What Makes a Good Team of Wikipedia Editors? A Preliminary Statistical Analysis

Leszek Bukowski[1], Michał Jankowski-Lorek[1], Szymon Jaroszewicz[2,3], and Marcin Sydow[1,3] *

[1] Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
[2] National Institute of Telecommunications
Szachowa 1, 04-894 Warsaw, Poland
[3] Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warsaw, Poland
bqpro@pjwstk.edu.pl,fooky@pjwstk.edu.pl,
s.jaroszewicz@ipipan.waw.pl, msyd@poljap.edu.pl

Abstract. The paper concerns studying the quality of teams of Wikipedia authors with statistical approach. We report preparation of a dataset containing numerous behavioural and structural attributes and its subsequent analysis and use to predict team quality. We have performed exploratory analysis using partial regression to remove the influence of attributes not related to the team itself. The analysis confirmed that the key issue significantly influencing article's quality are discussions between teem members. The second part of the paper successfully uses machine learning models to predict good articles based on features of the teams that created them.

Keywords: team quality, Wikipedia, dataset, statistical data mining

1 Introduction

This paper concerns the problem of quality of Wikipedia editor teams. More precisely, we report a recent work on a set of attributes computed on a real data concerning Wikipedia team quality.

We report *preparation of a dataset* containing numerous *behavioural* and *structural* attributes computed on a real collaboration network downloaded from Wikipedia. Some attributes were reported in other works before, but, up to our knowledge, the whole set of attributes reported here was not studied before in such a context. The presented dataset is a substantial extension of an earlier existing one and is to be made publicly available for other researchers at `http://wikitems.pjwstk.edu.pl/data`.

* This work is supported by Polish National Science Centre grant 2012/05/B/ST6/03364

Subsequently we report *preliminary statistical analysis* of this data and preliminary results concerning application of machine learning to *predict* the quality of a Wikipedia article.

Since there are available huge amounts of logged edit data concerning the work of contributors to Wikipedia articles it is interesting to apply data mining techniques to get some insights into this process. It is reasonable to expect that as a long-term goal of the preliminary research presented here , some phenomena observed in one dataset will be also present in other data, so that they would form some *universal laws*. Detecting such laws would be very valuable to improve the quality of social media and to understand some *sociological* aspects of the analysed processes.

The authors hope that the work presented in this paper would serve as one of the initial steps on a long way to achieve such goals.

One of the additional contributions of our paper is that we take into account correlations between variables such that the influence of variables strictly related to the team is isolated from variables influenced primarily by aspects such as article popularity. We show that such confounding variables can significantly distort the picture and present some techniques to tackle this problem in order to obtain more meaningful results.

1.1 Related Work

Social Network Analysis (SNA) have been used as a framework for study communities of Wikipedia editors. Behavioural social networks have been widely used to model the knowledge community of Wikipedia editors. Such networks are derived from observed behaviour that has been (in the case of Wikipedia) recorded in the edit history. Especially multi-dimensional (multi-layered) social network (MDSN) model [8] of Wikipedia knowledge community is a useful tool for practical applications, such as recommender systems of editors, admin candidates, and for specialised applications, such as conflict detection or evaluating article quality. For example in [12, 10, 4] researchers have been using network structure as a model of conflict and collaboration between editors. According to balanced networks theory [17] in such situations, when conflict arises in social network, actors shall form densely connected clusters of collaborators, who are conflicted with actors from other clusters.

Conflict and collaboration are not the only topics that have been analysed with SNA techniques. Another problem that attracts significant attention is the problem of coordination [9, 10]: one of the main phenomenon that is associated with Wikipedia is that all articles and entries have been created by teams of editors without any central authority. For example in [9] it has been shown that adding more editors to an article might improve article quality only when appropriate coordination techniques are used. Another work that examines problems surrounding coordination and conflict in Wikipedia is [11].

We are not aware of previous research that isolates team-related features from other confounding aspects such as article popularity.

In [16, 15], social network analysis was performed on the behavioural social network mined from the entire edit history since the inception of Polish Wikipedia in 2001. This paper partially builds on that work.

The machine learning approach to predict trust in social networks was studied in [2] and [3]. Similar analysis of usability of various attributes but in the domain of web spam prediction is presented in [13].

2 Data

We significantly extended our previous MDSN dataset mentioned earlier [15].

The MDSN data is a multi-graph, where the set of nodes A consists of the Wikipedia authors and the arcs fall into several behavioural categories (dimensions). : *co-edits*, *reverts* and *discussions*, that will be described later.

In this paper, we assume that for each Wikipedia article its *team* is the set of all the authors whose contributions are present in the final version of the article (at the moment of collecting the data). For each article, the nodes representing such authors induce a subgraph for which we computed various attributes. While one might argue that such a definition is a bit simplistic, it is quite natural and intuitive. [4]

When computing the attributes in this paper, the dimensions are never combined, i.e. each attribute is derived from exactly one of the dimensions.

In this paper, for each team we consider and compute a large number (almost 100) of attributes to be subsequently used in statistical analysis of the data. The attributes can be divided into two groups: behavioural and purely graph-based, that will be described in detail in next subsections. Some attributes are very technical and hard to be naturally interpreted, especially the "triad-based" ones. The approach of making the initial number of potential attributes large is typical in machine learning and makes it possible to avoid omitting any piece of information that may turn out to be useful in analysis or prediction. Subsequent analysis will indicate which attributes are actually useful, the remainder may be eventually dropped.

In addition, we assign each article a decision variable called `featured` that represents the fact whether it is a high-quality article. The variable is set to `TRUE` iff the article was marked as "featured" or "good" in Wikipedia. This decision variable will be treated as a gold-standard in our subsequent statistical analysis.

2.1 Behavioural Attributes

Chosen behavioral attributes utilize the available data in Wikipedia edit history. These attributes try to approximate significant social concepts, such as acquaintance, trust, distrust, that can have an impact on teamwork. For more detailed attribute explanation, see [15].

[4] Furtermore it was inherited from the initial dataset that is being extended in this paper. We plan to make the team definition more sophisticated in our ongoing work

For each pair of authors $(a_1, a_2) \in A^2$ in the same team, there potentially exist three directed arcs between them, each corresponding to one of the behavioural dimensions. The arcs have non-negative weights that are computed as follows.

Co-edits is defined as the amount of text (number of words) written by one author next to the text of other author invertedly weighted by the word distance between them.

$$coedits(a_1, a_2) = \sum_{(w_1, w_2)} 1/(wordDistance(w_1, w_2))$$

where the summation is taken over all pairs of words in the considered article (w_1, w_2) in each revision, where w_1 is added in the current revision by author a_1 and w_2 had been previously written by a_2 and the word distance is at most $maxWordDistance$, that after some experimentation, was set to 20. The threshold of 20 provides a good balance between computational efficiency and the relevance of the edition.

Reverts describes how many times one author reverted to a revision that is identical to a previous revision of the second author but not further away that $maxRecent$ revisions ago (that we set to 20).

The strength of a *discussion* dimension between two authors of a team counts the number of times when the first author wrote a word after the text of the second author in the same article or user talk page but not further away than `discussionDistance` (that we set to 20 words).

Abbreviations *disc* (or *discussion*) as well as *rev* or *edit* will be used as parts of attribute names later in the paper.

Next, for subgraph induced by each team T and for each dimension, we *aggregated* all the weights of arcs in this subgraph in three ways: as *sum*, sum normalised by the number of arcs in the subgraph (called *avge*) and sum normalised by the maximum potential number of edges in that graph $|T|(|T|-1)$ (called *avgv*). These abbreviations are used as suffixes in names of corresponding attributes mentioned later in this paper. For example, discussion_avge, etc.

We have created 300 069 subgraphs for all teams based on MDSN and then calculated 23 attributes for each MDSN dimension.

In addition, some other attributes have been computed for each team, for example: *edits* (total number of edits made to the article), *anon_edits* (number of edits made by anonymous users), *bot_edits* (number of edits when the username contained the "`bot`" as a suffix), etc.

2.2 Structural Attributes

Based on the MSDN network, we additionally computed numerous attributes that are based on pure structural properties of the underlying graph. The attributes are as follows. The abbreviations in brackets will be later used in attribute names mentioned later in this paper (e.g. *rev_nodc* stands for out-degree centrality computed on the *revisions* dimension, etc.)

- network in-degree centralisation (*nidc*)
- network out-degree centralisation (*nodc*)
- betweenness centralisation (*nbc*)
- number of week components (*nwcc*)
- size of the largest week component (*solwcc*)
- number of strong components (*nscc*)
- size of the largest strong component (*solscc*)
- triadic census (16 variables describing each triad tr_i for $i \in \{1, 2, 3...16\}$)

In-degree and *out-degree centrality* (also known as, simply, in-degree and out-degree) of a *single node* are the number of incoming and outcoming arcs, respectively.

E.g. the in-degree of some author x in the dimension of *co-edits* reflects the number of other authors that have edited the text near x's texts. Respectively the out-degree of x reflects the number of other authors near whose text x has edited. We also do such computations for reverts and discussion dimensions.

The *in-degree centralisation of a whole graph* G is the variation of in-degree centrality of nodes from G divided by the maximum possible variation of in-degrees in a network of the size G. It is given by the following formula:

$$C_G^{in} = \sum_{i=1}^{n} (d_{in-max}(G) - d_{in}(v_i))/(n-1)^2 \tag{1}$$

Where n is the total number of nodes in G; $d_{in}(v_i)$ is the in-degree of the i-th node in G and $d_{in-max}(G)$ is the highest observed in-degree in G. Metric C_G^{in} is always 1 for the networks where there is just one node to which all other nodes send ties – i. e. C_G^{in} is equal 1 in all networks, where $d_{in}(v^*) = (n-1)$ and for all other nodes $d_{in}(v_i) = 0$. The fact that maximal possible value of in-degrees variation in a network of size n is equal $(n-1)^2$ might be explained as follows: when all nodes send relations to just one node, then we have to $n-1$ times sum up $(n-1) - 0$, so it is just $(n-1)(n-1)$, which equals $(n-1)^2$.

C_G^{out} is defined analogously.

Betweenness centrality of a node v_i in a digraph is given by:

$$C_B(v_i) = \sum_{j<k} p_{jk}(v_i)/p_{jk} \tag{2}$$

Where $p_{jk}(v_i)$ is the number of the shortest paths between nodes j and k, that pass through node v_i and p_{jk} is the number of all the shortest paths between j and k. The maximum of equation (2) is $[(n-1)(n-2)]$, which is the number of all directed pairs in the network not including v_i. So in digraphs $C_B(v_i)$ is normalised as follows:

$$C'_B(v_i) = C_B(v_i)/[(n-1)(n-2)] \tag{3}$$

Based on the above, the *betweenness centralisation for a given network* G is defined as follows:

$$C_G^B = \sum_{i=1}^{n} (C'_B(v^*) - C'_B(v_i))/(n-1) \tag{4}$$

By $C'_B(v^*)$ we mean the highest betweenness centrality observed in G, normalised according to formula (3). The value of C_G^B is 1 in the networks where all the shortest paths, that would possibly pass through just one particular node, in fact pass trough that node and for all other nodes $C_B(v_i) = 0$, so there are no shortest paths that pass through them.

We also count the number of *connected components* for each network. A *weakly connected component* of a network G is a maximal such a subgraph of G so that for each pair of nodes in G there exists a path in G between them while the directions of the arcs can be ignored. A *strongly connected component* is defined similarly, but the path has to be directed.

2.3 Triads

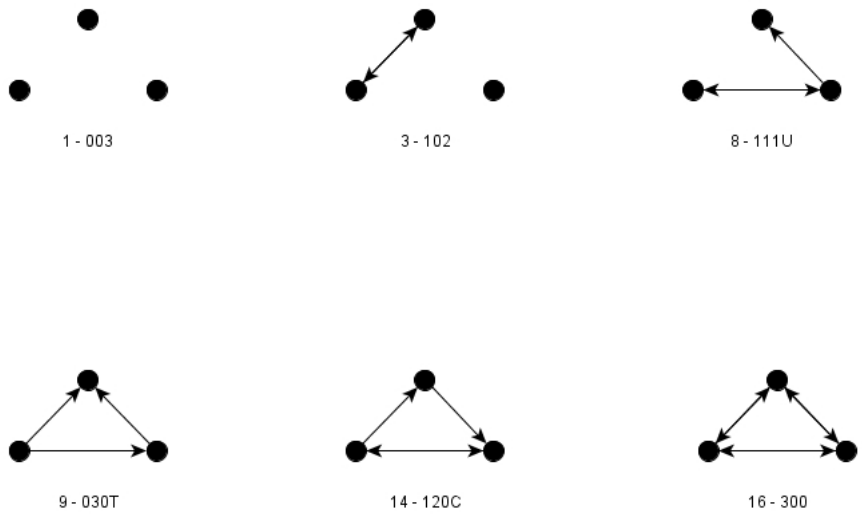

Fig. 1. Examples of triads

In case of digraphs, that represent some social networks, we can investigate of how many *triads* of a certain kind they include.

Triads are the triplets of actors who might interact somehow with each other. These interactions are represented by arcs. There are 16 possible configurations that any triple of actors from a digraph might be in. Starting from an empty triad (003), where there are no arcs between any two nodes and ending in a complete

triad, where all possible arcs are present (300). All triads might be identified by so called "M-A-N number" which is a three-digit number, sometimes supported by additional letter. The first digit indicates how many mutual dyads exist in a triad in question; that is how many pairs of nodes are in configurations where choices are reciprocated. The second digit indicates how many asymmetric dyads one can find in the triad under consideration. Finally, the last digit informs how many null dyads exist in the given triad – null dyads are pairs of actors between whom there are no arcs. For example, we can see on Figure 1, that the triad where there are no arcs between nodes is coded as 003. Respectively, the triad number 9 on the picture is coded as 030T: indeed it includes just three asymmetric dyads and it is transitive, so we have a letter "T".

We may test some structural hypothesis about networks in questions by investigating the frequencies of occurrence of some types of triads. The simplest situation is so called "balanced network", where all actors form just two clusters between which there are no interactions. Think for example about a conflicted group of people, who have formed two factions and they collaborate only with the members of their own group. For simplicity, assume that choices are always reciprocated. In such a network there would be only two types of triads – namely: 102 and 300. Indeed no matter which triplet of actors you choose from that network, they have to be in one of these two configurations: 102 or 300 – or it is the case that all three actors are in the same faction (300) or just two of them is in one party, and the third one is in the opposite faction (102).

Obviously, perfectly balanced networks are very rare in empirical studies. It is much more convenient to compare the empirical distributions of some types of triads with their expected distributions, which are computed under assumption, that the networks in question have been generated by some random, stochastic process. For example, if we investigate some social network, where arcs between actors represent trust, then we might expect, that there is a tendency towards transitivity in that network. So if actor i trusts j and j trusts k, then we might expect, that there is an arc from i to k. In such a network all triads, that are transitive – i. e. that contain a following configuration $i \rightarrow j \quad and \quad j \rightarrow k \quad and \quad i \rightarrow k$ – should be more frequent, than in a random network consisting of the same actors. In figure 1 all three bottom triads are transitive.

3 Exploratory Analysis

In this section we present an exploratory analysis of the data. We include all available variables and use statistical analysis to discover previously unknown relationships between those variables and the team quality.

As the pre-processing step, we removed teams with less than 15 members. The reason was that social network analysis measures such as triads are not meaningful for such small networks. Next, we replaced all attributes with their logarithms (one was added before taking logarithms to avoid taking logarithms of zero) in order to correct a significant skew present in the distributions of almost all variables. After the transformations, the data consisted of 38208 records with

89 attributes. Out of all articles, only 0.9% were marked as 'good' or 'featured', the class distribution is therefore highly imbalanced.

3.1 Initial Analysis

ROC curves [6] are a popular method of assessing the performance of classification models. It depicts the tradeoff between the percentage of positive cases identified by the model (y-axis) and the percentage of incorrectly labeled negative cases. The curve is frequently summarized with a single number: the Area Under the ROC Curve abbreviated AUC. The ROC curve is diagonal for a random model with AUC equal to 0.5, for a perfect predictor the curve passes through the point $(0, 1)$ and the AUC is equal to 1. AUC is equal to 0 for a model which always predicts the opposite class. Notice that such a model can easily be converted into a perfect predictor by reversing its scores.

We first look at how predictive each attribute is. Instead of using traditional measures for evaluating attribute usefulness, we use the AUC measure of single attribute models. This allows us to easily compare single attributes with more complicated models.

To this end, we treat each attribute as a predictive model and draw ROC curves for this model. For each attribute, its area under the ROC curve (AUC) is used to assess the predictiveness. Recall that highly predictive attributes have AUC close either to 0 or to 1 so we use $\max\{AUC, 1-AUC\}$ as the final measure.

In the following when we talk about AUC we are in fact referring to this quantity.

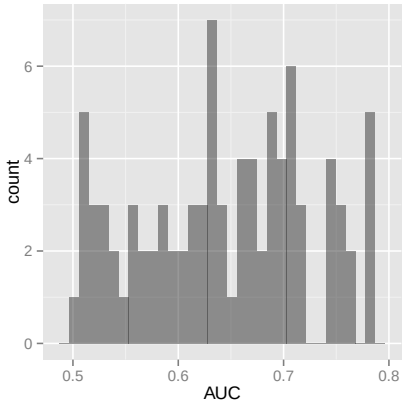

Fig. 2. Distribution of AUCs for attributes in the initial dataset.

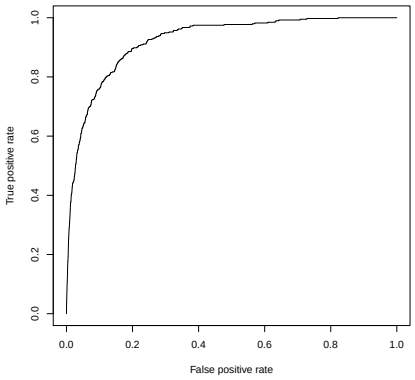

Fig. 3. ROC curve for logistic regression model built on the original variables.

Figure 2 shows the distribution of AUCs for all attributes. It can be seen than many attributes are predictive with the median AUC being 0.641.

The most predictive attribute was `disc_nidc` with an AUC of 0.785. Other most predictive attributes were also related to the discussion dimension.

Afterwards we computed a logistic regression model on all available variables. The model was highly predictive, its ROC curve is shown in Figure 3. Coefficients of several attributes had p-values close to 0 with the most significant attribute being the `edits` attribute: total number of editions made to the article.

This immediately shows the following problem with an approach that has been taken by most previous research: article quality depends on the amount of work devoted to it and this in turn is related to aspects such as article popularity which are *not* the qualities of the team per se. Moreover, the risk is that other attributes are predictive not because they measure team quality, but because they are correlated with the total amount of work devoted to the article. The correlation of `edits` with `disc_nidc` the variable with the highest AUC is in fact quite low (0.08), but its correlation with the second most predictive variable `disc_tr_16` was 0.53, indeed fairly high.

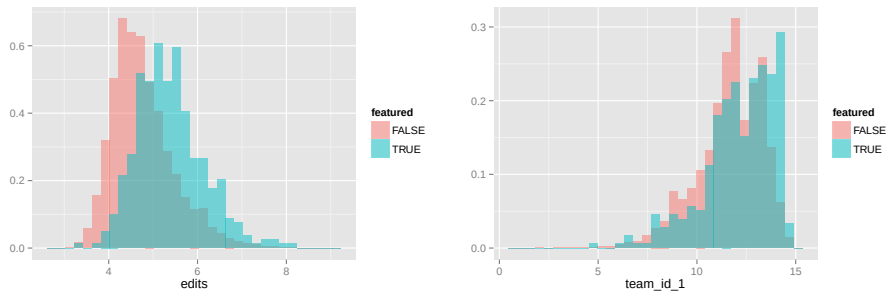

Fig. 4. Histograms within both classes for the `edits` and `team_id_1` attributes.

Figure 4 (left) shows the histogram of the logarithm of the total number of edits within both classes. One can see that higher number of edits favours the situation that the resulting article is "good". This attribute is a confounder since it is not a property of the team itself. Later, this will be further discussed.

An interesting observation is that the `team_id_1` attribute is also quite informative (AUC=0.602). The interpretation is that `team_id_1` is (invertedly) correlated to the age of the article, so the higher `team_id_1` the later first version of the article was created. Collective experience of all Wikipedia users is growing so that number of good articles is rising with time. Other interpretation might be that increased rate of good articles in time is due to the popularity of new topics and new way of promoting it for example through "article of the day".

3.2 Eliminating the Confounding Variables

There are several techniques available in statistical literature which allow for removing the effect of controlling variables [7]. The easiest is to keep a slice of the data small enough for the values of all such variables to be approximately constant. This amounts to conditioning on the confounding variables thus removing their influence. Typically, the analysis is performed by conditioning on several different values of the confounding variables. Special provisions exist in statistical packages e.g. for drawing regression plots for various conditioning values [7].

Note that simply removing the confounding variables is not sufficient, as they are correlated with the remaining attributes. For this reason, attribute selection is not sufficient, and we use partial regression techniques instead.

The advantage of conditioning is that it works correctly when the variables to be cancelled influence the remaining ones nonlinearly. Unfortunately, the method leads to severe data loss, and is thus in practice limited to conditioning on one or two variables. In the problem at hand the percentage of featured articles is quite small so the technique is not suitable.

Another approach, which we are going to use in this paper, are partial regression methods [7]. The idea is to build regression models which predict all variables in the dataset based on the confounding ones. The confounding variables are then removed and all remaining variables are replaced with *residuals* from regressions on the confounders. Recall that the residuals in a linear model are the differences between the true and predicted values, and that they are uncorrelated (see e.g. [14]) with the predictor variables (confounders in our case). As a result, after the transformation, the resulting variables are uncorrelated with the ones whose influence we are trying to eliminate.

The advantage of this method is that it avoids the data loss incurred by conditioning; the disadvantage, that the removal of confounding is only as good as the regression models used. The method is typically applied in the classical, linear regression setting. The differences between the two approaches are investigated (for the case of partial and conditional correlation) e.g. in [1].

Due to the problems with data loss incurred by conditioning we have used partial regression techniques to analyze our data. First, we remove the influence of the total number of edits applied to an article and the article id. To this end we build linear regression models from `edits` and `team_id_1` to all other variables in the dataset except the predicted variable `featured`. The two variables are then removed and all others replaced with residuals of their respective model. Notice that as a result all variables become *uncorrelated* with the total number of edits and article id. Recall that the class attribute has not been affected. Replacing it with appropriate residuals would lead to overly optimistic results as uncertainty caused by removing the confounders would also be removed.

After taking those steps we repeated the analysis from Section 3.1. Overall, the variables became much less predictive, the median AUC for separate variables went down from 0.641 to 0.587. This confirms that the true cause for predictiveness of several variables lied in them being correlated with the total number of work done on the article by the team.

After conditioning, the AUC of the most predictive attribute `disc_nidc` went down from 0.785 to 0.774, a very small decrease. However, as the next section shows, other confounding variables still need to be removed, which will more significantly decrease the predictive value of this attribute.

3.3 Anonymous Edits

Interestingly, the new most predictive attribute was `anon_edits` which gives the total number of anonymous edits. Its AUC was 0.812. Figure 5 (left) shows the histogram of the values of the attribute (actually the residuals obtained by projecting them onto the removed attributes).

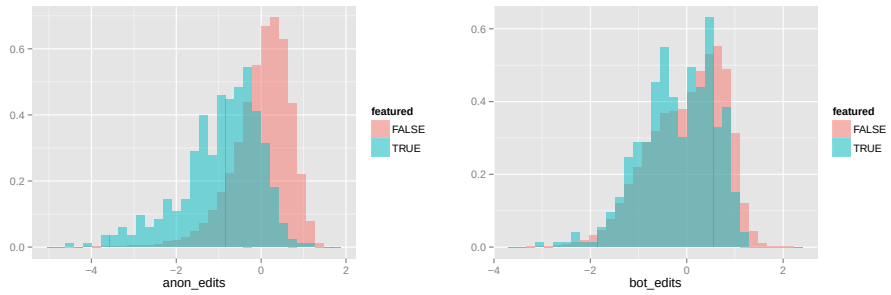

Fig. 5. Histograms of the total number of anonymous and bot edits conditioned on the total amount of work spent on an article. Note that the x axis represents the residuals from projecting the plotted variables onto `edits` and `team_id_1`.

One can observe that the general phenomenon that the more anonymous edits the worse quality of the article. It might be interpreted on a psychological background as follows. When a user makes an edit under an overt id she engages more and tries to do their best to make the edit of the best possible quality since it can be publicly assessed by the community and influences the overall user reputation. Anononymous edit, instead, does not put such a responsibility on the editing user.

The right-hand side of the figure shows the analogous histogram for the related `bot_edits` attribute which does not show such predictive power (AUC of 0.57). This confirms that anonymous edits are in a way special.

The number of anonymous edits may be partly related to the team itself, but we do not have any measures that would isolate such relationship from aspects such as how controversial a given topic is. We have thus decided to remove the attribute (and the `bot_edits` as well) by conditioning on it in the partial regression approach.

3.4 Variables Directly Related to the Team Performance

After conditioning on the numbers of anonymous and bot edits, the AUC of individual variables has decreased further, the median being equal to 0.540. Most variables are no longer predictive, which confirms that their correlations with the class variables were spurious and resulted mainly from being correlated with other, more predictive variables.

The five most predictive variables are `disc_nscc` (AUC=0.687), `disc_nidc` (AUC=0.681), `disc_nwcc` (AUC=0.671), `disc_nodc` (AUC=0.671) and `disc_avgv` (AUC=0.640). Figure 6 shows the histograms and ROC curves for those attributes. The plots are of course made *after* removing the influence of the total number of edits, team id, and the total numbers of anonymous and bot edits. Recall that when the AUC is less than 0.5 a better predictor can be obtained by reversing the scores. The first and third curves are thus reversed (this is the case since a large number of connected components means a weakly connected network).

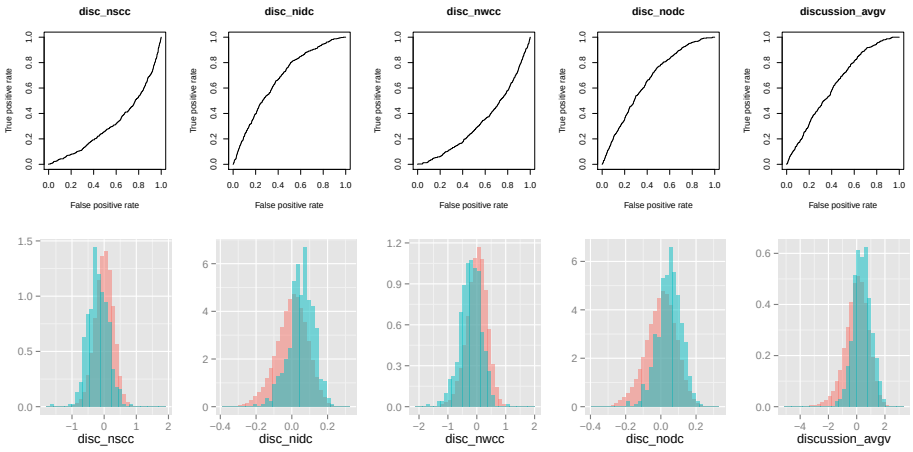

Fig. 6. Histograms and ROC curves for variables describing the discussion dimension of the Multidimensional Social Network. Note that the x axis of the histograms represents the residuals from projecting the plotted variables onto the removed variables.

It can be seen that the discussion dimension is very important. The average (thus corrected for the team size) number of discussions is highly, positively, correlated with article quality.

Unfortunately all those variables are correlated (minimum correlation between them is 0.46) thus it may be difficult to isolate the influence of specific, network-related, aspects of the discussion dimension. This is confirmed after removing the influence of the `disc_nscc` attribute in the next section.

3.5 Further Analysis

To see if any other dimensions influence article quality we added the `disc_nscc` attribute to the list of attributes on which we project in our partial regression procedure. All discussion-related variables lost their prominence and the most predictive variables were those related to the reverts dimension.

The most predictive attribute with AUC of 0.634 was `reverts_avgv` with negative values promoting better articles. It may be interpreted that if there is much mutual coordination (little reverts) between the authors then the resulting article is good.

After removing the influence of `reverts_avgv` the various attributes related to discussion, co-edits and reverts dimensions were the top predictors, but the picture became less clear and their AUCs were only about 0.6. We have thus decided that we have reached the limits of the current methods and stopped the exploratory analysis here.

4 Prediction

In this section we attempt to predict team quality based on all available variables after, however, removing the influence of the total number of edits, team id and the percentage of anonymous and bot edits. To this end we replace the original dataset with the dataset used in partial regression with controlling for those four attributes. Afterwards, several machine-learning models are built on the dataset and their predictive power is examined.

To accurately assess model peformance while taking overfitting into account we have split the data into two separate training and test sets of equal sizes. The models are built on the training set and evaluated on the test set.

We have used a representative selection of machine learning methods.

In particular, we have used the logistic regression model and the `CART` [5] decision tree learner implemented in the `R` statistical package in the `rpart` package. When building decision trees records from the more frequent class (not-featured) have been assigned lower weights to balance the distribution of the classes. Additionally, we used three machine learning methods from the `Weka` program: boosted decision trees (the `AdaBoost.M1` algorithm), Random Forest, and the naive Bayesian classifier with kernel estimation for numerical attributes. We used 50 iterations of the `AdaBoost.M1` and Random Forest algorithms.

Figure 7 shows the ROC curves for all those models. It can be seen that AdaBoost turned out to be the best model for the task. The area under the ROC curve for this model was 0.7965, dramatically higher than for any single attribute. Recall that in Section 3.4 the most predictive attribute `disc_nscc` achieved the AUC of only 0.687.

The results demonstrate that using different attributes as inputs to a predictive model can be a highly successful technique for team assessment. Recall that we have used partial regression techniques to remove the influence of attributes related mainly to the article and not the team itself (e.g. topic popularity).

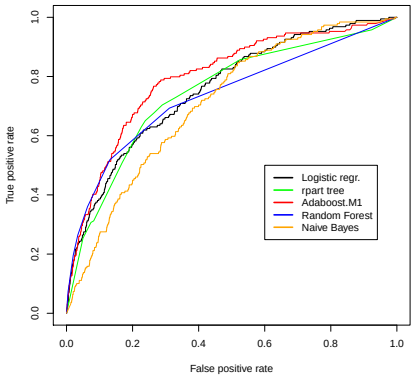

Fig. 7. ROC curves for models predicting the quality of teams of Wikipedia editors.

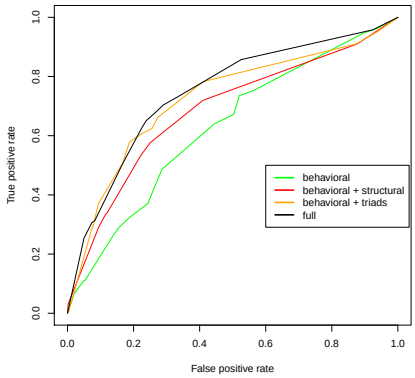

Fig. 8. ROC curves for CART decission tree models build on various subsets of variables.

4.1 Predictive power of various groups of attributes

Next we proceed to assessing predictive power of various types of variables in order to determine which factors are most useful in predicting team qaulity. To this end we have build predictive models on four subsets of variables and compared their performance. We began by using the AdaBoost model here as, in the previous experiment, it offered the best peformance.

The attributes were divided into three sets: behavioural (Section 2.1), structural excluding triads (Section 2.2) and triads (Section 2.3).

All models reported in this section included the behavioural attributes being the basic descriptors of features from which other attributes are derived. The Area Under the ROC Curve of an AdaBoost model containing only behavioural attributes was 0.6522, i.e. much lower than for the full model (0.7965). Next we added the structural attributes achieving AUC of 0.7254. A model using behavioural attributes and triads reached the AUC of 0.77 still lower than the full model. It is thus clear that network structure variables offer some potential predictive power beyond that offered by the behavioural variables introduced in [15] and using a combination of different types of team features in a predictive model is highly beneficial.

Next we have repeated the analysis for CART decision trees obtaining similar results. The basic behavioural variables achieved only AUC of 0.6273. Adding non-triad based structure variables increased the AUC to 0.6847. The model based on the basic behavioural variables and triads achieved an AUC of 0.7239. The full model including both types of attributes had AUC 0.7463. Figure 8 shows the ROC curves for the four models.

5 Conclusions

Reported results give some initial insights into the process of preparing a good Wikipedia article. It is demonstrated that applicaton of some appropriate statistical techniques helps to reveal some less obvious phenomena. Also, both behavioural and structural attributes seem to contribute to the resulting teamwork quality. The main contribution of this article is to demonstrate that while team size and number of edits have the strongest impact on Wikipedia article quality, it is possible to consider other, team related features that are also significant. We demonstrate this significance by using statistical techniques that allow to evaluate these weaker features in the presence of stronger ones. We find that attributes derived from the behavioral social network based on article discussion (talk pages) have a significant impact on article quality. In open teams, it is not always possible to increase the number of actively participating team members, or to increase the level of activity of existing members. Wikipedia editors are self-motivated. Yet, sometimes it may be possible to better organize the contributions of existing team members. This is the main future direction of our research.

References

1. Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.
2. Piotr Borzymek and Marcin Sydow. Trust and distrust prediction in social network with combined graphical and review-based attributes. In Piotr Jedrzejowicz, Ngoc Thanh Nguyen, Robert J. Howlett, and Lakhmi C. Jain, editors, *KES-AMSTA (1)*, volume 6070 of *Lecture Notes in Computer Science*, pages 122–131. Springer, 2010.
3. Piotr Borzymek, Marcin Sydow, and Adam Wierzbicki. Enriching trust prediction model in social network with user rating similarity. In Katarzyna Wegrzyn-Wolska Ajith Abraham, Vaclav Snasel, editor, *Proceedings of the 1st International Conference on Computational Aspects of Social Networks (CASoN 2009)*, pages 40–47, Los Alamitos, NY, USA, 2009. IEEE Computer Society.
4. Ulrik Brandes and Jürgen Lerner. Visual analysis of controversy in user-generated encyclopedias*. *Information Visualization*, 7(1):34–48, 2008.
5. L. Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and regression trees*. 1984.
6. T. Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.
7. J.Fox and S.Weisberg. *An R Companion to Applied Regression*. Sage, 2011.
8. Przemysaw Kazienko, Katarzyna Musial, Elbieta Kukla, Tomasz Kajdanowicz, and Piotr Brdka. Multidimensional social network: model and analysis. pages 378–387. ICCCI'11 Proceedings of the Third international conference on Computational collective intelligence: technologies and applications - Volume Part I, ICCI, 2011.
9. Aniket Kittur and Robert E Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46. ACM, 2008.

10. Aniket Kittur and Robert E. Kraut. Beyond wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 215–224, NY, USA, 2010. ACM.
11. Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: Conflict and coordination in wikipedia. pages 453–462, New York,, 2007. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI 2007, ACM.
12. Minh-Tam Le, Hoang-Vu Dang, Ee-Peng Lim, and Anwitaman Datta. Wikinetviz: Visualizing friends and adversaries in implicit social networks. In *Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on*, pages 52–57. IEEE, 2008.
13. Jakub Piskorski, Marcin Sydow, and Dawid Weiss. Exploring linguistic features for web spam detection: a preliminary study. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 25–28, New York, NY, USA, 2008. ACM.
14. C.R. Rao and H. Toutenburg. *Linear Models and Generalizations: Least Squares and Alternatives*. Springer, 2007.
15. Piotr Turek, Adam Wierzbicki, Radosaw Nielek, Albert Hupa, and Anwitaman Datta. Learning about the quality of teamwork from wikiteams. pages 17–24, Minneapolis,, 2010. Proceedings of the 2010 IEEE second international conference on social computing, SocialCom/IEEE international conference on privacy, security, risk and trust, PASSAT 2010.
16. Piotr Turek, Adam Wierzbicki, Radosaw Nielek, Albert Hupa, and Anwitaman Datta. Wikiteams: How do they achieve success? *IEEE Potentials 30(5)*, pages 2–7, 2011.
17. Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.