Classifying Wikipedia Article Quality With Revision History Networks

Narun Raman*
Carleton College
ramann@carleton.edu

Nathaniel Sauerberg*
Carleton College
sauerbergn@carleton.edu

Jonah Fisher
Carleton College
fisherj2@carleton.edu

Sneha Narayan
Carleton College
snarayan@carleton.edu

ABSTRACT

We present a novel model for classifying the quality of Wikipedia articles based on structural properties of a network representation of the article's revision history. We create revision history networks (an adaptation of Keegan et. al's article trajectory networks [7]), where nodes correspond to individual editors of an article, and edges join the authors of consecutive revisions. Using descriptive statistics generated from these networks, along with general properties like the number of edits and article size, we predict which of six quality classes (Start, Stub, C-Class, B-Class, Good, Featured) articles belong to, attaining a classification accuracy of 49.35% on a stratified sample of articles. These results suggest that structures of collaboration underlying the creation of articles, and not just the content of the article, should be considered for accurate quality classification.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**; **Wikis**.

KEYWORDS

Wikipedia, network analysis, quantitative methods, article quality, classification, collaboration

ACM Reference Format:
Narun Raman, Nathaniel Sauerberg, Jonah Fisher, and Sneha Narayan. 2020. Classifying Wikipedia Article Quality With Revision History Networks. In *16th International Symposium on Open Collaboration (OpenSym 2020), August 25–27, 2020, Virtual conference, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3412569.3412581

1 INTRODUCTION

Wikipedia has become the largest and most popular online reference encyclopedia in the world[1], and each of its articles is the product of collaboration between many volunteer editors working together to create a coherent and accurate resource [7][11]. Due to Wikipedia's widespread adoption, editors and researchers have long been interested in maintaining and investigating the quality of its content [4][6][12].

Editors and WikiProjects typically rely on assessments of article quality to focus volunteer attention on improving lower quality articles. This has led to multiple efforts to create classifiers that can predict the quality of a given article [3][4][18]. These classifiers can assist in providing assessments of article quality at scale, and help further our understanding of the features that distinguish high and low quality Wikipedia articles.

While many Wikipedia article quality classifiers have focused on assessing quality based on the content of the latest version of an article [1, 4, 18], prior work has suggested that high quality articles are associated with more intense collaboration among editors [8, 10, 21]. To this end, we introduce a novel model for article quality classification, inspired by the network-based article trajectory model proposed by Keegan et al. [7], in order to predict an article's quality by analyzing the structure of its collaboration network.

Using statistics generated from each article's revision history network, in conjunction with basic article characteristics like article size and edit count, we run a multinomial logisitic regression (MLR) and attain a classification accuracy of 49.35% on a stratified sample of 6000 articles.

2 COLLABORATION AND QUALITY

Structures of collaboration have been shown to affect outcomes in traditional organizations, as well as volunteer groups online [13]. Structural properties of internal collaborative networks can influence outcomes across a wide variety of contexts, from the creation of Broadway musicals [16], to open-source software development [15]. In particular, small-world networks (i.e. highly clustered networks with small characteristic path length [19]) are believed to enhance productivity and creativity [13, 15, 16].

On Wikipedia in particular, leadership behavior and an intermediate level of small-wordliness in the social networks of WikiProject talk pages have been found to be positively related to the efficiency and productivity of a project (measured in terms of edit count and edit longevity) [13]. Additionally, prior work has shown that article quality is positively impacted by more communication among editors on talk pages, as well as a more centralized collaboration structure [8]. While articles with many editors are more likely to be of higher quality than articles with fewer editors [8, 9, 20], the addition of editors to a page only seems to improve its quality when those editors are collaborating appropriately [8]. Preliminary results also suggest that structural parameters of collaboration

*Both authors contributed equally to this research.
[1]https://www.alexa.com/siteinfo/wikipedia.org

OpenSym 2020, August 25–27, 2020, Virtual conference, Spain

networks generated from article revision histories can help differentiate controversial articles from featured articles [2]. Given that these structures of collaboration have been related to individual article quality [2][8][10], as well as the overall productivity of WikiProjects [13], we believe that characteristics of the collaboration networks of individual articles may be beneficial in predicting the quality level of a Wikipedia article.

3 PREDICTING ARTICLE QUALITY ON WIKIPEDIA

Existing models for predicting article quality on Wikipedia have typically used features related to the content of the article at a particular moment in time [1, 4, 18]. One such feature is the length of the article. A model with article word count as its only predictor was able to achieve a 97.15% accuracy rate in completing a binary classification task of featured and non-featured articles, beating out several more complex models [1]. Warncke-Wang et al. then introduced the 'Actionable Model', which uses article size along with features like the number of internal links, headings, images and references to perform a more fine-grained classification task on all seven quality classes (including the now uncommon A-Class) with 42.5% accuracy on an imbalanced corpus of articles [18]. ORES, a machine-learning based web service that estimates Wikipedia article quality, uses a modified version of the Actionable Model with some additional predictor variables, including the number of "[citation needed]" templates and the number of "Main article" linking templates [4]. This version was able to achieve an accuracy of 62.9% on their own corpus of articles using the six current quality classes.

However, these models do not take into account how the article was developed by its editors and the way they collaborated with one another. In light of work outlined in the previous section that suggests a relationship between the structure of editor collaborations and article quality, we present a novel model that uses characteristics of an article's revision history network to predict the quality of an article.

4 MODELING APPROACH

4.1 Operationalizing Article Quality

In order to measure the baseline quality of an article, we make use of Wikipedia's content assessment project, which has provided ratings for over 6.1 million English-language articles that place articles into varying quality classes[2]. From lowest to highest quality, these classes are named Stub, Start, C-Class, B-Class, Good Articles (GA), and Featured Articles (FA). We omit A-Class articles, as other classifiers have done [4], as it is considered defunct. The quality assessments for a given article are typically made by members of WikiProjects, or groups of editors who are focused on articles about a particular topic[3]. Featured articles, considered to be the best Wikipedia has to offer, undergo an additional rigorous peer review process [17][21].

There are some limitations to using these classes as ground truth for article quality. First, despite specific criteria for each class, quality is subjective and classification judgments may vary among editors. Second, approximately 7.52% (as of June 13, 2020) of all Wikipedia articles have not been assessed [4]. This may impact the distribution of assessed Wikipedia articles. Finally, because Wikipedia is a constantly evolving platform, an article's quality class rating could be outdated because of new edits and revisions that occurred after its assessment.

Despite these limitations, we use them as the ground-truth for article quality in our classifier because of their frequent use in discussions of article quality on Wikipedia [5][10] as well as in similar classification models [4][6][14][18].

4.2 The Revision History Network Model

The page history of a Wikipedia article can be viewed as an ordered list of revisions, each of which stores the state of the article at a particular point in time. Each revision is associated with an editor, identified by a username or IP address, depending on whether the editor created an account.

We define the *revision history network* of an article as an undirected network whose nodes are the editors of the corresponding article, and whose edges join the editors of consecutive revisions. In other words, we can construct the revision history network of an article from its revision history by iterating over the revisions chronologically and, for each revision i after the first, creating an edge between the editors of revisions $i-1$ and i if one does not already exist (see Figure 1). Note that this holds even if an editor authors two consecutive revisions; in this case, we create an edge joining the corresponding node to itself. Our model is adapted from the article trajectory network model introduced in [7], which represents each article's revision history as a directed network with multiple edges.

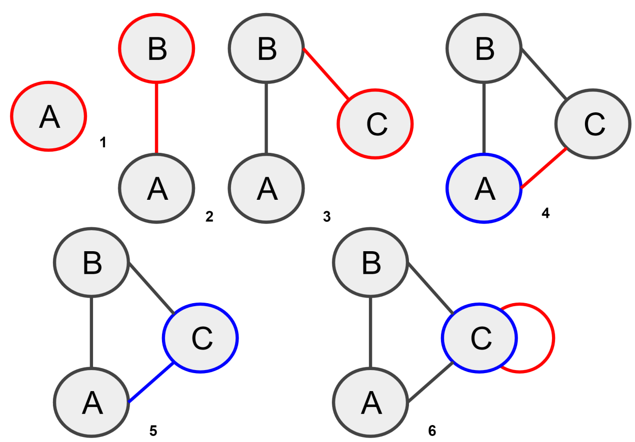


Figure 1: Example showing the sequential construction of a revision history network for an article where the first revision is made by A, followed by B, C, A, C, and C. The network numbered *i* corresponds to the state of the network after the first *i* revisions.

We interpret each edge as an indication of a collaborative interaction, and the full networks as representing the collaboration

[2]https://en.wikipedia.org/wiki/English_Wikipedia
[3]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment
[4]https://en.wikipedia.org/wiki/Wikipedia:Statistics

network of the authors. Our project investigates the extent to which the structures of collaboration present in these social networks relate to article quality.

4.3 Network Statistics

Every article on Wikipedia has an associated revision history network, and we expect several characteristics of these networks to be related to article quality. Therefore, we use a number of descriptive network statistics as features in our classifier.

Throughout the following section, we will use G to refer to a generic revision history network, with V and E denoting its node and edge sets. We use u, v, and w to refer to nodes, and refer to edges as pairs of nodes; for instance $uv \in E$ is an edge joining nodes u and v.

Basic Metrics. The most basic network statistics that we expect to be associated with article quality are n and m, the numbers of nodes and edges in the network, respectively. The number of nodes is the number of editors who have contributed to the article, while the number of edges corresponds to the number of unique collaborations between pairs of editors. We also consider the *density* of the network, which is the number of edges present in the network divided by the number possible [5]:

$$density(G) := m/\binom{n}{2} = \frac{m}{n(n-1)/2}.$$

We expect that all three of these metrics will help our classifier distinguish between higher and lower quality articles. Prior work has shown that high quality Wikipedia articles are associated with a greater number of contributors [8][9][20]. Additionally, when editors collaborate effectively, they synthesize their contributions and build a shared conception of the desired state of the article, leading to better organization and more consistent style [8][10]. Thus, we expect both the number and density of collaborations to correspond to higher quality articles.

Distance-Based Metrics. The *distance* between two nodes (denoted $dist(u, v)$) is defined to be the length of the shortest path between them in the network, where the length of a path is the number of edges it contains. The *eccentricity* of a node is the longest distance from it to any other node in the network. The *diameter* of the network is the maximum eccentricity of all its nodes, and the *radius* is the minimum eccentricity.

The *closeness centrality* of a node v is the reciprocal of the sum of the distances from v to all other nodes in the graph, normalized by the number of other nodes:

$$closeness(v) := \frac{n-1}{\sum_{u \in G, u \neq v} dist(v, u)}.$$

The *average closeness centrality* of a network is the average value of closeness centrality over all its nodes.

We use the *radius*, *diameter*, *average eccentricity* and *average closeness centrality* of the revision history network as features in our classifer. These metrics give us information about how far apart editors in the revision history network are. Low values for the first three metrics and high values for the fourth mean that the longest distances present in the network are small, which corresponds to low degrees of separation between editors. Short paths between editors are characteristic of small-world networks, and editors collaborating within such structures are shown to produce higher quality articles on Wikipedia [8].

Betweenness Centrality. The property of *betweenness centrality* captures the extent to which a node is a broker in the network. Formally, it is defined as follows, where $\sigma(s, t)$ is the set of shortest paths between s and t:

$$betweenness(v) := \sum_{s,t \in G} \frac{|\{p \in \sigma(s, t) : v \in p\}|}{|\sigma(s, t)|}.$$

In other words, the *betweenness centrality* of v is the proportion of shortest $s - t$ paths containing v, summed over all pairs of nodes s and t. Again, we take the average betweenness centrality over all nodes and use it as a network statistic in our analysis, which we refer to as *average betweenness.* High average betweenness might indicate the presence of long chains of nodes in the revision history network, which characterizes a relative lack of editors who repeatedly synthetize content across multiple revisions [7]. We thus expect high average betweenness to be an indicator of lower article quality.

Clustering Coefficients. Clustering is intended to capture the tendency of the collaborators of an editor to collaborate with each other. The clustering coefficient of a node is the proportion of its neighbors that are themselves joined by edges. If $N(v) := \{u : vu \in E\}$ is the set of neighbors of a node v, then

$$clustering(v) := \left|\{uw \in E : u, w \in N(v)\}\right| / \binom{|N(v)|}{2}.$$

We consider the *average clustering coefficient*, abbreviated as *average clustering*, as the average of all clustering coefficients of the nodes of the network, and include it as a feature in our classifier. We expect that average clustering is positively correlated with article quality because it is indicative of a larger collaboration network where an editor's collaborators are also working together.

Other Variables. In addition to our network statistics, we include two other independent variables that we expect are related to article quality, namely the *number of edits* made to the article, and the *size* of the latest version of the article in bytes. Both metrics are positively associated with high quality articles on Wikipedia [1][21].

5 DATA COLLECTION AND CLASSIFICATION STRATEGY

Each article in Wikipedia's Main namespace is assigned a page id value corresponding to its place in sequential order of creation. We created a corpus of articles by continuously polling the MediaWiki API for a random page ID and the ten preceding it. Within each tranche of page IDs polled, we added every article that had a content assessment score to our corpus until we had 1000 articles of each quality class (i.e. Start, Stub, C, B, GA or FA) for a final corpus containing 6000 articles. Since a large proportion of articles on Wikipedia are classified as either Start or Stub, and very few are

[5]Because we allow self-edges in the network, but they are not counted among the possible edges, it is possible for a graph to have density strictly greater than 1.

designated GA or FA, we chose to use a stratified sample to more accurately classifying high quality articles and prevent overfitting our model on low quality articles. We omitted A-Class articles from our corpus due to their scarcity[6] and omission from other classifier models [4]. We also collected the revision histories for each article in our corpus and used them to generate revision history networks. Using the Python package NetworkX, we calculated the network statistics to use as features within our model, along with the number of edits and the size of the article in bytes. We then standardized each variable by subtracting the mean and scaling to unit variance. We used multinomial logistic regression (MLR) as the classifier on this dataset, with a linear combination of the variables introduced in Section 4. Other classifiers such as ordinal logistic regression (41.53%), support vector (46.73%), K nearest neighbor (45.18%), random forest (47.26%), and decision tree classification (41.40%) algorithms performed worse in overall accuracy than MLR.

6 RESULTS

Trained on our dataset of statistics for the 6000 articles, we ran a 10-fold cross validation of the MLR to test the accuracy of our predictors. Overall, our classifier correctly predicts an article's quality 49.35% of the time, with 95% confidence interval: (0.4803, 0.5057). This is determined by the ratio of the articles our model correctly predicts (i.e. those on the diagonal of Table 3) to the entire dataset.

To examine how well our independent variables predict quality, we ran likelihood ratio tests on each of our predictors to find which ones improved our model fit (see Table 1). We determined that all our predictors except for radius, diameter and average eccentricity provided a statistically significant improvement to the goodness of fit.

Predictor	LR Chisq	Df	Pr(>Chisq)
article size	424.851	5	$< 2.2e\text{-}16^{***}$
editor count	31.955	5	$6.064e\text{-}6^{***}$
edit count	69.533	5	$1.282e\text{-}13^{***}$
density	53.384	5	$2.805e\text{-}10^{***}$
num edges	53.299	5	$2.920e\text{-}10^{***}$
betweenness	13.33	5	0.0204^{*}
clustering	34.071	5	$2.305e\text{-}6^{***}$
closeness	20.706	5	$9.21e\text{-}4^{***}$
diameter	5.547	5	0.353
radius	8.826	5	0.116
avg eccentricity	6.193	5	0.288

Table 1: Nested Log Likelihood Calculations. Significant predictors are bolded, with significance level indicated by asterisks.

In Table 2 we list the performance metrics of our classifier by article quality class. We see a few patterns emerge from the confusion matrix for our model seen in Table 3.

First, we see a stratification of guesses. The confusion matrix shows a clear distinction in the false positives of the classifier on low quality articles (Start and Stub) and higher quality articles (C-, B-, GA-, and FA-Class Articles). While our classifier often confuses high quality articles with each other, it is good at determining whether an article is of "good enough" quality or not; rarely does it classify a C-Class article or above to be Start or Stub, and vice versa. We ran a binomial logistic regression (BLR) on the same features to investigate this observation and got an accuracy score of 95.78%, validating the patterns seen in the MLR confusion matrix. With sensitivity and specificity values 0.974 and 0.923, respectively, the classifier guesses correctly and often. Accurately classifying articles with low quality is particularly impactful considering Start and Stub articles collectively make up 87% of all assessed articles in English Wikipedia[7].

However, much like other models that predict article quality [4, 18], our classifier has difficulty discerning between highly ranked articles. Although we omit A-Class from our analysis, the MLR struggles to classify GA-Class articles with high accuracy. One explanation for this confusion could be the backlog of potential Featured articles tagged GA; the classifier more often misclassifies a GA-Class article as Featured (319) than Featured as GA (173). Furthermore, not only does our classifier rarely guess GA (477 out of 1000), it categorizes GA-Class articles as FA-, B-, and C-Class over 79% of the time. It seems that the classifier on our statistics has difficulty determining exactly what defines an article that achieves a "good" rating. The relatively high true positive rate for FA (0.54) and skewed misclassification of GA as FA (versus B-Class) suggests that a queue of GA articles that could potentially become Featured might account for this difficulty.

Class	Precision	Sensitivity	False Positive Rate	False Negative Rate
SB	0.86	0.70	0.03	0.30
ST	0.54	0.64	0.09	0.36
C	0.54	0.41	0.10	0.59
B	0.32	0.33	0.14	0.67
GA	0.16	0.34	0.15	0.66
FA	0.54	0.46	0.09	0.54
Total	0.48	0.48	0.10	0.52

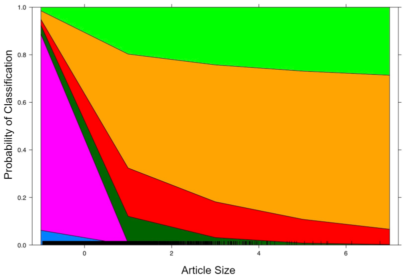
Table 2: Performance Metrics

In Figure 2, we see the effects of all the features in our model, represented as the likelihood of classification into a particular category given a value. These effects are represented as stacked plots, with a larger area indicating a higher likelihood of classification.

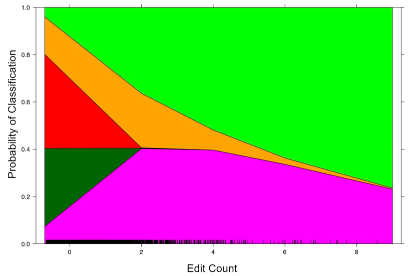
We highlight a few of these features to examine how they influence our classification. In part (k) of Figure 2, for instance, we see a positive relationship between *average closeness* and quality. As the value of closeness increases, the area of the Featured Article quality category increases; hence, the likelihood of classification as FA

[6] Only about .0004% of assessed articles have been rated A-Class, according to data from: https://en.wikipedia.org/wiki/Wikipedia:Statistics

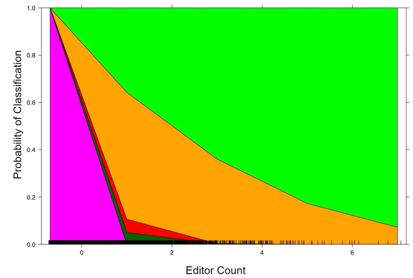
[7] https://en.wikipedia.org/wiki/Wikipedia:Statistics



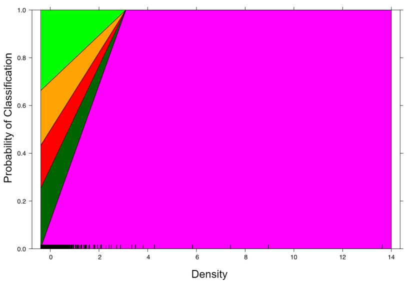
(a) Article Size



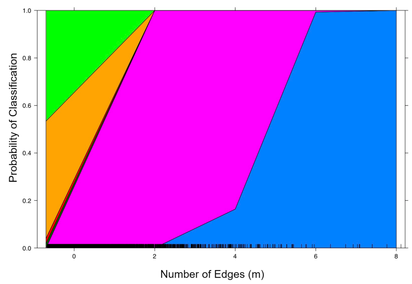
(b) Edit Count



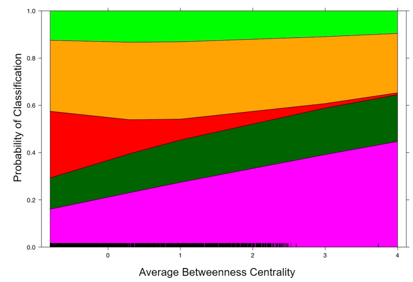
(c) Editor Count



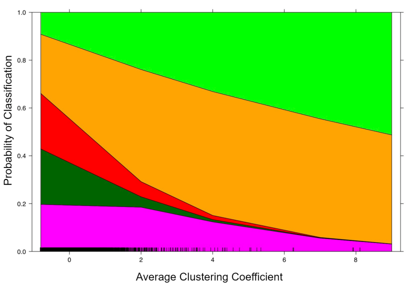
(d) Density



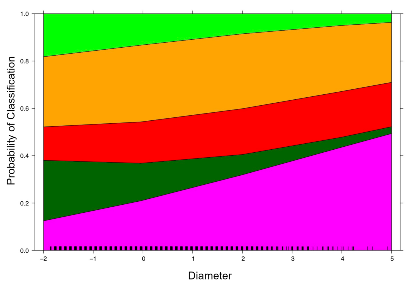
(e) Number of Edges



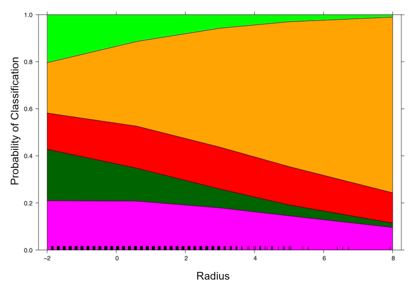
(f) Average Betweenness



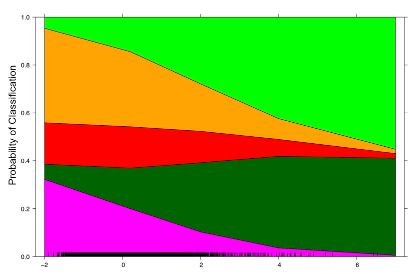
(g) Average Clustering



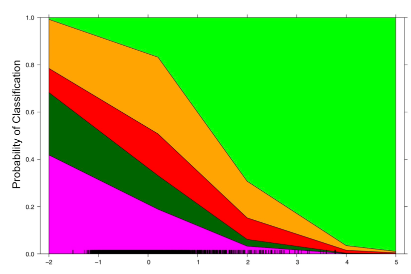
(h) Diameter



(i) Radius



(j) Average Eccentricity



(k) Average Closeness

Figure 2: Effects Plots on Quality Classification for all Features.

FA GA B-Class C-Class Start Stub

	Stub	**Start**	**C**	**B**	**GA**	**FA**	Total
Stub	859	351	9	0	1	0	1220
Start	137	542	75	29	47	16	846
C	3	82	535	371	222	98	1311
B	1	2	214	318	248	169	952
GA	0	4	54	83	163	173	477
FA	0	19	113	199	319	544	1194
Total	1000	1000	1000	1000	1000	1000	6000

Table 3: Confusion Matrix. Columns with bold headers show actual class, italicized rows show predicted class. The highlighted diagonal cells show the correctly classified articles.

increases with the average closeness of an article's revision history network. We see similar relationships exhibited by the *editor count* and *clustering* variables (parts (c) and (g) of Figure 2, respectively). Additionally, we find that consistent with our expectations, the larger the size of an article, the more likely it is to be categorized as either GA or FA (part (a) of Figure 2). We also see that as per our expectations, when *average betweenness* increases, the likelihood of classification into a lower quality class like Start increases. Contrary to our expectations, we find that as the *density* and *number of edges* in a revision history network increases, the more likely it is to be classified as Start or Stub respectively (parts (d) and (e) of Figure 2). This could be because articles with very small numbers of editors (characteristic of Start and Stub articles) are more likely to have all nodes adjacent to one another in their revision history networks, which leads to a high density.

Some of the effects are not consistent across all levels of quality. For instance, as the value of *edit count* increases (Figure 2b), the probability of classifying an article as B- or C-Class vanishes, while Start increases. However, classifying an article as FA increases far more, with probability greater than 0.6 given a standardized value of edit count greater than 6.

6.1 Comparing to the ORES Classifier

One of the more popular existing models for predicting article quality on Wikipedia is ORES, introduced in [4]. While both ORES and our model seek to classify articles by quality, our approach incorporates the structures of collaboration over time.

The features for the ORES classifier are taken directly from the WikiProject guidelines for quality. ORES uses predictors such as article size, number of headings, references, and broken links, which all correspond to parameters for quality as defined by the Wikipedia content assessment class criteria (e.g. broad coverage of a topic, presence of helpful section headers, and so on [8]). ORES claims an accuracy of 62.9% on their stratified dataset [4], however, we ran their model on our stratified corpus and attained an accuracy score of 52.20%.

Our model's predictors are based on structural traits of the article's revision history networks, rather than the content of the latest version of the article. Nevertheless, our model achieves an accuracy score (49.35%) quite close to ORES (52.20%) when run on our dataset. An MLR that combines our model's predictors with predicted probabilities from the ORES model increased accuracy to 60.29% on the same dataset.

[8]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

7 CONCLUSION

We put forward a novel model for predicting the quality of an article on Wikipedia based on network statistics derived from the article's revision history network. Our model performs comparably against those that are based on features of the article's content. While we do not demonstrate a causal relationship between our network statistics and article quality, the accuracy of our model indicates that structures of collaboration are highly related to article quality. The classifier's accuracy also suggests that revision history networks provide a meaningful, interpretable model of an article's underlying collaboration structure.

Our revision history network model is an abstraction of the complicated details of collaboration, and as such, it makes several simplifying assumptions. In particular, the model cannot account for non-collaborative interactions between editors. Vandalism and the corresponding reverts, for example, are not treated differently from collaborative interactions in our model.

The model also fails to account for temporal aspects of collaboration. For instance, the authors of consecutive revisions are assumed to be collaborators, even if weeks or longer occur between their revisions. Conversely, authors authors engaged in real-time collaboration will not be considered collaborators if they happen not to make consecutive revisions. Despite these limitations, we still find that metrics associated with revision history networks help distinguish between high and low quality articles in our corpus.

Future work could consider adding edge weights to the revision history networks in order to account for the frequency or volume of the collaboration between two editors. In our current model, multiple back-and-forth revisions between two editors are condensed into a single edge. To account for differential impacts on article quality of repeated collaboration as opposed to one-time collaboration, edge weights could depend on the number of consecutive revisions that occur between editors. Alternatively, edge weights could be designed to account for the size of individual revisions (in terms of bytes added or deleted). By evaluating the impact of these modifications to our model, we may be able to create an improved model that more successfully classifies the quality of a Wikipedia article by taking into account its internal network of collaboration.

ACKNOWLEDGMENTS

We thank Jeremy Foote, Nate TeBlunthuis, and Kaylea Champion for their feedback, and Addison Partida-Vásquez for his contributions on an earlier version of this project. We dedicate this paper to Addison and the Partida Family.

REFERENCES

[1] Joshua E. Blumenstock. 2008. Size matters: word count as a measure of quality on wikipedia. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*. ACM Press, Beijing, China. https://doi.org/10.1145/1367497.1367673

[2] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. 2009. Network Analysis of Collaboration Structure in Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web* (Madrid, Spain) *(WWW '09)*.

Association for Computing Machinery, New York, NY, USA, 731–740. https://doi.org/10.1145/1526709.1526808

[3] Baptiste de La Robertie, Yoann Pitarch, and Olivier Teste. 2015. Measuring Article Quality in Wikipedia using the Collaboration Network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*. ACM Press, Paris, France. https://doi.org/10.1145/2808797.2808895

[4] Aaron Halfaker. 2017. Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect. In *Proceedings of the 13th International Symposium on Open Collaboration - OpenSym '17*. ACM Press, Galway, Ireland. https://doi.org/10.1145/3125433.3125475

[5] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2009. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. 295–304.

[6] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W Lauw, and Ba-Quy Vuong. 2007. Measuring Article Quality in Wikipedia: Models and Evaluation. *Measuring Article Quality in Wikipedia: Models and Evaluation* (2007).

[7] Brian Keegan, Darren Gergle, and Noshir Contractor. 2012. Staying in the loop: structure and dynamics of Wikipedia's breaking news collaborations. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration - WikiSym '12*. ACM Press, Linz, Austria. https://doi.org/10.1145/2462932.2462934

[8] Aniket Kittur and Robert E Kraut. 2008. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, 37–46.

[9] Andrew Lih. 2004. Wikipedia as Participatory Journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature* 3, 1 (2004), 1–31.

[10] Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the Wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst.* 2, 2 (June 2011). https://doi.org/10.1145/1985347.1985352

[11] Jun Liu and Sudha Ram. 2018. Using big data and network analysis to understand Wikipedia article quality. *Data & Knowledge Engineering* 115 (May 2018). https://doi.org/10.1016/j.datak.2018.02.004

[12] Deborah L McGuinness, Honglei Zeng, Paulo Pinheiro Da Silva, Li Ding, Dhyanesh Narayanan, Mayukh Bhaowal, et al. 2006. Investigations into trust for collaborative information repositories: A Wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web*.

[13] Xiangju Qin, Pádraig Cunningham, and Michael Salter-Townshend. 2015. The influence of network structures of Wikipedia discussion pages on the efficiency of WikiProjects. *Social Networks* 43 (2015).

[14] Laura Rassbach, Trevor Pincock, and Brian Mingus. 2007. Exploring the feasibility of automatically rating online article quality. In *Proceedings of the 2007 International Wikimedia Conference (WikiMania), Taipei, Taiwan*, Vol. 66.

[15] Param Vir Singh. 2010. The small-world effect: The influence of macro-level properties of developer collaboration networks on open-source project success. *ACM Trans. Softw. Eng. Methodol.* 20, 2 (Aug. 2010). https://doi.org/10.1145/1824760.1824763

[16] Brian Uzzi and Jarrett Spiro. 2005. Collaboration and Creativity: The Small World Problem. *Amer. J. Sociology* 111 (Sept. 2005). https://doi.org/10.1086/432782

[17] Fernanda B. Viégas, Martin Wattenberg, and Matthew M. McKeon. 2007. The hidden order of Wikipedia. In *Proceedings of the 2nd International Conference on Online Communities and Social Computing (OCSC'07)*. Springer-Verlag, Beijing, China, 445–454.

[18] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell me more: an actionable quality model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration - WikiSym '13*. ACM Press, Hong Kong, China. https://doi.org/10.1145/2491055.2491063

[19] Duncan Watts and Steven Strogatz. 1998. Collective Dynamics of Small-World Networks. *Collective Dynamics of Small-World Networks* 393 (1998).

[20] Dennis M. Wilkinson and Bernardo A. Huberman. 2007. Assessing the Value of Coooperation in Wikipedia. *arXiv:cs/0702140* (Feb. 2007). http://arxiv.org/abs/cs/0702140

[21] Dennis M. Wilkinson and Bernardo A. Huberman. 2007. Cooperation and quality in Wikipedia. In *Proceedings of the 2007 international symposium on Wikis - WikiSym '07*. ACM Press, Montreal, Quebec, Canada. https://doi.org/10.1145/1296951.1296968