Revisionist History: Predicting Wikipedia Article Quality With Edit Histories

NARUN RAMAN, NATHANIEL SAUERBERG, ADDISON PARTIDA, and JONAH FISHER, Carleton College

ABSTRACT: We present a novel model for article quality classification based on structural properties of a network representation of the article's edit history. Inspired by Keegan et al. (2012), we create article trajectory networks, where nodes correspond to individual editors and edges join the authors of consecutive revisions. Using distance-, betweenness-, and clustering-based metrics generated from this model, along with general properties like the number of editors and article length, we predict which of six quality classes (Start, Stub, C-Class, B-Class, GA, FA) articles belong to, attaining a classification accuracy of 49.35% on a uniform sample of articles. This represents a similar level of accuracy to models that more directly align their predictors with Wikipedia quality class criteria, such as Warncke-Wang et al.'s "Actionable Model" (42.5% accuracy) [18] and Halfaker's ORES model (62.9% accuracy) [9]. These results suggest that structures of collaboration underlying the creation of articles, and not only characteristics of the current public version at a particular point in time, should be considered for accurate quality classification.

Additional Key Words and Phrases: Wikipedia, network analysis, graph theory, statistical modeling, article quality, quality classification

ACM Reference Format:
Narun Raman, Nathaniel Sauerberg, Addison Partida, and Jonah Fisher. 2020. Revisionist History: Predicting Wikipedia Article Quality With Edit Histories. 1, 1 (May 2020), 11 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Founded in 2001, Wikipedia has quickly grown to become the largest and most popular reference encyclopedia on the internet with over 6 million English-language articles and over 53 million articles in total [1][7]. Each of these articles is the product of a history of collaboration between many editors working together to create a coherent and accurate resource for public viewing.

The accuracy of Wikipedia's content has been questioned since its inception [11][17][20]. One frequently-discussed article from *Nature* found the quality of scientific articles on Wikipedia to be generally similar to that of Encyclopedia Britannica articles on the same topics [8]. This surprising display of quality has led some supporters to claim that is has successfully harnessed the "wisdom of the crowds" [13]. Yet others remain skeptical of Wikipedia's accuracy, pointing to the potential for vandalism and the crowd-sourced encyclopedia's susceptibility to hoaxes and misinformation [5].

Given the interest surrounding the question of quality on Wikipedia, various works have attempted to create models that can predict the quality of a given article [5][9][18]. These models not only can potentially further our understanding of the features that distinguish high and low quality Wikipedia articles, but they also provide two key benefits:

(1) The ability to provide Wikipedia users with a general sense of the quality of the content they are reading.
(2) The ability to direct Wikipedia editors to focus their time on articles that require the most improvement and attention.

Wikipedia quality classifiers have traditionally focused on predictor variables derived directly from the content on the current public version of an article rather than the structure of collaboration underlying that article's creation [5]. Given the often ignored importance of collaboration patterns underlying an article's development in determining its overall quality [16], we create a novel model, inspired by the graph-based revision history model proposed by Keegan et al. [12], in order to predict an article's quality by analyzing the structure of its collaborative network. We therefore focus on the following **Research Question**:

> Can we build a classifier that will accurately predict an article's quality using properties of revision history networks related to structures of collaboration?

To begin, we summarize the literature surrounding factors that determine a Wikipedia article's accuracy and quality. Next, we introduce the article trajectory model and network analysis metrics that we use to predict an article's quality class categorization. Finally, we discuss our results in context and their implications for assessing quality on Wikipedia, concluding with suggestions for future work.

2 RELATED WORK

Several works highlight features that explain the varying quality levels of Wikipedia articles. Lih [14] argues that rigor (total number of edits) and diversity (total number of unique authors) are positively correlated with article quality. He additionally demonstrated how citations from other established media can direct public attention to specific articles, ultimately resulting in increased quality. In general, articles with many editors are more likely to be higher quality than articles with fewer editors [6][14][19]. However, the addition of editors to a page only seems to improve its quality when those editors are collaborating appropriately [6].

The length of an article also seems to be correlated to quality, and a model with article word count as the 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 [4].

Other works have focused on the identity and roles of individual editors as an indicator of article quality. Zeng et al. models the trustworthiness of Wikipedia authors in a dynamic Bayesian network [20]. Adler and Alfaro determine the cumulative reputation of an author by how long his/her edited content survived in terms of time span (text survival) and number of revisions (edit survival) [3]. Through experiments on French and Italian Wikipedia, they

Authors' address: Narun Raman, ramann@carleton.edu; Nathaniel Sauerberg, sauerbergn@carleton.edu; Addison Partida, partidaa@carleton.edu; Jonah Fisher, fisherj2@carleton.edu, Carleton College, 300 N. College St., Northfield, Minnesota, 55057.

find that changes performed by authors with lower reputation are significantly more likely to display poor quality.

Predictive quality models have traditionally focused on assessments of the content and structure of fixed states of pages. Warncke-Wang et al. first introduced the “Actionable Model”, which uses five predictive variables: "Completeness"[1], "Informativeness"[2], Number of Headings, Article Length, and the number of references divided by article length [18].

This model was able to classify articles with 42.5% accuracy on a semi-uniform corpus of articles. ORES, a Wikipedia machine learning web service, provides 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 [9]. This version was able to achieve an accuracy of 62.9% on their own (nearly uniformly-distributed) corpus of articles.

However, these models do not take into account differences between editor interactions or the structure of their collaboration, factors that differentiate Wikipedia from typical encyclopedias [12][15][7]. Analysis of patterns of collaboration on Wikipedia has most often assessed the content contributed (or in some cases, erased) by an individual edit [11][15] rather than the quantity or structure of the coordination among editors. However, given that both the quantity [6] and quality [15] of collaboration between editors appear to affect an article’s overall quality, it seems highly important to examine the networked structure of article revision histories in order to understand common patterns in the way that quality articles are constructed and to create an accurate predictive quality classifier.

3 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 5.1 million English-language articles that place articles in to varying quality classes [2]. From lowest to highest quality, these classes are named Stub, Start, C-Class, B-Class, Good Articles (GA), A-Class, and Featured Articles (FA). Each class has its own set of specific criteria. For instance, criteria for a B-Class article include:

(1) The article is suitably referenced, with inline citations.
(2) The article reasonably covers the topic, and does not contain obvious omissions or inaccuracies.
(3) The article has a defined structure.
(4) The article is reasonably well-written.

The quality assessments for a given article are typically made by members of WikiProjects, groups of editors who are focused on articles about a particular topic. Some of the higher quality tiers, however, require special external designation: in order for an article to be ranked as Good (GA), it must be classified by an impartial reviewer (i.e. someone who has not personally contributed to the article) and in order for an article to reach featured status (FA) it must be assessed by a panel of impartial reviewers. The A-Class category is ignored by many WikiProjects, and these articles are therefore very scarce. For this reason, we ignore A-Class articles and instead focus on the other six quality classes.

There are some limitations to using these classes as a metric for article quality. First, despite specific criteria for each class, quality is subjective. Two different editors may, for instance, have completely different opinions on what it means for an article to be "well-written". Second, approximately 7%[3] of all articles on Wikipedia have not yet been assessed, a fact which may impact the quality distribution of assessed Wikipedia articles. Finally, 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, these classes remain the best and most widely-used measure for quality on Wikipedia and we therefore use them as the ground truth for our classifier.

4 ARTICLE TRAJECTORY GRAPH 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 author[4]. We apply and extend the article trajectory graph model of Keegan et al. [12], which represents each article revision history as a directed graph or network. The nodes of the graph are the editors of the article, and a directed edge joins the authors of consecutive revisions. We construct the article trajectory of an article from its revision history by iterating over the revisions and, for each revision i after the first, creating an edge from the author of revision $i-1$ to the author of revision i. Note that this holds even if an editor authors two consecutive revisions. In this case, we create a (self-)loop: an edge from the corresponding node to itself.

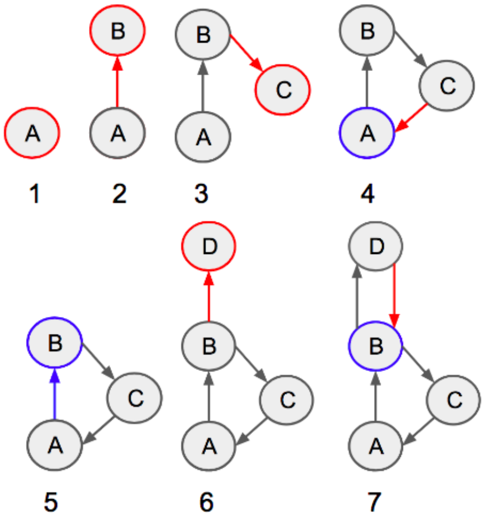

Fig. 1. Example showing the evolution of an article trajectory graph. The graph numbered i corresponds to the state of the graph after the first i revisions. Ordered Authors of Revisions: A B C A B D B.

The original model of Keegan et al. allows for parallel edges if multiple directed interactions occur between the same pair of editors.

[1] 0.4* Number of Broken Wikilinks + 0.4* Number of Wikilinks
[2] 0.6* InfoNoise + 0.3* Number of Images

[3] as of May 15, 2020
[4] The author is a username if the editor is logged in and otherwise the IP address from which the edit was made. We consider each unique IP address to be a single author in our model.

Because these parallel edges do not affect our network statistics or those used by [12], we do not include them in our model. Therefore, while Keegan et al. model each revision history with a directed multigraph with loops, our article trajectory model is a directed graph (digraph) with loops.

In addition to this directed article trajectory model, we also consider an undirected version. This is identical to our directed model except that edges are undirected and an edge exists between editors A and B if either A authored a revision directed after B authored one or B authored a revision directed after A did so. In other words, an edge AB in the undirected graph exists if at least one of the edges AB and BA exist in the directed graph for the same article.

4.1 Model Interpretation

We make some important assumptions in our choice of model and its interpretation. In particular, we interpret each edge as an indication of a (directed) collaborative interaction and the full networks as representing the social networks of the authors. Our project investigates the extent to which the structures of collaboration present in these social networks relate to article quality.

The model is an abstraction of the complicated details of collaboration, and as such we make several simplifying assumptions. In particular, the model cannot account for non-collaborative interactions between editors. For instance, vandalism and the corresponding revert (undo) revisions will be interpreted as collaborative. Similarly, if edges result from editors working on separate sections of an article simultaneously, our model will interpret them as collaborators, although the extent to which this constitutes collaboration is debatable.

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.

Throughout the paper, we analyze the directed and undirected versions of the article trajectory model. We consider the directed model to be our "base" model and the undirected model an extension. The undirected model is somewhat simpler than the original model: it has half the number of possible edges and so, in some sense, stores only half as much information. Therefore, if the undirected model is at least as successful at the directed model at predicting article quality, it should be considered a better representation of the revision history[5]. Such a result would suggest that any consecutive revisions constitute full collaboration between authors, regardless of whether or not two separate instances occurred, one with each other coming first. In other words, collaboration should be thought of as being symmetric, as it is usually considered to be in informal settings.

We will refer to the graphs as article trajectories, article trajectory graphs, and revision history graphs interchangeably.

5 NETWORK STATISTICS

A direct comparison of the different articles' trajectory graphs would require machine learning, but we take a more interpretable approach and measure attributes of the networks to capture characteristics we expect to be related to quality. We used the Python package NetworkX to compute all of our network statistics.

Keegan et al. use four network statistics in their analysis: *diameter*, *average closeness centrality*, *average betweenness centrality*, and *average clustering*. Their analysis is focused primarily on distinguishing "tight" and "loose" revision networks. Tight revision networks are characterized by many cycles[6] and few chains, as editors contribute multiple revisions to the article. Conversely, loose revision networks have many cycles and few chains, caused by editors who contribute exactly one revision. In other words, tight networks are highly bunched while loose networks contains many long induced paths. We incorporate and extend the ideas behind each of these metrics into our analysis and add some metrics based on ideas of our own.

Throughout the following section, we will use G to refer to a generic article trajectory graph and define E to be the edge set and V to be the vertex set of the graph. We will reference the edges as ordered pairs; for instance $uv \in E$ is a (directed) edge from vertex u to vertex v. We use u, v, and w for generic vertices or nodes. We will describe our metrics in terms of undirected networks, where they are more intuitive.

Basic Metrics. The most basic network statistics that we expect to be correlated with article quality are n and m, the numbers of nodes and edges in the graph, respectively. The number of nodes is the number of editor 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 graph, which is the number of edges present in the graph divided by the number possible [7]:

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

We hypothesize that all three of these metrics are positively correlated with article quality. It is clear why a greater number of contributors would be correlated with higher article quality. We expect that when editors collaborate, they synthesize their contributions and build a shared conception of the desired state of the article, leading to better organization and more consistent style. Therefore, greater numbers of edges should correspond to higher quality articles. Similarly, since *density* can be interpreted as the percentage of possible collaborations between editors that actually occurred, we expect greater density to correspond to higher quality articles.

Distance-Based Metrics. Density and number of edges (m) capture the number of and proportion of editors in direct collaboration. For

[5] On the other hand, the directed model could outperform the undirected model even if the direction of collaboration isn't important because the possibility of having zero, one, or two edges between a pair of editors acts as a heuristic for the amount of collaboration between the two editors. This slight increase in granularity relative to the binary undirected graph case might be helpful. We discuss the possibility of using weighted graphs to capture volume of collaboration between pairs of editors in the discussion.

[6] Keegan et al. refer to them as loops, but we interpret this to mean cycles.

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

the remaining editors, we expect that a low average degree of separation is correlated with high quality articles. In terms of network analysis, the degree of separation corresponds to the distance between two nodes and low degrees of separation means that short paths exist in the network. Keegan et al. consider two statistics based on the distances between nodes in the network, *diameter* and *average closeness*, and we additionally include *radius* and average eccentricity.

The distance between two nodes is defined to be the length of the shortest path in the network, where the length of a path is the number of edges it contains. We denote this $dist(u, v)$, and the notation reinforces that this is a property of pairs of nodes. The *diameter* of a network is defined as the largest distance present in the graph. This is generalized by the concept of *eccentricity*; the *eccentricity* of a node is the longest distance from it to another node in the graph. The *diameter* of the graph can then be redefined as the maximum eccentricity of its nodes. Similarly, the *radius* is defined as the minimum eccentricity value present in the graph. Note that *eccentricity* is a property of a particular node, while *radius* and *diameter* are properties of the network as a whole. In our analysis we consider *radius*, *diameter*, and *average eccentricity* (over all nodes in the network).

We hypothesize that all three are negatively correlated with article quality. These three metrics give us basic distribution information about the maximum degrees of separation of editors in the collaboration network: *radius* is the minimum, *diameter* the maximum, and *average eccentricity* the mean.

While the eccentricity of v measures the size of the largest distance from v to another node, the *closeness centrality* of v incorporates all distances from v. In particular, it 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:

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

To make this a network statistic, we consider *average closeness centrality*, which reflects a tendency of shorts paths to exist between arbitrary nodes in the network. We hypothesize that high *average closeness* is correlated with high quality articles because it corresponds to low typical degrees of separation between editors.

Betweenness Centrality. The distance metrics discussed previously consider shortest paths between editors. We theorize that the central editors present on many of these shortest paths are responsible for integrating the content and revisions made by the outlying editors. These editors are referred to as brokers. The property of *betweenness centrality* is intended to capture 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 of 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*. We expect high *average betweenness* to be associated with low article quality because it indicates that a few central authors are doing most of the synthesis work, and perhaps contributing most of the content as well. In contrast, low *average betweenness* indicates that this work is distributed among more editors and perhaps that more editors are invested in the article. Low *average betweenness* may also be associated with tighter revision networks [12], in which case there may be less synthesis necessary.

Clustering Metrics. Clustering is intended to capture the tendency of the collaborators of an editor to collaborate with each other, a property that real-world social network tend to exhibit [10]. It exists in both a local version as a node statistic, and as a global network statistic.

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 take the *average clustering coefficient* as a graph statistic, and sometimes abbreviate it as *average clustering*.

Meanwhile, the *global clustering coefficient*, or transitivity, of a graph is defined as the proportion of triads that are closed, where triad is a pair of of edges sharing a node and where closed triads are those that have an edge joining the other vertices of the two edges.

We hypothesize that local and global clustering are both positively correlated with article quality because high values indicate that the collaboration network of the editors is similar to real-world social networks.

Other Variables. In addition to our network statistics, we consider including other independent variables that could have a large impact on article quality. In particular, we consider *number of editors*, *number of edits*, and *article size*. The number of editors is represented in the revision graphs as the number of nodes n. This leaves us with two independent variables outside of the network statistics: *article size* and *number of edits*. While *number of edits* is not a graph statistic, it does fit with the big picture idea that article history, and not only the current state of the article, should be considered when predicting quality. Additionally, the importance of article size is well known in the quality literature as previously noted in our related work section [4].

6 DATA COLLECTION

Our corpus of articles for this project was randomly sampled from the Main namespace of Wikipedia – where all the encyclopedia articles reside[8]. Each article in Wikipedia's Main namespace is assigned a page id value corresponding to its place in sequential order of creation (i.e. Sideshow Bob with page id 64910 was created 230,791 articles before Joe Montana with page id 295701). We randomly sampled Wikipedia articles by continuously polling MediaWiki's API for a random page id and the ten preceding it. Within each

[8] as opposed to the talk, user, or other miscellaneous Wikipedia pages.

tranche of page id's polled, any article corresponding to the page id was marked with one of Start, Stub, C, B, GA, or FA as their quality score was added to our corpus. We omitted A-Class articles from our corpus due to their scarcity[9] and inconsistency of assessment. We took the first 1000 articles of each remaining quality class for a final corpus of 6000 articles. We then stored the revision histories for each article in our corpus in case of deletions or further edits to the articles after finalizing our corpus.

In addition to our uniform corpus, we wanted a sample of articles matching the distribution of quality in Wikipedia. We re-ran a modified data collection to gather 5000 random articles with page assessments from any one of our six categories in an effort to represent the distribution of article qualities in Wikipedia. Our final dataset is the statistics calculated on the revision histories for the 6000 articles in our uniform corpus and on the 5000 randomly sampled articles.

7 RESULTS

In the past, the literature on classifying quality in Wikipedia has mainly focused on binary classification, such as whether or not an article is Featured [4]. However, some recent models have attempted the multi-class classification problem, distinguishing the seven quality classes [18] or the six quality classes (excluding A-Class) [9]. This fine-grained classification is the problem that we are interested in, motivated by our desire to offer editors a tool to improve articles and readers an anchor point for quality determination. Unlike other quality classification models that observe only the current state of an article, we examine the impact of structures of collaboration within article trajectories on quality classification. Revisiting our research question, we will evaluate the performance of our model relative to models that look only at a specific state of an article.

First, we detail the classification model that we use. Next, we examine our hypotheses about the relationships between our statistics and quality. Then we evaluate the efficacy of a classifier using both the directed and undirected article trajectory networks Within each subsection, we discuss the inclusion of article size as an input parameter. Finally, we contextualize the performance of our models.

7.1 Classification Model

As quality is an ordered categorical variable, it might seem natural to use ordinal logistic regression (OLR). OLR is a multi-class logistic regression where the outcome variables are ordered, hence ordinal rather than nominal. However, OLR relies on the proportional odds assumption, whereby each independent variable must have an identical effect at each split of the ordinal dependent variable. To test this assumption, we conducted a Brant test for parallel lines seen in Table 15 in the Appendix[10], which assesses whether the observed deviations within the OLR model are larger than what could be attributed to chance alone. The test failed to confirm the assumption, so the OLR model is not applicable[11]. As a consequence, we use multinomial logistic regression for our analysis, which can also classify into two or more discrete outcomes. However, it does not treat these outcome categories as ordered and thus does not require the proportional odds assumption.

7.2 Assessing our Hypotheses

One way to assess the validity of our hypotheses is to examine the probability of the classifier choosing any given quality class versus a baseline given certain statistics. This ratio of the probability of choosing a specific value over the baseline, in this case Stub, is referred to as the relative risk ratio (RR). In order to determine if our data matches our hypotheses, we examine the table of relative risk ratios calculated from the coefficients outputted after training a multinomial logistic regression on our dataset. In particular, we look at the value of the log odds and the trend of the RR ratios. Each category of statistics is coded in Table 2 and we examine their relationship to quality below.

Edit Hypothesis: Basic metrics are positively correlated with quality.

As we can see in the first shaded region in Table 2, the RR ratios for *edit count* and *editor count* increase monotonically across all article quality values. While the RR ratios of $density$ do not increase monotonically, if we split on Start the relationship holds. Specifically, looking at the relationship within two disjoint contiguous subsets, i.e. {Stub, Start} and {C-Class, B-Class, GA, FA}, the monotonicity is evident. This pattern will reappear in further examinations of our hypotheses. Our hypothesis of positive correlation holds strongly on *edit count* and *editor count* and loosely on $density$.

Distance Hypothesis: Eccentricity metrics are negatively correlated with quality. Closeness is positively correlated.

Observing the RR ratios, $radius$ and $diameter$ do not follow our hypothesis. While $radius$ simply follows a positive correlation with quality, $diameter$ shows no monotonic trend in either direction. This suggests that while the maximum degree of separation of the most central editor increases with quality, the "spread out-ness" of the network somewhat remains somewhat constant. While higher quality articles tend to have tighter networks this is counteracted by their larger node count, which raises the maximum possible degree of separation. *Closeness*, much like *density*, follows our hypothesis if we split the data along Start articles. Lastly, *average eccentricity* generally follows our hypothesis with the exception of GA-Class articles, where the RR ratio deviates. Our hypothesis only loosely holds on *closeness* and *average eccentricity*.

Betweenness Hypothesis: Betweenness is negatively correlated with quality.

In the second non-shaded region of Table 2, the RR ratio for betweenness generally follows our hypothesis except for the GA-Class. Much like *average eccentricity*, our hypothesis of negative correlation holds loosely on $betweenness$.

Clustering Hypothesis: Clustering metrics are positively correlated with quality.

Finally, we turn to clustering. *Global clustering*, much like $diameter$ shows no monotonic trend across quality levels. In fact, it seems to alternate right around 1, where an RR ratio of 1 indicates no change

[9] Only 0.03% of assessed articles are considered "A-Class" compared to 0.1% Featured.

[10] $Density$ could not be evaluated when running the Brant test in R.

[11] This result also suggests that there is no monotonic relationship between the independent and the dependent variables. This is further evidenced by the results in Table 2, where we examine the ratio of the probability of choosing one outcome category over the probability of choosing the baseline category (Stub). As we can see, for each independent variable, the probability ratio of the outcome is not always increased or decreased by the existence of the independent variable.

in classification probability. *Clustering*, on the other hand, follows the same pattern as *density* and *closeness*, where the correlation is monotonic on specific subsets. In summary, our hypothesis does not hold on *global clustering*, and only loosely on *clustering*.

It seems the classifier has some trouble reconciling our hypotheses with GA quality articles; the RR ratios for *betweenness* and *average eccentricity* mostly follow our hypotheses across all quality levels other than GA. One explanation for this deviation could be the limitations of human classifiers and the subjective nature of quality, particularly on more "complete" articles. Note that while *m* and *global clustering* fail to satisfy our hypotheses, that does not point to their impact on classification. However, metrics that are significant predictors *and* have values consistent with our hypotheses are particularly meaningful to our research. In the sections below, we identify the statistics that are important and the classifier's accuracy on those statistics.

Table 1. Coefficients from Directed Statistics

	ST	C	B	GA	FA
edit count	−20.14	−1.99	−1.85	−0.47	0.02
editor count	−21.60	6.88	7.111	7.71	7.63
article size	−11.88	2.77	3.254	3.52	3.53
density	0.03	−10.61	−4.82	−3.65	−3.22
m	35.94	−4.01	−4.49	−6.28	−6.76
diameter	0.41	0.85	0.74	0.96	0.32
radius	0.65	0.78	1.20	1.25	1.47
avg eccentricity	−0.98	−1.65	−1.88	−2.07	−1.85
closeness	0.015	−1.26	−1.04	−0.82	−0.46
betweenness	0.041	−0.69	−2.16	−1.39	−2.42
clustering	0.01	−0.52	−0.33	0.18	0.405
global clustering	−0.06	0.0003	−0.22	0.059	−0.15

Table 2. Relative Risk Ratios from Directed Stats

	ST	C	B	GA	FA
edit count	1.79*e*-9	0.137	0.157	0.624	1.020
editor count	0.42*e*-9	0.97*e*-3	1.22*e*3	2.22*e*3	2.51*e*3
article size	6.93*e*-6	15.98	25.885	33.624	34.211
density	1.032	2.47*e*-5	0.008	0.026	0.040
m	4.04*e*15	0.018	0.011	0.002	0.001
diameter	1.502	2.328	2.093	2.612	1.379
radius	1.913	2.190	3.313	3.496	4.348
avg eccentricity	0.375	0.192	0.153	0.127	0.157
closeness	1.015	0.284	0.355	0.440	0.629
betweenness	1.042	0.501	0.115	0.249	0.089
clustering	1.014	0.594	0.717	1.195	1.500
global clustering	0.946	1.000	0.802	1.061	0.860

7.3 Evaluating the Directed Model

Classifying using MLR, we examine how well our independent variables predict quality. To do so, we ran likelihood ratio tests on the trained MLR to find our set of important predictor variables. From our full set of statistics that we calculated (Table 1), we explored the relevance of our important predictors for each level of article quality (Table 11 in the Appendix). On our directed model, we determine that *clustering*, *betweenness*, *radius*, *diameter*, *average eccentricity*, *editor count*, *edit count*, *density*, and *m* are important (the log likelihood results can be seen in Table 11 and Table 12 in the Appendix).

Training on our uniform corpus of 5000 articles, we ran a 5-fold cross validation[12] of our model to test the accuracy of our predictors. Overall, the classifier correctly predicts an articles quality 46.65% of the time. In table 4 we list the performance metrics of our classifier. Among these metrics is sensitivity and precision, which measure the proportion of actual positives identified correctly and correct positive identifications, respectively. From the confusion matrix for our model run on the metrics seen in Table 3, a few patterns emerge.

First, much like other models that try to predict quality [18], [9], 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. We saw some early evidence of this in section 7.2, where the trend of RR ratios for *betweenness* deviated on GA-Class articles. Furthermore, not only does our classifier rarely guess GA, it categorizes GA-Class articles as FA, B, and C over 86.8% of the time. It seems that the classifier on our statistics has difficulty determining exactly what defines an article that achieves a "good" rating.

Second, looking closer at Table 3, we see a stratification of guesses. Table 3 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. This behavior is likely related to the piecewise trends commonly observed in the relative risk ratio analysis in section 7.2. We ran a binomial logistic regression (BLR) to investigate this observation. From the BLR, we get an accuracy score of 78.3% validating the patterns seen in the MLR confusion matrix. Accurately classifying articles with low quality is particularly impactful considering Start and Stub articles collectively make up 80.5% of all articles in English Wikipedia.

7.3.1 *Adding in Article Length.*

In the previous section we examined the impact the statistics derived from an article trajectory network make on quality classification. However, our goal is to make the most accurate classifier possible. To that end, we include *article length* as an additional predictor. As demonstrated by previous models [4] (as well as our own verification), *article length* is an important predictor for article quality. An MLR run with only article length as a predictor correctly predicts an article's quality 43.3% of the time. When we include it as a predictor along with our important statistics, the classifier improves to 48.4% accuracy. In the appendix, in Table 9 and

[12] We ran folds rather than using distinct training and testing datasets because Featured Articles are scarce and so random sampling is time-intensive.

Table 3. Confusion Matrix On Uniform Dataset

Directed Model (without Article Size)

	SB	ST	C	B	GA	FA	Total
SB	480	191	78	24	44	26	843
ST	404	801	5	2	0	0	1212
C	86	8	461	315	202	89	1161
B	5	0	281	376	274	202	1138
GA	3	0	51	79	157	159	449
FA	22	0	124	204	323	524	1197
Total	1000	1000	1000	1000	1000	1000	6000

Table 4. Performance Metrics on Uniform Dataset

Directed Model (without Article Size)

Class	Sensitivity	Precision	Balanced Accuracy
SB	0.48	0.57	0.70
ST	0.80	0.66	0.86
C	0.46	0.39	0.66
B	0.38	0.33	0.61
GA	0.16	0.35	0.55
FA	0.52	0.44	0.69
Total	0.47	0.46	0.68

Table 10, we see the resulting confusion matrix and its corresponding performance metrics. If we repeat the binomial classification, our accuracy jumps from 78.3% to 95.7%. With sensitivity and specificity values 0.974 and 0.923, respectively, the classifier guesses correctly and often.

7.4 Evaluating our Undirected Model

We have examined the performance of the article trajectory model as created by Keegan et al. Now, we turn to our extension on that model, article trajectory networks with undirected edges. Here we analyze the performance in relation to our directed model; if the classifier performs at least as well using the statistics from the undirected model, then we can conclude that the notion of symmetric collaboration is a better predictor of quality.

Through similar processes as before, we determine our important predictors to be: *density*, *m*, *edit count*, *editor count*, *clustering*, and *closeness* (Table 12 in the Appendix). The classifier's performance on the undirected statistics outperforms the directed model, with an accuracy score of 46.8% compared to the directed model's 45.68%[13]. Thus, we conclude that the undirected article trajectory network is a better predictor of quality. In Table 5 and Table 6, we show the confusion matrix and the classifier's performance metrics. Unfortunately, none of the increase in performance can be attributed to a more accurate classification of GA-Class articles. In fact, sensitivity values decreased for GA classification. Instead, the increase in classification accuracy for the undirected model is in Start, B-, and FA-Class articles, while the other classes see little to no improvement.

Table 5. Confusion Matrix On Uniform Dataset

Undirected Model (without Article Size)

	SB	ST	C	B	GA	FA	Total
SB	454	179	81	23	47	26	810
ST	428	816	5	2	0	0	1251
C	83	5	460	314	204	88	1154
B	7	0	271	382	281	195	1136
GA	6	0	64	80	155	148	453
FA	22	0	119	199	313	543	1196
Total	1000	1000	1000	1000	1000	1000	6000

Table 6. Performance Metrics on Uniform Dataset

Undirected Model (without Article Size)

Class	Sensitivity	Precision	Balanced Accuracy
SB	0.45	0.56	0.69
ST	0.82	0.65	0.87
C	0.46	0.40	0.66
B	0.38	0.34	0.62
GA	0.16	0.34	0.55
FA	0.54	0.45	0.71
Total	0.47	0.46	0.68

Now that we have determined that the undirected model is a better predictor for quality, we want to optimize the classification accuracy of this model. To do so, we again add article length as a predictor variable. Much like in the directed model, this addition increases the classifier's accuracy to 49.35%. The confusion matrix and performance results are listed in tables 13 and 14. With or without the inclusion of article length, the classifier on our undirected model performs significantly better than on the directed model.

7.5 Representative Sample

Our uniform corpus of articles is not a representative sample of the population. Since there is not a uniform distribution of quality in Wikipedia articles, having data matching the underlying distribution is also critical. On our representative sample of articles, we calculated the statistics from the corresponding undirected article trajectory network. Running the MLR on this dataset, we get an accuracy score of 67%. In tables 7 and 8, the confusion matrix and corresponding performance metrics are listed. Note that the classifier never predicts an FA-Class article correctly, and much of the accuracy is from correctly predicting Stub articles. It confuses Start articles with Stub nearly 50% of the time. Moreover, the higher balanced accuracy of the undirected model with article length on the

[13] This difference is significant with p-value $< 2.2e - 16$.

uniform sample versus the representative sample, shows that our statistics alone offer a meaningful measurement of quality.

Table 7. Confusion Matrix On Representative Corpus

Undirected Model (with Article Size)

	SB	ST	C	B	GA	FA	Total
SB	2410	855	51	15	1	0	3332
ST	308	898	251	70	27	3	1557
C	3	20	26	18	3	0	70
B	0	7	5	9	1	4	26
GA	0	1	0	1	1	1	4
FA	0	0	0	1	1	0	2
Total	2721	1781	333	114	34	8	5000

Table 8. Performance Metrics on Representative Corpus

Undirected Model (with Article Size)

Class	n	Sensitivity	Precision	Balanced Accuracy
SB	2721	0.89	0.72	0.74
ST	1781	0.50	0.58	0.65
C	333	0.07	0.40	0.53
B	114	0.09	0.37	0.55
GA	34	0.03	0.33	0.51
FA	8	0.00	0.00	0.49
Total	5000	0.27	0.40	0.58

7.6 Comparing to Other Models in the Field

While this field is still relatively unsaturated, there are some existing models to predict the quality of Wikipedia articles. One such model is ORES, a decision tree classifier that takes a number of features as input, including article length and number of headings.

While both ORES and our model seek to classify articles by quality, they are fundamentally different approaches. ORES takes as input a specific revision id from an article, and examines features of that state of the article. In contrast, our model seeks to define quality through collaboration and interaction among Wikipedia contributors operationalized as network statistics of an article's revision history. ORES claims an accuracy of 62.9% on their nearly uniform dataset. However, we ran their model on our uniform corpus and found an accuracy score of 52%.

In addition, the parameters for ORES classification of quality are based on the WikiProject guidelines for quality. ORES inputs such as article length, number of headings, references, and broken links 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) [2]. Our model's predictors examine structural traits of article trajectories that are perhaps less clearly linked to the quality articles. Nevertheless, our undirected model (without the inclusion of *article length*) achieves an accuracy of 46.8%, thus performing nearly as well, while observing only the characteristics of graphs constructed via edits histories[14]. Our investigation therefore opens the possibility for similar graph analysis to be explored as a novel method of article quality prediction.

8 DISCUSSION

As demonstrated by our results, we have achieved our research goal of creating a novel method of article quality classification. Additionally, our model gives a “good” answer to our research question, performing similarly to ORES and the Actionable Model, although our results are not actionable. In the following sections we discuss the implications of our achievement and suggestions for future work.

8.1 Implications

As mentioned in our introduction, a predictive quality classifier is beneficial because of its ability to provide Wikipedia readers with a sense of the quality of a given article. While our model cannot replace human quality assessment and certainly does not fact-check the actual content of Wikipedia pages, it nevertheless can provide users with a general sense of whether a page, because of its underlying revision structure, is more or less likely to be of high quality. In addition, our model can be used to flag articles that might have been misclassified by WikiProjects editors or whose quality class might be outdated. For instance, if an article has been rated B-Class but possesses similar network characteristics to that of a Stub, we can direct editors to take a second look to ensure that the article is deserving of its quality rank. Conversely, if an article has been rated a Stub but has characteristics more similar to those of a B-Class article, we can notify editors that their time and effort might be better spent on other articles that more desperately require fixing.

In addition to these benefits, our unique model can begin to characterize particular collaboration structures or patterns that may result in higher quality articles. It is important to note that we do not prove causality between these variables. Still, the relative success of our model is indicative that these structural network elements are likely important. Furthermore, our model demonstrates similar accuracy in comparison to other models, such as the Actionable Model and ORES, that more directly measure the quality class criteria through predictors related to an articles current public state. This suggests not only that collaboration patterns are potentially important to article quality, but also that our graph model, inspired by Keegan et al., provides a valid and useful representation of an article's underlying collaboration structure and edit trajectory.

8.2 Suggestions for Future Work

Given that the predictive accuracy of our classifier model improved with the addition of non-network statistics like article length and number of revisions, future work should examine whether the inclusion of additional non-network predictors (such as those from Warncke-Wang's Actionable Model) would further increase the accuracy of our model.

[14] With *article length* we perform even better with a prediction accuracy of 49.35%

Future work could also add edge weights in order to quantify the frequency of the collaboration between two editors. In our current model, multiple back-and-forth revisions between two editors are only represented by a singular edge. To account for the potential impact of repeated collaboration as opposed to one-time collaboration on article quality, the weight of a given edge could potentially be dependent on the count of revisions that occur between editors. Alternatively, edge weights could be adjusted to account for the size of an individual revision in terms of the words added or deleted.

Finally, it may be worthwhile to re-examine our model's assumptions about what constitutes editor collaboration. Currently, our model treats all collaboration equally. This may not be the case if, for instance, an editor is simply reverting changes made by a previous editor who vandalized a section of the page. Additionally, a temporal requirement to collaboration could explored– the collaboration between authors of consecutive revisions that occur weeks or months apart could be argued to be qualitatively different that the collaboration between authors revising an article simultaneously. Furthermore, we could reconsider whether authors working on separate sections of an article simultaneously should be considered collaborators in our network.

By evaluating the impact of these modifications to our model, we could seek to improve our understanding of the characteristics of editor collaboration networks that correspond to quality on Wikipedia.

Dedicated to the Partida Family.

REFERENCES

[1] 2020. Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=958396108 Page Version ID: 958396108.

[2] 2020. Wikipedia:Content assessment. https://en.wikipedia.org/w/index.php?title=Wikipedia:Content_assessment&oldid=951373538 Page Version ID: 951373538.

[3] Thomas B. Adler and Luca de Alfaro. 2007. A Content-Driven Reputation System for the Wikipedia.

[4] 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, 1095. https://doi.org/10.1145/1367497.1367673

[5] 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, 464–471. https://doi.org/10.1145/2808797.2808895

[6] Kittur et al. 2007. He Says, She Says: Conflict and Coordination in Wikipedia. (2007).

[7] Li et al. 2015. Automatically Assessing Wikipedia Article Quality by Exploiting Article–Editor Networks. Vol. 9022. Cham.

[8] Jim Giles. 2005. Challenges of being a Wikipedian. *Nature* 438, 15 (2005), 2.

[9] 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, 1–9. https://doi.org/10.1145/3125433.3125475

[10] Paul W. Holland and Samuel Leinhardt. 1971. Transitivity in Structural Models of Small Groups. *Comparative Group Studies* 2, 2 (1971), 107–124. https://doi.org/10.1177/104649647100200201 arXiv:https://doi.org/10.1177/104649647100200201

[11] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W Lauw, and Ba-Quy Vuong. [n.d.]. Measuring Article Quality in Wikipedia: Models and Evaluation. ([n. d.]), 10.

[12] 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, 1. https://doi.org/10.1145/2462932.2462934

[13] Aniket Kittur and Robert E Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. (2008), 10.

[14] Andrew Lih. 2004. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. (2004), 31.

[15] 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), 1–23. https://doi.org/10.1145/1985347.1985352

[16] Jun Liu and Sudha Ram. 2018. Using big data and network analysis to understand Wikipedia article quality. *Data & Knowledge Engineering* 115 (May 2018), 80–93. https://doi.org/10.1016/j.datak.2018.02.004

[17] Deborah L McGuinness, Honglei Zeng, Li Ding, Dhyanesh Narayanan, and Mayukh Bhaowal. [n.d.]. Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study. ([n. d.]), 9.

[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, 1–10. https://doi.org/10.1145/2491055.2491063

[19] 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 arXiv: cs/0702140.

[20] Honglei Zeng, Maher A Alhossaini, Li Ding, Richard Fikes, and Deborah L McGuinness. [n.d.]. Computing Trust from Revision History. ([n. d.]), 8.

APPENDIX

Table 9. Confusion Matrix On Uniform Dataset

Directed Model (with Article Size)

	SB	ST	C	B	GA	FA	Total
SB	540	144	79	32	47	15	857
ST	349	851	5	0	1	0	1206
C	86	3	527	371	221	92	1300
B	3	0	209	312	239	179	942
GA	5	2	58	85	175	190	515
FA	17	0	122	200	317	524	1180
Total	1000	1000	1000	1000	1000	1000	6000

Table 10. Performance Metrics on Uniform Dataset

Directed Model (with Article Size)

Class	Sensitivity	Precision	Balanced Accuracy	Within-1 Accuracy
SB	0.54	0.63	0.74	88.9%
ST	0.85	0.71	0.89	99.8%
C	0.53	0.41	0.69	74.1%
B	0.31	0.33	0.59	76.8%
GA	0.18	0.34	0.55	73.1%
FA	0.52	0.44	0.69	71.4%
Total	0.49	0.48	0.69	80.68%

Table 11. Nested Log Likelihood Calculations, Undirected Model on Uniform Dataset

	LR Chisq	Df	Pr(>Chisq)
edit count	159.285	5	$< 2.2e\text{-}16$
editor count	24.107	5	0.0002
density	45.405	5	$1.20e\text{-}8$
m	55.659	5	$9.55e\text{-}11$
diameter	13.135	5	0.022
radius	17.191	5	0.004
avg eccentricity	13.451	5	0.019
closeness	9.670	5	0.085
betweenness	20.681	5	0.001
clustering	30.998	5	$9.37e\text{-}6$
global clustering	8.034	5	0.154

Table 12. Nested Log Likelihood Calculations of Undirected Model on Uniform Dataste

	LR Chisq	Df	Pr(>Chisq)
edit count	199.626	5	$< 2.2e\text{-}16$
editor count	30.736	5	$1.06e\text{-}5$
density	60.764	5	$8.448e\text{-}12$
m	75.951	5	$5.889e\text{-}15$
diameter	−1.834	5	1
radius	4.544	5	0.474
avg eccentricity	1.475	5	0.916
closeness	23.889	5	0.0002
betweenness	−7.151	5	1
clustering	19.335	5	0.002
global clustering	1.421	5	0.922

Table 13. Confusion Matrix of Undirected Model (with Article Size) On Uniform Dataset

	SB	ST	C	B	GA	FA	Total
SB	538	140	76	28	45	16	843
ST	351	856	7	0	1	0	1215
C	87	4	540	376	223	97	1327
B	2	0	202	309	246	171	930
GA	5	0	60	101	174	172	512
FA	17	0	115	186	311	544	1173
Total	1000	1000	1000	1000	1000	1000	6000

Table 14. Performance Metrics of Undirected Model (with Article Size) on Uniform Dataset

Class	Sensitivity	Precision	Balanced Accuracy	Within-1 Accuracy
SB	0.54	0.64	0.74	88.9%
ST	0.86	0.71	0.89	100%
C	0.54	0.41	0.69	74.9%
B	0.31	0.33	0.59	78.6%
GA	0.18	0.34	0.55	73.1%
FA	0.54	0.46	0.71	71.6%
Total	0.49	0.48	0.69	80.68%

Table 15. Brant Test Data – H_0: Parallel Regression Assumption Holds

Metric	χ^2	df	p-value
diameter	18.416	4	0.001
radius	−4.612	4	1
avg. eccentricity	6.268	4	0.180
closeness	142.203	4	9.53*e*-30
m	96.403	4	5.73*e*-20
edit count	180.652	4	5.40*e*-38
article size	−11.291	4	1
clustering	−11.303	4	1
global clustering	17.702	4	0.001
betweenness	416.146	4	9.02*e*-89