Wikipedia Controversial Pages Replication Project

By: Poonam Varkhedi

03/16/2020

Replication of: "Edit Wars in Wikipedia" by Yasseri et al 2012a

Wikipedia has grown to become the largest online encyclopedia. It has a global reach with users, content creators, and editors around the world. It has pages on topics ranging from celebrities such as Britney Spears to scientific theories on geological rock formations. Wikipedia, as defined in the Wikipedia page on Wikipedia, is an "open collaboration project." It has a community of people who maintain its pages and content. This has fostered the development of what has become known as an "edit war." Pages on topics that are controversial see a high number of editors making rapid changes in succession of one another. Oftentimes these edits contradict each other due to the controversiality of the topic. This is known as an edit war.

This research paper aims to quantify the level of controversiality of a Wikipedia Page based on tracking the edits wars, as well as analyzing the trends of the controversiality of a page over time.

Data/ Data Generation Process:

The data used in this research process was the record of all edits made to English Wikipedia pages. These records were collected as "edit dumps" and can be found on the website: https://dumps.wikimedia.org/enwiki The data was last updated on 01/20/2020. The data from this site includes information on each edit made to a page. The relevant information includes: which page, by which editor, at what time, if it was a revision to an existing edit or if it is an original addition to the page, and what text was changed or added.

The process of cleaning and transforming this data into a usable format is modelled by the "light dump" data created by Yasseri and found on the WikiWarMonitor website. The raw data was parsed into a file with a line per edit. Each line contains the timestamp of the edit, the editors username, a flag if it is a reversion or just an addition to the page, and the version number of original edits. If it is a reversion to a previous edition of the page, it contains the version number of the edit it matches with. The light dump format is displayed below:

Topic Title
[timestamp, 0/1, version number, username]
[timestamp, 0/1, version number, username]
[timestamp, 0/1, version number, username]
...
Topic 2 Title
[timestamp, 0/1, version number, username]
[timestamp, 0/1, version number, username]
...

The raw data does not contain direct information on whether or not an edit is a reversion as opposed to an addition to a page. This was extracted by matching the text data listed in the raw data. If two users edit the exact same text then they are both considered to be on the same side of the edit "war." And the previous editor, who's edit is being overridden, is on the other side. This is how the flag is generated in the light dump data.

Historical context:

Wikipedia has been around for 20 years. In this time it has grown

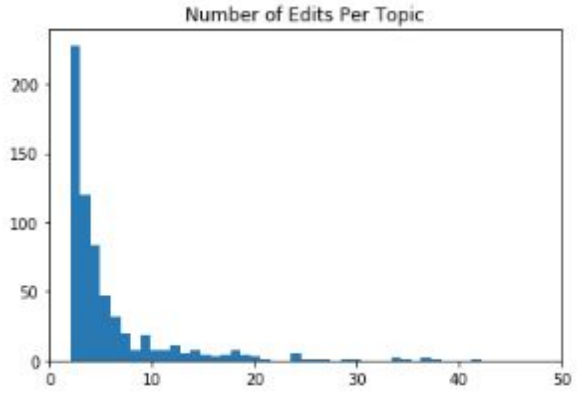
exponentially, amassing many new articles and even more edits. This means that the study done today on recent edit history data can be replicated on data going back 20 years. This can identify treads that are tracked through online footprints.

Exploratory Data Analysis:

The data used in this study can be classified as reliable. It is generated directly from Wikipedia and is simply a record of edits made.

The data cleaning process, as described in the previous section, was necessary. The process was not actually removing values but rather formatting the data into a way that can be more effectively used. This was necessary for two reasons. One, there had to be a way or recognizing mutual revert edits. And two, there had to be a cleaned form of the data that could be used to run analysis on.

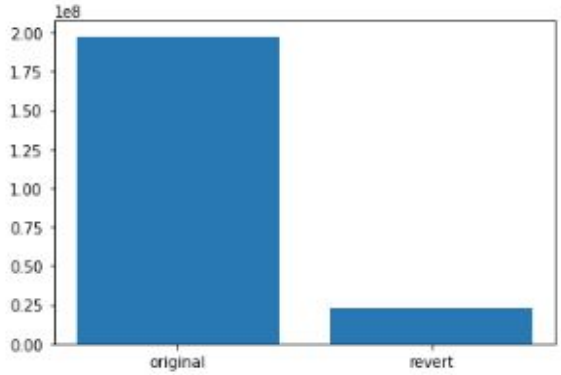
The following is a graph showing the number of edits per topic. This graph was generated using the light dump data.



Graph 1: Number of Edits Per Topic

The graph shows that the distribution of edits is highly skewed right. It appears as though many articles have few edits and it drastically drops as you get to a high number of edits per article. This makes sense if you think about the vast number of Wikipedia article topics. It is probably true that only a few of them get a heavy amount of editing. This could also mean that only a few topics are highly "controversial" if you use only the number of edits as an indication.

This next graph shows the distribution of edits that are additions to a page (original) versus a reversion (edit to an existing text on page):



Graph 2: Distribution of Edits

This shows that Wikipedia gets more new information than edits and that not many editors actually change other editors' contributions. This means that from the original dataset, the data that can be used in this research is a much smaller subset.

M-Statistic:

The "m-statistic" is a number that is calculated that is meant to quantify the controversiality of a Wikipedia page. The higher the m value the more controversial it is. The m-statistics was calculated only using the edit history data. In the explanation of the process a "revert" will be defined as an edit that was made that changes an existing version of the page back to what the editor had written before. This means that the editor of the revert does not agree with the editor that changed his or her original words. A "mutual revert" will be defined as two editors who have changed each others' edits back to their own opinion. In this manner the two editors have actively shown they disagree with each other.

The process of calculating the m-statistic of a Wikipedia page starts by collecting the number of edits a single editor makes throughout the edit history light dump data. This is a running list of all the editors that have made edits to a page and the number of times they have.

The next step is to collect all pairs of reverts. This means creating tuples of editors' username if they have made a revert edit. These tuples are (reverter, revertee). This means that the reverter does not agree with the revertee's edit and choose to change it back to an earlier version of the page.

The third thing that gets collected is the number of mutual reverts per page. This means revert tuples where (editor 1, editor 2) and (editor 2, editor 1) are both in the set of reverts.

Once all those have been collected the m-statistic is calculated as such:

$$M = len(mr) \times \Sigma min(a, b)$$

- mr = number of mutual reverting pairs
- a = number of edits made by editor 1 in the set of reverts for the page
- b = number of edits made by editor 2 in the set of reverts for the page

From the light dump data the following are the top and bottom 20 m-statistics. Please note that many articles had a score of 0 so the bottom 20 is just a subset of those:

TOP 20:

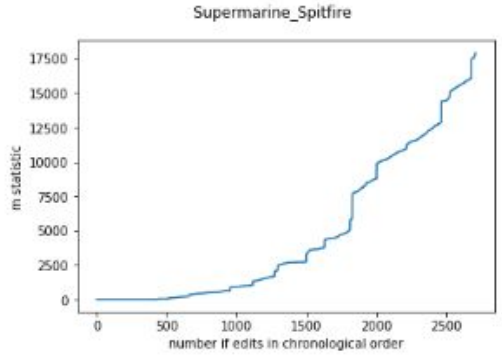
List of World Wrestling Entertainment Employees **87373860**
Michael Jackson **4986215**
Anarchism **33760584**
George W. Bush **32263428**
Neurofunk **22728084**
Muhammad **18140598**
Cathloic Church **17794712**
Global Warming **16442440**
RealMadrid C.F **15733025**
Barack Obama **14544492**
United States **13853504**
Race and Intelligence **13304166**
Circumcision **11948484**
Chevorlet Vega **11926584**
November 2004 **11332948**
Death in 2009 **9989298**
Falun Gong **9442162**
Jesus **9429321**
New York City **9270597**
Elvis Presley **8794566**

BOTTOM 20:

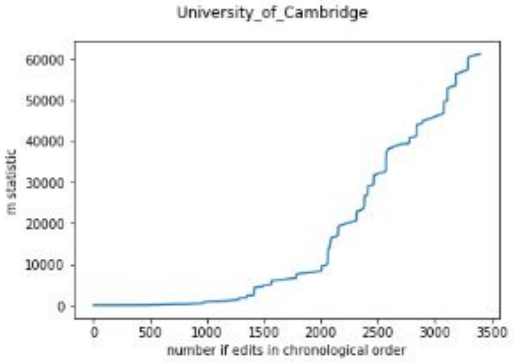
Blackboard **0**
Coimbatore Title Company **0**
The Moonstone 2 **0**
African Hunting Wasps **0**
Morgan Sheardown **0**
Hymenocallis **0**
Erica Barstein **0**
Ben Clarke **0**
British Association for Cognitive and Behavioral psychotherapies **0**
Ocportal **0**
British association for Cognitive and Behavioral psychotherapies **0**
One More Level **0**
Flake Ice Machine **0**
Andre Andrade **0**
Andrew K Skinner **0**
Lock Thief **0**
Enrico Scarampi **0**
Perfect Week **0**
Alexander Grant Ruthven **0**
Number Theory **0**

These m statistics were calculated using the entire edit history of the light

dump per each page. However in order to see trends over time the following graphs show the plot the m statistic that is calculated each time a new edit has been added:



Graph 3: M statistic over time of Supermarie Spitfire



Graph 4: M statistic over time of University of Cambridge

Graph 3 shows the change in the m statistic for an article with a relatively low m statistic and graph 4 shows one with a relatively high one. One similarity is that they both increase as new reverting edits are made. This makes sense in that more reverts means more disagreement on the information on the page and hence a higher m statistic.

One thing that can be extracted from the graphs is points of sudden jumps in the scores meaning sudden jumps in the controversiality of a topic. This is can seen in graph 3 at around 2100 number of edits and in graph 4 around 2000 edits. These can be considered times when the Wikipedia pages' topics were on topics that may have seen an increase in controversy based on real world events happening at that time. Otherwise known as "hot topics" in the media. A way to further investigate this is to see around when the 2100th and 2000th edits were made in the light dump data and see if there was an major real world event regarding those topics at the time.

Raw Data → M-Statistic:

The following will be a step by step explanation of going from the raw data to the light dump format to an m-statistic to finally a graph plotting it over time for two aWikipedia topics: Anarchism and Alaska.

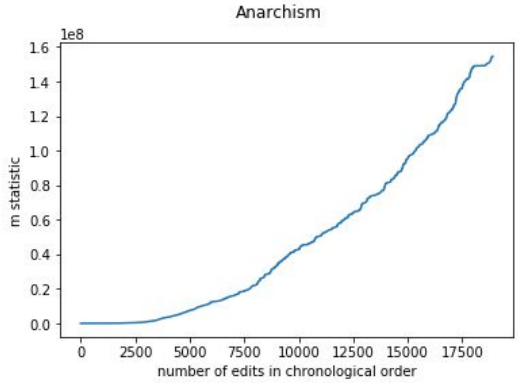
Anarchism:
The last 20 lines of the light dump created for this topic are:

Anarchism
^^^2019-12-22T22:28:39Z 1 15794 El C
^^^2019-12-22T22:25:57Z 0 15795
^^^2019-12-17T18:29:31Z 0 15794
InternetArchiveBot
^^^2019-12-17T06:51:59Z 0 15793 Davide King
^^^2019-12-16T06:18:04Z 0 15792
SurpriseandConquer
^^^2019-12-16T04:40:56Z 0 15791
^^^2019-12-14T18:53:27Z 0 15790
^^^2019-12-13T07:03:09Z 1 15788 Carl Tristan
Orense
^^^2019-12-13T07:02:35Z 0 15789
^^^2019-12-11T16:04:29Z 0 15788 ShannonBarill
^^^2019-12-11T16:03:27Z 0 15787 ShannonBarill
^^^2019-12-11T16:02:27Z 0 15786 ShannonBarill
^^^2019-12-10T11:33:17Z 1 15782
^^^2019-12-10T11:32:05Z 0 15785
^^^2019-12-10T11:30:32Z 0 15784
^^^2019-12-10T11:27:06Z 0 15783
^^^2019-12-10T11:02:47Z 0 15782 Oeqtte
^^^2019-12-09T08:23:10Z 0 15781 Cinadon36
^^^2019-12-09T08:22:26Z 0 15780 Cinadon36
^^^2019-12-09T07:00:54Z 0 15779 Oeqtte

As you can see the top line is a revert, indicated by the second value of "1." It also shows which edit version number it is reverting back to: 15794. This means that the editor "El C" agrees with "InternetArchiveBot" and disagrees with the editor of version number 15795. This would be added to this list of revert tuples.

The m statistic calculated from this is: 22084320. This number is different from the one from the "Top 20" because they were calculated using different light dump data.

The following is the graph of the change in m statistics:



Graph 5: M statistic over time of Anarchism

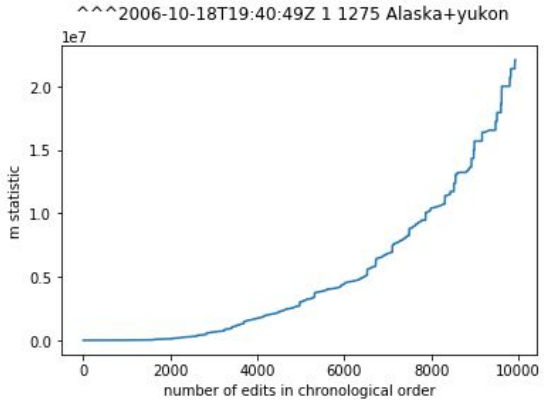
This graph does not show any drastic jumps in the m-statistic over time. It has a steady increase. This means that there were no periods of time where the topic "Anarchism" was more controversial than others. This means that edits were made at a constant rate. However, this does not mean this topic is not controversial; the m statistic values are high throughout this time span graphed here.

Alaska:
The last 20 lines of the light dump created for this topic are:

Alaska
^^^2020-01-01T07:16:33Z 0 7344 AnomieBOT
^^^2020-01-01T06:56:23Z 0 7343 Howcheng
^^^2019-12-19T00:44:30Z 1 7338 Youseefr 99
^^^2019-12-19T00:42:12Z 0 7342 Youseefr 99
^^^2019-12-14T03:12:01Z 1 7338 FlightTime
^^^2019-12-14T03:08:26Z 0 7341 Naulagmi
^^^2019-12-11T05:03:26Z 1 7338 Widjididji
^^^2019-12-11T04:57:28Z 0 7340 Widjididji
^^^2019-12-11T04:56:36Z 0 7339 Widjididji
^^^2019-12-07T15:42:48Z 0 7338 Demetrius Tremens
^^^2019-11-28T13:41:45Z 0 7337 Gog the Mild
^^^2019-11-19T04:26:20Z 0 7336 Pharexia
^^^2019-11-15T22:19:34Z 0 7335 InternetArchiveBot
^^^2019-11-11T17:26:09Z 0 7334 PanamanianBlanco
^^^2019-11-10T00:37:57Z 0 7333 Jonesey95
^^^2019-11-09T19:41:30Z 1 7331 Mmzx84mn
^^^2019-11-06T17:55:01Z 0 7332 CrystalBlacksmith
^^^2019-11-05T20:16:01Z 0 7331 Newscower
^^^2019-10-27T12:47:37Z 0 7330 Radom1967
^^^2019-10-15T02:17:06Z 0 7329 Anonymous from the 21st century

The m statistic calculated from this is: 22084320:

The following is the graph of the change in m statistics:



Graph 6: M statistic over time of Alaska

This graph is similar to that of Anarchism in that it is a steady increase in the m-statistic over time, hence similar conclusions can be drawn from it.

Critique of M-Statistic:

One main critique of this way of calculating the m-statistic is that it only takes into account mutual reverts. Mutual reverts only means the page was changed to an existing version of it. It does not necessarily mean that the change was to an opposing point of view. Calculating it this way can add superfluous and incorrect information as to the actual number of "wars" happening on the page. In order to fix this error the calculation of the m statistic should take in the actual text data being added/ changed and analyse that to see its sentiment.

Another critique is that many editors do not have a username. Hence large chunks of mutual reverting pairs will be either

counting a greater number or fewer number of revert pairs (depending on how the code handles null values for usernames). Since a significant number of edits do not have a username this will greatly negatively affect the results.

Conclusion:

This study created a way to measure the controversiality of a Wikipedia topic and ranked all the articles using this score. It also plotted a few articles to see where spikes in the controversiality of a page are, which can be mapped to certain periods of time in the real world.

This study found some surprising and some not so surprising Wikipedia pages that fall under "highly controversial." However these may not be very accurate due to flaws in the method of calculating the m-statistics.