Applied Network Science

RESEARCH Open Access

Relating Wikipedia article quality to edit behavior and link structure

Thorsten Ruprechter*, Tiago Santos and Denis Helic

*Correspondence:
ruprechter@tugraz.at
Institute of Interactive Systems and Data Science, Graz University of Technology, Inffeldgasse 16c, 8010 Graz AT Austria

Abstract

Currently, the relation between edit behavior, link structure, and article quality is not well-understood in our community, notwithstanding that this relationship may facilitate editing processes and content quality on Wikipedia. To shed light on this complex relation, we classify article edits and perform an in-depth analysis of editing sequences for 4941 articles. Additionally, we build a network of internal Wikipedia hyperlinks between articles. Using this data, we compute parsimonious metrics to quantify editing and linking behavior. Our analysis unveils that conflicted articles differ substantially from others in almost all metrics, while we also detect slight trends for high-quality articles. With our network analysis we find evidence indicating that controversial and edit war articles frequently span structural holes in the Wikipedia network. Finally, in a prediction experiment we demonstrate the usefulness of edit behavior patterns and network properties in predicting conflict and article quality. With our work, we assist online collaboration communities, especially Wikipedia, in long-term improvement of content quality by offering valuable insights about the interplay of article quality, controversies and edit wars, editing behavior, and network properties via sequence-based edit and network-based article metrics.

Keywords: Wikipedia, Edit behavior, Link structure, Article quality, Edit wars, Controversy, Conflict, Semantic edit types

Introduction

Editing behavior on Wikipedia has been a widely studied subject in previous research (Brandes et al. 2009; Flöck et al. 2017; Kittur et al. 2007; Yang et al. 2017; Yasseri et al. 2012). In particular, past studies investigated the misbehavior on Wikipedia including, among others, vandalism (Adler et al. 2011; Kumar et al. 2015), conflict or controversy (Borra et al. 2015; Yasseri et al. 2014), and so-called edit wars (Sumi et al. 2011; Yasseri and Kertész 2013). Edit wars are a type of behavior in which two or more opposing editors (or editor groups) override each other's content due to differences in opinion on a given subject. Prominent examples of such behavior include the Wikipedia pages on Nikola Tesla[1] or Barack Obama.[2] In controversial articles, disputes mainly arise because

[1] https://en.wikipedia.org/wiki/Nikola_Tesla
[2] https://en.wikipedia.org/wiki/Barack_Obama

their content represents controversial issues of our society, such as Evolution[3]. Scientifically, studying edit wars and controversial articles on Wikipedia has been and still is an important endeavor, as several studies show the connection between article quality and editing behavior on Wikipedia (Arazy et al. 2011; Editorial 2006; Lerner and Lomi 2019; Samoilenko et al. 2018). Practically, this research may help upgrade prevalent conflict prediction methods, such as those using mutual reverts (Gandica et al. 2014; Yasseri and Kertész 2013).

This paper extends our previous work, which investigates Wikipedia editing behavior in regard to the relation of editing and linking across article quality categories (Ruprechter et al. 2019). In this work, we expand our investigations towards low- and high-quality content, controversial issues, and edit war articles in the Wikipedia network. We further demonstrate the predictive potential of computed article metrics, especially along the dimensions of non-conflicted versus conflicted and low-quality versus high-quality articles. To that end, we are interested in the following research questions:

I) **Characterization.** How can we characterize editing behavior on Wikipedia?

II) **Relation existence and strength.** Is there a relation between editing behavior, article quality, and internal Wikipedia hyperlink (wikilink) network topology? If such a relation exists, how strong is it?

III) **Prediction.** Are editing behavior and wikilink network metrics predictive of edit war articles, controversy, or article quality?

To answer these research questions, we: (i) classify edit actions for 4 941 Wikipedia articles using a state-of-the-art machine learning approach, (ii) compute relative frequencies of edit actions and build first-order Markov chains from edit sequences for each of our Wikipedia articles, (iii) perform statistical significance tests on results to characterize differences in editing behavior, (iv) extract a wikilink network for our article sample, (v) compute and compare standard network metrics of sampled articles given article quality, and (vi)perform classification experiments showing the applicability of editing behavior and network metrics as prediction parameters.

We find that there is significant difference in editing behavior between edit wars or similar controversially edited content and higher quality articles, which corroborates previous results on edit wars (Sumi et al. 2011; Yasseri et al. 2012). Adding to those studies, we find that editors of conflicted articles make meaning-changing edits significantly more often while performing less formatting and potentially less link-editing operations, thus rendering factual content especially contested. Conversely, high-quality articles show the largest amount of format actions. This increased amount of formatting may relate to the polishing necessary to meet the standards of high-quality content on Wikipedia. We conjecture that the differing editing behavior of conflicted content influences link structure, and vice versa. Consequentially, we find that edit war articles are clear outliers in several standard network measures, with controversial articles also bearing clear differences to low- or high-quality articles. Particularly, edit war articles have, on average, significantly higher in-degree, out-degree, PageRank, and k-core in contrast to a lower reciprocity, clustering coefficient, and average path length. In addition, betweenness results of disputed articles suggest their location in structural holes of the wikilink network, supposedly taking on

[3] https://en.wikipedia.org/wiki/Evolution

a brokerage role between densely clustered article subgroups, topics, and corresponding editor communities. Furthermore, the distribution of network metric quantities differs substantially from non-conflicted articles in both moments and shape, with shape frequently taking on non-typical forms for such a Web-based editing processes. Overall, computed metrics suggest considerable differences of controversial and edit war articles in comparison to low- and high-quality content on Wikipedia.

With our work, we provide practical contributions for Wikipedia's community. Firstly, we depict a clear trend in semantic edit intentions and the occurring disparities in the underlying link structure of different article categories. Secondly, we demonstrate the usefulness of these article features as indicators of quality, controversy, and edit wars via logistic regression. These results may inform further development of existing Wikipedia content-assessment tools such as Huggle,[4] Contropedia (Borra et al. 2015), and ORES (Halfaker et al. 2018). Moreover, we facilitate new solutions for problems such as editor role identification in Wikipedia or other online collaboration systems through our findings on semantic edit labels. Additionally, our approach quantitatively outmatches similar studies which apply semantic labeling of Wikipedia edits (Daxenberger and Gurevych 2012; Yang et al. 2016). On top of that, we suggest a novel way of interpreting the attraction of editors to contentious content on Wikipedia based on network metrics. Altogether, this work proposes a practical framework which combines Wikipedia editing behavior, article quality, edit wars, controversial articles, and wikilink structure to extend these separate lines of inquiry. Our base methods can be readily applied to similar domains, as we make our code available on GitHub.[5]

Related work

Controversy, conflict, and edit wars on Wikipedia. In the context of Wikipedia, the terms controversy, conflict, edit wars, and vandalism are sometimes used interchangeably, despite their actual meaning being fairly different. Firstly, conflict and controversy do not inherently resemble destructive behavior on Wikipedia (Kittur et al. 2007). In fact, reasonable controversies over contested knowledge have quite the opposite effect. Diverse points of view seem to foster knowledge generation and even increase quality of articles (Shi et al. 2019). Recently, researchers studied the relation between editing behavior, content quality, and conflict via editor collaboration patterns (Arazy et al. 2011; Lerner and Lomi 2019; 2020). Within this scope, (Lerner and Lomi 2020) analyzed positive and negative relations between editors to investigate the creation process of controversial articles.

In order to further quantify effects of controversy and conflict, previous work introduced corresponding metrics (Borra et al. 2015; Flöck et al. 2017; Yasseri et al. 2014). Contrarily to conflict and controversy, vandalism and edit wars distinctively represent destructive behavior on Wikipedia (Adler et al. 2011; Potthast et al. 2008). Automatically detecting vandalistic contributions and performing counter-vandalism actions on articles presents a well-developed research area (Adler et al. 2011; Halfaker et al. 2018; Kumar et al. 2015). Similarly, broad research exists about edit war prediction in Wikipedia articles (Sumi et al. 2011; Yasseri et al. 2012). Multiple authors proposed using mutual reverts of Wikipedia revisions as detectors of this behavior (Gandica et al. 2014; Yasseri and Kertész

[4] https://en.wikipedia.org/wiki/Wikipedia:Huggle
[5] https://github.com/ruptho/editlinkquality-wikipedia

2013). Mutual reverts occur when two editors revert each other's revisions. Controversy metrics based solely on mutual reverts detect vandalism and edit wars very well.

In contrast to mutual reverts, we use granular edit actions as well as network metrics to analyze and predict controversial and edit war articles. We believe that considering all edit actions could improve prevalent detection mechanisms and extend existing understanding of controversy and edit wars on Wikipedia. In addition, predicting potential future conflicts arising due to an article's location in the Wikipedia hyperlink network could be important for quality assurance and conflict moderation. Considering these conjectures, we explore the role of arbitrary edit actions and network properties in controversies and edit wars on Wikipedia.

Wikipedia networks. In past research, researchers frequently analyzed collaboration and social structure in Wikipedia editor networks, as well as their effect on article quality (Brandes et al. 2009; De La Robertie et al. 2015; Li et al. 2015; Liu and Ram 2018; Platt and Romero 2018). In studies on Wikipedia hyperlink networks, authors investigated link structure to leverage semantic (Milne and Witten 2008), topical (Coursey and Mihalcea 2009), or categorical (Suchecki et al. 2012) information. Dimitrov et al. (2017) explored link success and found that users seem to frequently choose links leading to target Wikipedia articles less prominent than the source article, or to one with a similar topic. In another line of work, authors discovered that featured articles are "more central" than others in specific Wikipedia language editions, depicted by their lower clustering coefficient and shorter average path length (Ingawale et al. 2013). This finding supports the general theory by Burt (2001) that nodes spanning structural holes of a network act as brokers of information and profit from their position between different communities or groups. Such nodes benefit from access to information and knowledge which are not universally accessible, as resources circulate between groups through structural holes (Burt 1992). By the same token, (Granovetter 1973) argues that weak ties between nodes carry novel information, while strong ties most likely only transfer information familiar to the involved nodes. Although (Burt 1992) and (Granovetter 1973) base these theories on social networks, we argue that similar effects are applicable to information networks such as Wikipedia, as articles spanning structural holes connect topical subgroups. Other authors suggested that in the Wikipedia network out-degree correlates with quality, in-degree with popularity, and PageRank with importance for the Portuguese Wikipedia — although correlation was generally weak for quality and importance, while being moderate for popularity (Hanada et al. 2013). In 2009, a study provided information about general network metrics for Wikipedia, claiming that median values for article in- and out-degree were 4 and 12, while average degree was 20.63 (Kamps and Koolen 2009). On a completely different note, article pages contain text links created by human editors as well as links generated by templates. To solely focus on editor-created links, (Consonni et al. 2019) proposed the WikiLinkGraphs dataset, which only includes wikilinks in article texts and excludes automatically generated links.

Extending those studies, we provide novel findings about internal Wikipedia links to unravel relations between article hyperlinks, quality, and potential conflicts.

Wikipedia editing behavior. When investigating editing behavior in a general context, researchers formerly proposed multiple edit label taxonomies, considering both semantic and syntactic changes. Early works differentiate edits which either change ("Text-Base") or preserve ("Surface") the meaning of texts (Faigley and Witte 1981). Later, authors introduced more sophisticated label taxonomies, adapted to the context of Wikipedia (Antin et al. 2012; Daxenberger and Gurevych 2012; Habernal et al. 2016; Yang et al. 2016) or other online collaboration systems such as StackOverflow (Yang et al. 2014). In 2016, the Wikimedia foundation deployed an experimental three-level taxonomy for article edits,[6] which is structured into 14 semantic intentions, 18 syntactic elements, and three editor actions (Yang et al. 2017). We apply an edit action classification approach which adapts this taxonomy's 14 semantic edit labels.

Altogether, we extend existing literature on editing behavior and edit wars on Wikipedia by analyzing the triplet of semantic edit actions, Wikipedia network structure, and article quality.

Materials and methods

Background and preliminaries

Namespaces. Wikipedia organizes pages into numbered namespaces. In this work, we focus on analyzing pages in namespace 0 ("ns0") of the English Wikipedia, which represents all article pages. We ignore content in other namespaces, for example article talk ("ns1") or user pages ("ns2").

Revisions, edits, wikilinks. Human editors generate most article contents on Wikipedia. Both registered and unregistered editors perform changes to articles via revisions (Sage Ross 2014). Each revision consists of edits which either insert, modify, or delete content of articles. While editing, editors create hyperlinks to other Wikipedia articles (i.e., wikilinks) as well as external pages. Wikilinks add useful context for readers and connect articles into a Wikipedia hyperlink network, enabling readers to follow the flow of topical or categorical information on Wikipedia.

Content assessment on wikipedia. Wikipedia establishes article ratings using a well-defined content assessment system[7]. This system allows for assessment of articles according to quality and importance. In this work, we focus on article quality as the main distinguishing factor for articles. Wikipedia defines concise guidelines for assessing article quality. Quality assessments of articles range from highest to lowest, including: Featured (FA), A-class (A), Good (GA), B-Class (B), C-class (C), Start, and Stub articles. Start and Stub pages usually represent newly created or very short articles. For our analysis, we ignore such articles due to their very low quality and in many cases short revision history. In addition, Featured List (FL) and List pages exist, in which listed items represent links to articles and include supplemental information. We also exclude FL and List articles from our analysis because of their peculiar content structure, which is atypical for regular articles.

Wikipedia derives quality ratings from tags applied by users of separate WikiProjects.[8] While articles usually have different ratings across various WikiProjects, the final assessment for an article is determined by its best rating over all WikiProjects. For example, as of June 2019 the Wikipedia article about the Austrian city Graz[9] is rated C by "WikiProject Cities" but B by "WikiProject Austria", thus producing a B rating overall. However, the highest-quality content on Wikipedia, such as GA and FA, must be approved separately through a review process. During this review, several editors assess factors such as

[6] https://en.wikipedia.org/wiki/Wikipedia:Labels/Edit_types/Taxonomy
[7] https://en.wikipedia.org/wiki/Wikipedia:Content_assessment
[8] https://en.wikipedia.org/wiki/Wikipedia:WikiProject
[9] https://en.wikipedia.org/wiki/Graz

Table 1 Number of Rated Articles by Quality as of 7th June 2019. The quality categories relevant to our analysis are FA, A, GA, B, and C. Articles in other categories can be divided into 3 346 793 Stub, 1 792 928 Start, 1 956 FL, 249 008 List, and 509 940 unassessed articles

Quality	FA	A	GA	B	C	Others	**Total**
#Articles	6 705	1 874	32 759	127 107	308,621	5 900 625	6 377 691

accuracy, neutrality, and completeness according to a particular set of criteria[10]. Out of the 6 377 691 articles in the English Wikipedia, there are 6 705 FA, 1 874 A, 32 759 GA, 127 107 B, 308 621 C, 1 792 928 Start, 3 346 793 Stub, 1 956 FL, 249 008 List, and 509 940 articles without assessment. Table 1 summarizes Wikipedia quality assessment data as of June 2019.

Dataset

We utilize the Wikimedia API RevScoring[11] to process revisions. RevScoring is an automatic revision scoring system developed as the library powering the Wikimedia service ORES (Halfaker et al. 2018). ORES is a tool developed by the Wikimedia foundation which enables automatic assessment of quality and importance of Wikipedia pages as well as individual article revisions.

For our dataset, we retrieve and process a total of 4 941 articles. First of all, we collect revision histories of articles deemed to be especially controversial or contain edit wars. We therefore sample a total of 1 000 articles from the following sources:

I) **"Most conflicted" articles.** A sample of 450 articles deemed most conflicted by metrics calculated using the TokTrack dataset (Flöck et al. 2017).

II) **"Most controversial" articles**. Yasseri et al. (2014) collected 100 of the most controversial articles for ten Wikipedia language editions based on mutual reverts of revisions, from which we gather those in the English Wikipedia.

III) **"Lamest Edit Wars" on Wikipedia.**[12] We extract 50 articles from this short list of prominent edit war articles on Wikipedia, including edit wars about ethnicity and nationality, politics and religion, or spelling.

IV) **"List of Controversial Issues" on Wikipedia.**[13] These articles are monitored because they are subject to constant circular re-editing due to their controversial content. We retrieve 400 articles from this list.

After removal of duplicates (article names present in multiple lists), 941 articles remain: 401 "most conflicted" articles, 94 "most controversial" articles, 50 "Lamest Edit Wars", and 396 articles from the "List of Controversial Issues". The first three lists contain 545 articles that have been classified as conflicted either via edit-based metrics such as mutual reverts (Lists I and II) or manual detection of edit warring (List III). We combine these three article lists into a category which we term "Edit Wars" (EW). Besides, we treat List IV, which represents articles about controversial issues (CI), as a separate category. These 396 manually labeled articles contain contentious topics, which embody societally divisive issues instead of merely disputes between specific editors (Lerner and Lomi 2020). Afterwards, we retrieve and process revision histories of a sample of 800 articles per Wikipedia quality

[10] https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria
[11] https://github.com/wikimedia/revscoring
[12] https://en.wikipedia.org/wiki/Wikipedia:Lamest_edit_wars
[13] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

category (FA, A, GA, B and C) with more than 50 revisions. We combine FA, A, and GA into a collection of high-quality articles (HQ), while merging B and C into a low-quality category (LQ). For our analysis, we consider LQ, HQ, EW, and CI articles, thus resulting in a dataset of 4 941 revision histories.

Labeling edit actions on Wikipedia

Edit actions. To label edit actions, we adopt a multi-label classification approach. For robust labeling of the edit actions we start with the 14-label[14] taxonomy proposed by Yang et al. (2016) and combine those 14 labels into three super-labels: Content, Format, and WikiContext. Firstly, the Content label captures all edits aiming towards modifying actual information on the article page. It combines fact-updates, simplification, elaboration, clarification, increasing verifiability, and establishing a neutral point of view. Secondly, the Format label describes all editing actions not changing the meaning of texts, facts, or contained information. This includes refactoring, copy-editing, manipulating wikilinks and wiki markup, as well as link disambiguation. Lastly, the WikiContext label captures all wiki-specific interactions such as modifying processing tags, vandalism, counter-vandalism, or other intentions.

Classification of edit actions. We retrieve and build the revision feature set for our multi-label classifier by utilizing most of the framework published alongside the 14-label taxonomy. We derive features as well as training set from previous work (Yang et al. 2017). Per revision, we use RevScoring to retrieve 163 base features. These features consist of text features (e.g., differences in words, punctuation, and numbers), revision data (e.g., timestamp and comment), as well as editor information (e.g., user registration date). We further process this feature set to produce 207 final features for classification. Through this, we extend the 163 base features by several more advanced text features such as the number of spelling errors, markup changes, differences in stemmed text in comparison to the preceding revision, amount of relocated text, and information about special content such as templates, files, or references.

Our classification training set consists of 5 777 manually labeled revisions (Yang et al. 2017). We transform this dataset to fit our three-category taxonomy through aggregation of edit labels. For this, we transform training labels into corresponding super-labels (Content, Format, and WikiContext) according to the aforementioned label combinations. Furthermore, we collect additional revision samples for the WikiContext label via retrieving edits performed by the anti-vandalism bot ClueBotNG.[15] The bot's edits are categorized as examples of counter-vandalism, while their preceding revisions are classified as vandalism (0.1% false positive rate). By combining these samples with the training dataset, we accumulate 6 670 multi-label revisions which contain 3 497 Format, 2 346 Content, and 1 641 WikiContext edits.

Finally, to determine a feasible configuration for a multi-label Random Forest classifier given our training set and features, we apply grid search using a 80–20 train/test split with 10-fold cross validation. We find a suitable configuration (weighted F1 score of 0.8153) with a parameter setting of 750 estimators, a maximum tree depth of 25, and 50% of features considered for finding the best split.

[14] Copy Editing, Clarification, Simplification, Point of View, Refactoring, Fact Update, Elaboration, Verifiability, Link-Disambiguation, Wikification, Vandalism, Counter-Vandalism, Process, and Other Intentions
[15] https://en.wikipedia.org/wiki/User:ClueBot_NG

Modeling edit action sequences

We characterize editing behavior in Wikipedia articles using two metrics: relative label frequency and label transition probability.

Relative edit label frequency. We compute relative edit label frequency for each category by macro-averaging the relative frequencies of their articles. Accordingly, we first calculate relative frequency of the automatically created edit labels for all articles. After that, averaging the article values in each quality category produces the per-category macro-averages.

Edit label transition probability. For our investigation of label transitions, we compute first order Markov chains from the automatically labeled article revision histories. Similar to the computation of relative label frequencies explained above, we accumulate macro-averaged transition probabilities for all quality categories. Firstly, transition probabilities are computed from edit label sequences in article revision histories. Subsequently, we average article results for each quality category to generate per-category label transition probabilities.

Characterizing differences in categories. To assess statistical significance of differences in label frequency and transition probability we perform pair-wise permutation tests (Chandrasekharan et al. 2017; Vautard et al. 1990) between all article categories. These tests compare the distance of observed category means to category means of randomly permuted assignments of articles to categories. The null hypothesis for our test states that values for article subsets drawn from different categories stem from the same probability distribution.

Network of wikilinks

Wikilink graph. We generate a wikilink graph by employing the framework used to create the WikiLinkGraphs dataset (Consonni et al. 2019). This framework enables graph generation solely from wikilinks in article texts extracted from Wikipedia XML dumps.[16] We thereby exclude automatically generated links (e.g., via Wikipedia templates) and only include wikilinks which were intentionally modified by editors. In the resulting graph, each article is a node and wikilinks between articles are edges. We generate a graph from Wikipedia dumps for June 2019. In addition, we execute our own post-processing pipeline, which removes nodes modeling redirects between Wikipedia articles (i.e., nodes with a single outgoing link) and resolves duplicate article titles. For the English Wikipedia article namespace, the graph contains 5 879 005 articles connected by 163 526 307 wikilinks. The largest strongly connected component holds 5 241 679 nodes and 155 056 731 edges.

The resulting WikiLinkGraphs dataset misses 59 (3 LQ, 20 HQ, 11 EW, 25 CI) of our 4 941 investigated articles, caused by invalid redirects or broken wikilinks.

Network metrics. We investigate our wikilink network via typical parsimonious network metrics and calculate empirical complementary cumulative distribution functions (CCDF) of in-degree (deg^-), out-degree (deg^+), PageRank (*PR*) (Page et al. 1999), reciprocity (*r*) (Garlaschelli and Loffredo 2004), directed clustering coefficient (*CC*) (Watts and Strogatz 1998), and *k*-core (Shin et al. 2016).

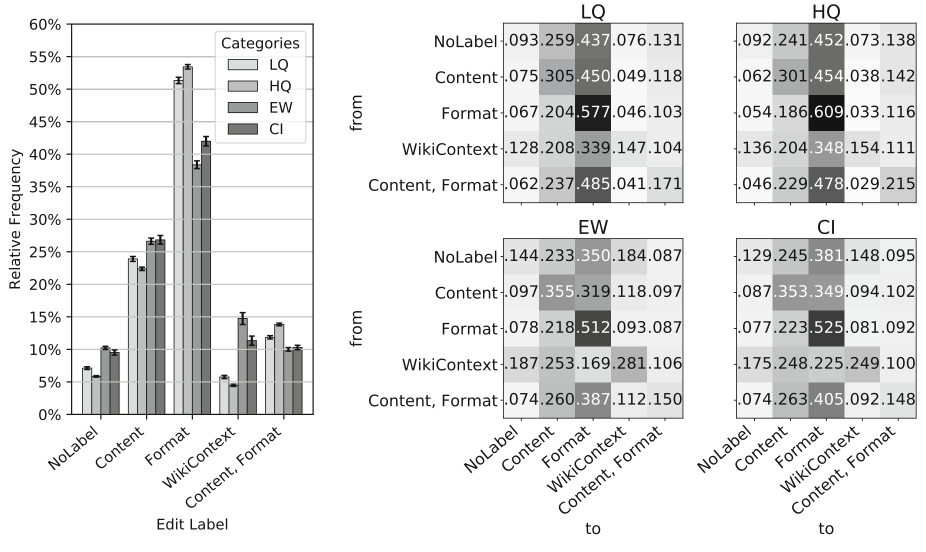
[16] https://dumps.wikimedia.org/

Results and discussion

Relative edit label frequency

Results. We show results for category-wise relative label frequencies in Fig. 1a. We exclude results for any label combination besides Content and Format, due to their extremely low relative frequency (< 0.0004). Furthermore, we introduce a label class termed “NoLabel” to encompass revisions where our classifier could either not assign a definite label, which were deleted, or where other inconsistencies occurred.

We find several substantial disparities between various article categories. Relative edit label frequency for EW deviates considerably from LQ and HQ, while also exhibiting slight differences to CI for some labels. Revisions for articles in EW contain on average 38.4% Format (which includes wikilink manipulations), 26.6% Content, 14.8% WikiContext, 10% Content and Format, and 10.2% NoLabel edits. CI article histories mostly show similar edit label frequencies, although they generate moderately fewer WikiContext (11.3%), which includes vandalism, in favor of more Format edits (42.1%). Alternatively, HQ (53.4%) is the overall highest-scoring category for Format, somewhat ahead of LQ (51.4%). As opposed to the lower Format score, LQ articles are subject to relatively greater amounts of Content and NoLabel edits than HQ (23.9% and 7.1% versus 22.3% and 5.8%, respectively). Considering WikiContext edits, EW and CI revisions contain significantly more such modifications than non-conflicted categories ($< 6\%$ for LQ and HQ). On top of that, EW and CI exhibit substantially greater amounts of Content (26.6% and 26.8%) and NoLabel (10.2% and 9.5%) editing. Overall, we conclude that relative edit label frequencies for EW and CI differ considerably from LQ and HQ. Results for EW and CI deviate



LQ	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.093	.259	.437	.076	.131
Content	.075	.305	.450	.049	.118
Format	.067	.204	.577	.046	.103
WikiContext	.128	.208	.339	.147	.104
Content, Format	.062	.237	.485	.041	.171

HQ	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.092	.241	.452	.073	.138
Content	.062	.301	.454	.038	.142
Format	.054	.186	.609	.033	.116
WikiContext	.136	.204	.348	.154	.111
Content, Format	.046	.229	.478	.029	.215

EW	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.144	.233	.350	.184	.087
Content	.097	.355	.319	.118	.097
Format	.078	.218	.512	.093	.087
WikiContext	.187	.253	.169	.281	.106
Content, Format	.074	.260	.387	.112	.150

CI	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.129	.245	.381	.148	.095
Content	.087	.353	.349	.094	.102
Format	.077	.223	.525	.081	.092
WikiContext	.175	.248	.225	.249	.100
Content, Format	.074	.263	.405	.092	.148

(a) Relative frequency (b) Transition probability

Fig. 1 Relative frequency and transition probability for edit labels. In Fig. 1a, we visualize relative label frequencies by article category (bootstrapped 95% confidence intervals). EW articles are subject to the least amount of Format edits and are most likely to receive WikiContext edits, marking them significantly dissimilar from non-conflicted LQ and HQ articles. Results for CI yield similar effects as EW, although differences to LQ and HQ are not as considerable. Figure 1b shows transition probabilities between edit labels per article category. Most notably, EW and CI article histories contain fewer successive Format edits and less formatting in general than those in LQ or HQ. EW and CI also exhibit a greater probability for consecutive Content or WikiContext edits

slightly from each other for Format and WikiContext. At the same time, we observe scarce dissimilarities between LQ and HQ.

Permutation tests for relative frequencies reveal statistically significant differences across all article categories ($p < 0.01$ after Bonferroni correction), except for Content as well as combined Content and Format edits for EW versus CI.

Discussion. Relative frequencies of edit actions in conflicted EW and CI articles follow significantly different patterns than non-conflicted articles in LQ and HQ. The increased frequencies for Content, WikiContext, or NoLabel indicate editors regularly rewriting EW and CI articles. This behavior possibly hints towards the high potential contentiousness of the content. In general, NoLabel edits signal that a revision contains atypical content, such as ASCII art,[17] or that it was removed by administrators. Therefore, we argue that EW's high number of NoLabel edits is a by-product of increased vandalism. When comparing articles according to quality, HQ's peak in formatting frequency exemplifies editor efforts to create content which complies with the standards of high-quality articles on Wikipedia. LQ shows similar patterns as HQ across all edit labels, although differences are still significant. Altogether, our results confirm that content is significantly more conflicted in EW and CI. On the one hand, conflicts about facts can lead to increased article quality. On the other hand, frequent editor conflict might also raise the possibility of destructive behavior. In contrast to frequent content editing, we detect a substantially lower amount of formatting actions in EW and CI than LQ and HQ, potentially indicating less link-editing operations.

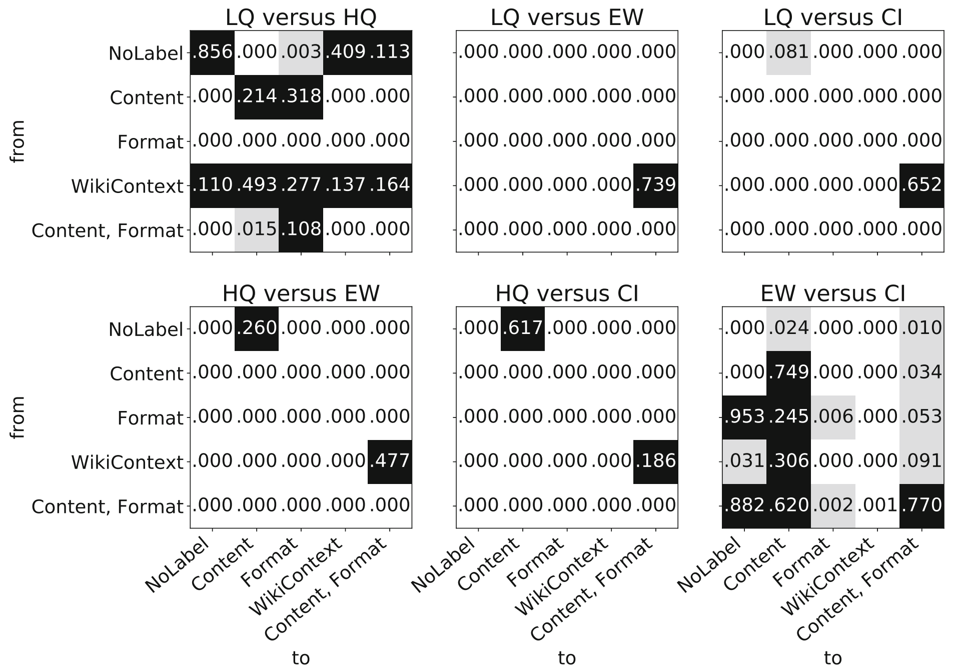
Label transition probabilities

Results. We present label transition probabilities in Fig. 1b, which unveils that probabilities for LQ and HQ are rather different from EW and CI. Particularly, results for HQ show the highest probability of consecutive Format actions (0.609), followed by LQ (0.577). Successive Format edits occur less frequently in EW (0.512) and CI (0.525). In general, EW and CI produce fewer transitions involving Format than LQ and HQ. By contrast, EW and CI more strongly lean towards sequences containing Content, NoLabel, and WikiContext, making their revision histories highly dissimilar from LQ and HQ. The most notable difference between EW and CI is marked by articles in EW generating slightly more transitions ending in WikiContext edits. On the whole, these findings principally corroborate the difference of EW and CI to LQ and HQ as depicted in the relative label frequencies.

Permutation test results in Fig. 2 largely confirm a statistically significant difference of EW and CI to LQ and HQ ($p < 0.002$ after Bonferroni correction). Test results also highlight the similarity between EW and CI, as well as LQ and HQ.

Discussion. Transition probabilities for LQ and HQ suggest that editors more strongly focus on content formatting in these non-conflicted articles. As for quality, the considerably higher probabilities for consecutive Format edits in HQ may stem from such articles representing some of the highest-quality Wikipedia content. Due to Wikipedia's policies for high-quality content, these articles often undergo phases of intensive formatting when considered for promotion to featured or good articles, which we collected in HQ. This process further explains why LQ articles contain less Format transitions than those in HQ. Contrarily, EW and CI are characterized by higher Content, WikiContext, or NoLabel

[17] https://en.wiktionary.org/wiki/ASCII_art



LQ versus HQ

	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.856	.000	.003	.409	.113
Content	.000	.214	.318	.000	.000
Format	.000	.000	.000	.000	.000
WikiContext	.110	.493	.277	.137	.164
Content, Format	.000	.015	.108	.000	.000

LQ versus EW

	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.000	.000	.000	.000	.000
Content	.000	.000	.000	.000	.000
Format	.000	.000	.000	.000	.000
WikiContext	.000	.000	.000	.000	.739
Content, Format	.000	.000	.000	.000	.000

LQ versus CI

	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.000	.081	.000	.000	.000
Content	.000	.000	.000	.000	.000
Format	.000	.000	.000	.000	.000
WikiContext	.000	.000	.000	.000	.652
Content, Format	.000	.000	.000	.000	.000

HQ versus EW

	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.000	.260	.000	.000	.000
Content	.000	.000	.000	.000	.000
Format	.000	.000	.000	.000	.000
WikiContext	.000	.000	.000	.000	.477
Content, Format	.000	.000	.000	.000	.000

HQ versus CI

	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.000	.617	.000	.000	.000
Content	.000	.000	.000	.000	.000
Format	.000	.000	.000	.000	.000
WikiContext	.000	.000	.000	.000	.186
Content, Format	.000	.000	.000	.000	.000

EW versus CI

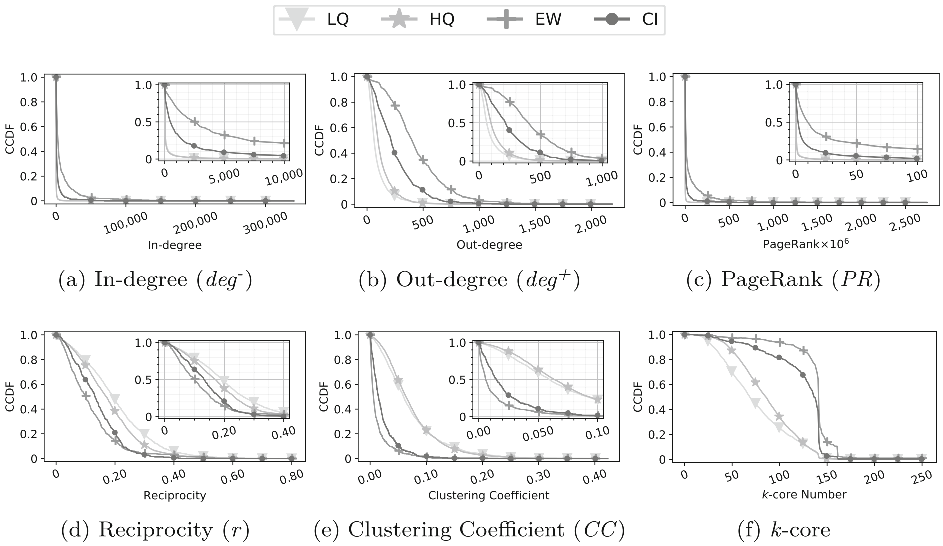
	NoLabel	Content	Format	WikiContext	Content, Format
NoLabel	.000	.024	.000	.000	.010
Content	.000	.749	.000	.000	.034
Format	.953	.245	.006	.000	.053
WikiContext	.031	.306	.000	.000	.091
Content, Format	.882	.620	.002	.001	.770

Fig. 2 Results of permutation tests for transition probabilities. We visualize *p*-values of permutation tests for label transition probabilities of our article categories, with box colors signaling the significance of results. White boxes highlight statistically significant results for $p < 0.002$ (after Bonferroni correction). Black represents non-significant differences which we operationalize as $p > 0.1$. Gray signals *p*-values of $0.002 \leq p \leq 0.1$. Transition probabilities for LQ and HQ show similarities except for transitions originating in Format. Results for EW versus CI mostly demonstrate similarity besides transitions leading to WikiContext. All other category comparisons indicate substantial differences in transition probabilities

probabilities, which possibly infer content disputes among editors. Such detection of conspicuous interaction patterns could potentially be adopted to warn administrators of conflicts arising in specific articles. Besides, our approach of collecting behavioral information could be leveraged for editor role identification or similar applications benefiting from semantic edit labels.

Network of wikilinks

Results. As we visualize in Fig. 3, EW and CI exhibit fairly different CCDFs than LQ and HQ for multiple standard network metrics. At the same time, EW articles produce the most extreme values, usually followed closely by CI, with HQ and LQ trailing behind. Our results for EW indicate that mean deg^-, deg^+, *PR*, and *k*-core are substantially higher, while *r* and *CC* are lower than for other categories. Firstly, non-conflicted articles in LQ and HQ rarely have a deg^- of more than 5 000, while over 30% of EW articles do (Fig. 3a). Secondly, we also observe a higher average deg^+ for EW, although the effect is not as extreme as for deg^- (Fig. 3b). Thirdly, the majority of our non-conflicted articles has extremely low *PR*, leading to hardly any articles reaching a value of $5 \cdot 10^{-5}$, while over 20% of those in EW and close to 10% in CI have a higher *PR* than that (Fig. 3c). Next, even though mean *r* is lower for EW than for other categories the difference is small, especially in comparison to CI (Fig. 3d). LQ and HQ results depict somewhat higher average *r*. Furthermore, over 95% of EW and CI articles have a *CC* lower than 0.075 (Fig. 3e). In contrast, LQ and HQ reach considerably higher *CC* with CCDFs being quite alike for both categories.



(a) In-degree (deg^-)

(b) Out-degree (deg^+)

(c) PageRank (PR)

(d) Reciprocity (r)

(e) Clustering Coefficient (CC)

(f) k-core

Fig. 3 CCDFs of parsimonious network metrics across article categories, with insets highlighting relevant chart areas. In all CCDFs, results for EW stand out, with CI being most similar to EW. Both categories differ substantially from LQ and HQ. We find that, on average, deg^- (Fig. 3a), deg^+ (Fig. 3b), and PR (Fig. 3c) are considerably higher, while r (Fig. 3d) and CC (Fig. 3d) are seemingly lower for EW and CI than for LQ and HQ. Additionally, we attribute the higher k-core numbers (Fig. 3f) to their correlation with degree (Shin et al. 2016)

Moreover, computed k-core numbers indicate that a large number of EW articles (150) is grouped in a connected subgraph in which all vertices have a degree of at least 141 — a higher amount than for all other categories (Fig. 3f). Altogether, we conclude that EW articles show substantially different results for the considered network metrics. Results for CI generally follow similar trends as EW and mostly constitute values in between results of EW and LQ or HQ. In addition, high-quality articles regularly seem close to low-quality articles in terms of CCDFs.

We consequently perform pairwise Mann-Whitney U tests between all categories and metrics for the difference in median. We report statistically significant differences of EW and CI to LQ and HQ for all metrics after Bonferroni correction ($p < 8.33 \cdot 10^{-4}$). EW and CI show significant differences for all metrics besides C_B^{ego}. LQ and HQ test results are significant for all metrics but deg^-, CC, and C_B^{loc}.

Discussion. EW articles exhibit, on average, higher deg^- and PR, signaling frequent referral from other (or more prominent) articles. CI also shows higher values for these metrics than LQ and HQ, although results are not as protruding as for EW. Consequentially, EW and CI articles possibly "attract more attention" than others due to them being in the spotlight. Increased public exposure may lead to a boost in popularity and, as a result, edit wars and controversy. A rapid increase in in-links might therefore signal an article gaining traction due to a recent event or current news. This could possibly be interpreted as a "warning signal" by administrators. It might be feasible to start monitoring or even semi-protect[18] such content to potentially prevent edit wars. Furthermore, we observe lower CC for EW and CI, which could be the consequence of such content being relevant to multiple diverse, closer connected article subgroups. For example, controversial articles such as "United States", "Vladimir Putin", and "World War II" connect different topical

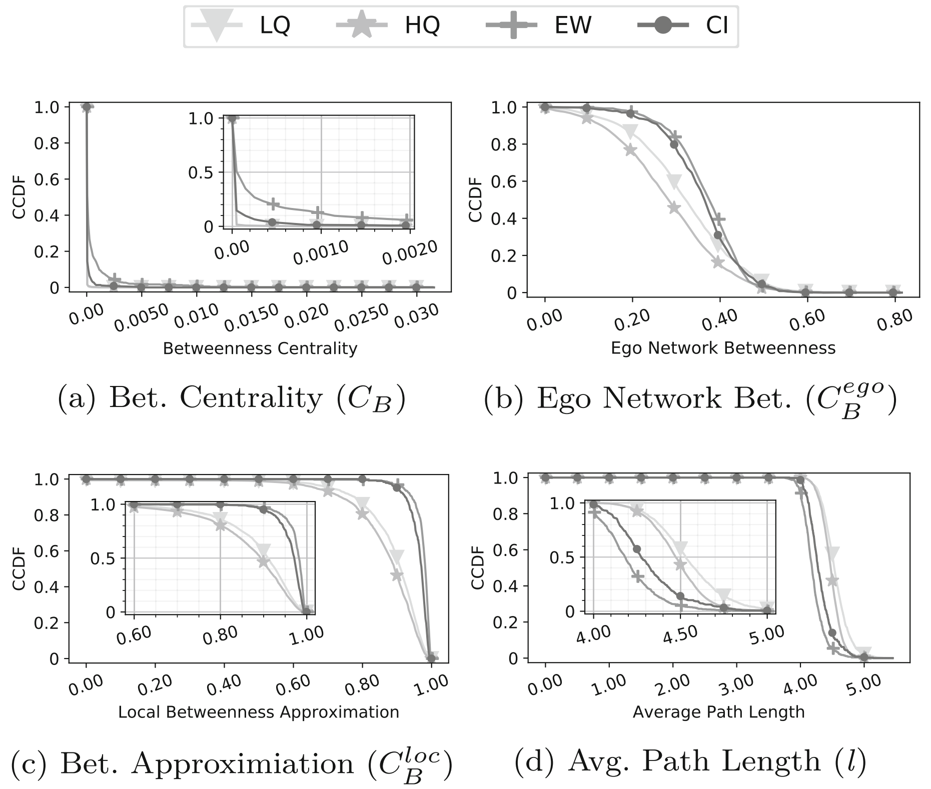
[18] https://en.wikipedia.org/wiki/Wikipedia:Protection_policy#semi

categories on Wikipedia, acting as information brokers between different communities interested in these articles. Considering previous findings about conflict on Wikipedia, this might be a peculiarity of the English language version, where broader topics are more contested (Yasseri et al. 2014). Apparently, the English version merges contributions of people with differing origin and background, thus generating conflict and edit wars — an effect which is not as prevalent for non-English Wikipedia. However, this may also be explained by the negative correlation of clustering coefficient and degree (Ravasz and Barabási 2003).

To shed light on the role of EW and CI articles as connectors of topical subgroups on Wikipedia, we further investigate these articles' network properties. We argue that conflicted articles might frequently be located in structural holes of the wikilink network. Theoretically, articles which span structural holes (Burt 2001) or resemble local bridges (Granovetter 1973) act as brokers of information, thus fitting the characteristics of EW and CI that we hypothesize. We assess these characteristics by computing further network metrics for our 4 941 articles, in particular average path length (l) (Albert and Barabási 2002) as well as three betweenness metrics: betweenness centrality in the whole graph (C_B) (Brandes 2001), betweenness centrality of the article ego network $\left(C_B^{ego}\right)$, and a local approximation of betweenness centrality $\left(C_B^{loc}\right)$. For each article a, C_B^{loc} measures the ratio of the number of shortest paths between neighbors of a that actually go through a to the total number of paths between neighbors going through a.

Figure 4 visualizes CCDFs for the computed centrality metrics, seemingly supporting our assumptions about controversial and edit war articles' possible location in structural holes. Firstly, EW articles are more central than articles in other categories according to the substantially higher C_B (Fig. 4a). CI articles on average also show moderate C_B, thus assuming a more central position than LQ and HQ, which exhibit extremely low values. Although following a different distribution, the CCDF for C_B^{ego} shows about 80% of EW and CI articles holding a higher value than 0.3, with LQ and HQ bearing lower values (Fig. 4b). Interestingly, LQ has a higher average C_B^{ego} than HQ, suggesting that high-quality articles are the least central. This finding is supported by results for C_B^{loc}, which acts as a counterpart to CC. HQ articles' lower C_B^{loc} indicates that high-quality articles lie on less shortest paths between their neighbors in comparison to LQ or EW and CI (Fig. 4c). Furthermore, C_B^{loc} seems to confirm centrality of EW and CI articles. Lastly, l is somewhat lower for EW (4.20) and CI (4.31) than HQ (4.47) and LQ (4.54) in Fig. 4d.

Collectively, centrality findings indicate that controversial and edit war articles are indeed frequently located in structural holes of the wikilink network, acting as brokers of information between topical subgroups and different-minded communities. Previous research of structural holes in Wikipedia attributed such a brokerage role to high-quality articles (especially featured articles), arguing that their ability to broker information increases article quality (Ingawale et al. 2013). Our findings tend to object this characteristic of high-quality content, suggesting considerably higher centrality and shorter average path length for CI and EW articles. However, since we combine good, A-class, and featured articles in HQ, the characteristics might still be prevalent for featured articles alone. Therefore, it appears that brokerage of information across structural holes in a Wikipedia network could be regarded as a double-edged sword. On the one hand, quality of articles in structural holes benefits from a large variety of information sources, because of



(a) Bet. Centrality (C_B)

(b) Ego Network Bet. (C_B^{ego})

(c) Bet. Approximiation (C_B^{loc})

(d) Avg. Path Length (l)

Fig. 4 CCDFs of metrics depicting betweenness across multiple article categories, with insets highlighting relevant chart areas. CCDFs for C_B (Fig. 4a), C_B^{ego} (Fig. 4b), and C_B^{loc} (Fig. 4c) are substantially higher for EW and CI than for LQ and HQ. In addition, l is slightly lower for EW and CI articles (Fig. 4d)

their connections to multiple strongly tied article subgroups. On the other hand, content of such article subgroups usually concerns specific topical or cultural information, thus engaging corresponding communities and increasing polarization (Shi et al. 2019). Either way, our results suggest that controversial and edit war articles hold a particular role as brokers of information between communities in the Wikipedia network.

Relation of network properties, quality, and conflict. In previous work, authors frequently utilized mutual reverts (or similar metrics) as indicators of conflict and edit wars. However, our dataset suggests that such conflict metrics appear to favor articles with a larger number of revisions, considerably differing from those in Wikipedia lists of edit wars and controversial topics. While articles listed on Wikipedia as "Lamest edit wars" and in the "List of controversial issues" average about 5 000 and 4 300 revisions, the most conflicted articles according to researchers' metrics average approximately 9 300 and 14 200 revisions (Flöck et al. 2017; Yasseri et al. 2014). This disparity is important for our analysis, since higher revision counts typically increase article length, which in turn correlates with network properties such as degree (Lamprecht et al. 2016).

To account for these correlations and to better estimate underlying relations between network metrics, quality, controversy, and edit wars we estimate a logistic regression model at article level with article category as the dependent and the given network metrics as independent variables. For this, we perform binary one-versus-one classification experiments for all category combinations with the category labels as dependent variables. Additionally, we include deg^{-}, deg^{+}, due to the prevalent correlation with network metrics, for example clustering coefficient (Ravasz and Barabási 2003). We apply logistic regression with robust variable scaling and L1 regularization.

Table 2 lists the coefficients of network metrics for this logistic regression. Firstly, all coefficients for LQ versus HQ exhibit statistical significance ($p < 0.05$), marking low-quality and high-quality articles significantly different in all investigated network aspects. Furthermore, HQ and CI show significant differences for the regression intercept, deg^{+}, n_{rev}, k-core, and C_B^{loc}, which is aligned with the highest network metric coefficient. When comparing HQ to EW, we identify significance for the network metrics deg^{-}, deg^{+}, n_{rev}, C_B, and C_B^{loc}. Finally, EW appears to be significantly different from CI for deg^{+} and n_{rev}. This might support the supposition that edit war articles may be of greater length due to a higher revision count, and thus contain more outgoing links. We note that further correlation of specific network metrics, for example CC and C_B^{loc} (Spearman's $\rho = -0.93$), could skew p-value results. Accordingly, we remove C_B^{loc} from the regression and repeat our experiments. This renders CC a statistically significant predictor ($p < 0.05$) for regressions in which the coefficient for C_B^{loc} was significant before. We omit the recomputed result table for the sake of brevity. Note that we counter other multi-collinearity issues with L1 regularization. Overall, this logistic regression illustrates the relation of network metrics, article quality, controversy, and edit wars.

Prediction of article category via computed metrics

We now illustrate a practical application of our empirical results via a series of logistic prediction experiments. These experiments utilize the computed editing and network features to predict our article categories, which take the dimensions of article quality and conflict into account. For the experiments, we divide our collected article features into five feature sets:

Table 2 Logistic Regression Results: Network Metric Coefficients. Coefficients for network metrics of one-versus-one logistic regression models predicting article category. Boldface ($p < 0.05$) and italics ($p < 0.1$) denote the statistical significance level of coefficients. Abbreviations resemble the following metrics: in-degree (deg^{-}), out-degree (deg^{+}), number of revisions (n_{rev}), PageRank (PR), reciprocity (r), clustering coefficient (CC), betweenness centrality (C_B), ego network betweenness $\left(C_B^{ego}\right)$, local betweenness approximation $\left(C_B^{loc}\right)$, and average path length (l)

Metrics	**LQ vs HQ**	**LQ vs CI**	**LQ vs EW**	**HQ vs CI**	**HQ vs EW**	**EW vs CI**
deg^{-}	**-0.085**	0.032	0.145	0.027	**0.057**	-0.182
deg^{+}	**0.347**	0.048	**0.764**	**-0.210**	**0.532**	**-1.370**
n_{rev}	**0.058**	**0.839**	**2.551**	**0.397**	**1.075**	**-1.906**
PR	**0.017**	**-0.040**	-0.026	-0.010	-0.007	0.134
r	**-0.258**	**-0.518**	-0.006	-0.206	0.115	-0.085
CC	**-0.545**	-0.107	-0.167	-0.166	-0.093	0.108
k-Core	**0.556**	**1.315**	1.017	**0.634**	0.434	-0.015
C_B	**0.011**	**0.021**	0.004	0.001	**-0.014**	-0.021
C_B^{ego}	**-0.271**	*0.064*	0.039	-0.126	-0.162	-0.022
C_B^{loc}	**-1.245**	**1.059**	0.299	**3.088**	**1.956**	0.175
l	**-0.278**	-0.133	-0.622	0.009	-0.085	-0.227
Intercept	**0.230**	-3.107	**-4.472**	**-3.360**	**-4.438**	**-0.373**

- **Number of revisions ($\mathbf{n_{rev}}$).** Total number of revisions to an article.
- **Control Variables (CV).** Control variables from the previous experiment, namely deg^{-}, deg^{+}, and n_{rev}.
- **Network Metrics (NM).** PR, r, CC, k-core, C_B, C_B^{ego}, C_B^{loc}, and l.
- **Relative Edit Label Frequencies (LF).** Relative frequency of Content, Format, WikiContext, as well as combined Content and Format edits. We exclude results for edits which were not classified with a definite label (NoLabel).
- **Label Transition Probabilities (LT).** Transition probabilities for all possible label combinations besides transitions from and to article edits which could not be definitely classified (NoLabel).

Using these feature sets, we perform binary classification via logistic regression with robust variable scaling and L1 regularization. We carry out binary one-versus-one as well as one-versus-all classification for all category combinations. We evaluate our regression results using 10-fold cross validation and report the ROC-AUC score.

For one-versus-one classification (Table 3), using only n_{rev} as a prediction variable performs exceptionally well for all logistic regressions, especially for those involving EW. Regressions for LQ or HQ versus EW which utilize only n_{rev} achieve respective ROC-AUC scores of 0.968 and 0.961. We explain these particularly strong results by the previously elaborated bias of conflict metrics towards the number of revisions. Extending these predictors with all other features improves classification performance even further (0.996 and 0.997), marking EW easily distinguishable from LQ and HQ. ROC-AUC for CI versus LQ or HQ is slightly lower (0.973 and 0.984). In addition, ROC-AUC scores are somewhat lower for CI versus EW (0.879) and LQ versus HQ (0.805) when using all feature sets. Consequently, our experiments suggest that LQ and HQ are the hardest categories to differentiate. We explain this fact by the overlap of general characteristics between non-conflicted articles. The lower ROC-AUC for CI versus EW may therefore imply a similar overlap for controversial and edit war articles. Although these findings seem to indicate a partial inability to predict quality, they also demonstrate that non-conflicted articles are adequately distinguishable from conflicted ones via regression. When employing single feature sets, all regressions using only CV (deg^{-}, deg^{+}, and n_{rev}) perform remarkably. However, solely utilizing NM, LF, or LT also yields ROC-AUC scores reaching over 75% of the best respective scores. Furthermore, combining editing behavior features (LF and LT) outperforms network features for one-versus-one regression.

From Table 4 we conclude that one-versus-all classification results follow similar trends as those for one-versus-one classification. Most notably, the regression for EW versus all other articles exhibits the highest ROC-AUC (0.986), followed by CI (0.888), HQ (0.853), and LQ (0.823). This further highlights EW's difference to the rest of the investigated content. CI appears to be fairly distinguishable from other categories as well, although classification performance is moderately worse. All in all, these experiments demonstrate the prediction of article categories characterized by quality, controversy, and edit wars via editing behavior and network features.

Conclusion

In this work, we showed that controversial and edit war articles significantly differ from non-conflicted low- and high-quality articles on Wikipedia, both in regard to editing

Table 3 Logistic Regression Results: One-Versus-One Classification. ROC-AUC scores for one-versus-one logistic regression experiments for all category combinations. Feature sets consist of the number of revisions (n_{rev}), control variables (CV), network metrics (NM), label frequencies (LF), and label transition probabilities (LT). Utilizing all feature sets performs best for all experiments

Features	**ROC-AUC**					
	LQ vs HQ	**LQ vs CI**	**LQ vs EW**	**HQ vs CI**	**HQ vs EW**	**CI vs EW**
n_{rev}	0.577	0.870	0.968	0.849	0.961	0.810
CV	0.652	0.923	0.989	0.883	0.978	0.837
NM	0.707	0.861	0.938	0.871	0.937	0.710
LF	0.640	0.807	0.880	0.881	0.928	0.665
LT	0.713	0.825	0.907	0.884	0.933	0.740
CV + NM	0.727	0.930	0.987	0.919	0.978	0.833
CV + LF	0.677	0.947	0.992	0.951	0.991	0.850
CV + LT	0.736	0.952	0.994	0.947	0.992	0.874
NM + LF	0.742	0.906	0.967	0.950	0.980	0.738
NM + LT	0.775	0.923	0.974	0.949	0.978	0.788
LF + LT	0.744	0.892	0.956	0.953	0.980	0.769
NM + LF + LT	0.799	0.947	0.986	0.979	0.993	0.813
CV + NM + LF	0.755	0.954	0.991	0.967	0.992	0.840
CV + NM + LT	0.783	0.959	0.994	0.963	0.993	0.863
CV + NM + LF + LT	0.805	0.973	0.996	0.984	0.997	0.879

as well as linking behavior. We further observed a strong similarity between edit war articles and articles about controversial issues for most of our measurement metrics. First of all, our findings for relative label frequencies and label transition probabilities indicated substantial editor disputes over content in conflicted articles. We argue that this makes editing behavior in such articles vastly different from regular content, which instead exhibits higher amounts of formatting actions. Though our study may be more representative than previous work, there is an opportunity to extend it to larger data

Table 4 Logistic Regression Results: One-Versus-All Classification. ROC-AUC scores for one-versus-all logistic regression experiments for all categories. Feature sets consist of the number of revisions (n_{rev}), control variables (CV), network metrics (NM), label frequencies (LF), and label transition probabilities (LT). Utilizing all feature sets performs best for all experiments

Features	**ROC-AUC**			
	LQ vs All	**HQ vs All**	**CI vs All**	**EW vs All**
n_{rev}	0.675	0.605	0.778	0.950
CV	0.720	0.702	0.789	0.970
NM	0.745	0.722	0.809	0.915
LF	0.617	0.729	0.789	0.885
LT	0.691	0.762	0.793	0.898
CV + NM	0.775	0.767	0.821	0.970
CV + LF	0.737	0.754	0.814	0.980
CV + LT	0.782	0.792	0.811	0.982
NM + LF	0.759	0.793	0.858	0.951
NM + LT	0.788	0.806	0.861	0.956
LF + LT	0.709	0.809	0.852	0.948
NM + LF + LT	0.796	0.843	0.887	0.970
CV + NM + LF	0.788	0.809	0.861	0.980
CV + NM + LT	0.814	0.825	0.866	0.981
CV + NM + LF + LT	0.823	0.853	0.888	0.986

samples. Additionally, considering more sophisticated classification methods or label taxonomies could improve results. Nonetheless, we demonstrated feasibility of deriving semantic editing behavior from article revision histories, which may prove helpful to existing issues such as editor role identification on Wikipedia. In view of the discovered disparities in wikilink structure of different article categories, we proposed utilization of network metrics for detection of conflict or controversy. Our computed network metrics also depicted significant differences in the link structure of low- and high-quality articles. Notwithstanding that we addressed apparent relations between network metrics, further debiasing existing correlations may improve classification performance. Moreover, we elaborated the effect that articles containing edit wars or controversial issues are not only more frequently referred to from other articles, but may also act as connectors between topical subgroups due to their occupation of structural holes in the Wikipedia network. Momentousness of this effect could be explored in more detail by employing the Wikipedia clickstream dataset,[19] which provides wikilink browsing data extracted from Wikipedia logs. Finally, we combined findings about editing behavior and network metrics to highlight the potential of distinguishing low-quality articles, high-quality articles, articles about controversial issues, and edit war articles using a logistic regression classifier. As opposed to the English Wikipedia, the exact nature of quality, controversy, and conflict in general subjects in other language editions is still an open research question, thus making a similar analysis of non-English Wikipedia an interesting avenue for future research.

Abbreviations

Wikilink: Link between Wikipedia articles; FA: Featured article; GA: Good articles; A: A-class articles; B: B-class articles; C: C-class articles; FL: featured lists; EW: Conflicted (determined via edit-based metrics) and edit war articles; CI: Articles about controversial issues, LQ: Low-quality articles (B and C), HQ: High-quality articles (FA, A, and GA); CCDF: Complementary cumulative distribution functions; *PR*: PageRank; deg^-: In-degree; deg^+: Out-degree; *r*: Reciprocity; *CC*: Clustering coefficient; C_B: Betweenness centrality; C_B^{ego}: Ego network betweenness centrality; C_B^{loc}: Local betweenness centrality approximation; n_{rev}: Number of revisions per article; ROC-AUC: Area under the receiver operating characteristic curve, CV: Control variables; NM: Network metrics; LF: Relative label frequencies; LT: Label transition probabilities

Acknowledgements

We thank the anonymous reviewers for their valuable feedback on the manuscript.

Authors' contributions

TR, TS, and DH designed the study. TR supplied the data and carried out analysis. TR, TS, and DH wrote, reviewed, and approved the manuscript.

Funding

Tiago Santos was a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the "Relating Wikipedia Article Quality to Edit Behavior and Link Structure" repository, https://zenodo.org/record/3908191. The code repository for this paper can be found at https://github.com/ruptho/editlinkquality-wikipedia.

Competing interests

The authors declare that they have no competing interests.

Received: 19 March 2020 Accepted: 11 August 2020
Published online: 09 September 2020

References

Adler BT, De Alfaro L, Mola-Velasco SM, Rosso P, West AG (2011) Wikipedia vandalism detection: combining natural language, metadata, and reputation features. In: CICLing. Springer, Cham. pp 277–288

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47

[19] https://dumps.wikimedia.org/other/clickstream

Antin J, Cheshire C, Nov O (2012) Technology-mediated contributions: Editing behaviors among new wikipedians. In: CSCW. ACM, New York. pp 373–382

Arazy O, Nov O, Patterson R, Yeo L (2011) Information quality in wikipedia: the effects of group composition and task conflict. J Manag Inf Syst 27(4):71–98

Borra E, Weltevrede E, Ciuccarelli P, Kaltenbrunner A, Laniado D, Magni G, Mauri M, Rogers R, Venturini T (2015) Societal controversies in wikipedia articles. In: SIGCHI. ACM, New York. pp 193–196

Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25(2):163–177

Brandes U, Kenis P, Lerner J, Van Raaij D (2009) Network analysis of collaboration structure in wikipedia. In: WWW. ACM, New York. pp 731–740

Burt RS (1992) Structural holes: the social structure of competition. Harvard University Press, Cambridge

Burt RS (2001) Structural holes versus network closure as social capital. Soc Capital Theory Res 1:30–56

Chandrasekharan E, Pavalanathan U, Srinivasan A, Glynn A, Eisenstein J, Gilbert E (2017) You can't stay here: the efficacy of reddit's 2015 ban examined through hate speech. HCI 1(CSCW):31–13122

Consonni C, Laniado D, Montresor A (2019) Wikilinkgraphs: a complete, longitudinal and multi-language dataset of the wikipedia link networks. In: ICWSM, vol 13. AAAI, Palo Alto. pp 598–607

Coursey K, Mihalcea R (2009) Topic identification using wikipedia graph centrality. In: NAACL HLT. ACL, Boulder. pp 117–120

Daxenberger J, Gurevych I (2012) A corpus-based study of edit categories in featured and non-featured wikipedia articles. In: COLING. ACL, Mumbai. pp 711–726

Daxenberger, J, Gurevych I (2013) Automatically classifying edit categories in wikipedia revisions. In: EMNLP. ACL, Seattle, WA. pp 578–589

De La Robertie B, Pitarch Y, Teste O (2015) Measuring article quality in wikipedia using the collaboration network. In: ASONAM. IEEE, New York. pp 464–471

Please capture the below reference here.

Dimitrov D, Singer P, Helic D, Strohmaier M (2015) The Role of Structural Information for Designing Navigational User Interfaces. In: HT. ACM. pp 59–68

Dimitrov D, Lemmerich F, Singer P, Strohmaier M (2017) What makes a link successful on wikipedia? In: WWW. ACM, New York. pp 917–926

Editorial (2006) Britannica attacks. Nature 440(7084):582

Faigley L, Witte S (1981) Analyzing revision. Coll Compos Commun 32(4):400–414

Flöck F, Erdogan K, Acosta M (2017) TokTrack: a complete token provenance and change tracking dataset for the english wikipedia. In: ICWSM. AAAI, Palo Alto. pp 408–417

Gandica Y, dos Aidos FS, Carvalho J (2014) The dynamic nature of conflict in Wikipedia. EPL 108(1):18003

Garlaschelli D, Loffredo MI (2004) Patterns of link reciprocity in directed networks. Phys Rev Lett 93(26):268701

Granovetter MS (1973) The strength of weak ties. Am J Socioly 78(6):1360–1380

Habernal I, Daxenberger J, Gurevych I (2016) Mass collaboration on the web: textual content analysis by means of natural language processing. In: Mass Collaboration and Education. Springer, Cham. pp 367–390

Halfaker A, Geiger RS, Morgan JT, Sarabadani A, Wight A (2018) ORES: Facilitating re-mediation of Wikipedia's socio-technical problems. Wikimedia Research, San Francisco

Hanada R, Cristo M, Pimentel MdGC (2013) How do metrics of link analysis correlate to quality, relevance and popularity in wikipedia? In: WebMedia. ACM, New York. pp 105–112

Ingawale M, Dutta A, Roy R, Seetharaman P (2013) Network analysis of user generated content quality in Wikipedia. Online Inf Rev 37(4):602–619

Kamps J, Koolen M (2009) Is wikipedia link structure different? In: WSDM. ACM, New York. pp 232–241

Kittur A, Suh B, Pendleton BA, Chi EH (2007) He says, she says: conflict and coordination in wikipedia. In: SIGCHI. ACM, New York. pp 453–462

Kumar S, Spezzano F, Subrahmanian V (2015) VEWs: a wikipedia vandal early warning system. In: SIGKDD. ACM, New York. pp 607–616

Lamprecht D, Dimitrov D, Helic D, Strohmaier M (2016) Evaluating and improving navigability of wikipedia: A comparative study of eight language editions. In: OpenSym. ACM, New York. pp 1–10

Lerner J, Lomi A (2019) The network structure of successful collaboration in wikipedia. In: Proceedings of the 52nd Hawaii International Conference on System Sciences. ScholarSpace, Honolulu

Lerner J, Lomi A (2020) The free encyclopedia that anyone can dispute: An analysis of the micro-structural dynamics of positive and negative relations in the production of contentious wikipedia articles. Soc Networks 60:11–25

Li X, Tang J, Wang T, Luo Z, De Rijke M (2015) Automatically assessing wikipedia article quality by exploiting article-editor networks. In: European Conference on Information Retrieval. Springer, Cham. pp 574–580

Liu J, Ram S (2018) Using big data and network analysis to understand Wikipedia article quality. Data Knowl Eng 115:80–93

Milne D, Witten IH (2008) An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: AAAI. AAAI, Palo Alto

Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: Bringing order to the Web. In: WWW. ACM, New York. pp 161–172

Platt EL, Romero DM (2018) Network structure, efficiency, and performance in wikiprojects. In: ICWSM. AAAI, Palo Alto. pp 251–260

Potthast M, Stein B, Gerling R (2008) Automatic vandalism detection in Wikipedia. In: Advances in Information Retrieval. Springer, Berlin. pp 663–668

Ravasz E, Barabási A-L (2003) Hierarchical organization in complex networks. Phys Rev E 67(2):026112

Ruprechter T, Santos T, Helic D (2019) On the relation of edit behavior, link structure, and article quality on wikipedia. In: Complex Networks and Their Applications VIII. Springer, Cham. pp 242–254

Sage Ross (2014) Editing Wikipedia, a print guide for new contributors. https://w.wiki/86W. Accessed 09 Apr 2019

Samoilenko A, Lemmerich F, Zens M, Jadidi M, Génois M, Strohmaier M (2018) (Don't) mention the war: a comparison of wikipedia and britannica articles on national histories. In: WWW. ACM, New York. pp 843–852

Shi F, Teplitskiy M, Duede E, Evans JA (2019) The wisdom of polarized crowds. Nat Hum Behav 3(4):329–336
Shin K, Eliassi-Rad T, Faloutsos C (2016) Corescope: graph mining using k-core analysis - patterns, anomalies and algorithms. In: ICDM. IEEE, Barcelona. pp 469–478
Suchecki K, Salah AAA, Gao C, Scharnhorst A (2012) Evolution of Wikipedia's Category Structure. Adv Compl Syst 15:1250068
Sumi R, Yasseri T, et al (2011) Edit wars in wikipedia. In: PASSAT/SocialCom. IEEE, Boston. pp 724–727
Vautard R, Mo KC, Ghil M, Vautard R, Mo KC, Ghil M (1990) Statistical Significance Test for Transition Matrices of Atmospheric Markov Chains. J Atmos Sci 47(15):1926–1931
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world'networks. Nature 393(6684):440
Yang D, Halfaker A, Kraut R, Hovy E (2016) Edit categories and editor role identification in wikipedia. In: LREC. ELRA, Portoroz. pp 1295–1299
Yang D, Halfaker A, Kraut R, Hovy E (2017) Identifying semantic edit intentions from revisions in wikipedia. In: EMNLP. ACL, Copenhagen. pp 2000–2010
Yang J, Hauff C, Bozzon A, Houben G-J (2014) Asking the right question in collaborative Q&A systems. In: HT '14. ACM, New York. pp 179–189
Yasseri T, Kertész J (2013) Value production in a collaborative environment. J Stat Phys 151(3):414–439
Yasseri T, Spoerri A, Graham M, Kertész J (2014) The most controversial topics in wikipedia. Glob Wikipedia 25:25–48
Yasseri T, Sumi R, Rung A, Kornai A, Kertész J (2012) Dynamics of conflicts in wikipedia. PloS ONE 7(6):1–12

Publisher's Note

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