Detecting Controversial Articles in Wikipedia

Joy Lind
Department of Mathematics
University of Sioux Falls
Sioux Falls, SD 57105

Darren A. Narayan
School of Mathematical Sciences
Rochester Institute of Technology
Rochester, NY 14623-5604

Abstract

In this paper, we apply graphical models to facilitate quantitative and qualitative investigations into the edit history of articles posted on Wikipedia. Quantitatively, we use the models to measure controversy arising from Wikipedia articles. Qualitatively, we use the models to provide insights into the distribution of editor roles associated with articles. The paper includes exercises that can be integrated into an undergraduate course in graph theory or discrete mathematics.

Wikipedia (wikipedia.org) is a website that contains content similar to what might be found in a traditional encyclopedia. However, unlike a traditional encyclopedia, Wikipedia's content is provided by millions of contributors, called *editors* [10]. This broad spectrum of contributions leads to questions regarding the correctness of the content. While the correctness of traditional encyclopedias is verified by many experts before they are printed, Wikipedia boasts that it is "self-correcting", meaning that over time, articles improve from a multitude of contributions. In many cases, this does indeed happen. However, sometimes these multiple contributions are contradictory in nature and spiral into an ongoing controversy known as a "Wikipedia edit war"[3],[4],[5]. An edit war "occurs when editors who disagree about the content of a page repeatedly override each other's contributions, rather than trying to resolve the disagreement by discussion"

[3]. Editors can add new content or delete existing content. (Replacing content is viewed as deleting then adding.) Inherent to the collaborative approach to writing Wikipedia articles is the possibility of disagreement between editors who wish to contribute to the same article. For each article, Wikipedia provides the reader the capability to view the article's history. Some histories reveal instances of editors having reverted the work of other editors [10]; one can see from the histories that certain articles are more controversial than others.

Conflict in Wikipedia has generated a significant amount of interest and consequently literature on the topic. In [8], Kittur et al. quantify the distribution of conflict among Wikipedia articles in 10 different topic areas, finding that the topics of Religion and Philosophy are most contentious. Several articles, including [13], explore the relationship between conflict and language complexity; in [13], Yasseri et al. conclude that conflict or controversy has the effect of reducing language complexity. Many articles focus on different ways of measuring controversiality in Wikipedia. One simplistic approach is to count the number of "mutual reverts," that is instances of pairs of editors having reverted each other in the article. In [12], Yasseri and Kertész develop a more sophisticated model based on mutual reverts. Another approach is the *Contributor Count Model,* which uses the number of editors of an article as a measure of its controversy [11]. In [7], Kittur et al. develop an automated way of identifying the properties that make an article high in conflict using "machine learning techniques and simple, efficiently computable metrics." Their list of metrics includes the number of revisions, number of anonymous edits, and number of reverts. In [9], Rad and Barbosa compare five different methods for modelling and identifying controversy in Wikipedia articles. One method in scope of their review is the *Basic Model*, that we examine in detail in the next section of this paper.

Measuring Controversy in Articles

Recall that the Contributor Count Model uses the number of editors of an article as a measure of its controversy. In this section, we focus on an improvement to the Contributor Count Model that was developed by Vuong et al. (see [11]); this is known in the literature as the *Basic Model.* This model is based on bipartite graphs and is quite tractable for students with only minimal exposure to graph theory.

First, we review some elementary terminology from graph theory. A *graph* is composed of a set of *vertices* and *edges*, where each edge connects two vertices. A *bipartite graph* is a graph whose vertex set can be partitioned into two subsets so that no edge of the graph extends between vertices of the same subset. A *directed* graph is a graph where each edge has a direction. A *weighted* graph is a graph where each edge is assigned a numeric weight. The Basic Model for measuring article controversiality is based on information from two directed weighted bipartite graphs; here, we use the notation of [11]. The first graph we construct is the *contribution graph*. For this graph, the vertex set consists of a set of editors $\{u_i\}$ and a set of articles $\{a_k\}$. A directed edge with weight o_{ik} exists from vertex u_i to vertex a_k if editor u_i contributed o_{ik} words to article a_k in that article's history.

Exercise 1: Given the information in Table 1, students could be asked to construct the associated contribution graph. Entry in row u_i, column a_k is o_{ik}, the number of words editor u_i contributed to article a_k in that article's history.

	a_1	a_2	a_3	a_4
u_1	412	56	314	0
u_2	619	37	117	0
u_3	0	0	603	152

Table 1.

Solution:

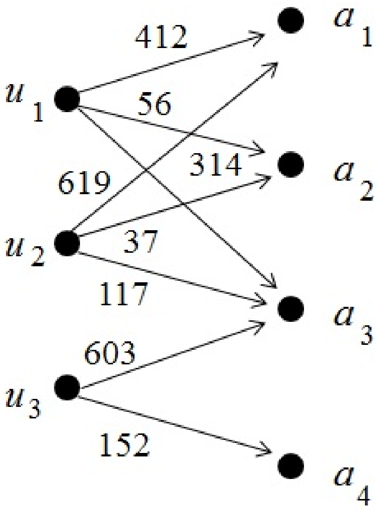

Figure 1. A contribution graph

The second relevant graph is a *dispute graph.* We say that a *dispute* exists between a pair of Wikipedia editors of a particular article if at least one of the editors has deleted words from the other according to the article's edit history. A dispute graph is constructed as a weighted directed bipartite graph as follows: the vertex set consists of a set of ordered pairs $\{(u_i, u_j)\}$ of editors and a set of articles $\{a_k\}$. A directed edge with weight w_{ijk} exists from vertex (u_i, u_j) to vertex a_k if there is a dispute between editors u_i and u_j in article a_k, and the number of editor u_j's words deleted by editor u_i in the complete edit history of article a_k is w_{ijk}.

Exercise 2: Given the information in Table 2, students could be asked to construct the associated dispute graph. Entry in row (u_i, u_j), column a_k is w_{ijk}, the number of editor u_j's words deleted by editor u_i in article a_k.

	a_1	a_2	a_3	a_4
(u_1, u_2)	221	0	0	0
(u_2, u_1)	103	0	0	0
(u_1, u_3)	0	0	26	0
(u_3, u_1)	0	0	0	0
(u_2, u_3)	0	0	17	0
(u_3, u_2)	0	0	0	0

Table 2.

Solution:

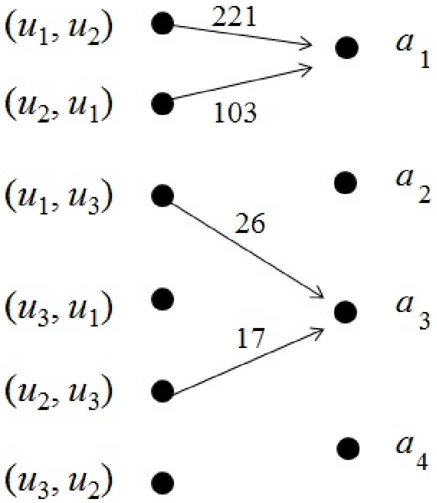

Figure 2. A dispute graph

Exercise 3: Which editor was the largest contributor, in that he/she contributed the most number of words to the set of articles?

Solution: The answer can be obtained from the contribution graph. For each editor (vertex) u_i, we sum the weights of the edges leaving u_i. Editor u_1 is the largest contributor, having contributed a total of 782 words.

Exercise 4: Which article had the most number of words contributed?

Solution: Again, the answer can be obtained from the contribution graph. For each article (vertex) a_k, we sum the weights of the edges entering a_k. Article a_3 had the most number of words contributed, namely 1034.

Exercise 5: Which article is the longest, that is, has the most number of words?

Solution: We include this question to highlight the difference between it and the question posed in Exercise 4; the key is that the weights in the contribution graph do not take into account the number of words deleted. For this exercise, we require information from both the contribution and dispute graphs. For article a_1, we see that editor u_1 contributed 412 words, but 103 were deleted, and editor u_2 contributed 619 words, but 221 were deleted; this means the length of article a_1 is $412 - 103 + 619 - 221 = 707$ words. Article a_2 had no deletions, and had contributions of $56 + 37 = 93$ words. We can similarly calculate the lengths of the remaining articles: a_3 has a length of $314 + 117 + 603 - 26 - 17 = 991$ words and article a_4 has a length of 152 words. Hence, article a_3 is longest.

Exercise 6: Does the data suggest any possible instances of edit wars, that is, pairs of editors who repeatedly reverted each other?

Solution: From the dispute graph, we see that editors u_1 and u_2 reverted each other in article a_1. This might suggest the possibility of an edit war between u_1 and u_2 with a_1 being a controversial article. However, note that the graph does not capture the number of distinct interactions between those editors. Perhaps each reverted the other only once, or perhaps the changes happened repeatedly, over a series of interactions. We would need to consult the article's edit history in Wikipedia for this additional level of detail.

Exercise 7: Recall that one simplistic model for detecting controversial articles is the Contributor Count Model, which uses the number of editors of an article as a measure of its controversy. According to this model, which article is most controversial?

Solution: We would conclude that article a_3 is the most controversial since it had three distinct editors.

At this point, students might be asked to consider whether a larger number of editors should necessarily imply that the article is more controversial. In fact, this leads to a significant drawback to the Contributor Count Model, namely that it may “wrongly classify heavily edited high quality articles to be controversial" [11].

The Basic Model improves upon the Contributor Count Model by taking into account the amount of dispute in each article. In the Basic Model, the controversy C_k of an article a_k is measured as:

$$C_k = \frac{\sum_{i,j} w_{ijk}}{\sum_i o_{ik}}$$

Recall that w_{ijk} is the number of words from editor u_j to article a_k that are deleted by editor u_i, and o_{ik} is the number of words that editor u_i contributed to article a_k in that article's history. Hence, the sum in the numerator can be considered a quantification of the amount of dispute on article a_k, which has been normalized by dividing by the total contribution to article a_k.

Exercise 8: For each of the four articles in the exercises above, calculate its controversy measure C_k. According to this measure, which article is most controversial?

Solution: $C_1 = 0.314, C_2 = 0, C_3 = 0.042$, and $C_4 = 0$. We therefore conclude that article a_1 is the most controversial. Note the contrast with the findings from the Contributor Count Model, where article a_3 was identified as the most controversial.

We clarify here that Vuong et al. have developed more sophisticated models than the Basic Model. They compare the effectiveness of several models, including Contributor Count and Basic (where unsurprisingly, their metrics show that the Basic Model outperforms Contributor Count). Rad and Barbosa also compare several models, including the Basic Model, in their paper. We refer the interested reader to [11] and [9] for details.

Further Explorations into Edit Histories

In all of the graphical models from the previous section, both articles and editors were represented by vertices. Alternatively, we can construct a graph that has as vertices only the editors, with edges used to encode how authors contribute to the articles, and how editors interact with each other while editing them.

One approach is to construct a colored, weighted directed graph as follows: label the vertices with the editors, and associate each article with a color. Construct an edge from vertex u to v, colored with color c, and assigned a weight d if editor u deletes d words of editor v's work on article c.

Exercise 9: Given the information in the table in Exercise 2, students could be asked to construct the associated graph, using this approach.

Solution:
Note that in place of colors, we use dashed and non-dashed edges to differentiate between articles.

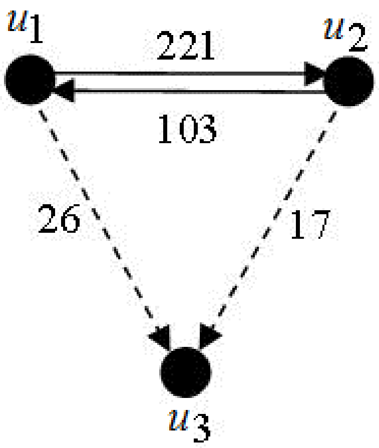

Figure 3. A graph showing interaction between editors

The above example illustrates a benefit of this method over the approach pursued in Exercise 2, namely that it reduces the number of vertices, which facilitates the modeling of larger problems.

In [2], Brandes et al. present a modified version of this model, that focuses on just one article (thereby reducing the complexity introduced by the multiple colors) but that captures different types of interactions among editors of the article. They call their model an *edit network*; below, we record the method from [2] for constructing an edit network as the tuple $G = (V, E, A)$:

1. The vertices V of the graph (V, E) correspond to the editors that have done at least one revision on the article.
2. The directed edges $E \subseteq V \times V$ of the graph (V, E) encode the interaction among editors. A particular pair of editors $(u, v) \in V \times V$ is in E, if u performed one of the following three actions with respect to v:

   (a) u deletes text that has been written by v;

   (b) u undeletes text that has been deleted by v (and written by a potentially different editor w);

   (c) u restores text that has been written by v (and deleted by a potentially different editor w).

   Since editors may revise text written by themselves, loops (i.e. edges connecting an editor with himself/herself) are allowed.

3. A is a set of attributes on edges, that encode how much text that editors add, delete, or restore. (We note that the article [2] includes a more extensive set of attributes than those we give here.) For each pair of editors $(u, v) \in V \times V$:

   (a) $\mathit{delete}(u, v)$ denotes the number of words deleted by u that had been written by v;

   (b) $\mathit{undelete}(u, v)$ denotes the number of words restored by u, that had been previously deleted by v (and written by a potentially different editor w);

   (c) $\mathit{restore}(u, v)$ denotes the number of words restored by u, written by v and deleted by a potentially different editor w.

Large values of the weights $\mathit{delete}(u, v)$ and $\mathit{undelete}(u, v)$ imply a negative relationship between editors u and v. The sum of these weights $\mathit{delete}(u, v) + \mathit{undelete}(u, v)$ can be interpreted as a measure of how strongly u disagrees with v. Conversely, a large value of $\mathit{restore}(u, v)$ suggests a positive relationship from u to v since u is defending v's contributions against deletions, and can therefore be interpreted as a measure of how strongly u

agrees with v [2].

Exercise 10: Students could be asked how they might use a model constructed in this way to address the following questions:

1. Who are the most active editors for a given article?
2. Are there editors who are particularly focused on deletion of content? Restoration of content?

These questions can be more readily answered by visually enhancing the graphical model to facilitate a qualitative exploration of an article's edit history. Brandes et al. suggest the following approach: make the size of the vertices proportional to editor activity; color the vertices to distinguish between editors that are mostly deleters (shaded darker) and those that mostly add or restore text (shaded brighter); and alter the vertex shapes so that flattened vertices mostly revise others and broad vertices mostly get revised. One example configuration is shown in the figure below. Note that this configuration exhibits a high degree of *bipolarity,* essentially decomposing into two groups that mutually undo the edits of the other group [2]. We can conclude from this that we have two groups of editors that have contradicting opinions on the article's subject.

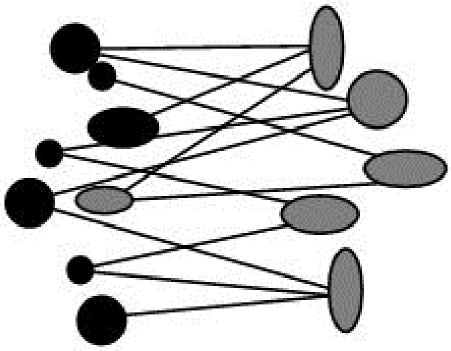

Figure 4. A graph for quantifying an article's edit history

If instructors are interested in implementing exercises that are similar in spirit to the ones above but that are based on more extensive data sets, we note that the Stanford Network Analysis Project (SNAP) provides large collections of data, including one of the complete Wikipedia edit history from its inception through January 2008. This data is available for free download from [6].

Acknowledgements:

This work is supported by a NSF-CCLI Grant #1019532. We would also like to thank an anonymous referee for their valuable comments.

References

[1] U. Brandes and J. Lerner, Visual Analysis of Controversy in User-Generated Encyclopedias, Information Visualization, 7:34-48, 2008.

[2] U. Brandes, J. Lerner, P. Kenis, and D. van Raaij, Network Analysis of Collaboration Structure in Wikipedia, 2009 World Wide Web Conference, 20-24 April 2009, pp. 731-740, Madrid, Spain.

[3] http://en.wikipedia.org/wiki/The_Three_Revert_Rule

[4] http://en.wikipedia.org/wiki/Wikipedia:Edit_warring

[5] http://en.wikipedia.org/wiki/File:Editwar.png

[6] http://snap.stanford.edu/data/index.html

[7] A. Kittur, B. Suh, B. Pendleton, E. Chi, He Says, She Says: Conflict and Coordination in Wikipedia, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 29 April 2007, pp. 453-462.

[8] A. Kittur, E. Chi, B. Suh, What's in Wikipedia?: Mapping Topics and Conflict Using Socially Annotated Category Structure, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 8 April 2009, Boston, MA, USA, pp. 1509-1512.

[9] H.S. Rad and D. Barbosa, Identifying controversial articles in Wikipedia: a comparative study, Proceedings of the 8th International Symposium on Wikis and Open Collaboration, WikiSym'12, August 2009, Linz, Austria.

[10] R. Sumi, T. Yasseri, A. Rung, A. Kornai, and J. Kertész, Edit wars in Wikipedia, 2011 IEEE Third International Conference on Social Computing (SocialCom), 9-11 Oct. 2011, pp. 724-727, Boston, MA.

[11] B. Vuong, E. Lim, A. Sun, M. Le, H. Lauw, and K. Chang, On Ranking Controversies in Wikipedia: Models and Evaluation, Proceedings of the International Conference on Web Search and Web Data Mining, WSDM '08, pp. 171-182, New York, NY, 2008.

[12] T. Yasseri and J. Kertész, Value Production in a Collaborative Environment: Sociophysical Studies of Wikipeda, Journal of Statistical Physics, 151: 414-439, 2013.

[13] T. Yasseri, A. Kornai, and J. Kertész, A Practical Approach to Language Complexity: A Wikipedia Case Study. PLoS ONE 7(11): e48386, November 2012.