365 Dots in 2019: Quantifying Attention of News Sources

Alexander C. Nwala
Old Dominion University
Norfolk, Virginia, USA
anwala@cs.odu.edu

Michele C. Weigle
Old Dominion University
Norfolk, Virginia, USA
mweigle@cs.odu.edu

Michael L. Nelson
Old Dominion University
Norfolk, Virginia, USA
mln@cs.odu.edu

ABSTRACT

We investigate the overlap of topics of online news articles from a variety of sources. To do this, we provide a platform for studying the news by measuring this overlap and scoring news stories according to the degree of attention in near-real time. This can enable multiple studies, including identifying topics that receive the most attention from news organizations and identifying slow news days versus major news days. Our application, StoryGraph, periodically (10-minute intervals) extracts the first five news articles from the RSS feeds of 17 US news media organizations across the partisanship spectrum (left, center, and right). From these articles, StoryGraph extracts named entities (`PEOPLE`, `LOCATIONS`, `ORGANIZATIONS`, etc.) and then represents each news article with its set of extracted named entities. Finally, StoryGraph generates a news similarity graph where the nodes represent news articles, and an edge between a pair of nodes represents a high degree of similarity between the nodes (similar news stories). Each news story within the news similarity graph is assigned an attention score which quantifies the amount of attention the topics in the news story receive collectively from the news media organizations. The StoryGraph service has been running since August 2017, and using this method, we determined that the top news story of 2018 was the *Kavanaugh hearings* with attention score of 25.85 on September 27, 2018. Similarly, the top news story for 2019 was *AG William Barr's release of his principal conclusions of the Mueller Report*, with an attention score of 22.93 on March 24, 2019.

KEYWORDS

news similarity, attention score, top news, NLP, graph theory

1 INTRODUCTION AND BACKGROUND

It is natural to ask "what were the top news stories of 2019?" A partisanship study might ask, "how often do news stories from different partisan media organizations overlap?" A retrospective study might ask, "when did *Hurricane Harvey* begin to receive serious coverage?", or "how did the attention given to *Hurricane Harvey* by the media differ from hurricanes that occurred in similar timeframes (but different locations) such as *Irma* or *Maria*?" Addressing these questions requires the fundamental operation of measuring overlap, or similarity, of news topics across different news sources.

We developed a method of measuring the similarity among news articles in near-real time and quantifying the level of attention the topics in the news stories receive. Specifically, we created a service called *StoryGraph* (http://storygraph.cs.odu.edu/) that creates a news similarity graph from 17 left, center, and right news media organizations. StoryGraph quantifies the level of attention the topics in the news stories receive by assigning each an *attention score*. Major breaking news stories are often reported by multiple different news organization within the same time period. Similarly, a major news story is characterized by a high degree of similarity between different pairs of news stories from different news organizations. For example, below is a list of headlines showing a high degree of similarity among news reports collected on October 24, 2018, at 5:34 PM EST from four news organizations, following the incident in which mail bombs were sent to multiple Democratic public figures.

- **Vox**: Explosive devices sent to Clintons, Obamas, CNN: what we know [12]
- **FoxNews**: FBI IDs 7 'suspicious packages' sent to Dem figures containing 'potentially destructive devices' [10]
- **CNN**: Bombs and packages will be sent to FBI lab for analysis [11]
- **Breitbart**: Live Updates: Democratic Leaders Receive Mail Bombs [9]

The prerequisite for deriving the attention score is calculating the similarity between documents (e.g., news articles). This problem has been studied extensively. Methods that represent documents as vectors [7, 8, 15] often use the Cosine Similarity vector-based metric to quantify similarity between pairs of documents. Methods that represent documents as sets [19, 20] often use set-based metrics such as the Jaccard similarity or the Overlap coefficient metric to quantify the similarity between a pair of documents. In this work, we represent each news article as a set of named entities, and utilize a set similarity measure (Section 2, Step 4) to quantify the degree of similarity between a pair of news documents.

Computation + Journalism Symposium, March 20–21, 2020, Boston, MA
.

Table 1: List of 17 left (blue), center (purple), and right (red) news media RSS feeds from which StoryGraph extracts news stories. The list of media sources was derived from Faris et al.'s [13] list of popular media sources.

LEFT
http://www.politicususa.com/feed
https://www.vox.com/rss/index.xml
http://www.huffingtonpost.com/section/front-page/feed
http://www.msnbc.com/feeds/latest
http://rss.nytimes.com/services/xml/rss/nyt/HomePage.xml
http://feeds.washingtonpost.com/rss/politics
CENTER
http://rss.cnn.com/rss/cnn_topstories.rss
http://www.politico.com/rss/politics.xml
http://abcnews.go.com/abcnews/topstories
http://thehill.com/rss/syndicator/19109
http://feeds.feedburner.com/realclearpolitics/qlMj
RIGHT
http://www.washingtonexaminer.com/rss/news
http://feeds.foxnews.com/foxnews/latest
http://feeds.feedburner.com/dailycaller
http://conservativetribune.com/feed/
http://feeds.feedburner.com/breitbart
http://www.thegatewaypundit.com/feed/

Our investigation into measuring near-real time news similarity and quantifying the attention of news sources has resulted in the following contributions. First, we proposed the attention score, a transparent method for quantifying attention given to a news story by different news sources. The attention score facilitates finding the top news stories for a given day, month, or year. This enabled us to show the top stories of 2018 and 2019 (Table 2). Second, we introduced the StoryGraph service, which has been running for over two years (since August 8, 2017), generating news similarity graphs every 10 minutes from 17 news organizations across the left, center, and right partisanship spectrum. Third, we showed that the StoryGraph service and dataset provides a platform for multiple longitudinal studies (Section 3). The code for StoryGraph is publicly available [3, 4], and the entire StoryGraph dataset are available upon request.

2 METHODOLOGY

The StoryGraph process has four steps (Fig. 1) outlined below. First, StoryGraph collects the first five news articles from the RSS feeds of 17 news media organizations (Table 1). Second, StoryGraph dereferences the URLs of the news articles and extracts plaintext after removing the HTML boilerplate [1]. Third, StoryGraph utilizes the Stanford CoreNLP Named Entity Recognizer [5, 14] to extract seven entity classes – `PERSONS`, `LOCATIONS`, `ORGANIZATIONS`, `DATES`, `TIME`, `MONEY`, and `PERCENT` from the news documents. In addition to these entity classes, we created and extracted text that belong to two additional classes: `TITLE` and `TOP-K-TERM`. The `TITLE` class represents title terms from the news articles, while the `TOP-K-TERM` class represents the top k (we set $k = 10$) most frequent terms. All text that does not belong to one of the entity classes is discarded. Subsequently, each news article is represented as a set of entities extracted from the article. Fourth, StoryGraph creates a graph where the nodes (set of entities) represent news articles, and an edge between a pair of nodes represents a similarity score beyond some threshold between the nodes (similar news stories). Finally, the attention scores of the connected components of the recently generated graph is calculated. Formally, consider a *news similarity graph* G in which the nodes represent news articles, and an edge between a pair of nodes represents a high degree of similarity (Section 2, Step 4) between the nodes (similar news stories). Consider the set of G's connected components C, such that $\forall c_i \in C$, the nodes (news articles) in c_i originate from multiple news sources. The attention score of a news story represented by a connected component is simply the average degree of the connected component.

Table 2: StoryGraph: Top news stories of 2018 [2] & 2019 [6]

Rank	Attn. score	Date (MM-DD)	News story
Section 1: Top News Stories of 2018			
1	25.85	09-27	Kavanaugh and Christine Blasey Ford testify before congress
2	18.81	02-02	Nunes memo released
3	18.15	06-12	Trump and Kim Jong Un meet in Singapore
4	17.03	10-24	Bombs mailed to Clinton, Obama, etc.
5	16.32	03-17	Ex-FBI Deputy Director Andrew McCabe fired
Section 2: Top News Stories of 2019			
1	22.93	03-24	AG William Barr releases Mueller Report's principal conc.
2	18.60	09-24	House Speaker Pelosi announces formal impeachment inquiry
3	18.18	11-19 11-20	Impeachment inquiry public testimony
4	17.19	01-19	Mueller: BuzzFeed Report 'Not Accurate'
5	15.39	07-31	2019 Democratic debates

Step 1: News article extraction

StoryGraph extracts the URLs of the first five news articles from each of the 17 RSS feeds (Table 1). Next it dereferences each URL yielding 85 HTML documents.

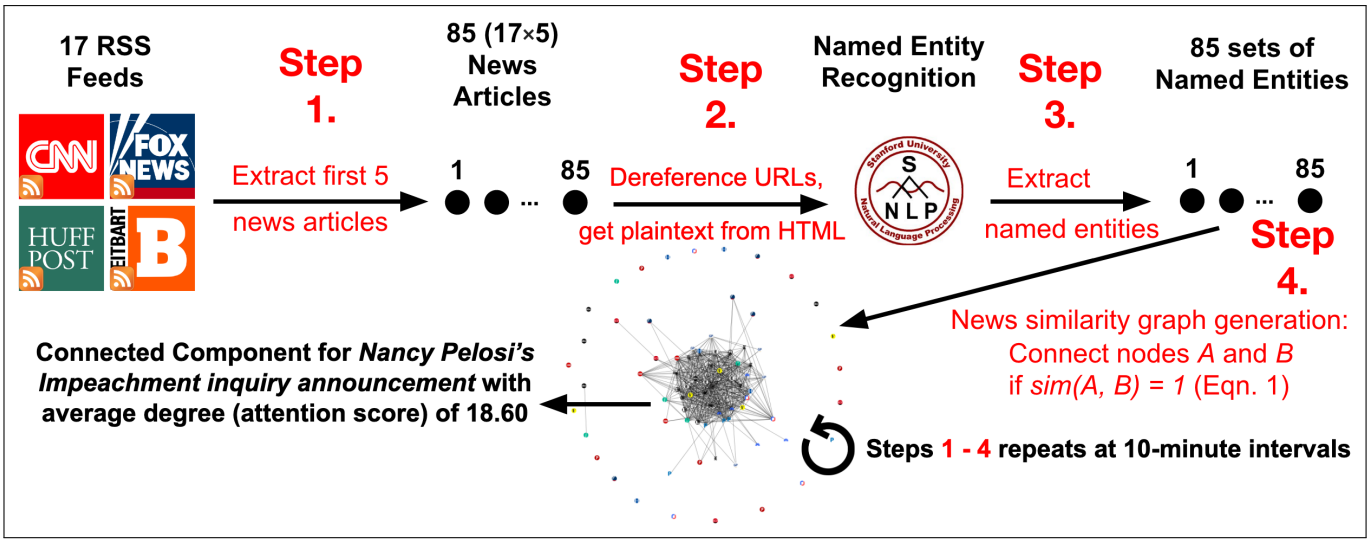

Figure 1: Overview of StoryGraph illustrating the process of generating a *news similarity graph* is four primary steps.

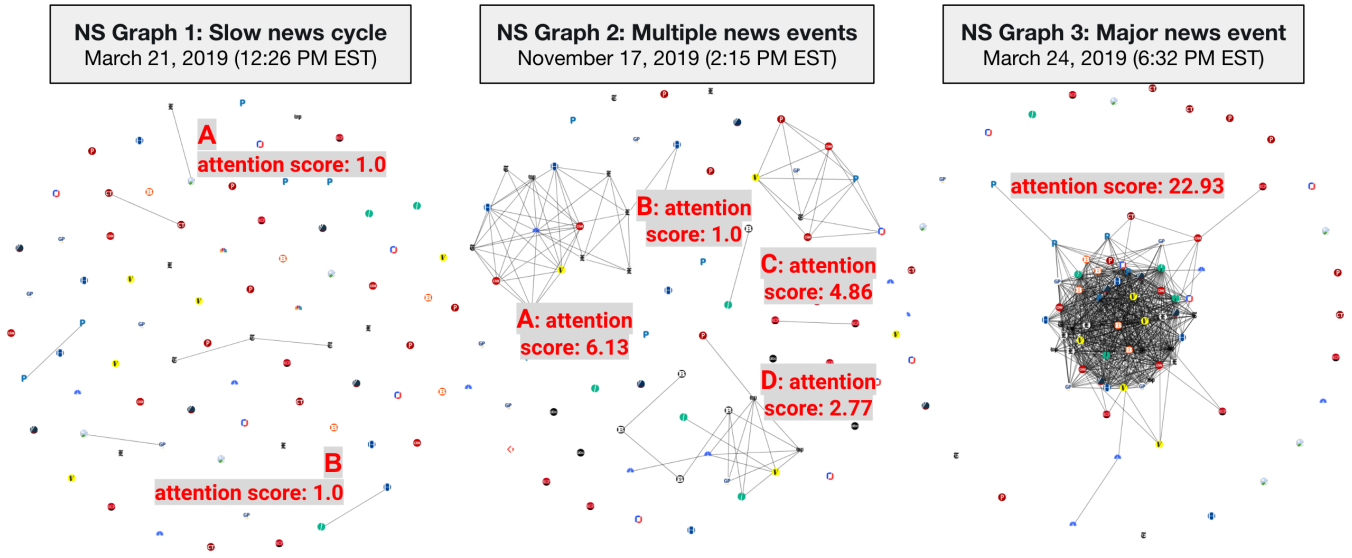

Figure 2: Three News Similarity (NS) graphs illustrating the dynamics of the news cycle. In these graphs, a single node represents a news article, a connected component (multiple connected nodes) represents a single news story reported by the connected nodes. The first (NS Graph 1 [18]) shows what is often referred to as a slow news day; low overlap across different news media organizations resulting in a low attention score (1.0) for news stories (connected components *A* and *B*). The second graph (NS Graph 2 [17]) shows a scenario where the attention of the media is split across four different news stories (connected components *A* – *D*). The third graph (NS Graph 3 [16]) for the *AG William Barr's release of his principal conclusions of the Mueller Report* story shows a major news event; high degree of overlap/connectivity across different news media organizations, resulting in a high attention score of 22.93

Step 2: Plaintext extraction

The HTML boilerplates from the 85 documents from Step 1 are removed [1], yielding 85 plaintext documents.

Step 3: Named entities extraction

The 85 plaintext documents from Step 2 are passed into the Stanford CoreNLP Named Entity Recognizer [5, 14], yielding 85 different sets of named entities.

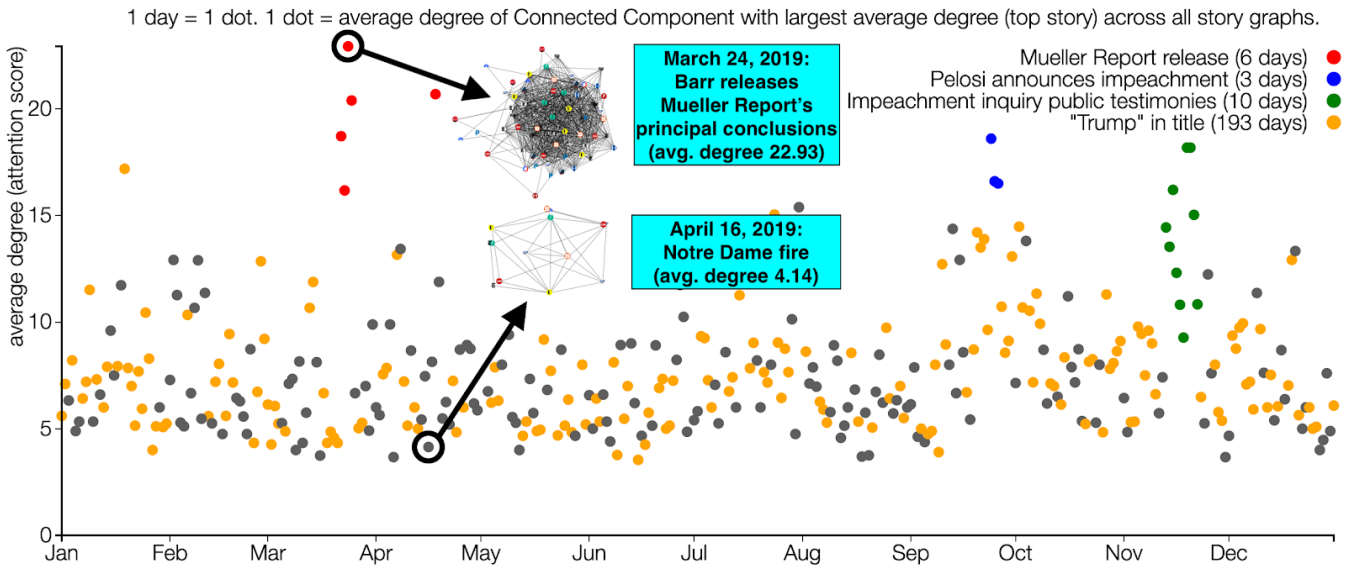

Figure 3: 365 dots in 2019 [6]: Top news stories for 365 days in 2019. Each dot represents the highest attention score across 144 story graphs for a given day.

Step 4: News similarity graph generation

Given a pair of news articles represented by their respective set of named entities A and B, the weighted Jaccard-Overlap similarity $sim(A, B)$ is given by Eqn. 1, where β is the coefficient of similarity, defining the threshold two documents must reach to be considered similar ($sim(A, B) = 1$). This threshold was empirically derived from a gold-standard dataset and set to $\beta = 0.27$. An edge is formed between nodes for which $sim(A, B) = 1$.

$$sim(A, B) = \begin{cases} 1 & \text{, if } \alpha J(A, B) + (1 - \alpha)O(A, B) \geq \beta \\ 0 & \text{, otherwise} \end{cases} \quad (1)$$

$J(A, B)$ is the Jaccard index of both documents, $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, and $O(A, B)$ is the Overlap coefficient of both documents,
$O(A, B) = \frac{|A \cap B|}{min(|A|, |B|)}$.

StoryGraph has been running since August 8, 2017, generating a news similarity graph once every 10 minutes. Since then, the application has generated 120,663+ graphs. For a given day, the connected component with the highest average degree (attention score) from 144 candidate graphs maps to the top news story of the day. Similarly, for a given month, the connected component with the highest attention score maps to the top news story for the month. And for a given year, the top k news stories is derived by finding k connected components with the highest attention scores. Specifically, the top k (e.g., $k = 10$) news stories is the first k connected components from the sorted (in descending order by attention score) list of all news similarity graphs.

3 RESULTS AND DISCUSSION

We can now quantify "slow news days" vs. major news, as well as show that the *Mueller report* was 2019's top story.

Slow news day vs. Major news

Fig. 2 illustrates how the attention score (average degree) of the connected components in a news similarity graph helps characterize different news cycle scenarios. All news graphs in this section refer to Fig. 2. Most of the nodes (news articles) in *NS Graph 1* are isolated with few connected components; the news sources mostly report on divergent topics. Consequently, no single news story (e.g., connected component A or B) receives attention from more than two different news sources.

The second news similarity graph (*NS Graph 2*), unlike the first, shows attention split among four primary news stories. The attention score of each news story represented by the connected component indicates the magnitude of attention given to the news story. For example, connected component A (attention score: 6.13) represents the story *Poll: Pete Buttigieg becomes the presidential frontrunner in Iowa - Vox* and B (1.0) - *Colin Kaepernick Skips NFL Organized Workout, Wears Shirt Likening Himself to a Slave - Breitbart.*

The third news similarity graph (*NS Graph 3*) indicates a major news event, characterized by a giant connected component (news story) with a high attention score (22.93) indicating a high degree of overlap among news sources. This

indicates a scenario when most news sources report on the same story (*AG William Barr's release of his principal conclusions of the Mueller Report*).

The top news stories of 2019

Stories surrounding the release of the Mueller Report (red dots in Fig. 3, Table 2 Section 2, No. 1) received the most attention in 2019. On March 22, 2019, Robert Mueller submitted his report to AG William Barr (attention score: 18.72). Two days later, AG William Barr released his summary (principal conclusions) of the report. This story received the most attention (attention score: 22.93) in 2019. AG William Barr's principal conclusions of the Mueller report was received with skepticism by the Democrats who claimed the conclusions were highly favorable to President Trump. In contrast, the Republicans claimed the summary exonerated the President from any wrongdoing. The next top story in 2019 (blue dots in Fig. 3, Table 2, Section 2, No. 2) with attention score of 18.60 was Speaker Nancy Pelosi's announcement of an official impeachment inquiry (September 24, 2019) four days after the whistleblower's report. Similarly, at rank three (green dots in Fig. 3) were stories chronicling the public testimonies of the impeachment inquiry.

4 FUTURE WORK AND CONCLUSION

StoryGraph has been generating one news similarity graph every 10 minutes since August 2017. A single graph file includes the URL of the news articles, plaintext, entities, publication dates, etc. In this paper, we only reported the result of two studies. The first studies the dynamics of the news cycle (*slow news cycle* vs. major news event). The second utilized attention scores to facilitate finding top stories.

StoryGraph provides the opportunity for further study beyond the two presented here. For example, a study focused on the coverage of mass shootings can utilize StoryGraph to approximate how much attention the 2018 Parkland, Florida shooting received compared to the 2019 Dayton, Ohio and El Paso, Texas mass shootings. A different study could narrowly apply news similarity to focus on a single news organization, e.g., FoxNews, in order to identify the news stories where they focus the most attention, or compare the attention span of different events. Therefore, we believe the StoryGraph process of quantifying news similarity and the attention of news sources provides a valuable means for studying news.

ACKNOWLEDGMENTS

This work was supported in part by IMLS LG-71-15-0077-15. We also appreciate the help of Sawood Alam in the deployment of StoryGraph.

REFERENCES

[1] Alexander Nwala. 2017. A survey of 5 boilerplate removal methods. https://ws-dl.blogspot.com/2017/03/2017-03-20-survey-of-5-boilerplate.html.
[2] Alexander Nwala. 2018. 365 dots in 2018. http://storygraph.cs.odu.edu/stats/2019-03/365-dots-in-2018/.
[3] Alexander Nwala. 2018. Github repository for StoryGraph graph generator. https://github.com/anwala/storygraph-grapher.
[4] Alexander Nwala. 2018. Github repository for StoryGraph graph visualizer. https://github.com/anwala/storygraph-web.
[5] Alexander Nwala. 2018. Installing Stanford CoreNLP in a Docker Container. https://ws-dl.blogspot.com/2018/03/2018-03-04-installing-stanford-corenlp.html.
[6] Alexander Nwala. 2019. 365 dots in 2019. http://storygraph.cs.odu.edu/stats/2019-12/365-dots-in-2019/.
[7] Grant C Atkins, Alexander Nwala, Michele C Weigle, and Michael L Nelson. 2018. Measuring News Similarity Across Ten US News Sites. *arXiv preprint arXiv:1806.09082* (2018).
[8] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. paper recommender systems: a literature survey. *International Journal on Digital Libraries* 17, 4 (2016), 305–338.
[9] BreitbartNews. 2018. Live Updates: Democratic Leaders Receive Mail Bombs. https://www.breitbart.com/politics/2018/10/24/live-updates-mail-bombs-to-multiple-democratic-leaders.
[10] Chamberlain, Samuel and Gibson, Jake. 2018. FBI IDs 7 'suspicious packages' sent to Dem figures containing 'potentially destructive devices'. https://www.foxnews.com/politics/fbi-ids-7-suspicious-packages-sent-to-dem-figures-containing-potentially-destructive-devices.
[11] CNN. 2018. Bombs and packages will be sent to FBI lab for analysis. https://www.cnn.com/politics/live-news/clintons-obama-suspicious-packages/h_f80b306c7567a4169228c9262f62f06e.
[12] Coaston, Jane and Emily, Stewart and Kirby, Jen. 2018. Explosive devices sent to Clintons, Obamas, CNN: what we know. https://www.vox.com/policy-and-politics/2018/10/24/18018256/explosive-device-bomb-clinton-obama-cnn-soros.
[13] Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. *Berkman Klein Center Research Publication* 6 (2017).
[14] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of Association for Computational Linguistics (ACL 2005)*. 363–370.
[15] Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis.. In *IJcAI*, Vol. 7. 1606–1611.
[16] StoryGraph. 2019. Major news event graph. http://storygraph.cs.odu.edu/graphs/polar-media-consensus-graph/#cursor=135&hist=1440&t=2019-03-24T22:32:21.
[17] StoryGraph. 2019. Multiple news events graph. http://storygraph.cs.odu.edu/graphs/polar-media-consensus-graph/#cursor=115&hist=1440&t=2019-11-17T19:15:38.
[18] StoryGraph. 2019. Slow news cycle graph. http://storygraph.cs.odu.edu/graphs/polar-media-consensus-graph/#cursor=98&hist=1440&t=2019-03-21T16:26:25.
[19] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. 2000. Impact of similarity measures on web-page clustering. In *Workshop on artificial intelligence for web search (AAAI 2000)*, Vol. 58. 64.
[20] Tuan Tran, Andrea Ceroni, Mihai Georgescu, Kaweh Djafari Naini, and Marco Fisichella. 2014. Wikipevent: Leveraging wikipedia edit history for event detection. In *International Conference on Web Information Systems Engineering*. 90–108.