal page, but their total scores is below a set threshold, suggesting that scholarly expertise *may* be mentioned on the page.
High match	Expertise-related terms found on the personal page and their total scores exceed the threshold, suggesting that scholarly expertise is *likely* mentioned on the page.
ii. Tracking of relevant scholarly publications history:	
No match	No candidate scholar profile found on *Google Scholar*, possibly because the editor's real name is not specified.
Name mismatch	Candidate scholar profile(s) found, but the editor and researcher names are not sufficiently similar.
Topical mismatch	None of the candidate scholar profiles show topical relatedness to the focus article.
Low topical match	There exists a scholar profile whose publications are marginally related to the edited Wikipedia article, suggesting that the editor is possibly an SME.
High topical match	There exists a scholar profile whose publications are well-related to the edited Wikipedia article, indicating that the editor is a likely an SME.

Table 8. The Results of Automatic Screening of All of the Editor-article Pairs in the Reference Corpus, Assessing Potential Evidence of Scholarly Expertise on the Editor's Personal Page and Relevant Publication History Outside of Wikipedia

Scholar/Personal page	High Match	Low match	No match	No page	Total
High topical match	0.05%	0.06%	0.33%	0.10%	0.55%
Low topical match	0.05%	0.08%	0.43%	0.12%	0.67%
Topical mismatch	0.13%	0.28%	1.23%	0.34%	1.98%
Name mismatch	0.89%	1.46%	9.46%	2.61%	14.43%
No match	2.96%	6.56%	54.08%	18.78%	82.38%
Total	4.07%	8.45%	65.53%	21.95%	100.00%

The categories are defined in Table 7.

Table 9. The Ratio of Examples Manually Annotated per Cell in Table 8, and the Ratio of Examples Positive Labeled as SMEs Out of the Example Sampled per Cell (in Brackets)

Scholar/Personal page	high Match	Low match	No match	No page	Total
High topical similarity	100% (69%)	100% (43%)	14% (15%)	8% (27%)	30% (42%)
Low topical similarity	100% (36%)	100% (16%)	6% (7%)	8% (6%)	23% (20%)
Topical mismatch	100% (30%)	8% (0%)	7% (0%)	5% (0%)	13% (15%)
Name mismatch	100% (12%)	7% (1%)	6% (0%)	7% (0%)	12% (6%)
No match	15% (17%)	6% (1%)	6% (0%)	6% (0%)	6% (2%)
Total	39% (17%)	8% (7%)	6% (0%)	6% (0%)	8% (4%)

effort focused on editor-article pairs for which expertise-related terms on the editor's personal page, or a possibly matching scholar profile, were found. A high ratio of verified SMEs (69%) was found among those pairs that were assigned high scores along both of these dimensions. Only a minority of the screened examples met these strict conditions, however. We therefore further sampled editor-article pairs for which weaker indications of expertise were tracked automatically, e.g., pairs for which a scholar page has been found, but the topical similarity between the scholar's articles and the focal article was low; or no relevant scholar was found, with expertise-related words tracked on the editor's personal user page. As shown in Table 9, for those cells that were annotated in full, the ratio of pairs labeled as SMEs was substantial, in the range of 12%–43%. In contrast, the proportion of verifiable subject-matter expertise where no evidence was detected by the preliminary automatic screening was negligible (0–1%).

ACKNOWLEDGMENTS

The authors are grateful to Michael Zajko for early discussions that sparked this research project, Johannes Daxenberger for his assistance with data processing, and Revital Zilonka for advice on inclusive language. The authors also thank the Editor in Chief, the Associate Editor, and anonymous reviewers of this journal for their constructive feedback on earlier revisions of this paper.

REFERENCES

[1] Aleksi Aaltonen and Giovan Francesco Lanzara. 2015. Building governance capability in online social production: Insights from Wikipedia. *Organiz. Stud.* 36, 12 (2015), 1649–1673.

[2] Mark S. Ackerman. 1998. Augmenting organizational memory: A field study of answer garden. *ACM Trans. Inf. Syst.* 16, 3 (1998), 203–224.

[3] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge sharing and Yahoo answers: Everyone knows something. In *Proceedings of the 17th International Conference on World Wide Web.* ACM, 665–674.

[4] B. Thomas Adler and Luca De Alfaro. 2007. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web.* ACM, 261–270.

[5] B. Thomas Adler, Luca De Alfaro, Ian Pye, and Vishwanath Raman. 2008. Measuring author contributions to the Wikipedia. In *Proceedings of the 4th International Symposium on Wikis and Open Collaboration.* ACM, 15.

[6] Denise Anthony, Sean W. Smith, and Tim Williamson. 2007. The quality of open source production: Zealots and good Samaritans in the case of Wikipedia. *Rationality and Society* (2007), 32.

[7] Ofer Arazy, J. Daxenberger, H. Lifshitz-Assaf, O. Nov, and I. Gurevych. 2016. Turbulent stability of emergent roles: The dualistic nature of self-organizing knowledge co-production. *Inf. Syst. Res.* 27, 4 (2016), 792–812.

[8] Ofer Arazy and Ian Gellatly. 2013. Corporate Wikis: The effects of owners' motivation and behavior on group members' engagement. *J. Manag. Inf. Syst.* 29, 3 (2013), 87–116.

[9] O. Arazy, I. Gellatly, E. Brainin, and O. Nov. 2016. Motivation to share knowledge using Wiki technology and the moderating effect of role perceptions. *Journal of the Association for Information Science and Technology* 67, 10 (2016), 2362–2378.

[10] O. Arazy, H. Lifshitz-Assaf, O. Nov, J. Daxenberger, M. Balestra, and C. Cheshire. 2017. On the how and why of temporal dynamics of emergent roles in Wikipedia. In *Proceedings of the 20th Conference on Computer-supported Cooperative Work and Social Computing (CSCW'17).*

[11] Ofer Arazy, Aron Lindberg, Shakked Lev, Kexian Wu, and Alex Yarovoy. 2020. Emergent routines in peer-production: Examining the temporal evolution of Wikipedia's work sequences. *ACM Trans. Soc. Comput.* 3, 1 (2020), 1–24.

[12] Ofer Arazy, Aron Lindberg, Mostafa Rezaei, and Michele Samorani. forthcoming. The evolutionary trajectories of peer-produced artifacts: Group composition, the trajectories' exploration, and the quality of artifacts. *Manag. Inf. Syst. Quart.* (forthcoming).

[13] Ofer Arazy, Oded Nov, and Felipe Ortega. 2014. The Wikipedia world is not flat: On the organizational structure of online production communities. In *Proceedings of the 22nd European Conference on Information Systems.*

[14] Ofer Arazy, Oded Nov, Raymond Patterson, and Lisa Yeo. 2011. Information quality in Wikipedia: The effects of group composition and task conflict. *J. Manag. Inf. Syst.* 27, 4 (2011), 71–98.

[15] Ofer Arazy, Felipe Ortega, Oded Nov, Lisa Yeo, and Adam Balila. 2015. Functional roles and career paths in Wikipedia. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing.* ACM, 1092–1105.

[16] O. Arazy, L. Yeo, and O. Nov. 2013. Stay on the Wikipedia task: When task-related disagreements slip into personal and procedural conflicts. *J. Amer. Soc. Inf. Sci. Technol.* 64, 8 (2013), 1634–1648.
[17] Çiğdem Aslay, Neil O'Hare, Luca Maria Aiello, and Alejandro Jaimes. 2013. Competition-based networks for expert finding. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 1033–1036.
[18] N. Auray. 2012. *Online Communities and Governance Mechanisms.* Cambridge University Press.
[19] Samy A. Azer. 2014. Evaluation of gastroenterology and hepatology articles on Wikipedia: Are they suitable as learning resources for medical students? *Eur. J. Gastroent. Hepatol.* 26, 2 (2014), 155–163.
[20] Martina Balestra, Lior Zalmanson, Coye Cheshire, Ofer Arazy, and Oded Nov. 2017. It was fun, but did it last? The dynamic interplay between fun motives and contributors' activity in peer-production. *Hum.-comput. Interact.* 1 (2017), 1.
[21] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.: Series B (Methodol.)* 57, 1 (1995), 289–300.
[22] Yochai Benkler. 2002. Coase's Penguin, or, Linux and "The nature of the firm."*Yale Law J.* 112, 3 (2002), 369–446. Retrieved from http://www.jstor.org/stable/1562247.
[23] Yochai Benkler. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* Yale University Press. Retrieved from http://cyber.law.harvard.edu/wealth_of_networks/Main_Page.
[24] Carl Bereiter and Marlene Scardamalia. 1993. *Surpassing Ourselves: An Inquiry into the Nature and Implications of Expertise.* Open Court.
[25] I. Beschastnikh, T. Kriplean, and D. W. McDonald. 2008. Wikipedian self-governance in action: Motivating the policy lens. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'08).* AAAI.
[26] Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. 2008. Identifying authoritative actors in question-answering forums: The case of Yahoo! answers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 866–874.
[27] John G. Breslin, Uldis Bojars, Boanerges Aleman-Meza, Harold Boley, Malgorzata Mochol, Lyndon J. B. Nixon, Axel Polleres, and Anna V. Zhdanova. 2007. Finding experts using Internet-based discussions in online communities and associated social networks. In *Proceedings of the 1st International ExpertFinder Workshop*, Vol. 11. 38–47.
[28] Oliver Burkeman. 2006. Wikipedia: Jimmy Wales. *The Guardian, International edition.* Retrieved from https://www.theguardian.com/technology/2006/nov/04/wikipedia.web20.
[29] B. Butler, E. Joyce, and J. Pike. 2008. Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems (Session: Shared Authoring).* ACM, 1101–1110.
[30] Brian Butler, S. Sproull, Sara Kiesler, and Robert E. Kraut. 2007. *Community Effort in Online Communities: Who Does the Work and Why.* Lawrence Erlbaum Associates, London, UK, 171–194.
[31] Christopher S. Campbell, Paul P. Maglio, Alex Cozzi, and Byron Dom. 2003. Expertise identification using email communications. In *Proceedings of the 12th International Conference on Information and Knowledge Management.* ACM, 528–531.
[32] Nicholas Carr. 2005. The amorality of Web 2.0. [Blog post]. Retrieved from http://www.roughtype.com/?p=110.
[33] Xavier Carreras, Lluís Màrquez, and Jordi Girona Salgado. 2001. Boosting trees for anti-spam email filtering. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'01).*
[34] Imogen Casebourne, Chris Davies, Michelle Fernandes, and Naomi Norman. 2012. Assessing the accuracy and quality of Wikipedia entries compared to popular online encyclopaedias: A comparative preliminary study across disciplines in English, Spanish and Arabic. *Epic, Brighton, UK. Accedido o* 9, 10 (2012), 2012.
[35] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 785–794.
[36] Anna T. Cianciolo, Cynthia Matthew, Robert J. Sternberg, and Richard K. Wagner. 2006. Tacit knowledge, practical intelligence, and expertise. *The Cambridge Handbook of Expertise and Expert Performance.* Cambridge University Press, 613–632.
[37] D. V. Cicchetti and R. Heavens. 1981. A computer program for determining the significance of the difference between pairs of independently derived values of kappa or weighted kappa. *Educ. Psychol. Meas.* 41, 1 (1981), 189–193.
[38] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI'03 Workshop on Information Integration on the Web.* 73–78.
[39] Zoe Corbyn. 2011. Wikipedia wants more contributions from academics. *The Guardian* (29 Mar. 2011). Retrieved from https://www.theguardian.com/education/2011/mar/29/wikipedia-survey-academic-contributions.
[40] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2006. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 1037–1046.

[41] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. 2007. SuggestBot: Using intelligent task routing to help people find work in Wikipedia. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*. ACM, 32–41.
[42] Nick Craswell, David Hawking, Anne-Marie Vercoustre, and Peter Wilkins. 2001. P@ noptic expert: Searching for experts not just for documents. In *Ausweb Poster Proceedings, Queensland, Australia*, Vol. 15. 17.
[43] J. Daley, Barbara. 1999. Novice to expert: An exploration of how professionals learn. *Adult Educ. Quart.* 49, 4 (1999), 133–147.
[44] Bhavana Dalvi, Einat Minkov, Partha P. Talukdar, and William W. Cohen. 2015. Automatic gloss finding for a knowledge base using ontological constraints. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. ACM, 369–378.
[45] Quang-Vinh Dang and Claudia-Lavinia Ignat. 2016. Measuring quality of collaboratively edited documents: The case of Wikipedia. In *Proceedings of the 2nd IEEE International Conference on Collaboration and Internet Computing (CIC'16)*. IEEE, 266–275.
[46] Gianluca Demartini. 2007. Finding experts using Wikipedia. In *Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics*, Vol. 290. 33–41.
[47] H. L. Dreyfus, S. E. Dreyfus, and T. Athanasiou. 2000. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Simon and Schuster.
[48] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Trans. Inf. Syst.* 29, 2 (2011), 8.
[49] K. Anders Ericsson. 2006. *An Introduction to the Cambridge Handbook of Expertise and Expert Performance: Its Development, Organization, and Content*. Cambridge University Press.
[50] K. Anders Ericsson, Ralf T. Krampe, and Clemens Tesch-Römer. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* 100, 3 (1993), 363.
[51] K. Anders Ericsson and Jacqui Smith. 1991. *Prospects and Limits of the Empirical Study of Expertise: An Introduction*. Cambridge University Press.
[52] Robert D. Finn, Paul P. Gardner, and Alex Bateman. 2012. Making your database available through Wikipedia: The pros and cons. *Nucleic Acids Res.* 40, D1 (2012), D9–D12.
[53] J. L. Fleiss. 1981. *The Measurement of Interrater Agreement* (2nd ed.). John Wiley, New York, 212–236.
[54] Andrea Forte, Vanessa Larco, and Amy Bruckman. 2009. Decentralization in Wikipedia governance. *J. Manag. Inf. Syst.* 26, 1 (2009), 49–72.
[55] Wikimedia Foundation. 2011. Wikipedia editors do it for fun: First results of our 2011 editor survey. [Blog post]. Retrieved from https://diff.wikimedia.org/2011/06/10/wikipedia-editors-do-it-for-fun-first-results-of-our-2011-editor-survey/.
[56] Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07)*, Vol. 7. 1606–1611.
[57] Jay R. Galbraith. 1974. Organization design: An information processing view. *Interfaces* 4, 3 (1974), 28–36.
[58] Paul P. Gardner, Jennifer Daub, John Tate, Benjamin L. Moore, Isabelle H. Osuch, Sam Griffiths-Jones, Robert D. Finn, Eric P. Nawrocki, Diana L. Kolbe, Sean R. Eddy, and Alex Bateman. 2011. Rfam: Wikipedia, clans and the "decimal" release. *Nucl. Acids Res.* 39, suppl-1 (2011), D141–D145.
[59] Luis Garicano. 2000. Hierarchies and the organization of knowledge in production. *J. Polit. Econ.* 108, 5 (2000), 874–904.
[60] Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature* 438 (2005), 900–901. DOI: https://doi.org/10.1038/438900a
[61] Eric Goldman. 2010. Wikipedia's labor squeeze and its consequences. *J. Telecommun. High Technol. Law* 8, 1 (2010), 157–184.
[62] Shane Greenstein and Feng Zhu. 2018. Do experts or crowd-based models produce more bias? Evidence from Encyclopædia Britannica and Wikipedia. *MIS Quarterly* 42, 3 (2018), 945–959.
[63] Guy J. Groen and Vimla L. Patel. 1990. *Professional and Novice Expertise in Medicine*. Erlbaum, Hillsdale, NJ, 35–44.
[64] Aaron Halfaker, Aniket Kittur, Robert Kraut, and John Riedl. 2009. A jury of your peers: Quality, experience and ownership in Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*. ACM, 15.
[65] Todd Holloway, Miran Bozicevic, and Katy Börner. 2007. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity* 12, 3 (2007), 30–40.
[66] A. Kittur, B. Suh, and E. H. Chi. 2009. What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In *Proceedings of the 27th Annual CHI Conference on Human Factors in Computing Systems (CHI'09)*.
[67] Sotiris B. Kotsiantis, I. Zaharakis, and P. Pintelas. 2007. Supervised machine learning: A review of classification techniques. *Emerg. Artif. Intell. Applic. Comput. Eng.* 160 (2007), 3–24.

[68] T. Kriplean, I. Beschastnikh, D. W. McDonald, and S. A. Golder. 2007. Community, consensus, coercion, control: CS*W or how policy mediates mass participation. In *Proceedings of the Conference on Supporting Group Work*. ACM, 167–176.

[69] Jie Li, Harold Boley, Virenda Bhavsar, and Jing Mei. 2006. Expert finding for eCollaboration using FOAF with RuleML. In *Montreal Conference of eTechnologies*, Vol. 2006. 53–65.

[70] Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the Wikipedia and their impact on article quality. *ACM Trans. Manag. Inf. Syst.* 2, 2 (2011), 11.

[71] Darren W. Logan, Massimo Sandal, Paul P. Gardner, Magnus Manske, and Alex Bateman. 2010. Ten simple rules for editing Wikipedia. *PLoS Comput. Biol.* 6, 9 (2010), e1000941.

[72] Omid Madani, Manfred Georg, and David A. Ross. 2012. On using nearly-independent feature families for high precision and confidence. In *Proceedings of the Asian Conference on Machine Learning*. 269–284.

[73] Ann Majchrzak and Arvind Malhotra. 2019. *Unleashing the Crowd: Collaborative Solutions to Wicked Business and Societal Problems.* Springer Nature.

[74] Ann Majchrzak, Christian Wagner, and Dave Yates. 2013. The impact of shaping on knowledge reuse for organizational improvement with wikis. *MIS Quart.* 37, 2 (2013), 455–469.

[75] Thomas W. Malone. 2004. *The Future of Work: How the New Order of Business Will Shape Your Organization, Your Management Style, and Your Life.* Harvard Business School Press, Boston, MA.

[76] David Wright Mcdonald. 2000. *Supporting Nuance in Groupware Design: Moving from Naturalistic Expertise Location to Expertise Recommendation.* PhD Thesis, University of California, Irvine.

[77] Harald A. Mieg. 2001. *The Social Psychology of Expertise: Case Studies in Research, Professional Domains, and Expert Roles.* Psychology Press.

[78] Einat Minkov, Richard Wang, Anthony Tomasic, and William Cohen. 2006. NER systems that suit user's preferences: Adjusting the recall-precision trade-off for entity extraction. In *Proceedings of the Human Language Technology Conference of the NAACL*.

[79] Dana Movshovitz-Attias, Yair Movshovitz-Attias, Peter Steenkiste, and Christos Faloutsos. 2013. Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 886–893.

[80] Yiftach Nagar. 2012. What do you think: The structuring of an online community as a collective-sensemaking process. In *Proceedings of the ACM Conference on Computer Supported Collaborative Work*. ACM, 393–402.

[81] Sabine Niederer and José van Dijck. 2010. Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media Soc.* 12, 8 (2010), 1368–1387.

[82] Joel Nothman, Hanmin Qin, and Roman Yurchak. 2018. Stop word lists in free open-source software packages. In *Proceedings of the Workshop for NLP Open Source Software (NLP-OSS'18)*. 7–12.

[83] Mathieu O'Neil. 2010. Shirky and Sanger, or the costs of crowdsourcing. *J. Sci. Commun.* 9, 1 (2010), 1–6.

[84] Tim O'Reilly. 2007. What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies* 1, 65 (2007), 17–37.

[85] Julian Edgerton Orr. 1996. *Talking about Machines: An Ethnography of a Modern Job.* Cornell University Press, Ithaca, NY.

[86] Aditya Pal, F. Maxwell Harper, and Joseph A. Konstan. 2012. Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Trans. Inf. Syst.* 30, 2 (2012), 10.

[87] Vimla L. Patel and Guy J. Groen. 1991. *The General and Specific Nature of Medical Expertise: A Critical Look.* Cambridge University Press, 93–125.

[88] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (2011), 2825–2830.

[89] Ronaldo C. Prati, Gustavo E. A. P. A. Batista, and Diego F. Silva. 2015. Class imbalance revisited: A new experimental setup to assess the performance of treatment methods. *Knowl. Inf. Syst.* 45, 1 (2015), 247–270.

[90] Xiang Ren, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, and Jiawei Han. 2016. Label noise reduction in entity typing by heterogeneous partial-label embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM, 1825–1834.

[91] Oleg Rokhlenko and Idan Szpektor. 2013. Generating synthetic comparable questions for news articles. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 742–751.

[92] Karol G. Ross, Jennifer L. Shafer, and Gary Klein. 2006. *Professional Judgments and "Naturalistic Decision Making."*

[93] Larry Sanger. 2004. Why Wikipedia Must Jettison Its Anti-Elitism [Blog post]. Retrieved from https://web.archive.org/web/20200807044910/https://larrysanger.org/2004/12/why-wikipedia-must-jettison-its-anti-elitism/.

[94] Lawrence M. Sanger. 2009. The fate of expertise after Wikipedia. *Episteme* 6, 1 (2009), 52–73. DOI: https://doi.org/10.3366/e1742360008000543

[95] Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. 2017. Building automated vandalism detection tools for Wikidata. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1647–1654.

[96] Andreas Schroeder and Christian Wagner. 2012. Governance of open content creation: A conceptualization and analysis of control and guiding mechanisms in the open content domain. *J. Amer. Soc. Inf. Sci. Technol.* 63, 10 (2012), 1947–1959. DOI: https://doi.org/doi:10.1002/asi.22657

[97] Clay Shirky. 2008. *Here Comes Everybody: The Power of Organizing without Organizations*. Penguin Press.

[98] Stefan Siersdorfer, Sergej Sizov, and Gerhard Weikum. 2004. Goal-oriented methods and meta methods for document classification and their parameter tuning. In *Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM'04)*. ACM, 59–68.

[99] Ivan Srba and Maria Bielikova. 2016. A comprehensive survey and classification of approaches for community question answering. *ACM Trans. Web* 10, 3 (2016), 18.

[100] B. Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. 2008. Information quality work organization in Wikipedia. *J. Amer. Soc. Inf. Sci. Technol.* 59, 6 (2008), 983–1001. DOI: https://doi.org/10.1002/asi.v59:6

[101] Olof Sundin. 2011. Janitors of knowledge: Constructing knowledge in the everyday life of Wikipedia editors. *J. Document.* 67, 5 (2011), 840–862. DOI: https://doi.org/doi:10.1108/00220411111164709

[102] C. R. Sunstein. 2006. *Infotopia: How Many Minds Produce Knowledge*. Oxford University Press.

[103] J. Surowiecki. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, New York.

[104] D. Tapscott and A. D. Williams. 2006. *Wikinomics: How Mass Collaboration Changes Everything*. Portfolio.

[105] Dario Taraborelli, Daniel Mietchen, Panagiota Alevizou, and Alastair Gill. 2011. Expert participation on Wikipedia: Barriers and opportunities. Wikimania 2011, 4-6 Aug 2011, Haifa, Israel. Retrieved from http://oro.open.ac.uk/32619/.

[106] Neil Thompson and Douglas Hanley. 2018. *Science is Shaped by Wikipedia: Evidence from a Randomized Control Trial.* MIT Sloan School Working Paper 5238-17.

[107] Jeffrey W. Treem and Paul M. Leonardi. 2016. *Expertise, Communication, and Organizing*. Oxford University Press.

[108] Reuth Vexler and Einat Minkov. 2016. Multi-source named entity typing for social media. In *Proceedings of the 6th Named Entities Workshop (NEWS'16), collocated with ACL*.

[109] F. Viegas, M. Wattenberg, and M. McKeon. 2007. The hidden order of Wikipedia. In *Proceedings of the 2nd International Conference on Online Communities and Social Computing (OCSD'07), held as part of HCI International*. Springer, 445–454.

[110] Jimmy Wales. 2005. Jimmy Wales: The birth of Wikipedia [Video file]. Retrieved from https://www.ted.com/talks/jimmy_wales_the_birth_of_wikipedia.

[111] Samuel S. Wineburg. 1991. Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *J. Educ. Psychol.* 83, 1 (1991), 73–87.

[112] W. R. Rice. 1989. Analyzing tables of statistical test. *Evolution* 43 (1989), 223–225.

[113] J. Yan, L. Du, X. Yao, and Li Shen. 2016. Machine learning in brain imaging genomics. In *Machine Learning and Medical Imaging*. Elsevier, 411–434.

[114] Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who did what: Editor role identification in Wikipedia. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*.

[115] Jun Zhang, Mark S. Ackerman, and Lada Adamic. 2007. Expertise networks in online communities: Structure and algorithms. In *Proceedings of the 16th International Conference on World Wide Web*. ACM, 221–230.

[116] Jianhan Zhu, Alexandre L. Goncalves, Victoria S. Uren, Enrico Motta, and R. Pacheco. 2005. Mining web data for competency management. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE, 94–100.

Received May 2019; revised April 2020; accepted August 2020