Edit Cascades

CAS CS 565: Algorithmic Data Mining

Isaac Hu and Olivia Ma

May 4, 2025

1 Introduction

Wikipedia is one of the largest collaborative knowledge bases on the internet. Understanding how Wikipedia articles evolve can reveal the patterns of how information spreads. Inspired by work on social media reshare cascades, **we hypothesize that Wikipedia articles experience similar cascade-like editing behaviors, particularly with politically controversial and sensitive topics**.

The dataset we used is the complete Wikipedia edit history up to January 2008, sourced from a large XML dump and a parsed metadata set. This includes all article revisions, timestamps, and hyperlinks, allowing us to reconstruct article relationships and temporal editing patterns.

To manage this volume of data, we developed a suite of Python scripts to process, filter, and convert the edit histories into graph structures. We pruned irrelevant connections, constructed neighborhood trees, and calculated cascability metrics at scale.

Our goal is to **investigate if edits to one article trigger subsequent edits to its related (hyperlinked) articles, forming information cascades** that mirror the way content spreads on platforms like Facebook or Twitter. To quantify this, we created a metric called **"cascadability", defined as a ratio between two probabilities: the average probability that a neighbor (hyperlinked) article was edited within a week (P_{avg}), and the conditional probability that a neighbor is edited in the same week that the central article is edited (P_{cond_avg}). A ratio greater than one suggests a possible correlation in edit behavior between related pages.**

Beyond identifying cascades, we aim to explore **what drives cascadability**. Specifically, we examine whether articles that are edited more frequently or that cover controversial topics are more likely to generate cascading edit activity. First, we analyzed five of the most frequently edited articles on Wikipedia up to 2008–George W. Bush, Britney Spears, India, Christianity, and Anarchism–building two-layer hyperlink trees and evaluating the editing activation probabilities of their neighbors. We then repeated the experiment on a random sample of 100 articles to establish a baseline. We followed this up with a targeted study of 500 articles related to the Iraq War to explore whether political or emotionally charged content tends to exhibit higher cascadability.

As a longer-term goal, we hope this research can shed light on the dynamics of online knowledge formation and the potential pathways through which misinformation may spread within open editing systems like Wikipedia.

2 Prior Research and Inspiration

We read over 3 main papers:

1. **Cheng et al., "Can Cascades be Predicted?" (2014)**

   This paper investigates the predictability of content resharing cascades via Facebook photos. By observing the first k reshares, the authors show it is possible to predict if a cascade will double in size, with temporal features (how quickly reshares happen) proving to be more predictive than

structural ones. The researchers found that broad, early cascades (many direct reshares) tend to grow more than deep, narrow ones.

Cheng's paper inspired our definition of "cascability" and our focus on temporal clustering of edits in Wikipedia. We borrowed the idea of measuring activation in a time window to detect early cascade growth, which could correlate to editing chains.

2. **Bellomi & Bonato, "Network Analysis for Wikipedia" (2005)**

   This study analyzes Wikipedia's hyperlink network structure in 2005 using algorithms like PageRank and HITS. It found that different ranking algorithms highlight different biases (HITS favors modern, Western, and entertainment-related topics, while PageRank favors historically central articles).

   Since our data comes from the mid-2000s, this analysis helped us understand the structural context of Wikipedia during that time. It also motivated us to look into what types of articles are more likely to act as cascade hubs.

3. **Formisano et al., "Counter-Misinformation Dynamics: The Case of Wikipedia Editing Communities during the 2024 US Presidential Elections" (2024)**

   This paper looks into Wikipedia editing behavior during the 2024 U.S. election, showing that edit spikes cluster around political events and figures. The researchers found that edits with misinformation were relatively rare and handled well by Wikipedia moderation tools.

   We were inspired by this paper's approach with event-based analysis, prompting us to apply a similar lens to the Iraq War.

From these papers, we hypothesize that politically charged pages, with high edit counts or controversial content, are likely to exhibit cascade effects in their editing activity.

3 Methodology

The goal in our experiments was to test whether edit activity propagates through Wikipedia's hyperlink networks in a manner similar to the information cascades discussed in the paper "Can Cascades be Predicted". To do so we designed a pipeline that could extract our massive dataset containing all edit data for Wikipedia from January 1st 2001 to January 1st 2008, would build depth 2 hyperlink based graphs around specific subsets of pages, and compute metrics to describe their cascading dynamics.

We analyzed data from the Snap dataset from Stanford. Data collection required significant cleanup, as the original dataset contained billions of entries with large amounts of additional information, totaling almost 200 gigabytes. We extracted only the article titles and the edit timestamps, as they would be all we needed for the brunt of our analysis. From there, we created programs that would create depth 2 trees for the Wikipedia articles by digging through hyperlinks using BFS, adding only neighbors who already existed as an entry in our edits datasets (since the dataset comes from 2008, it is quite likely there were a significant number of neighbor nodes that did not exist yet and therefore wouldn't appear in the dataset).

To get a visual representation of the activations going on in each of the trees for the articles we extracted, we also created a basic program that would visualize each graph over a time lapse, lighting up nodes whenever they were edited. What we found from this was that oftentimes, nodes would light up around the same time as the center node in each of these trees. We therefore derived a metric, measuring the likelihood of a node being activated given that the center node was active, within the time span of a week, divided by likelihood of being within a week of the node's activation.

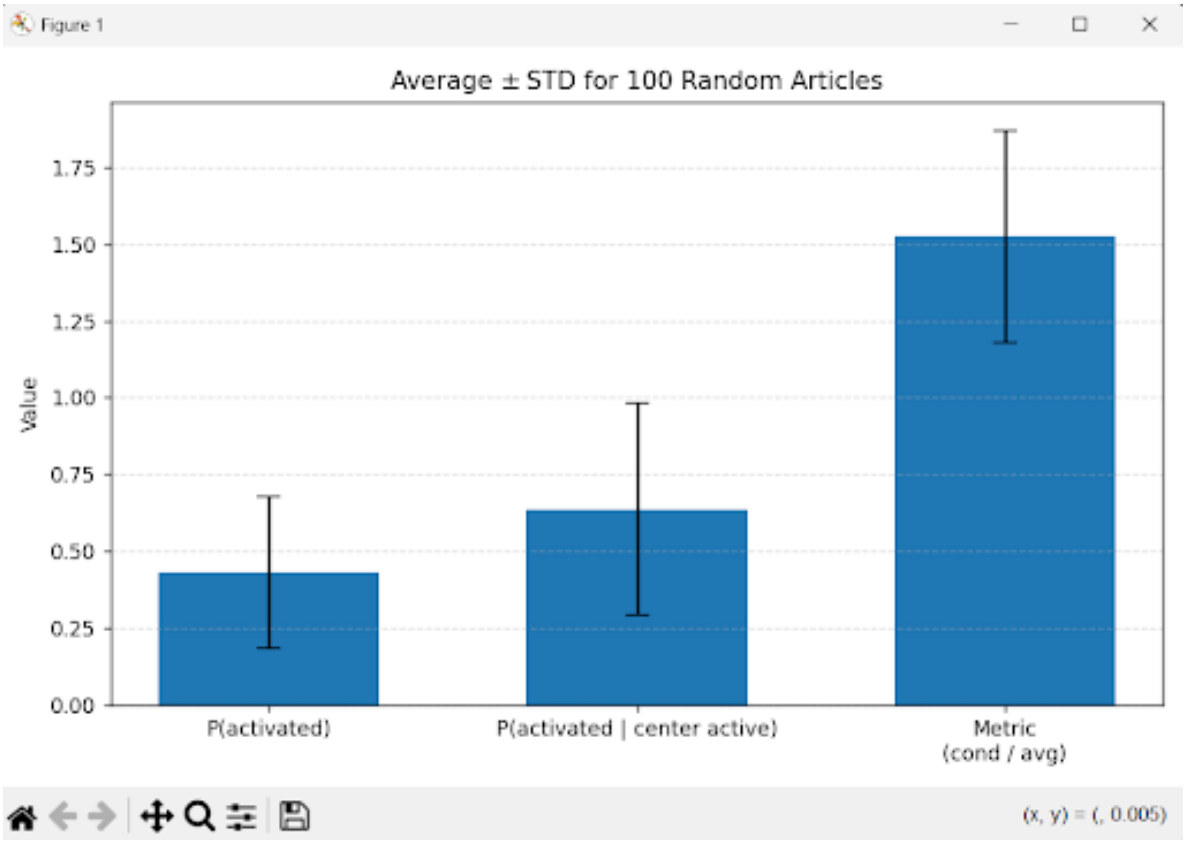

Figure 1: Edit behavior of articles' hyperlinked articles with our cascadability metric.

4 Results

While computing metrics across our various article data subsets, we found several compelling patterns.

Using our metric for cascadability– the ratio between condition activations (the probability that a neighbor is edited within a given week that its central node is activated) and the average neighbor edit rate over the entire period, we found some unexpected results.

We began by randomly sampling 100 articles from our dataset of unique Wikipedia articles. For each sampled article, we constructed its neighborhood tree and measured the edit behavior of its hyperlinked articles with our cascadability metric. 1

The average cascadability metric for this random sample was approximately 1.5236.2

We then computed the same metric for the top 100 most edited articles.

Surprisingly, the average cascadability for this group was significantly lower than that of the random sample, at 1.2039.3 This suggests that a high number of total edits does not necessarily lead to greater

0	Late_Night_Tales	0.5375	0.9520	1.7712	NaN
1	Ross_Malinger	0.7255	1.0000	1.3784	NaN
2	Sanity_testing	0.4927	0.6671	1.3540	NaN
3	The_Courtship_of_Princess_Leia	0.4424	0.7972	1.8020	NaN
4	"Sonora,_California"	0.0000	0.0000	NaN	NaN
..	...	...	...	...	...
97	"Nepalese_legislative_election,_1994"	0.0000	0.0000	NaN	NaN
98	Bonnie_J._Dunbar	0.4370	0.6795	1.5549	NaN
99	Polyglot_(computing)	0.6584	0.8569	1.3015	NaN
100	NaN	0.4332	0.6370	1.5236	AVERAGE
101	NaN	0.2469	0.3450	0.3444	STD

Figure 2: The cascadability metric for 100 randomly selected articles.

	article	P_avg	P_cond_avg	Metric (P_cond_avg / P_avg)
0	George_W._Bush	0.4852	0.5548	1.1434
1	Britney_Spears	0.4598	0.5292	1.1509
2	India	0.5609	0.6504	1.1596
3	Christianity	0.5399	0.6192	1.1469
4	Anarchism	0.5108	0.5845	1.1443
..	...	...	...	...
94	Afghanistan	0.5266	0.5971	1.1339
95	Isaac_Newton	0.5327	0.6248	1.1729
96	Industrial_Revolution	0.5483	0.6486	1.1829
97	AVERAGE	0.5160	0.6204	1.2039
98	STD	0.0732	0.0932	0.0931

Figure 3: The cascadability metric for the top 100 edited articles.

	article	P_avg	P_cond_avg	Metric (P_cond_avg / P_avg)
0	"Abdul_Illah,_Regent_of_Iraq"	0.0000	0.0000	NaN
0	"Abdul_Illah,_Regent_of_Iraq"	0.0000	0.0000	NaN
1	"Al-Faw,_Iraq"	0.0000	0.0000	NaN
2	"Al_Kut,_Iraq"	0.0000	0.0000	NaN
0	"Abdul_Illah,_Regent_of_Iraq"	0.0000	0.0000	NaN
0	"Abdul_Illah,_Regent_of_Iraq"	0.0000	0.0000	NaN
0	"Abdul_Illah,_Regent_of_Iraq"	0.0000	0.0000	NaN
0	"Abdul_Illah,_Regent_of_Iraq"	0.0000	0.0000	NaN
1	"Al-Faw,_Iraq"	0.0000	0.0000	NaN
2	"Al_Kut,_Iraq"	0.0000	0.0000	NaN
3	"Alleged_human_rights_abuses_by_UK_troops_in_I...	0.0000	0.0000	NaN
4	"Arbil,_Iraq"	0.0000	0.0000	NaN
..	...	...	...	...
537	Withdrawal_of_U.S._troops_from_Iraq	0.0000	0.0000	NaN
538	Worker-Communist_Party_of_Iraq	0.3045	0.4752	1.5606
539	World_Tribunal_on_Iraq	0.4189	0.6520	1.5565
540	AVERAGE	0.4336	0.6070	1.4517
541	STD	0.2270	0.3151	0.5638

Figure 4: The cascadability metric for all articles related to the Iraq War.

influence on surrounding articles. In fact, the inverse relationship hints that heavily edited pages may act more as "sinks" of attention than as sources of cascading edits. This raises the possibility that article connectivity, rather than sheer edit volume, plays a more critical role in cascade dynamics. A more robust sampling over a larger pool of random articles would help determine whether our initial random average holds at scale.

Lastly, we tested our metric on a set of politically charged articles, specifically focusing on those related to the Iraq War from January 1, 2001, to January 1, 2008.

These articles showed a noticeable increase in cascadability, settling around 1.4597.4 This may lend credence to our hypothesis that controversial topics are more likely to trigger cascading edits across related pages, as editors could be responding to real-world events or attempting to shape narrative consistency across the network.

However, it must still be noted that both metrics still fall below the score obtained from just randomly sampling, which may indicate that edit quantity and our particular topic are not strong indicators of cascadability.

5 Limitations of Our Work

One major issue that we faced was the scale and scope of our dataset. The original Wikipedia dump was over 200 gigabytes, containing not just metadata but the full content of every edit. Due to resource constraints, we filtered out all the articles with no edits, and from there, extracted a reduced version, around 20 gigabytes, retaining only the article titles and their associated edit timestamps. With this, we could conduct temporal and structural analysis, however, we had to leave behind the semantic substance of the edits themselves.

Our analysis could not account for the type of edits being made–whether they reflected vandalism, bias, corrections, or factual updates. Understanding the nature of edits would require the full dataset, with a verification or clarification method. We wanted to see how misinformation cascades, but had to leave this idea behind for another time.

We observed a correlation between editing on central nodes and subsequent edits on their linked articles, but lacked a sense of cause, we could only make guesses based on the timing of edits! Unlike resharing behavior in social networks, where paths of influence are clear, Wikipedia edits do not come with an indicator of influence. It is plausible that articles related by topic are just being edited around the same time, rather than each edit triggering another in a cascade-like fashion.

6 Future Work

We explored the transmission of edits, so a natural next step would be to analyze the content of the edits themselves. A future goal could be identifying whether certain types of information, such as biased or misleading content, spread across Wikipedia, and trying to reproduce similar results to the 2024 U.S. election paper. This looks beyond when and where cascades occur, but what kind of narratives propagate through Wikipedia.

We also observed that both sparsely edited and heavily edited articles tend to have unusually high average neighbor edit rates. This may suggest that what drives a cascade is not the volume of edits, but the article's connectivity (how many hyperlinks it has to other pages in the network). Further exploration can be done about hyperlink centrality, degree distribution, and article clustering. This could lead to insights on whether certain structural positions make articles more susceptible to initiating or participating in edit cascades.

WorksCited

1. Bellomi, Francesco, and Roberto Bonato. *Network Analysis for Wikipedia.* Department of Computer Science, University of Verona, presented at Wikimania 2005.

2. Cheng, Justin, et al. "Can Cascades be Predicted?" *Proceedings of the 23rd International Conference on World Wide Web*, ACM, 2014, pp. 925–936. https://doi.org/10.1145/2566486.2567997.

3. Formisano, Giuliano, et al. "Counter-Misinformation Dynamics: The Case of Wikipedia Editing Communities during the 2024US Presidential Elections." Sept. 2024, pre-print.

4. Leskovec, Jure, and Andrej Krevl. "Wikipedia Articles and MetaData (Wiki-Meta) Dataset." *Stanford Large Network Dataset Collection*, Stanford U., Oct. 2014, https://snap.stanford.edu/data/wiki-meta.html.

5. Moyer, Daniel. *Wikipedia During Elections: A Case Study in the Limits of Consensus.* Unpublished paper, 2022.