Understanding the Signature of Controversial Wikipedia Articles through Motifs in Editor Revision Networks

James R. Ashford  
AshfordJR@cardiff.ac.uk  
School of Computer Science and Informatics, Cardiff University, UK

Liam D. Turner  
TurnerLD9@cardiff.ac.uk  
School of Computer Science and Informatics, Cardiff University, UK

Roger M. Whitaker  
WhitakerRM@cardiff.ac.uk  
School of Computer Science and Informatics, Cardiff University, UK

Alun Preece  
PreeceAD@cardiff.ac.uk  
Crime and Security Research Institute, Cardiff University, UK

Diane Felmlee  
Dhf12@psu.edu  
Dept. of Sociology Criminology, Pennsylvania State University, USA

Don Towsley  
Towsley@cs.umass.edu  
College of Information and Computer Sciences, University of Massachusetts Amherst, USA

ABSTRACT

Wikipedia serves as a good example of how editors collaborate to form and maintain an article. The relationship between editors, derived from their sequence of editing activity, results in a directed network structure called the revision network, that potentially holds valuable insights into editing activity. In this paper we create revision networks to assess differences between controversial and non-controversial articles, as labelled by Wikipedia. Originating from complex networks, we apply motif analysis, which determines the under or over-representation of induced sub-structures, in this case triads of editors. We analyse 21,631 Wikipedia articles in this way, and use principal component analysis to consider the relationship between their motif subgraph ratio profiles. Results show that a small number of induced triads play an important role in characterising relationships between editors, with controversial articles having a tendency to cluster. This provides useful insight into editing behaviour and interaction capturing counter-narratives, without recourse to semantic analysis. It also provides a potentially useful feature for future prediction of controversial Wikipedia articles.

KEYWORDS

Network Motifs, Wikipedia, Controversy, Complex Networks

ACM Reference Format:


1 INTRODUCTION

Wikipedia has become a tremendous platform for crowdsourcing knowledge, representing a cornerstone of the World Wide Web [7]. It allows the “wisdom of the crowd” [24] to potentially emerge, providing intelligence on a vast range of topics [4]. However, complex dynamics support the emergence of content, since the formation of Wikipedia articles involves both human cooperation and human conflict, based on the extent of convergent and divergent views. Narrative and counter-narrative frequently jostle for presence in an article, representing a source of friction that is seen through editor interaction [22] and in the semantics of article content [20]. Wikipedia conveniently provides a list of controversial content that are labelled by the Wikipedia community themselves1.

In an age of misinformation [1, 17, 26], understanding characteristics of controversial articles has increased in importance. Because of the controversial nature of some topics, the narrative in a Wikipedia article may contain misleading information that stops a neutral consensus emerging. Prior work in this area has established insights such as the predictability of controversy from editor behaviour [20], such as deletions, reversions, and statistics from the collaboration network, prediction of article quality taking insights

from multiple models [29], and interactions between users, bots, admin and pages [9]. There has also been a number of different types of network developed to assess Wikipedia articles, including collaboration networks [6] that capture the positive or negative relationship between editors, edit networks that capture “undoing” of edits by a third party [12] and affiliation networks [10].

Our focus in this paper is to further understand the relationship between small groups of editors, as induced by their editing sequences, by using a revision network. This does not require information on the nature of the editing undertaken - it simply captures the ordering in which editing occurs and is therefore a simple metric to infer. Editors are represented by nodes and a directed edge from node A to B indicates that “Editor A edits the article after Editor B” (see Figure 1). From this, we seek to determine the extent to which controversial articles exhibit a distinctive signature relative to those that are deemed non-controversial.

There has already been some consideration of revision networks in the literature [8, 11], where more recently the emphasis [27] has been to combine them with other network representations. However, given the fundamental nature of revision networks, it is interesting to question the extent to which they hold sufficient information to characterise controversial Wikipedia articles. Currently this is not well-understood, and motivates our work.

**Hypotheses**

We hypothesise that interaction differences between small subgroups of Wikipedia editors is sufficient to distinguish between controversial articles and non-controversial articles. To address this we consider the extent to which revision networks of Wikipedia articles have different local induced substructures based on their having controversial classifications. Our approach is based on techniques from complex networks [15, 16], that have been successful in classifying diverse and complex biological networks based on their latent induced subgraphs.

To achieve this at scale, and in contrast to previous literature [9, 20, 22], we assess a relatively large sample of Wikipedia articles, involving over 21,000 Wikipedia articles, by determining their subgraph ratio profiles. Each such profile represents the under and over representation of induced triads in the revision network of a Wikipedia article using 13 dimensions of connected triads, while also normalising for differences in network size.

Understanding the relationship between Wikipedia articles is important, as it allows us to determine the extent to which they may cluster. The sample we are considering represents 21,631 articles each represented in 13 dimensions. Therefore we perform dimensionality reduction, and project the subgraph ratio profiles into lower dimensional spaces. This allows us to examine the relationship both within and between controversial and non-controversial articles.

The results draw attention to distinctive clustering patterns concerning controversial Wikipedia articles. We further assess the results by examining the correlation with a range of variables, allowing to understand the role of substructures. The findings reaffirm that the sequence of editing provides an important mechanism to understand Wikipedia articles, independent of an article’s topic and without recourse to semantic analysis.

## 2 RELATED WORK

Understanding the content of crowd-sourced platforms such as Wikipedia and the behaviour of their contributors is of wide research interest [24]. Wikipedia represents a dynamic network of articles with a structure resembling that of the World Wide Web [30], whereby dominant articles act as connectivity hubs. Dynamics also exist within the formation and maintenance of individual Wikipedia articles, through open and collaborative editing.

Interactions between editors range from positive to negative, where debates and arguments lead to different patterns of revision (e.g., [14, 20]), capturing behaviours such as vandalism [19] and the propagation of disinformation [13]. Characterising articles and contributors through revisions provides a means for Wikipedia to manage and review its content. This is potentially labour intensive and has led to interest in creating and exploiting methods to detect issues (e.g., [2, 23]).

Controversial articles have become an increasing point of focus, and characterised as such by Wikipedia. Automated methods for classifying articles have received much attention
Wikipedia’s web API

We collected revision history logs and article meta-data (e.g., [21]). This has ranged from global features such as the basis for comparison. These were taken from an original sample of 23,000 articles (20,000 non-controversial and 3,000 controversial), from which articles were removed if they did not contain sufficient information for motif analysis.

Network Construction

For each article, we construct a revision network (in the same manner as [11, 27]) where nodes represent unique editors and directed edges show that an editor added a revision after another editor. We traverse the revision log list to build a network that spans the article’s lifetime, adding nodes and edges as they appear in each event. Specifically let the revision network of a Wikipedia article be defined by $G = (V, E)$, where each editor is represented by a node $v \in V$. An edge $(v_i, v_j) \in E$ indicates that editor $v_i$ edits the article after editor $v_j$. This excludes self-loops and editor $v_i$ editing after editor $v_j$ multiple times does not result in a multi-edge.

A simple example can be found in Figure 1, which describes how the network is constructed. Two extreme examples from the dataset shown in Figures 3a and 3b. This form of representation is potentially useful as large articles do not typically follow a linear or incremental structure. For example, it is highly likely that users will restore work back to an earlier revision should a revision become vandalised or irrelevant. Editors are likely to refer back to previous editor’s work. These behaviours result in complex sequential patterns that are captured through revision networks.

Network Motif Analysis

Network motif analysis focuses on determining the under or over representation of induced subgraphs [5, 16], as compared to an alternative sample of graphs (i.e., a null-model that acts a relevant baseline for comparison). Dyads, triads and tetrads are often considered as induced subgraphs such over and under-represented induced subgraphs are called network motifs or anti-motifs. In this analysis, we examine each article’s revision network using triads, representing how all possible triples of editors may sequentially interact. Triads are sufficiently large enough to capture both direct and indirect reciprocity between editors, while not being of a scale that is impeded by combinatorial complexity - there are 13 possible connected triads, shown in Figure 2. The coded names listed in Figure 2 are provided as part of the convention used in the triad census algorithm [3].

![Figure 2: 13 possible combinations of connected triads in directed networks.](image-url)
For each article, we calculate the subgraph ratio profile (SRP) as defined by Milo et al [15]. This accounts for variations in network size. This is achieved by determining the relative abundance of each type of triad compared to random graphs generated by the null-model. For each type of triad $i$, we firstly calculate $\Delta_i$:

$$\Delta_i = \left( \frac{N_{real,i} - \langle N_{rand,i} \rangle}{N_{real,i} + \langle N_{rand,i} \rangle} \right) + \epsilon$$  \hspace{1cm} (1)$$

where $N_{real,i}$ is the number of such triads observed in the graph under observation, and $N_{rand,i}$ is the average frequency of such triads seen across the sample of networks composed by the null model. In this case the null model uses 100 random graphs with the same number of nodes and edges as the graph under observation. The value of $\epsilon$ is set to 4 [15, 25] to prevent the result from being misleadingly large when a subgraph rarely appears. This process is repeated for each triad $i$ and normalised across triads to form the subgraph ratio profile (SRP) for a given network. The $i^{th}$ SRP, denoted $SRP_i$, denotes the extent of under or over representation of the triad $i$, and is defined by:

$$SRP_i = \Delta_i / (\sum \Delta_i^2)^{1/2}.$$

Article comparison and dimensionality reduction

The SRP composed for each article provides a 13-dimensional vector whose components indicate the extent of triad representation relative to networks in the null model. To assess these collectively, we use principal component analysis (PCA) that allows the SRPs to be considered in a lower dimensional space. We apply three and two dimensional PCA. Finally, we make comparison with a number of external variables (number of editors/nodes, age of article, and revision rate) to understand potential correlations with motifs.

4 RESULTS

Subgraph Ratio Profiles

We first examine the SRPs that arise from both controversial and non-controversial articles in isolation (Figures 4 and 5 respectively). To find significant triads, we use a cutoff value of +0.3 and -0.3. We determine that controversial articles are strongly represented by triads $111D$, $111U$ and $201$, which attain average SRP scores of 0.382, 0.375, 0.372, with relatively low dispersion (SDs of 0.136, 0.149 and 0.124 respectively). Interestingly these represent a chain of three nodes, where one edge is reciprocated, with the other edge covering all possible directional types (i.e., reciprocated, directed in, directed out).

In contrast, the results for the non-controversial articles provide a different profile. Here $021D$ and $021U$ are significantly under represented (average SRP scores of -0.511, -0.485), albeit with higher standard deviations present (SDs of 0.192, 0.2). Interestingly these anti-motifs ($021D$ and $021U$) relate to a lack of subgraphs where directed edges either emanate from or are received by a single node in the triad. Such configurations relate to the role of a mediating editor...
that may be presented with or respond to the editing of others. In other words, such mediators have a reduced role in non-controversial articles.

The comparison between these subgraph ratio profiles is shown in Figure 6. These profiles are quite distinct. We also calculate the Pearson correlation coefficient for each distinct pair of articles in three groups - controversial articles, non-controversial articles and all articles. Controversial articles provide the greatest correlation to each other (M=0.41193, SD=0.41262). Non-controversial articles have a lower mean correlation (M=0.37498, SD=0.36749) which is similar to the result when considering all articles together (M=0.37569, SD=0.36844).

Principal Component Analysis

We analyse the 21,631 subgraph ratio profiles to determine the relationship in terms of relative clustering. We apply PCA in order to reduce the 13 dimensions of the SRPs down to a more manageable form. We initially project the SRPs into 3-dimensional space for clarity, as seen in Figure 7. This presents a distinctive region where controversial articles are dominant. This provides evidence for a distinction between the controversial and non-controversial articles, consistent with the variation in motifs identified in the previous section.

The PCA coefficients (Table 1) that define the three dimensions reveals that principal component one (x axis in Figure 7) primarily depends on triads 111D and 111U. Principal component two (y axis in Figure 7) primarily depends on triad 021C. The third principal component primarily depends on triads 021C and 030C. However, we also note that when represented in the three dimensional space, the revision networks have limited dispersion in the third dimension (i.e., vertical dimension as plotted).

Calculating the percentage of explained variance by principal component confirms that the first principal component produces 53.7% of the shared variance, the second produces 22.3% and the third produces the least with 6.7%. This confirms that the third principal component provides a limited contribution to representation of the total variance across the significant ratio profiles. This supports representation through two principal components, as plotted in Figure 8, with the relative composition of each principal component being near identical to PC-1 and PC-2 in Table 1. As anticipated, this is similar dependency on the first and second primary components in the three dimensional representation.

Representation in two dimensions further clarifies the distinction between controversial and non-controversial revision networks. In particular, from Figure 8 we note that both classes of article exhibit a similar maximum and minimum range against principal component two, which is primarily defined by the linear path between three nodes (021C). However, it is the variation in the first principal component, dominated by 111D and 111U, which represent linear paths with reciprocation on one edge, that differentiate the non-controversial from controversial. High values in principal component one correlate with controversial articles - in other
Table 1: PCA coefficients displaying the strongest triads

<table>
<thead>
<tr>
<th>PC-1</th>
<th>PC-2</th>
<th>PC-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>021D</td>
<td>0.332</td>
<td>0.427</td>
</tr>
<tr>
<td>021U</td>
<td>0.321</td>
<td>0.438</td>
</tr>
<tr>
<td>021C</td>
<td>-0.126</td>
<td>0.643</td>
</tr>
<tr>
<td>111D</td>
<td>0.509</td>
<td>-0.278</td>
</tr>
<tr>
<td>111U</td>
<td>0.522</td>
<td>-0.242</td>
</tr>
<tr>
<td>030T</td>
<td>0.145</td>
<td>0.207</td>
</tr>
<tr>
<td>030C</td>
<td>0.205</td>
<td>-0.242</td>
</tr>
<tr>
<td>201</td>
<td>0.318</td>
<td>0.104</td>
</tr>
<tr>
<td>120D</td>
<td>0.078</td>
<td>-0.052</td>
</tr>
<tr>
<td>120U</td>
<td>0.079</td>
<td>0.059</td>
</tr>
<tr>
<td>120C</td>
<td>0.204</td>
<td>0.060</td>
</tr>
<tr>
<td>210</td>
<td>0.117</td>
<td>0.047</td>
</tr>
<tr>
<td>300</td>
<td>0.065</td>
<td>0.074</td>
</tr>
<tr>
<td>021D</td>
<td>0.145</td>
<td>0.074</td>
</tr>
<tr>
<td>021U</td>
<td>0.205</td>
<td>0.047</td>
</tr>
<tr>
<td>021C</td>
<td>0.318</td>
<td>0.074</td>
</tr>
<tr>
<td>111D</td>
<td>0.509</td>
<td>0.060</td>
</tr>
<tr>
<td>111U</td>
<td>0.522</td>
<td>0.047</td>
</tr>
<tr>
<td>030T</td>
<td>0.145</td>
<td>0.117</td>
</tr>
<tr>
<td>030C</td>
<td>0.205</td>
<td>0.074</td>
</tr>
<tr>
<td>201</td>
<td>0.318</td>
<td>0.060</td>
</tr>
<tr>
<td>120D</td>
<td>0.078</td>
<td>0.047</td>
</tr>
<tr>
<td>120U</td>
<td>0.079</td>
<td>0.074</td>
</tr>
<tr>
<td>120C</td>
<td>0.204</td>
<td>0.060</td>
</tr>
<tr>
<td>210</td>
<td>0.117</td>
<td>0.074</td>
</tr>
<tr>
<td>300</td>
<td>0.065</td>
<td>0.074</td>
</tr>
</tbody>
</table>

Additional Metrics

We examine the relationship between revision networks and primary external variables (number of editors/nodes, age of article, and revision rate) using motifs. Specifically, using the dimensions of two-dimensional PCA analysis, we examine the correlation with the external variables, and how this differs between controversial and non-controversial articles. The results are shown in Figures 9, 10 and 11.

The greatest differences between controversial and non-controversial articles occur with respect to article age (Figure 10). Here, controversial articles with high age cluster against high values of principal component one, and to some extent this occurs for principal component two. This contrasts against the clustering seen for non-controversial articles.

Figure 10 also shows that controversial articles have a tendency to be older. The high values in principal component one and two which align with dense clustering of controversial articles show that while such articles accumulate the linear revision path between authors (021C which dominates principal component two), controversial articles also accumulate instances of linear paths where one edge is reciprocated (i.e., 111D and 111U which dominate principal component one).

5 DISCUSSION

Motif analysis of revision networks gives insight into how the temporal editing relationship between small groups of
Wikipedia authors create signatures that allow controversial articles to be distinguished. In contrast to previous work, we have investigated this using a relatively large sample of Wikipedia articles, where distinct patterns emerge. This provides strong support for our hypothesis, and reaffirms the importance of the revision network as a simple but fundamental element in editing Wikipedia.

Through motif analysis, we identify that reciprocation on linear paths among triads of editors in the revision network is over represented in controversial Wikipedia articles. These motifs are defined by the triads 111D, 111U and 201. In contrast, the revision networks from non-controversial articles exhibit two anti-motifs, involving the under representation of triads involving two directed edges either arriving at or emanating from a mediating node (triads 021D and 021U). These motifs and the underlying subgraph ratio profiles represent an unusual and distinctive profile that we believe represent distinctive "super-families" beyond those seen in other technologically related networks, such as the world wide web [30].

Performing dimensionality reduction upon the subgraph ratio profiles from each revision network allows us to further understand the relationships between Wikipedia articles. Our analysis shows that the structure of the data is amenable to reduction to two dimensions, where the principal components are dominated by triads 111D and 111U in the first component, and mainly 021C in the second component, but with lesser contributions from triads 021D and 021U.

The results from two-dimensional principal analysis are illuminating - the dominant triads in both components, as listed above, correspond to linear paths, i.e., open triads which represent sequences of editing without indirect reciprocity. The extent and format of reciprocated (i.e., bidirectional) edges on these open triads is sufficient to define the two principal components. The dominant triads in the first principal component each involve reciprocation on one edge, where as interestingly, in the second principal component, the dominant triads are open triads with no reciprocated edges. From this we deduce that short paths, rather than short loops of editing that represent indirect reciprocity, are important features in characterising Wikipedia revision networks.

We also observe through two-dimensional principal component analysis that it is the first principal component that strongly distinguishes between the revision networks of controversial and non-controversial articles. The dominant triads defining this capture the extent of direct reciprocation being present. Finally, through consideration of the principal components against additional external variables, we find in particular that article age plays a role in distinguishing the controversial articles. High values in both principal components aligns with strongest clustering of controversial articles, which is not the case for non-controversial articles.

6 CONCLUSION

The analysis has given insights into the structure underlying revision networks from Wikipedia articles, and has shown that a relatively small number of features, in terms of substructures in revision networks, characterise controversial Wikipedia articles. The results have identified key clusters of editorial interactions to this effect, in support of the hypothesis. These are distinctive and indicate that the revision networks for controversial and non-controversial Wikipedia articles have differentiated subgraph ratio profiles. Our study gives understanding as to how prediction or classification of articles can be enhanced using the latent structures relating to editor behaviour. This also reaffirms the importance of the revision network as a simple but useful representation for assessment of Wikipedia articles.

ACKNOWLEDGEMENTS

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.
REFERENCES


