Users roles identification on online crowdsourced Q&A platforms and encyclopedias: a survey

Akrati Saxena¹ · Harita Reddy²

Received: 19 September 2020 / Accepted: 17 May 2021 / Published online: 8 June 2021
© The Author(s) 2021

Abstract

Online informal learning and knowledge-sharing platforms, such as Stack Exchange, Reddit, and Wikipedia have been a great source of learning. Millions of people access these websites to ask questions, answer the questions, view answers, or check facts. However, one interesting question that has always attracted the researchers is if all the users share equally on these portals, and if not then how the contribution varies across users, and how it is distributed? Do different users focus on different kinds of activities and play specific roles? In this work, we present a survey of users' social roles that have been identified on online discussion and Q&A platforms including Usenet newsgroups, Reddit, Stack Exchange, and MOOC forums, as well as on crowdsourced encyclopedias, such as Wikipedia, and Baidu Baike, where users interact with each other through talk pages. We discuss the state of the art on capturing the variety of users roles through different methods including the construction of user network, analysis of content posted by users, temporal analysis of user activity, posting frequency, and so on. We also discuss the available datasets and APIs to collect the data from these platforms for further research. The survey is concluded with open research questions.

Keywords Social roles · Role modeling · Discussion and Q&A · Crowdsourced encyclopedia

Introduction

In the modern era of the Internet, online question-answering (Q&A) and knowledge-sharing platforms have become an important source of gaining knowledge, discussing, and clearing doubts related to a variety of fields ranging from literature

✉ Akrati Saxena
a.saxena@tue.nl

¹ Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

² National Institute of Technology Karnataka, Mangalore, India

Table 1 Number of registered users on top-20 English Q&A stack exchange websites [87] (as on May 10, 2020)

Website	Users	Visits/day	Questions	Answers	% Answered (%)
Stack overflow	12 m	10 m	19 m	29 m	70
Super user	889 k	775 k	443 k	633 k	65
AskUbuntu	813 k	496 k	348 k	440 k	65
Mathematics	655 k	305 k	1.2 m	1.7 m	77
Meta stack exchange	569 k	11 k	93 k	138 k	87
Server Fault	462 k	321 k	289 k	468 k	77
Unix and Linux	370 k	349 k	183 k	272 k	76
Software engineering	308 k	56 k	56 k	163 k	94
English language and usage	292 k	387 k	112 k	260 k	94
Ask different	284 k	230 k	105 k	161 k	86
Cross validated	235 k	119 k	159 k	156 k	64
Android enthusiasts	221 k	79 k	53 k	64 k	58
Physics	197 k	79 k	160 k	233 k	80
Code review	194 k	41 k	66 k	103 k	90
Electrical engineering	192 k	93 k	136 k	234 k	87
Information Security	190 k	55 k	58 k	104 k	91
TeX–LaTeX	178 k	126 k	196 k	255 k	92
Database administrators	175 k	116 k	83 k	112 k	81
Web applications	171 k	77 k	30 k	42 k	77
Arqade	157 k	376 k	91 k	150 k	93

to programming. Quora, Yahoo Answers, Reddit, Stack Overflow are some of the popular websites that provide a platform where users can post questions and obtain answers from other users of the community who claim to have knowledge about the specific topics. One of the most popular collections of online Q&A communities is Stack Exchange (launched on September 2009), which has a total of 173 websites (as per May 2020¹) related to technology, culture, arts, science, profession, and business [87]. As per the data available on 10 May 2020, about 12 Million users are registered on Stack Overflow, i.e., the most popular Q&A website of Stack Exchange for programmers. Table 1 shows the details of top-20 Q&A websites associated with Stack Exchange based on the number of registered users.

On online Q&A platforms and discussion forums, users can choose their areas of interest and the latest discussions will be displayed on their feed based on their interest. For example, Stack Overflow has tags for the questions to indicate the associated sub-topics and the users can choose to be notified of the activity related to specific tags that they are interested in. On Reddit, users can subscribe to *subreddits*

¹ More details including the number of monthly visits, page-views, questions asked, submitted answers, upvotes, downvotes, and comments are available at <https://stackexchange.com/about>.

dedicated to specific topics of their choice. The websites may also suggest a user to write an answer for a particular question based on the topics of users' interest and answering history.

These Q&A websites enable users to get information from other users by asking specific questions. A strength of these websites is that multiple users can answer the same question, providing more than one explanation from different perspectives. Moreover, many of these websites have the provision for upvoting (if you like or agree with the answer) or downvoting (if you dislike or do not agree with the answer) to indicate the correctness and worthiness of the answer and to show the confidence in the knowledge of the answerer who might not be an expert in the given domain, unlike the formal educational system. These websites may also employ badge system where badges are rewarded to users as a testimony to their contributions to the community, and these badges can be awarded for the specific types of contributions users can make [4]. Badges are incentives for users to increase their reputation and motivate them to answer and ask more questions [105]. A study on the effect of the badges performed on Stack Overflow has suggested that not only the badges increase users' participation, but may also have a positive impact on users' participation in activities which are not considered for awarding a badge [58]. Some platforms also provide monetary benefits to reward users' contributions, for example, Quora's Partner Program, which was started with an intention of paying contributors based on the revenue generated from the ads that came on their content [80].

In online knowledge-sharing platforms, people contribute content, respond to others' content, and discuss a topic of interest as per their comfort. Researchers were always interested in understanding the dynamics of users' interactions and knowledge sharing among people. Initial work in this direction was done by identifying the user roles on Usenet newsgroups, which were one of the earliest discussion platforms using the Internet [17, 28, 39, 47, 90, 94]. The study of the content posted by various users, the frequency of posting, and the overall activity of users has helped in identifying the differences between diverse types of users on the newsgroups, including prolific users who contribute heavily to the community and are likely to be known by many users, and *Lurkers* who just prefer to read the posts and do not make any significant contribution to it [39]. However, some users just ask questions, whereas other users mainly answer the questions posted on the forums and adopt the role of an *Answer Person* [90]. There exist experts who are selective in answering questions and answer only the most difficult ones, whereas some users try to answer as many questions as possible in order to build up their reputation [105]. Some users like to indulge in discussions or debates; such users often have a dense network of interactions with other users on threads, and they tend to contribute multiple messages within a particular discussion thread [18, 90, 91, 94]. There are also users who try to disrupt the discussion forums by posting irrelevant content, bringing up fruitless debates, and making aggressive statements [39, 90, 91]. In this work, we will survey different types of social roles obtained by the users on several Q&A websites.

Apart from Q&A platforms, there exist websites that serve as repositories of articles related to different topics, and some of these websites use crowdsourcing (allowing users from across the world to edit articles) for generating the content. In our discussion, we will refer them as *Online Crowdsourced Encyclopedia*. Some

examples of such crowdsourcing websites are Wikipedia, WikiHow (an open database containing a number of how-to-guides), Baidu Baike (a Chinese language encyclopedia owned by Baidu), and so on. These crowdsourced platforms differ from the discussion and Q&A platforms in the way that the primary goal of the editors is to collaboratively build an artifact [93]. Similar to Q&A forums, their contributors also adopt a variety of roles. For example, some are experts in their area and focus on editing pages related to that, often discussing in the content talk pages, while some are technical editors who focus on issues related to formatting and grammar [93, 102]. The user roles can be differentiated on the basis of types of edits they make, how their edits are distributed across various articles, the structure of their networks, their article editing history, and their interaction on the talk pages [63, 93, 102].

Learning and knowledge sharing has become highly decentralized with the help of these platforms. Users acquire different roles on different platforms based on their interests, expertise, and behavior. In this work, we discuss social roles of users on online crowdsourced Q&A platforms and encyclopedias. As per the best of our knowledge, this is the first systematic survey of its kind. The paper is structured as follows: In the next section, we discuss how the roles are defined in a society and online community. “[Discussion and Q&A platforms](#)” and “[Online crowdsourced encyclopedias](#)” provide a comprehensive survey of the research on identifying users roles on online discussion and Q&A platforms, including Usenet Newsgroups, Reddit, Stack Exchange, MOOCs, and crowdsourced encyclopedias including Wikipedia, Baidu Baike, respectively. In “[Datasets](#)”, we discuss the available datasets which are used in the considered studies and the APIs that can be used to collect the dataset. The paper is concluded in “[Conclusion](#)” with future directions.

Definitions of social role

People acquire different roles in various tasks of day-to-day life. Roles may, therefore, be referred to as ‘patterned human behavior’ [14]. A common belief connects roles to social positions or statuses, which are the identities that are associated with groups of people [14]. The well-known roles based on social positions which are recognized formally are either obtained as a profession, such as the role of a doctor and a lawyer, or assigned legally, such as the director of a university and the vice president of a company. Another kind of role is defined based on personal relationships, for example, mother, father, siblings, friends, etc. Certain role theorists argue that people are stimulated to play a certain role due to the expectations of others as well as oneself about their behavior [14]. Those who manifest a particular role exhibit a set of behaviors that are expected out of them; for example, a school teacher interacts with students and tries to help them in studies because this is expected from the teacher. Golder and Donath [39] summarize a social role in terms of two aspects: (1) a person’s skills, responsibilities and privileges, and (2) the social context surrounding the person which includes the people’s expectations about her behavior. Social roles also have been studied in online social networks using users’ behavior, their position with respect to other users, and their virtual online identities [29].

The concept of social role is not only limited to society, but it also applies to users' interactions and behavior on online crowdsourced portals. Electronic media has facilitated several online discussion and Q&A platforms which have given rise to several online communities for users. For example, a person learning computer programming who mostly posts questions and comments on Stack Overflow may acquire the role of *Information Seeker*, while there are few users with expertise in their subject who focus on providing answers to questions and are referred to as *Experts* [74]. In an online crowdsourcing encyclopedia, such as Wikipedia, the work of a user is mainly seen as making edits to the articles; however, the organization of the platform leads to different types of user roles ranging from copy editing to protect the pages from vandalism [102]. The study of structural roles in online communities brings attention to the existing structure enforced by the online platform, including access privileges of users [8]. With these structural roles, users will have certain actions and responsibilities. On the other hand, modeling of emergent roles on such online communities breaks free the dependency on platform-enforced user privileges, expectations, norms, and hierarchies, and focuses on the self-organizing nature of the online knowledge communities which involves users choosing what to do, when to do, and how to do on the platform [7, 9]. Next, we discuss various user roles observed on online Question-and-Answer (Q&A) platforms. We begin with the heavily researched Usenet Newsgroups and continue with other popular platforms including Reddit, Stack Exchange, and discussion forums on Massive Open Online Courses (MOOCs).

Discussion and Q&A platforms

Online Q&A platforms and discussion boards enable users to ask questions, post answers or content and vote on the content, which can evolve into discussions. We club discussion and Q&A platforms under the same section as their structure is very similar; discussion forums include posting of questions and users answering to them, and Q&A platforms also have intense discussion component.

Online discussion and Q&A platforms have some commonalities and differences. Stack Exchange is a question and answering oriented platform where users seek an answer to a particular question. In contrast, Reddit and Usenet Newsgroups are mainly discussion platforms. MOOC forums can be designed to be either Q&A oriented or discussion oriented. The Reddit platform allows users to post content that includes text, images, and videos without seeking knowledge, and these posts can lead to further discussion threads. The modern platforms, including Reddit and Stack Exchange, provide the feature of users voting for answers to enable the good quality content to be brought to the top and make it more visible, while the lower quality and incorrect content will lose the visibility.² These platforms also implement some form of point-scoring for the users based on their activity and contributions, such as the *points-system* on Stack Exchange and *karma* on Reddit.

² <https://stackoverflow.com/help/why-vote>.

Different platforms have different ways to organize questions or discussions on specific topics. For instance, Reddit allows the creation of different sub-communities called *subreddits*, which are dedicated to particular topics. Usenet Newsgroups are organized hierarchically based on the topics, with the sub-topic newsgroups falling below the main topic. Stack Exchange is a collection of multiple websites, where each website is dedicated to a particular field, and the questions within each website are tagged with the respective sub-topics.

In this section, we will discuss social roles that users obtain on Q&A platforms and discussion boards.

Usenet Newsgroups

Usenet Newsgroups were one of the earliest discussion platforms that allowed users to post their views on a variety of topics using the internet. Usenet was highly popular until the arrival of the World Wide Web (WWW), with which Usenet lost its importance to other discussion forums that had an edge due to the ease of access through browsers. Generally, one newsgroup focuses on one particular topic of interest; the users' published messages on the newsgroup are referred to as *posts* [99]. All the users do not show the same levels of participation- a small segment of users contribute a large share of posts whereas many users post scarcely [97]. An observational study of features like the content that users post, posting frequency, and users' activity spanning across different groups helps in identifying the typical user roles in the newsgroups [39]. Unlike Celebrities, *Newbies* are the newly joined users who usually lack the proficiency in communication required for a newsgroup, may not have the knowledge of newsgroup-specific acronyms, and are extra-attentive to make their posts acceptable to the community. Factors like high participation and good communicative competence enable a user to acquire positive feedback from the newsgroup community [50]. The positive feedback can propel the user to become highly influential in the community and attain the role of a *Celebrity* [39]. There also exist users known as *Lurkers*, who do not post anything but only read others' posts, which could be due to a number of reasons including not liking to post publicly [39, 75]. Sun et al. [88] classified the lurking reasons into four main categories, (1) environmental influence, (2) personal preference, (3) individual-group relationship, and (4) security consideration. They further proposed strategies, such as a better user-friendly environment, motivating users for participation, and guidance for new users, for de-lurking the user behavior and their active participation in online communities.

Analysis of directed ego networks in different types of newsgroups showed that the network structure of authors was dependant on the nature of the group [28]. In a question and answer based newsgroup, for example, in a technical group, a large share of replies of active answer people are to users with less outgoing links. However, newsgroups with intense heated discussions have users who reply mostly to other well-connected users having a high density of connections and high reciprocity of ties. In newsgroups that are created with the intention to offer social support, the majority of users show good participation in terms of incoming and outgoing

links. Welser et al. [94] focused on identifying the role signature of answer people, i.e., the users who provided responses on threads. The authors identified users' contribution volume through metrics like the number of threads the user posts to and the number of messages posted in each thread. In a question-and-answer oriented newsgroup like *Server General*, the most active participant was found to be an *Answer Person*, whose ego network had many outgoing ties to nodes with a low degree, resembling a star. On the other hand, in a discussion-oriented newsgroup like *Kites*, the *Discussion Person* role was found to be prominent, with her network being dense with connections to many other well-connected alters.

Prolific users in a community, including celebrities, play different kinds of roles in the process of community building [39]. The users who have spent a significant amount of time and energy in a newsgroup community often play the role of protecting the community. These users may remove their newsgroup community from threads on irrelevant cross-posts, i.e., some posts which are posted to multiple newsgroups without the approval of the members of those newsgroups. They may also keep a rein on free-riders, i.e., those users who only post questions and get the required knowledge from the community members' responses, but do not make any knowledge contribution of their own to the community [39]. There are some well-established users who try to expand the scope of the conversations in the newsgroup community to include some other connected topics apart from the primary discussion topic. Some users also play a negative role, in contrast with the users who contribute a lot to the community and have a positive impact on its success [39]. Golder and Donath [39] note the existence of *Flamers*, who tend to wander from one group to another and make confrontational statements against the members of the group. On the other hand, *Trolls* pretend to be a legitimate participant by exhibiting good communication proficiency but actually bring up fruitless arguments, thus wasting others' time. Somewhat similar to Trolls, *Ranters* try to bring up fruitless debates and post frequently in their particular area of interest in which they have a particular agenda.

Social networks constructed based on the replies by the users to others' messages can be used to visualize user roles in the Usenet newsgroups [90]. The analysis of Turner et al. [90] was not only based on the users' networks but also their posting behavior, including the number of threads they initiate or contribute to. They observed that an *Answer Person* gives many responses to others and has a dense network, while a *Questioner* has limited connections as she replies very less and mostly focuses on posting questions. Similar to [39], they also identified *Trolls* who exhort people into pointless debating. *Spammers* start a large number of discussions and post content irrelevant to the topic of discussion of the newsgroup. The authors also try to underline the difference between a *Conversationalist*, who enjoys engaging in discussion of ideas, and a *Flame Warrior*, who merely intends to emerge victorious in debates. Another network-based analysis approach by Himelboim et al. [47] identified the users who take the initiative of starting discussions in the newsgroups. They analyzed political discussions in the newsgroups to identify the role of a *Discussion Catalyst*, who has dense incoming ties in the directed network constructed on the basis of replies. As expected, these incoming ties indicate the replies received by the

discussion catalyst, who starts discussions. Further content analysis shows that the discussion catalysts' root posts in threads often contain content from other websites, thus highlighting them as *content importers*.

Users' activity across time and posting frequency per thread can also help in revealing the underlying author's roles through visualization. While both a *Question Asker* and an *Answer Person* have less number of posts per thread, an *Answer Person* is usually active for more number of days compared to a *Question Asker* [91]. *Debaters*, on the other hand, have a high number of posts per discussion thread, possibly due to their tendency to involve in prolonged discussions. *Bursty Contributors* are active for a lesser number of days but their number of posts in the threads they contribute is higher than average. While both *Answer Person* and *Debater* prominently reply to threads started by others, a *Balanced Conversationalist* takes a balanced approach by starting discussion threads as well as replying to threads started by other users. *Spammers* mostly start threads but do not post further on the initiated thread.

Turner and Fisher [89] identified four social types among Usenet users through a combination of manual observation of participants by the authors, interviews with leading contributors, and meta-data of the newsgroups. These roles are, (1) *Members* who are active or passive participants of newsgroups including *Posters* who regularly post content and *Lurkers* who mostly read content posted by others, (2) *Mentors*, who are the leading users, often rewarded for their constant activity, especially in providing answers, (3) *Managers*, who enforce certain rules with regards to how users should conduct themselves in the newsgroup, and (4) *Moguls*, who are the technical experts who answer difficult questions which others are unable to answer. Self-reporting of roles by users can be combined with other analysis methods for a better understanding. The approach of self-reporting through user surveys was used by Brush et al. [17], who clustered the users participating in the survey into *Key Contributors*, *Low Volume Repliers*, *Questioners*, *Readers*, and *Disengaged Observers*. They also validated these roles by analyzing the user behavioral data obtained through the users' self-reported Usenet handles. The behavioral data included the number of newsgroups the user was participating in, the number of messages posted by a user, the number of replies obtained, and the number of threads initiated by the user. For instance, the users who reported in the survey that they were key contributors were observed to have posted more messages on the newsgroups based on the quantitative analysis.

Usenet newsgroups were thus analyzed in the early research on defining user roles on discussion platforms. Roles have been mapped based on different kinds of user behaviors on the platform, including users who ask questions, users who prefer to give answers, users who prefer not to contribute but lurk, and users who contribute negatively to the groups. Some users take up leadership roles and play an important role in the development of the community. In recent times, several popular discussion forums have arisen, still, the modeling of roles on newsgroups often forms the basis for research on other forums.

Reddit

Reddit is a discussion website where users can post their content, rate, and comment on others' content, and form sub-communities dedicated to specific topics, known as *subreddits*. Examples of some subreddits include r/movies, r/LifeProTips, and r/ask-science. In subreddits, users can post their questions that eventually may evolve into discussion threads. Some users largely answer others' questions and engage less in discussions, for instance, scientists. Users with this *Answer Person* role have a large number of outgoing connections to low-degree users; their 1.5° degree ego network has a hub and spoke structure [18]. This is similar to the *Answer Person* role proposed by Welser et al. [94] for Usenet newsgroups.

Some users prefer to mainly ask questions while others prefer to answer questions. Metrics capturing question posting and answering behaviors, as well as network out-degree in a directed and weighted network of subreddit for MS Excel discussion (r/excel) highlight six social roles: *Frequent Questioner*, *Occasional Questioner*, *Occasional Answerer*, *Community Activist*, *Elaborative Questioner*, and *Experienced Answerer* [59], with 60.4% of users belonging to the category of questioners and around 25% of users being *Occasional Questioners*. Question-posting behaviors are captured by (1) how often a user posts questions, (2) how well the questions are backed up using formulas/references, and (3) other users' ratings of the questions' value. Replying behaviors, in this case, would be captured based on metrics like average length of replies and the number of replies to the base post in the discussion. Liang and Introne [59] used Gaussian Mixture Models (GMMs) [76] to cluster the users based on the metrics related to the directed-network structure which include out-degree and difference between in-degree and out-degree, question posting, and reply behaviors. GMMs are a generalization of K-means technique for clustering and are used to model data as a function of multiple finite Gaussian distributions [59, 76]. Based on aggregated edges (messages) between different role clusters in a directed network, the messages in which *Community Activists* are involved account for the largest fraction of traffic. Though there is an existence of a variety of roles, questioners constitute the majority of the users while active answerers form only a small proportion of the community.

Discussion threads are also used to capture user roles through users' replying behaviors. Network growth models can quantify the probability of a given node in a graph to become a parent of another node [44, 65], and these models are applicable to discussion trees in terms of the probability of a comment in a tree to receive a reply. The probability of a comment in an online discussion thread to elicit a reply might depend on its (1) *popularity* in terms of the number of replies it currently has (popular posts attract more replies due to *preferential attachment*), (2) *novelty* of the comment which decays with time, and (3) *root bias*, the bias of the users to add a comment under the base post of the discussion rather than replying to the comments on it [40]. Lumberras et al. [65] used the growth-model along with an Expectation-Maximization algorithm to estimate parameters for user roles based on the reply behavior on Reddit discussion threads. Choi et al. [22] studied the behavior of different types of users in viral comment trees, including (1) *Initiators* who initiate discussions, (2) *Commentators* who actively add comments to the threads,

(3) *Attractors* who get many replies from other users, and (4) *Translators* who are involved in multiple communities. The comment trees are characterized based on the number of nodes in the tree, rate at which comments are added to the tree, and cascade *virality*, which is measured through Wiener Index [21]. As discussion threads become viral, the share of replies attracted by the initiators reduces while that of the other three types increases.

Users despite having a high engagement, may receive low levels of replies or reciprocity to their content. Through egocentric reply-graphs on subreddits, Morrison et al. [72] extracted features including in-degree, out-degree, submitted posts, comments per posts in which a particular user participates, comments by the user that received at least one reply, forum entropy of a user (high if a user participates in multiple forums), the ratio of peers with whom there is bidirectional interaction, and the ratio of posts in which the user participated that have bidirectional interaction. They mapped four of the roles, defined in the work on general discussion forums by Chan et al. [20], to the clusters obtained using these features. The analysis showed that a high number of users belong to *Ignored* and *Lurkers* category in subreddits. *Ignored* users are highly engaged on the forum but are unable to get replies from other users, while *lurkers* have low engagement overall. Next, they applied the Fruchterman Reingold method [31] that is a force-directed algorithm for visualizing graphs. The Fruchterman Reingold method highlighted the core of the graph prominently containing *Contributors*, who contribute high utility content which attracts more replies from other users.

It is also interesting to understand the patterns of users' participation in multiple communities or *subreddits*. A particular user often does not post actively in two communities with dissimilar topics of interest, say science and music [18]. *Translators* are the users who get involved in multiple subreddits, and help in spreading ideas across different communities; they may be identified through *Subreddit Entropy* [22]. Users playing this role tend to attract more comments in viral discussion threads, which can be modeled through comment trees. Are these Reddit user roles static? Grayson and Greene [42] analyzed the temporal changes in user roles in Reddit networks using struc2vec [81] method to create user role embeddings. The struc2vec method generates the network embedding in a d-dimensional space based on the local network structure of the nodes using word2vec algorithm. They used cosine distance between the role embeddings of individuals at different points of time. The preliminary analysis suggested that loyal and vagrant users, as defined by Hamilton et al. [43] in terms of their loyalty to their subreddit, show different extents of variations in their roles, and vagrant users show a greater tendency to change roles over time. However, on a community level, the roles seem to remain more static in nature.

Reddit provides a platform for forming communities for discussion on emerging fields on which there are not many traditional sources of knowledge. Thus, platforms like Reddit are a valuable source of information for such areas of interest and enable learners to clarify their doubts. Kou et al. [53, 101] studied a User Experience (UX) subreddit by using open coding to extract information about how each author described and presented their knowledge and experience in the field from their subreddit posts and comments. They finally obtained 5 major roles: (1) *Knowledge*

Brokers brought in new knowledge by sharing links, (2) *Translators* made posts presenting their academic insights, (3) *Experienced Practitioners* shared their mastery obtained through their long experience in the industry, (4) *Conversation Facilitators* initiated discussions, and (5) *Learners* were newcomers who intended to obtain knowledge from experts. They also constructed a weighted social network, with the edges between users based on commenting on a common thread. The network showed that most of the high centrality actors were experienced practitioners due to their active participation in commenting, whereas knowledge brokers had lesser degrees as their shared links did not garner many comments.

Future research on Reddit can focus on using features like the post-sharing behavior of the users, the tendency of a user to participate in highly popular threads, the tendency of a user to enforce social conduct or to advise other users to improve their content, and user's experience on the platform. One can further measure how roles vary with gender and what is the correlation of roles with the geographical location of the users. It will be important to understand if users from different geographies can acquire important roles in Reddit communities.

Stack exchange

Stack exchange is a collection of Q&A websites covering topics in diverse fields where each site allows users to post questions, and answers to them on a specific area managed by that website. Some of the most popular websites hosted by Stack exchange include Stack overflow (stackoverflow.com), Ask Ubuntu (askubuntu.com), Super User (superuser.com), and Mathematics Stack Exchange (math.stackexchange.com).

Stack Overflow is a popular question and answer platform for developers where users can not only post questions and answers related to computer programming but also vote, edit, and comment on others' answers. Users attempt to increase their reputation on the website by contributing high-quality questions and answers through which they receive badges, like *Socratic*, *Guru*, etc³. The incentive for advancing one's reputation drives the most active users, known as *Sparrows* who focus on answering easy questions—they constitute the most active users who provide the answers to the majority of questions. *Owls*, on the other hand, engage comparatively less but write high-quality answers to challenging and popular questions [105]. The popularity of a question is ascertained based on view count, whereas difficulty is ascertained based on the time taken for the questioner to mark an answer as an acceptable answer.

On Community Q&A websites like Stack Exchange, user roles can be defined based on the probabilistic distribution of actions taken by a user who is assuming a particular role [37]. Geigle et al. [37] defined 15 actions as outlined in Table 2. They proposed a Mixture of Dirichlet-multinomial Mixtures (MDMM) model to determine user roles, where the roles are modeled as a categorical distribution over a set

³ <https://stackoverflow.com/help/badges>

Table 2 15 Actions defined on Stack Exchange by Geigle et al. [37]

Category	Actions
Questioning	Ask a question
Answering	Answer one's own question, Answer others' question
Commenting	Comment on one's own question, Comment on others' question, Comment on own answer to one's own question, Comment on own answer to others' question, Comment on others' answer to one's own question, Comment on others' answer to others' question
Editing	Edit one's own question, Edit others' question, Edit one's own answer, Edit others' answer
Moderation	Vote on a moderating action, Moderate some post

of user actions. The ascertained roles include *Eager Asker*, *Careful Asker*, *Answerer*, *Clarifier* and *Editor/Moderator*. The distinction between the first two roles is that while an Eager Asker tends to comment on the answers to her question, a Careful Asker comments on the discussions on the question and is more likely to edit the question. An Editor/Moderator often edits others' questions. An Answerer's engagement is more with her own answer by commenting on it and updating it, whereas a Clarifier also comments on others' answers frequently.

Cluster analysis of metrics based on the number and quality of questions, answers, and comments contributed by the users on five popular websites of Stack Exchange: Super User, Mathematics, Server Fault, Programmers and Ask Ubuntu highlights the existence of ten user profiles among the contributors [33]. Hierarchical clustering [92] was used to find optimal cluster count and cluster centers before the application of non-hierarchical k-means clustering [45]. The most active user profiles extracted include *Q-A Activists*, *Answer Activists*, and *Hyperactivists*; these contributors are the most likely to provide quick answers to posts. Surprisingly, *Expert Answerers*, *Expert Questioners*, and *Expert Commenters* show less activity, which suggests that they might be selective in the discussions they choose to engage in. Not only this, but experts also show a higher tendency to leave the site. Users who provide poor quality answers (*Unskilled Answerers*) are very likely to get demotivated and eventually leave the site.

The user signatures on a Q&A website can be identified using how frequently a user answers questions. Such an accounting of monthly user activity on Stack Overflow gives four signatures: (1) *Lurkers* and *Visitors*, which includes users who do not ask or answer questions and unregistered users, (2) *Low Profile Users*, who have intermittent activity, (3) *Shooting Stars*, i.e., users who have high activity at some points of time and after that transition into low activity mode, and (4) *Community Activists*, i.e., the users who display high activity for long [67]. The user categories like Lurkers and Visitors, Low-Profile users, and Community Activists already have been observed in previous works [89]; however, Shooting Stars was a new observation revealed by the data. Santos et al. [82] used the temporal records of users' activity to identify user activity archetypes through time series clustering. Clustering was done through the k-means algorithm and the similarity between the time series was measured using the Euclidean distance. Based on the analysis of the clusters' activity patterns, the *Non-Recurring Activity Archetype* includes users with a single peak in activity and forms a majority of Stack Exchange users (88.4%). Users belonging to *Sporadic Activity Archetype*

have comparatively more activity than the former and have few isolated activity peaks on the platform over time; they constitute 10.1% of users in a Stack Exchange instance. The most active users are the rarer ones (0.2%) and belong to the *Permanent Activity Archetype*, who constantly show high levels of activity. Users belonging to *Frequent Activity Archetype*, whose proportion is 1.3% of users in a Stack Exchange instance, show regular activity but the levels of activity may vary over time.

Identification of typical types or *archetypes* of users can be upgraded using the users' behavior data at different stages. Narang et al. [74] proposed a Gaussian Hidden Markov Model based on the representation of a user's trajectory of evolution in terms of the distribution of the activities (posting an answer, posting a question, adding a comment) within each session, where a session stands for a *fundamental temporal unit* for the purpose of analysis, which in the case of Stack Exchange is one visit to the platform by a user. On the application of the model to Stack Overflow, they discovered four types of user archetypes: (1) *Experts*, who remain dedicated to the cause of answering questions over time, (2) *Information Seekers*, who join mainly to ask questions and add comments but also contribute a bit by answering questions for short periods of time, (3) *Enthusiasts*, who begin with asking questions but eventually also end up answering questions based on the knowledge gained and (4) *Facilitators*, who start with asking questions, then start answering questions and eventually restrict themselves to only adding comments. A user who initially starts as a learner in a particular topic can evolve gradually to play the role of an *Answerer* on that topic due to an increase in knowledge. Such changes in user roles could be traced by using the time-aware model (TRM) for tracing the evolution of user roles on question and answer websites including Stack Overflow based on the user, content posted by the user with a timestamp, and the role of the user in each post, which formed the base for formulating a question recommendation system for such platforms [32]. Fu [32] tested this model on Stack Overflow and compared the role distributions over time with the ground truth using distance measures including Kullback–Leibler divergence [98] and Residual Sum of Squares. The models outperformed other baseline models in terms of accuracy of measuring user-role evolution for each time period. A drawback highlighted in the model is considering time as discrete slices of intervals, rather than continuous.

Stack Exchange hosts some of the most active Q&A platforms and can be used for the further analysis of various role dynamics. One can analyze whether a user can have multiple roles across different topics, for instance, an expert in one domain can be highly active but not have expertise in another. In future work, researchers can also focus on comparing the similarities and differences in user roles across different stack exchange websites, for example, Stack Overflow, which is used for programming, with Movies and TV website. The comparison can also be made between websites with similar topics but possibly varying levels of professional competency of the users.

An interesting research direction that we suggest is to study the correlation between user roles and gender. This could be extremely insightful in fields like programming which have high gender disparities. Previous works have suggested that the mean reputation score of men is higher than women on StackOverflow,

and this is possible due to the differences in their behavior on the website; women showed a greater tendency of asking questions, while men inclined towards answering [69]. Additionally, men on StackOverflow showed a greater tendency to complete their bio and provide links to their Github and other accounts. These gender-based differences pertaining to participation may be used to correlate the distribution of roles between men and women. The difficulty that arises when analyzing websites like StackOverflow on the basis of gender is that gender information is not publicly available. However, there have been studies exploring ways of determining users' gender based on their name, information derived from other online social networks where gender is recorded, user image processing, and using information from their GitHub accounts [61].

Massive open online courses (MOOCs)

MOOCs are becoming a popular source of online learning in recent years. MOOCs not only provide a platform for video lectures but also provide forums for discussion with peers on the subject. These discussion threads consist of posts that seek as well as give information [46]. The semantic analysis based on users' participation in common threads and their network structure can be used to compute user similarity [64, 96]. The use of blockmodeling [25] based on social and semantic similarity of users on a Coursera MOOC highlighted three user roles [46]. The first role comprises of the most engaged *core* users who participate in the forums to both gain and furnish information. The second role constitutes those who only gain information and thus have only incoming ties, i.e., *Information Seekers*, whereas the third role comprises of users who only give information and have a certain expertise in the topic, i.e., *Information Providers*.

More fine-grained role-modeling based on structural patterns in two Coursera engineering MOOCs shows the existence of *One-time Help Seekers* and *One-time Help Givers* [15]. It is also observed that *Active Participants* are able to benefit from the forum discussions and score grades comparable to *One-time Help Givers*, who are students with greater expertise in the subject. Some students on these MOOC discussion forums are *Superposters*, i.e., the students posting the most persistently on the forums [48]. These students are found to perform well and exhibit high engagement in all the courses they enroll in. They post useful content and are essential for maintaining the health of the forums. Role identification based on the metrics related to collaboration, communication, and activity of students in MOOCs can also help in predicting the performance of teams [104]. Network structure and network-based metrics like centrality measures [83] have been used to predict students' achievements on MOOCs; it can be used further for role modeling [84].

Other discussion platforms

Several other discussion forums have been analyzed to further understand user roles and their dynamics. *Answer Person* and *Discussion Person* are commonly observed user roles that can be identified using answer person score and discussion person score; these scores can be computed based on the number of threads in which the user has contributed posts, an average of the number of posts made in each thread, and density of a user's local network [13]. We will now briefly discuss social roles identified on some other discussion platforms.

İnci Sözlük platform More active users in terms of contributing content and motivating others to do so tend to have more information flowing through them. Akar and Mardikyan [2] analyzed İnci Sözlük, i.e., a Turkish online discussion forum, using betweenness centrality [30] that measures the importance of a node based on how often that particular node lies on the shortest path between any two nodes in a network. The authors highlighted the greater betweenness centrality of *Content Generators* and *Socializers* compared to the *Passive Members* and *Visitors*. The users are categorized into four sub-communities, where the most active sub-community is comprised of *Socializers*, who contribute a large amount of content, as well as interact with others [3]. *Content Generators* also often visit the community and are next to *Socializers* in terms of the volume of content they generate. *Visitors* only visit the forums and show less activity in terms of initiating topics or contributing content. The *Passive Members* could be considered somewhat similar to *Lurkers* identified in the Usenet newsgroups' analysis [39, 75]; they rarely initiate topic discussions or generate content.

Boards.ie discussion boards White et al. [95] differentiated user profiles on the Irish discussion Boards called Boards.ie (<https://www.boards.ie>), that is a discussion forum in Ireland for a variety of topics, through the features of their egocentric network, including in-degree, out-degree, and reciprocity. Through their mixed membership models, they accounted for different user roles having some characteristics in common. All users do not provide and receive the same amount of reciprocity for the content on the communities. It is noted that *Problem Solvers* usually exhibit reciprocity in their interactions and end discussions in the thread through their advice. The network science metrics, user popularity, user's tendency to initiate discussions, and discussion persistence of the user have been used to determine the user roles on Boards.ie platform [5, 20]. Some of the metrics include in-degree ratio based on the proportion of the users replying to that particular user, the proportion of user's posts that received a reply, proportion of threads initiated by the user, and the average number of posts per thread by the user. A *Popular Initiator* has a high in-degree ratio and high thread initiation, while an *Elitist* attracts a lesser proportion of replies. A *Joining Conversationalist* initiates fewer threads but has a high average number of posts per thread.

Salon24 Among the influential users, some try to improve their position in the online network by contributing posts and adding comments to only their posts, while some

others play an active role in commenting on others' posts as well. Gliwa et al. [38] analyzed such user roles on a blogosphere named *Salon24*⁴ by collecting statistical data of posts and comments by the users. They differentiated between *Selfish Influential User*, who is influential but mainly focuses on posting content in the context of her own posts, and a *Social Influential User*, who contributes posts and comments in the context of other users' posts as well. Among the leading roles in the blogosphere, *Influential Commentators* have a major share. Influential Commentators actively post comments which have high impact; the impact is measured based on the number of users who reference the author's comments. Another study [41] identified *Superparticipants*, which includes *Super Posters*, *Agenda Setters*, and *Facilitators*, who despite being dominant in terms of quantity of posts, play a positive role on the forums and promote inclusive discussions by helping and empathizing with other participants, which is shown through qualitative analysis.

New to Java Forum New to Java is a discussion forum for developers that especially focus on newcomers as information seekers or new learners are the users who mainly get benefits from the online technical forums. Analysis of newcomers to *New to Java* forum based on their activity like posting questions, clarifications, and responses showed that the majority of newcomers in the forum are *One-time Help Seekers*, who do not follow up on their question or seek further clarifications on the answers provided [86]. On the contrary, a *Regular Help Seeker* is active in posting questions and asking for clarifications and is an example for a newcomer seeking knowledge but not willing to post answers. *In-transition Help Seekers* are neither completely new nor an expert and take part in both asking and responding to questions on the forums.

Answerbag Q&A platform Q&A communities often have certain regulations regarding the type of questions that should be posted on the forum. Asking homework questions directly is often discouraged and the content analysis of answers to such questions on *Answerbag* Q&A community suggested that users often ask the questioner to give a roadmap of how the question would be solved, and sometimes the users may even furnish wrong answers on purpose to reproach the questioner for directly posting answers of assignment questions [35]. Users can thus be differentiated based on their intention behind posting of questions, (1) *Sloths*, who post their homework questions directly and do not engage in the community further, and (2) *Seekers*, who actively interact with the community through discussions. Another analysis on the *Answerbag* community [34] outlined the presence of *Specialists* and *Synthesists* among the answerers. *Specialists* answer questions related to their area of expertise purely based on their knowledge and experience, and are highly valued in legal and financial fields. *Synthesists* on the other hand, are not experts and explicitly refer to various sources of information in their answer content.

⁴ Salon24 is a Polish blogging website, <https://www.salon24.pl>.

Gazeta.pl Forum Morzy [73] analyzed the data of discussion forums on bicycles and banks from <https://www.Gazeta.pl> website. Directed egocentric sociograms constructed for users can help in differentiating the star patterns obtained for experts and trolls. The authors noted that the star pattern of egocentric networks obtained for *Experts* who answer many questions have the user radially connected to many other isolated nodes. The edges in the case of experts are outgoing edges, as the experts answer others' questions. On the other hand, *Trolls*, who initiate controversial and heated threads have an inverted star, i.e., there are many incoming edges from other users who replied to their threads. The work also identified *Newbies*, who have empty egocentric sociograms, *Observers*, who have sparse sociograms as they participate more regularly than newbies, and *Commentators*, who comment and answer questions, though with average post length less than that of experts.

Online support communities In online communities created for health advice and other support, there are more emotional aspects to consider when modeling roles, for example, encouraging or motivating other users, and making them feel welcomed. A user can play multiple roles in a session, and to account for this, Yang et al. [103] used Gaussian mixture model [70] for identifying user role clusters in the *Cancer Survivors Network*⁵ based on behavioral features associated with four aspects of a social role: goal, interaction, expectation, and context. Features related to the goals of users included providing or seeking information or emotional support. Interaction implies the actions of users' in the community and behaviors related to this aspect were analyzed through psycho-linguistic lexicon LIWC [78] and users' reply-network. The third aspect, i.e., the expectation associated with the role was modeled based on the interaction of users with moderators and the usage of modality in content in order to advise or suggest other users. The context was modeled on the basis of whether conversations were public or private. The 11 roles obtained by them are mentioned below as follows:

- (i) *Emotional Support Provider*, who provides support, appreciation or encouragement in the community,
- (ii) *Welcomer*, who interacts with newcomers and also encourages them,
- (iii) *Informational Support Provider*, who provides informational support and advice to others,
- (iv) *Story Sharer*, who shares personal experiences,
- (v) *Informational Support Seeker*, who requests for information,
- (vi) *Private Support Provider*, who provides support to others through private conversations,
- (vii) *Private Communicator*, who furnishes or seeks support only in private chats,
- (viii) *All Round Expert*, who engages in both public and private discussions,
- (ix) *Newcomer*, who is a newly joined member and asks questions to get information or support,
- (x) *Knowledge Promoter*, who provides links and references,

⁵ Extracted from <https://csn.cancer.org/>.

- (xi) *Private Networker*, who heavily participates in private conversations.

The analysis of user role transitions showed that newly joined users often adopt the role of *Informational Support Seeker* and *Story Sharer*. With time, they move on to roles like *Emotional Support Provider* and start welcoming newcomers to the community. The Markov process based modeling of these transitions showed that the users eventually moved from seeking to sharing roles, as expected in online platforms. In terms of frequency of occurrence, the most frequent role is *Emotional Support Provider* (33.3%), followed by *Welcomer* (15.9%), *Informational Support Provider* (13.3%) and *Story Sharer* (10.2%). Similar work on an online support-community for aged people through social network and content analysis showed that there exists a social role of a *Moderating Supporter*, who displays a heavy interest in the community through posting of both self-disclosure and support-related content as well as welcoming newcomers [79]. Users playing such roles often form the backbone of the community.

Online Q&A and discussion platforms enable dissemination of knowledge by enabling users to ask questions, initiate discussions, provide answers and opinions. A common inference that can be made from the literature is that all users do not show similar participation in these websites. A small number of users are often responsible for a high proportion of content. Users can be separated into different role clusters based on parameters, such as the type of activities they involve in on the platforms- the most commonly studied activities are asking questions, answering questions and adding comments, ego network structure based on user interaction, user statistics, reply behavior analysis, the variance of user participation with time, answer content and its quality, user expertise, and cross-community participation. Some of the selected work on role identification on Q&A platforms, covering all kinds of roles, are explained in Table 3.

Online crowdsourced encyclopedias

In Q&A forums, users seek help regarding the specific questions they have, while in online crowdsourced encyclopedia, users contribute to create a comprehensive article on a specific topic. Initially, encyclopedias such as *Encyclopædia Britannica*⁶ [16], *Brockhaus Enzyklopädie*,⁷ *Cyclopædia: or, An Universal Dictionary of Arts and Sciences*⁸ were written and maintained by the experts in the area, but online crowdsourced encyclopedias changed the way the knowledge is gathered where any user can contribute more or less based on the capacity and expertise. Now-a-days, users frequently refer these encyclopedias to gain knowledge or clear doubts on a topic. Some of the popular crowdsourced article writing platforms are Wikipedia,

⁶ More details are available at https://en.wikipedia.org/wiki/Encyclopædia_Britannica.

⁷ Further details can be found at https://en.wikipedia.org/wiki/Brockhaus_Enzyklopädie.

⁸ More details are available at https://en.wikipedia.org/wiki/Cyclopædia_or_an_Universal_Dictionary_of_Arts_and_Sciences.

Table 3 User roles on discussion and Q&A platforms (selected works)

Platform	User Roles/Key Findings	Metrics & Method	Ref
Usenet	Celebrity, Newbie, Lurker, Flamer, Troll, Ranter	Observational study of posts' content, no. of posts, activity across groups	[39]
	Answer Person, Questioner, Troll, Spammer, Flame Warrior	Analysis of posting behavior and user networks. Metrics like initiation of threads, no. of threads contributed to, reply patterns used	[90]
	Question people and answer people, User network depends on the nature of newsgroups	Out-degree, interconnectivity, and reciprocity in egocentric directed networks	[28]
	Answer Person, Discussion Person	Neighbourhood and degree distribution in directed egocentric network	[94]
	Discussion Catalysts (they mainly import content from other websites and receive many replies.)	Posting behavior, content and network in terms of reply and replier share, success ratio, info source for posts, original content in posts, no. of ties	[47]
	Answer Person, Debater, Bursty Contributor, Newcomer, Spammer, Balanced Conversationalist	Visualization of users' activity over time. Metrics used like avg. posts per thread, no. of days user is active	[91]
Reddit	Answer Person and Non-Answer Person	Network density and degree distribution in egocentric networks	[18]
	Frequent Questioner, Occasional Questioner, Elaborative Questioner, Occasional Answerer, Elaborative Answerer, Community Activist	Metrics include Out-degree, In-degree in weighted directed network, and a no. of metrics to capture posting and replying behavior including no., length and score of questions posted, depth and avg. time ranking of posted replies. Gaussian mixture model for clustering	[59]
	Knowledge Broker, Translator, Experienced Practitioner, Learner, Conversation Facilitator	Used content analysis, open coding, and axial coding. Analysed no. of posts, no. of comments, and posts' score.	[53]
Stack Exch.	Low-activity user, Occasional user, Unskilled answerer, Expert answerer, Expert questioner, Expert commenter, Q-A activist, Unskilled commenter, Answer activist, Hyperactivist	Used Ward hierarchical and K-means clustering. Metrics include no. of questions, answers, comments per user, and their utility	[33]
	Sparrows (highly active users trying to build reputation), Owls (experts who answer difficult questions)	Evaluation of expertise, preference of questions, and temporal changes. Authors analysed the debatleness of questions and utility of answers.	[105]
MOOC Discussions	Core, Information Seekers, Experts	Blockmodeling on social and semantic similarities. Metrics include participation of users in similar threads, and regular similarity	[46]

Baidu Baike, Hudong Baike, and so on. Wikipedia is the most popular online encyclopedia which runs on the contributions of 37 million registered users across the world as on 7 December 2019 [100]. It therefore has attracted researchers to analyze its data collection process and users' roles. We will mainly discuss users' roles on Wikipedia, followed by other platforms.

Wikipedia

Wikipedia, hosted by Wikimedia foundation, is a free-to-use online encyclopedia that runs on the open collaboration of users across the world. Users can contribute content to Wikipedia articles distinctively or anonymously. These voluntary contributors are considered a part of the Wikipedia community and are commonly known as *Wikipedians*. Contributions comprise of a variety of recognized work

on the platform: (1) editing of articles that ranges from just copy-editing to citing appropriate sources, (2) actions for social and community support that include welcoming and guiding newcomers, (3) border patrol which includes fighting vandalism and copyright infringements on the articles' content, (4) administrative work, for example, granting certain privileges to users and mediating conflicts, (5) collaborative actions, such as helping other editors comply with the platform's policies, and (6) undifferentiated work [54]. Information on Wikipedia pages is published through collaborative writing of different types of users and the users who contribute to the article are known as editors.

Structure of Wikipedia

Wikipedia does not have a flat organization, instead, it has a hierarchical structure of roles; the hierarchy can be noted through the superior-subordinate relationships between the user roles obtained through the consolidation of different access privileges [11]. Wikipedia privileges can be categorized into 12 role classes in a hierarchical structure by grouping of users' access privileges through exploratory factor analysis, with the Benevolent Dictator at the top of the hierarchy, and unregistered users are considered at level 0, as shown in Fig. 1 [11, 12]. Users with roles high in the hierarchy contribute more to talk pages for community building activities, whereas the lower-level roles focus more on their specific domain [12]. Initially when Wikipedia was introduced, most of the members of the Wikipedia community were at level 1; the number of members at levels 2 and 3 started increasing eventually due to the need for filtering and organizing the massive amount of content being generated, as well as fighting vandalism [12]. It has been noted through qualitative analysis of vertical transitions that promotions to administrative roles are not only based on the active contributions in terms of edits, but also to other namespaces including talk pages.

Wikipedia editors can be assigned roles based on their activities and behavior on the platform, and not only based on their functional roles acquired through privileges. In the following subsections, we discuss how various works have identified users' roles on Wikipedia based on (1) the type of edits made by the users, (2) distribution of a user's edits across various Wikipedia namespaces, (3) collaborating/edit network structures, (4) users' editing behavior over time, and (5) user behavior on talk page discussions.

User roles based on types of edits

Several types of edits can be done on Wikipedia and they are either meaning-preserving or meaning-changing [23]. Based on the types of edits made by users and their activity in various namespaces, Latent Dirichlet Allocation (LDA) is used to identify 8 editor roles in Wikipedia: *Social Networker*, *Fact Checker*, *Substantive Expert*, *Copy Editor*, *Wiki Gnome*, *Vandal Fighter*, *Fact Updater*, and *Wikipedian* [102]. While Substantive Experts contribute content and insert references in articles, a Fact Checker focuses on the removal of content and

Fig. 1 Hierarchy of Wikipedia roles where power decreases from top to bottom [11, 12]

references. Liu and Ram [63] identified different types of contributors in terms of a single article based on ten types of actions on it: Creating sentence, Modifying sentence, Deleting sentence, Creating link, Modifying link, Deleting link, Creating reference, Modifying reference, Deleting reference and Reverting. By using repeated K-Means clustering [62], they obtained 6 clusters corresponding to 6 user roles. While *All-Round Editors* perform a variety of actions, *Watch-dogs* perform a majority of reverts. *Content Justifiers* add links and references, whereas *Starters* only create sentences. Incorrect content (sentences, links, and references) is removed by *Cleaners*.

Another interesting aspect is how the proportions of user roles change as a Wikipedia article transitions into the maturity stage from the point of creation. Arazy et al. [7] built upon the research on the classification of edits in Wikipedia [6, 10, 54] to come up with the following 12 types of edits:

- (i) Moving/creation of article
- (ii) Addition of substantive content (meaning is changed)
- (iii) Deletion of substantive content (meaning is changed)
- (iv) Fixing typographical and grammatical mistakes
- (v) Rephrasing/re-structuring of text (meaning not changed)
- (vi) Addition/deletion/change hyperlink
- (vii) Addition/deletion/change references
- (viii) Addition/deletion/change of Wiki Markup
- (ix) Reorganization of text (change article's structure)

- (x) Insertion of Vandalism (malicious content)
- (xi) Removal of Vandalism
- (xii) Miscellaneous edits

Multiple-labeling was allowed, and features of the edits including textual features and meta data [7] were used for classifying edits based on the manually labeled training set. K-Means clustering was then used to identify the emergent user roles based on the user profiles constructed using the number and type of edits made by the users. They obtained seven user roles: (1) *All-Round Contributors*, (2) *Copy-Editors*, (3) *Content Shapers*, (4) *Quick-and-Dirty Editors*, (5) *Layout Shapers*, (6) *Vandals*, and (7) *Watchdogs*, which were observed to be stable roles over time through comparison of clusters obtained on Wikipedia data belonging to different time periods. By applying the ANOVA model [26], Arazy et al. [7] observed that in the beginning phase of a particular article, *All-Round Contributors*, and *Content and Layout Shapers* constitute a high proportion of roles, however, it declines as the article matures, and the proportion of *Quick-and-Dirty Editors* increases. The increase in the share of *Vandals* over time is accompanied by the increase in *Watchdogs* whose actions intend to protect the article from vandal edits. Vandals may not be inhibited by the reverting of their edits through anti-vandalism efforts and may continue to vandalize other pages, which when crosses a certain limit, leads to them getting blocked [36].

User roles based on distribution of edits in various namespaces

The distribution of users' edits across the different Wikipedia namespaces can reveal the underlying user roles on the platform. Welser et al. [93] studied the average distribution of edits across the namespaces and the network structure based on the user-relationships on Wikipedia's talk pages to differentiate user roles. About half the edits of *Substantive Experts* are to the content space; these experts also post significantly to content talk, explaining and accounting for their edits on the content pages. *Technical Editors* contribute many small edits to the content, especially edits related to grammar and spellings. *Social Networking Editors* contribute less to the content and mainly spend time on their user pages, user talk pages, and on Wikipedia namespace pages. *Counter-Vandalism Editors* place warnings on user talk pages to combat vandalism and also make edits to other users' pages for blocking vandals.

Receiving more access privileges in Wikipedia may bring a change in the editing activity of users across the various namespaces. Arazy et al. [8] studied the users' edits and noted that obtaining access privileges like *Rollbacker* and *Reviewer* leads to a reduction in the users' edits in the talk namespaces.

User roles based on network structure

How are the user roles associated with their position in the Wikipedia collaboration network? The network constructed using the edits on users' talk pages, with directed edge going from the user who edits to the user whose talk page has been edited, suggested that *Technical Editors* mainly focus on issues independent of page content

like formatting and grammar, and have outgoing connections to editors who are not much connected to each other [93]. On the other hand, *Substantive Experts* focus on a smaller specific set of pages related to their area of expertise and may have a denser local network.

Analysis of the network on Chinese Wikipedia showed the association between three centrality measures and user roles: (1) a higher degree centrality indicates a content-expert or specialist role, who mostly concentrates on their articles, (2) a central user in terms of closeness centrality plays a generalist role who edits articles belonging to diverse areas, as well as gets help from a diverse set of people, and (3) a user with greater betweenness centrality plays a bridging role between diverse editors in the form of a content extender [106]. Temporal analysis of the centrality measures also showed that initially there were a few highly central editors, but with time, the network becomes relatively decentralized with new editors joining and contributing.

When a user is a part of a social network, her ties and the nature of those ties shape the user's social role. Doran [24] obtained two Wikipedia roles based on the k-means clustering of censuses of conditional triads in users' ego networks: *Specialists* and *Generalist Attractors* [19]. A directed link was added between two users if the first user does text editing, change reversion or vote for an action on the article made by the second user. *Generalist Attractors* tend to review and suggest corrections in terms of grammar, hyperlinks, etc., and they are somewhat equivalent to copy editors discussed before [102]. They are embedded in a highly interconnected and dense network because the variety of articles they edit are also edited by the number of specialists as well as other technical editors.

Editing changes by users on an article can be converted into a collaboration network by constructing a directed edge between two editors if they have consecutive edits [51]. Featured or highly rated articles may be constructed through successful coordination between different editors supported by a category of active users called *Coolfarmers*. This category contains two kinds of users: (1) *Mediators*, who help in reconciling varying viewpoints of different editors and have more conciliatory conversations, and (2) *Zealots*, who add fuel to discussions on highly controversial topics. Analyzing the communication network of four such potential coolfarmers based on editing on a controversial article, Iba et al. [51] noted that mediators have a high number of discussions with small groups of users thus leading to a higher fluctuation in their betweenness centrality, whereas zealots tend to have fewer conversations, with a small number of one-to-one fights.

User roles based on article editing history

Prolific editors have different behaviors based on the range of articles they edit. Some editors mostly spend their energy in editing a particular article and its related articles, whereas some others keep changing focus from one article to another, thus playing the role of a generalist as opposed to a specialist in one topic [52]. Keegan [52] studied the trajectories of different prolific editors by modeling their edits through a directed network; there exists an edge between artifacts A and B if the user takes an action on A, immediately followed by an action on B. While one

type of generalist works on a set of related articles and then changes to a different subject, another type of generalist editor shows a tendency to go back to his earlier articles after working on a variety of articles.

Arazy et al. [9] also studied the motivational factors and transitions between the roles identified in [7]. For analyzing the user trajectories in terms of roles, they created one vector for each user-year-article combination obtained through the temporal data of user activity and linked it with a role. They identified four classes in users' transitioning behavior: (1) users having a single role on a single article, (2) users having a single role in many articles, (3) users having multiple roles in a single article, and (4) users having multiple roles across many articles. Users belonging to class (3) tend to have a persistent Wikipedia career in terms of activity and are fueled by all the four motivational factors considered in the study: peer-approval, reputation, fun, and friendship [56, 85], while users falling in category (2) are driven majorly by peer-approval. For comparison of the above-mentioned user categories and their motivational factors for contributing, ANOVA [26] and Least Significant Difference (LSD) tests were used.

Statistical data based on the number of edits and the time gaps between user activity on Romanian and Danish Wikipedia gives four clusters of users, which can be interpreted by Principal Component Analysis (PCA) [55]. One of the clusters contains users who show short bursts of high activity, while another cluster contains users who regularly contribute to the encyclopedia. The other two clusters include casual contributors and top contributors. An important profile of editors on Wikipedia is the *Wikipedians*, i.e., the editors who create the most value for Wikipedia content [77]. Through the visualization of users' edit frequency over time, how long the users' edits survived subsequent revisions and their activity in community work related namespaces, Panciera et al. [77] noted that these prolific editors tend to maintain their levels of activity over time compared to an ordinary contributor. They produce a better quality of content, measured by how long their edits persist over time and cite community norms or conventions more often.

User roles based on talk page discussions

Wikipedia talk pages are an important platform for discussions related to edits in the articles. Based on talk page discussions, Ferschke et al. [27] proposed a Role Identification Model (RIM) to match users and roles using maximum weighted matching [1] on a bipartite graph. They assigned each post by the editors on talk pages in the selected data a label indicating the type of post. For example, one category of labels was *article criticism*, which included posts that point out issues with the articles. The behavior vectors for the participants were constructed based on the post label frequencies, and four user roles were identified: (1) *The Doer*, who makes a plethora of contributions, (2) *Critiquer*, whose posts mainly focus on critiquing, (3) *Encourager*, whose posts generally have a positive attitude, and (4) *Manager*, whose actions involve requesting edits and reporting actions.

The limitations of the above discussed model include the assumption that a particular role is played by only one user and that a conversation contains all the

defined roles. To overcome such limitations, Maki et al. [66] proposed a probabilistic graphical model to identify five roles: (1) *Moderator*, who mostly engages in summarizing and shaping the flow of discussions, (2) *Architect*, whose focus is on the hierarchical organization of pages, (3) *Policy Wonk*, who highlights the policies pertaining to fairness and copyright, (4) *Wordsmith*, who focuses on the spellings, pronunciations and other grammatical aspects, and (5) *Expert*, which is the most content-centric role.

Other online crowdsourced encyclopedias

Baidu Baike, the famous Chinese alternative of Wikipedia has a different working model as compared to Wikipedia. Submissions made by contributors of the platform should go through a review by the employees of Baidu, whereas on Wikipedia such reviews are done by volunteers across the world [60]. In the case of Baidu Baike, the *Power Users*, or the users with certain extra privileges are selected by Baidu employees, whereas in the case of Wikipedia, there is rarely any intervention by the Wikimedia Foundation employees.

While on Wikipedia, even anonymous users may contribute to articles, on Baidu Baike only registered users can edit articles. In order to analyze the behavior of editors in terms of their contributions, Huang et al. [49] used a time series model, with data points being the editor contribution measured in terms of the number of articles created and edited in a certain period of time. Using k-means clustering, they obtained four classes of users on Baidu Baike, (1) *Dropouts*, who stop editing after contributing heavily in the second month, (2) *Delayers*, who highly contribute in the initial two months and then their contribution dwindles, (3) *Testers*, who constitute 85.4% of the users and stop contributing after making a few edits initially, and (4) *Stickers*, who maintain a certain level of contributions even after a five month period and thus are important because of their continued contributions.

Online crowdsourced encyclopedias like Wikipedia may have a hierarchical structure of privilege-based roles. However, we can classify editors' social roles on the platform through their editing behavior and interactions. All users do not exhibit similar behavior. Users may focus on different types of editing activities ranging from copy-editing to fact-checking, contribute differently to the content and discussion pages, have differing collaboration and editing network structures, post different kinds of content on the discussion pages, and exhibit different editing behavior over time in terms of the topics, number of edits and time gaps between editing activity. These aspects have been taken into consideration by the works on Wikipedia social roles. Some of the selected works covering different kinds of user roles have been summarized in Table 4.

Datasets

In this section, we discuss the available datasets that have been used in previous research studies, and the available APIs that can be used to collect your own dataset.

Discussion and Q&A platforms

Usenet Newsgroups

- Majority of authors [28, 90, 94] have collected newsgroups' related posting and its network data using Microsoft Research's Netscan.
- Mitchell [71] has also provided the dataset of 20 newsgroups containing 20000 messages; this dataset is available at UCI's machine learning website: <https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups>. This is a potential dataset for role modeling.

Reddit

- Reddit provides an API (<https://www.reddit.com/dev/api>) which can be used to get posts and comments on Reddit communities. PRAW⁹ is a Python wrapper package that can be used to access Reddit API. There is a publicly available toolkit to crawl reddit communities and obtain statistical data related to posting behaviors of users at <https://github.com/cbuntain/redditResponseExtractor> [18].
- Another publicly available dataset, that has been used for user roles research in [65], contains Reddit content in a downloadable format from December 2005. This dataset is available at <http://files.pushshift.io/reddit/comments/>.

Stack Exchange

- The publicly available Stack Exchange data dump has been one source of data for research (used in [37, 105]). It contains anonymized user-content on the Stack Exchange websites, including information like posts, votes, comments etc, and is available at <https://archive.org/details/stackexchange>.
- Stack Exchange also provides data access using their API. More details are available at <https://api.stackexchange.com/docs>.

Online crowdsourced encyclopedias

- WikiMedia provides a dump of Wikipedia content for education and research purpose; the dataset is available at https://meta.wikimedia.org/wiki/Data_dumps (used in [93]).
- The dataset for classification of Wikipedia edits into different categories is available at, https://www.informatik.tu-darmstadt.de/ukp/research_6/data/_textual_revisions/wikipedia_edit_category_corpus/index.en.jsp [23].
- A dataset [57] containing the complete edit history till January 2008 including all revisions and all pages of Wikipedia is available at <http://snap.stanford.edu/data/wiki-meta.html>.

⁹ <https://github.com/praw-dev/praw>.

- Massa [68] extracted the dataset of Wikipedia user talk pages, and this is available at https://sonetlab.fbk.eu/data/social_networks_of_wikipedia/. This is another potential dataset for the purpose of role analysis.
- A list of Wikimedia datasets and available tools is available at, <https://meta.wikimedia.org/wiki/Datasets>.

Conclusion

Online discussion and Q&A platforms provide means for users to share knowledge and opinions, thus paving the way for learning. In this process of knowledge sharing, distinctive user behaviors can be observed, which can be modeled into social roles. The shaping of user roles may also be influenced by the badges and other incentives provided by the studied online platform as well as the platform-specific access privileges. Hence, the generalization of roles across different online knowledge production communities becomes difficult. In this paper, we have provided a detailed description of social roles obtained by users on different online platforms. However, this area has various open directions that should be looked into further. Studies providing insights into whether a user's perception of roles matches with the roles obtained through quantitative and qualitative analysis of data can help us understand if users are aware of the roles they are playing in the knowledge production process. The proposed methods for identifying roles can be validated with users' perceptions about their roles as well as others' roles.

The correlation of roles with different user features, such as gender, age, race, country, and other human characteristics is still not well understood. Are the

Table 4 Modeling roles on online crowdsourced encyclopedias (selected works)

Platform	User Roles/Key Findings	Metrics/Method	Ref
Wikipedia	All-Round Editors, Watchdogs, Content Justifiers, Starters, Cleaners, Casual Contributor	K-Means clustering on 10 types of edits on an article (Type of Edits)	[63]
	Substantive Experts, Technical Editors, Social Networking Editors, Counter-Vandalism Editors	Distribution of Edits across content space, content talk, user pages, user talk pages (Distribution of Edits)	[93]
	Specialist and Generalist Attractors	K-means clustering of conditional triads' censuses in users' ego networks (Network Structure)	[24]
	The Doer, Critiquer, Encourager, Manager	Bipartite matching of users and roles based on types of posts by user on talk pages (Talk Page Discussions)	[27]
	Casual Contributors, Regular Contributors, Top Contributors, Contributors on a Mission	Hierarchical clustering and partitioning algorithms on features like no. of edits, time gaps between user activity (Article Editing History)	[55]
Baidu Baike	Dropouts, Delayers, Testers, Stickers	K-means clustering on temporal no. of articles created/edited by user	[49]

important roles in terms of generating content restricted to users of certain belonging to certain geographical regions? It will be interesting to understand if the social roles of the users on these platforms vary with the gender or geographies of users. These studies will help in customizing the platforms so that all people can contribute more. One can further analyze how a single user's role varies across different discussion topics and communities on the website. Does there exist a difference in professional competencies of users on different websites; it will be interesting to compare how user roles and their distributions are different on such websites.

Future research can also focus on understanding how gamification initiatives like badges on Q&A websites can (1) lead to role transitions for a particular user on the same platform as the user acquires privileges/badges, which can help us understand the changes in the user behavior due to the incentives provided, and (2) differences in user roles between platforms having these initiatives and the ones which do not. For MOOCs, the study of role transitions of successful students right from their first enrolled course can help us understand the trajectory of such students, which can be used to recommend students to take up appropriate activities to improve their grades. On currently active Q&A websites in general, it will be interesting to understand how the user's badges correlate to the tendency of the user to advise or correct other users and enforce some norms related to the community.

The roles are organized on Wikipedia based on the privileges given to the users as well a variety of other parameters including the type of edits the users make, the distribution of their edits across the various namespaces, their network structure, their article editing history and their behavior on talk page discussions. One can further pursue to identify the overlapping of different roles adopted by Wikipedia users as not much research has been done on identifying multiple roles that users play in a Wikipedia setting, and how likely it is for two defined roles to be played by one user. Future research can focus on this aspect to define and model a user's role from multiple perspectives.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1988). *Network flows*. Cambridge, Massachusetts: Alfred P. Sloan School of Management.
2. Akar, E., & Mardikyan, S. (2018). User roles and contribution patterns in online communities: A managerial perspective. *SAGE Open*, 8(3), 2158244018794773.
3. Akar, E., Mardikyan, S., & Dalgic, T. (2018). User roles in online communities and their moderating effect on online community usage intention: An integrated approach. *International Journal of Human Computer Interaction*, 35, 495–509.
4. Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2013). Steering user behavior with badges. In: Proceedings of the 22nd international conference on World Wide Web, pp. 95–106. ACM
5. Angeletou, S., Rowe, M., & Alani, H. (2011). Modelling and analysis of user behaviour in online communities. In: International semantic web conference, pp. 35–50. Springer
6. Antin, J., Cheshire, C., & Nov, O. (2012). Technology-mediated contributions: editing behaviors among new wikipedians. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, pp. 373–382. ACM
7. Arazy, O., Daxenberger, J., Lifshitz-Assaf, H., Nov, O., & Gurevych, I. (2016). Turbulent stability of emergent roles: The dualistic nature of self-organizing knowledge coproduction. *Information Systems Research*, 27(4), 792–812.
8. Arazy, O., Lifshitz-Assaf, H., & Balila, A. (2019). Neither a bazaar nor a cathedral: The interplay between structure and agency in wikipedia's role system. *Journal of the Association for Information Science and Technology*, 70(1), 3–15.
9. Arazy, O., Lifshitz-Assaf, H., Nov, O., Daxenberger, J., Balestra, M., & Cheshire, C. (2017). On the how and why of emergent role behaviors in wikipedia. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, pp. 2039–2051. ACM
10. Arazy, O., & Nov, O. (2010). Determinants of wikipedia quality: the roles of global and local contribution inequality. In: Proceedings of the 2010 ACM conference on computer supported cooperative work, pp. 233–236. ACM
11. Arazy, O., Nov, O., & Ortega, F. (2014). The [wikipedia] world is not flat: On the organizational structure of online production communities. In: ECIS 2014 proceedings—22nd European conference on information systems. Association for Information Systems
12. Arazy, O., Ortega, F., Nov, O., Yeo, L., & Balila, A. (2015). Functional roles and career paths in wikipedia. In: Proceedings of the 18th ACM conference on computer supported cooperative work and social computing, pp. 1092–1105. ACM
13. Barash, V. D., Smith, M., Getoor, L., & Welser, H. T. (2009). Distinguishing knowledge vs social capital in social media with roles and context. In: Third international AAAI conference on weblogs and social media
14. Biddle, B. J. (2013). *Role theory: Expectations, identities, and behaviors*. Academic Press.
15. Boroujeni, M. S., Hecking, T., Hoppe, H. U., & Dillenbourg, P. (2017). Dynamics of mooc discussion forums. In: Proceedings of the seventh international learning analytics and knowledge conference, pp. 128–137. ACM
16. Britannica, E., et al. (2008). *Britannica concise encyclopedia*. Encyclopaedia Britannica Inc.
17. Brush, A., Wang, X., Turner, T. C., & Smith, M. A. (2005). Assessing differential usage of usenet social accounting meta-data. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 889–898. ACM
18. Buntain, C., & Golbeck, J. (2014). Identifying social roles in reddit using network structure. In: Proceedings of the 23rd international conference on world wide web, pp. 615–620. ACM
19. Burt, R. S. (1990). Detecting role equivalence. *Social Networks*, 12(1), 83–97.
20. Chan, J., Hayes, C., & Daly, E. M. (2010). Decomposing discussion forums and boards using user roles. In: Fourth international AAAI conference on weblogs and social media
21. Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014). Can cascades be predicted? In: Proceedings of the 23rd international conference on World wide web, pp. 925–936
22. Choi, D., Han, J., Chung, T., Ahn, Y. Y., Chun, B. G., & Kwon, T. T. (2015). Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In: Proceedings of the 2015 acm on conference on online social networks, pp. 233–243. ACM

23. Daxenberger, J., & Gurevych, I. (2013). Automatically classifying edit categories in wikipedia revisions. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 578–589
24. Doran, D. (2014). Triad-based role discovery for large social systems. In: International conference on social informatics, pp. 130–143. Springer
25. Doreian, P., Batagelj, V., & Ferligoj, A. (2005). *Generalized blockmodeling* (Vol. 25). Cambridge University Press.
26. Edwards, A. L. (1985). *Multiple regression and the analysis of variance and covariance*. WH Freeman/Times Books/Henry Holt & Co.
27. Ferschke, O., Yang, D., & Rosé, C. P. (2015). A lightly supervised approach to role identification in wikipedia talk page discussions. In: Ninth international AAAI conference on web and social media
28. Fisher, D., Smith, M., & Welsch, H. T. (2006). You are who you talk to: Detecting roles in usenet newsgroups. In: Proceedings of the 39th Annual Hawaii international conference on system sciences (HICSS'06), vol. 3, pp. 59b–59b. IEEE
29. Forestier, M., Stavrianou, A., Velcin, J., & Zighed, D. A. (2012). Roles in social networks: Methodologies and research issues. *Web Intelligence and Agent Systems: An international Journal*, 10(1), 117–133.
30. Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), 215–239.
31. Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software Practice and Experience*, 21(11), 1129–1164.
32. Fu, C. (2019). Tracking user-role evolution via topic modeling in community question answering. *Information Processing and Management*, 56(6), 102075.
33. Furtado, A., Andrade, N., Oliveira, N., & Brasileiro, F. (2013). Contributor profiles, their dynamics, and their importance in five q&a sites. In: Proceedings of the 2013 conference on Computer supported cooperative work, pp. 1237–1252. ACM
34. Gazan, R. (2006). Specialists and synthesists in a question answering community. *Proceedings of the American Society for Information Science and Technology*, 43(1), 1–10.
35. Gazan, R. (2007). Seekers, sloths and social reference: Homework questions submitted to a question-answering community. *New Review of Hypermedia and Multimedia*, 13(2), 239–248.
36. Geiger, R. S., & Ribes, D. (2010). The work of sustaining order in wikipedia: the banning of a vandal. In: Proceedings of the 2010 ACM conference on computer supported cooperative work, pp. 117–126. ACM
37. Geigle, C., Dev, H., Sundaram, H., & Zhai, C. (2019). A generative model for discovering action-based roles and community role compositions on community question answering platforms. In: Proceedings of the international AAAI conference on web and social media, vol. 13, pp. 181–192.
38. Gliwa, B., Zygmunt, A., & Koźlak, J. (2013). Analysis of roles and groups in blogosphere. In: Proceedings of the 8th international conference on computer recognition systems CORES 2013, pp. 299–308. Springer
39. Golder, S. A., & Donath, J. (2004). Social roles in electronic communities. *Internet Research*, 5(1), 19–22.
40. Gómez, V., Kappen, H. J., Litvak, N., & Kaltenbrunner, A. (2013). A likelihood-based framework for the analysis of discussion threads. *World Wide Web*, 16(5–6), 645–675.
41. Graham, T., & Wright, S. (2014). Discursive equality and everyday talk online: The impact of superparticipants. *Journal of Computer Mediated Communication*, 19(3), 625–642.
42. Grayson, S., & Greene, D. (2019). Temporal analysis of reddit networks via role embeddings. arXiv preprint. [arXiv:1908.05192](https://arxiv.org/abs/1908.05192)
43. Hamilton, W. L., Zhang, J., Danescu-Niculescu-Mizil, C., Jurafsky, D., & Leskovec, J. (2017). Loyalty in online communities. In: Eleventh international AAAI conference on web and social media
44. Hand, D. J. (2010). Statistical analysis of network data: Methods and models by Eric D. Kolaczyk. *International Statistical Review*, 78(1), 135–135.
45. Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 28(1), 100–108.
46. Hecking, T., Chounta, I. A., Hoppe, H. U. (2016). Investigating social and semantic user roles in mooc discussion forums. In: Proceedings of the sixth international conference on learning analytics and knowledge, pp. 198–207. ACM

47. Himelboim, I., Gleave, E., & Smith, M. (2009). Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of Computer Mediated Communication*, 14(4), 771–789.
48. Huang, J., Dasgupta, A., Ghosh, A., Manning, J., Sanders, M. (2014). Superposter behavior in mooc forums. In: Proceedings of the first ACM conference on Learning@ scale conference, pp. 117–126. ACM
49. Huang, L., Zhu, Q., Du, J. T., & Lee, B. (2016). Exploring the dynamic contribution behavior of editors in wikis based on time series analysis. *Program Electronic Library and Information Systems*, 50(1), 41–57.
50. Hymes, D. (2013). *Foundations in sociolinguistics: An ethnographic approach*. Routledge.
51. Iba, T., Nemoto, K., Peters, B., & Gloor, P. A. (2010). Analyzing the creative editing behavior of wikipedia editors: Through dynamic social network analysis. *Procedia Social and Behavioral Sciences*, 2(4), 6441–6456.
52. Keegan, B. C. (2015). Emergent social roles in wikipedia's breaking news collaborations. *Roles, trust, and reputation in social media knowledge markets* (pp. 57–79). Springer.
53. Kou, Y., Gray, C. M., Toombs, A. L., & Adams, R. S. (2018). Understanding social roles in an online community of volatile practice: A study of user experience practitioners on reddit. *ACM Transactions on Social Computing*, 1(4), 17.
54. Kriplean, T., Beschastnikh, I., McDonald, D. W. (2008). Articulations of wikiwork: Uncovering valued work in wikipedia through barnstars. In: Proceedings of the 2008 ACM conference on computer supported cooperative work, pp. 47–56. ACM
55. Krishna, S., Billot, R., Jullien, N. (2018). A clustering approach to infer wikipedia contributors' profile. In *Proceedings of the 14th international symposium on open collaboration* (pp. 1–5).
56. Lakhani, K. R., Wolf, R. G. (2003). *Why hackers do what they do: Understanding motivation and effort in free/open source software projects*. MIT Sloan working paper.
57. Leskovec, J., Huttenlocher, D., Kleinberg, J. (2010). Governance in social media: A case study of the wikipedia promotion process. In Fourth international AAAI conference on weblogs and social media.
58. Li, Z., Huang, K. W., Cavusoglu, H. (2012). Can we gamify voluntary contributions to online Q&A communities? Quantifying the impact of badges on user engagement. In *Proceedings of 2012 workshop on information systems and economics (WISE 2012)*.
59. Liang, Y., Introne, J. (2019). Social roles, interactions and community sustainability in social q&a sites: A resource-based perspective. In *Proceedings of the 52nd Hawaii international conference on system sciences*
60. Liao, H. T. (2015). Harnessing the power of collaborative filtering: Comparing the network gate-keeping of baidu baike and Chinese wikipedia. *China Perspectives*, 2015(2015/4), 35–49.
61. Lin, B., Serebrenik, A. (2016). Recognizing gender of stack overflow users. In: Proceedings of the 13th international conference on mining software repositories, pp. 425–429
62. Liu, H., & Kešelj, V. (2007). Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests. *Data and Knowledge Engineering*, 61(2), 304–330.
63. Liu, J., Ram, S. (2009). Who does what: Collaboration patterns in the wikipedia and their impact on data quality. In: 19th Workshop on information technologies and systems, pp. 175–180
64. Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1(1), 49–80.
65. Lumbreras, A., Jouve, B., Velcin, J., & Guégan, M. (2017). Role detection in online forums based on growth models for trees. *Social Network Analysis and Mining*, 7(1), 49.
66. Maki, K., Yoder, M., Jo, Y., Rosé, C. (2017). Roles and success in wikipedia talk pages: Identifying latent patterns of behavior. In: Proceedings of the eighth international joint conference on natural language processing (Vol. 1: Long Papers), pp. 1026–1035
67. Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., Hartmann, B. (2011). Design lessons from the fastest q&a site in the west. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 2857–2866. ACM
68. Massa, P. (2011). Social networks of wikipedia. In: Proceedings of the 22nd ACM conference on hypertext and hypermedia, pp. 221–230
69. May, A., Wachs, J., & Hannák, A. (2019). Gender differences in participation and reward on stack overflow. *Empirical Software Engineering*, 24(4), 1997–2019.

70. McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 84). Marcel Dekker.
71. Mitchell, T. M., et al. (1997). *Machine learning*. New York: McGraw-hill.
72. Morrison, D., Hayes, C. (2013). Here, have an upvote: Communication behaviour and karma on reddit. In: INFORMATIK 2013—Informatik angepasst an Mensch, Organisation und Umwelt
73. Morzy, M. (2009). On mining and social role discovery in internet forums. In: 2009 International workshop on social informatics, pp. 74–79. IEEE
74. Narang, K., Chung, A., Sundaram, H., Chaturvedi, S. (2019). Discovering archetypes to interpret evolution of individual behavior. arXiv preprint. [arXiv:1902.05567](https://arxiv.org/abs/1902.05567)
75. Nonnecke, B., Preece, J. (2001). Why lurkers lurk. In: AMCIS 2001 proceedings, p. 294
76. Pal, A., Chang, S., Konstan, J. A. (2012). Evolution of experts in question answering communities. In: Sixth international AAAI conference on weblogs and social media
77. Panciera, K., Halfaker, A., Terveen, L. (2009). Wikipedians are born, not made: a study of power editors on wikipedia. In: Proceedings of the ACM 2009 international conference on supporting group work, pp. 51–60. ACM
78. Pennebaker, J. W., Boyd, R. L., Jordan, K., Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report
79. Pfeil, U., Svangstu, K., Ang, C. S., & Zaphiris, P. (2011). Social roles in an online support community for older people. *International Journal of Human Computer Interaction*, 27(4), 323–347.
80. Quora. Quora Partner Program. <https://www.quora.com/q/quorapartnersupdates/A-guide-to-partner-program-compensation>. Accessed 04 Apr 2020
81. Ribeiro, L. F., Saverese, P. H., Figueiredo, D. R. (2017). struc2vec: Learning node representations from structural identity. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 385–394. ACM
82. Santos, T., Walk, S., Kern, R., Strohmaier, M., & Helic, D. (2019). Activity archetypes in question-and-answer (q&a) websites—a study of 50 stack exchange instances. *ACM Transactions on Social Computing*, 2(1), 4.
83. Saxena, A., Iyengar, S. (2020). Centrality measures in complex networks: A survey. arXiv preprint. [arXiv:2011.07190](https://arxiv.org/abs/2011.07190)
84. Saxena, A., Saxena, P., Reddy, H., Gera, R. (2019). A survey on studying the social networks of students. arXiv preprint. [arXiv:1909.05079](https://arxiv.org/abs/1909.05079)
85. Schroer, J., & Hertel, G. (2009). Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it. *Media Psychology*, 12(1), 96–120.
86. Singh, V., Johri, A., Mitra, R. (2011). Types of newcomers in an online developer community. In: Proceedings of the ACM 2011 conference on computer supported cooperative work, pp. 717–720. ACM
87. StackExchange. StackExchange all sites. <https://stackexchange.com/sites#users>. Accessed 10 May 2020
88. Sun, N., Rau, P. P. L., & Ma, L. (2014). Understanding lurkers in online communities: A literature review. *Computers in Human Behavior*, 38, 110–117.
89. Turner, T. C., Fisher, K. E. (2006). The impact of social types within information communities: Findings from technical newsgroups. In: Proceedings of the 39th annual Hawaii international conference on system sciences (HICSS'06), vol. 6, pp. 135b–135b. IEEE
90. Turner, T. C., Smith, M. A., Fisher, D., & Welser, H. T. (2005). Picturing usenet: Mapping computer-mediated collective action. *Journal of Computer Mediated Communication*, 10(4), JCMC1048.
91. Viégas, F. B., Smith, M. (2004). Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. In: Proceedings of the 37th annual Hawaii International conference on system sciences, p. 10 IEEE
92. Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.
93. Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., Smith, M. (2011). Finding social roles in wikipedia. In: Proceedings of the 2011 iConference, pp. 122–129. ACM
94. Welser, H. T., Gleave, E., Fisher, D., & Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, 8(2), 1–32.
95. White, A. J., Chan, J., Hayes, C., Murphy, B. (2012). Mixed membership models for exploring user roles in online fora. In: Sixth international AAAI conference on weblogs and social media

96. White, D. R., & Reitz, K. P. (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5(2), 193–234.
97. Whittaker, S., Terveen, L., Hill, W., Cherny, L. (2003). The dynamics of mass interaction. In: From Usenet to CoWebs, pp. 79–91. Springer
98. Wikipedia. Kullback-Leibler Divergence. https://en.wikipedia.org/wiki/Kullback-Leibler_divergence. Accessed 19 Apr 2020
99. Wikipedia. Usenet Newsgroups. https://en.wikipedia.org/wiki/Usenet_newsgroup. Accessed 04 July 2019
100. Wikipedia. Wikipedia:Wikipedians. <https://en.wikipedia.org/wiki/Wikipedia:Wikipedians>. Accessed 07 Dec 2019
101. Wuetherick, B. (2010). Basics of qualitative research: Techniques and procedures for developing grounded theory. *Canadian Journal of University Continuing Education*. <https://doi.org/10.21225/D5G01T>.
102. Yang, D., Halfaker, A., Kraut, R., Hovy, E. (2016). Who did what: Editor role identification in wikipedia. In: Tenth international AAAI conference on web and social media
103. Yang, D., Kraut, R. E., Smith, T., Mayfield, E., Jurafsky, D. (2019). Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In: Proceedings of the 2019 CHI conference on human factors in computing systems, p. 344. ACM
104. Yang, D., Wen, M., Rose, C. (2015). Weakly supervised role identification in teamwork interactions. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Vol. 1: Long Papers), pp. 1671–1680
105. Yang, J., Tao, K., Bozzon, A., Houben, G. J. (2014). Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In: International conference on user modeling, adaptation, and personalization, pp. 266–277. Springer
106. Zhang, X., & Wang, C. (2012). Network positions and contributions to online public goods: The case of chinese wikipedia. *Journal of Management Information Systems*, 29(2), 11–40.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.