A System for High Performance Mining on GDELT Data

Konstantin Pogorelov
Simula Research Laboratory
Fornebu, Norway
konstantin@simula.no

Daniel Thilo Schroeder
Simula Metropolitan
Center for Digital Engineering
Oslo, Norway
Technical University of Berlin
Berlin, Germany
daniels@simula.no

Petra Filkukova
Simula Research Laboratory
Fornebu, Norway
petrafilkukova@simula.no

Johannes Langguth
Simula Research Laboratory
Fornebu, Norway
langguth@simula.no

***Abstract*—We design a system for efficient in-memory analysis of data from the GDELT database of news events. The specialization of the system allows us to avoid the inefficiencies of existing alternatives, and make full use of modern parallel high-performance computing hardware. We then present a series of experiments showcasing the system's ability to analyze correlations in the entire GDELT 2.0 database containing more than a billion news items. The results reveal large scale trends in the world of today's online news.**

***Keywords*-GDELT, Data mining, High Performance Computing, Misinformation, Publishing**

I. Introduction

In the wake of the 2016 US presidential elections, the topic of online misinformation, often referred to as *fake news*, has gained attention of the scientific community. Of particular concern is the phenomenon of *digital wildfires*, i.e. fast-spreading misinformation with the potential for serious harm in the real world, which was identified as one of the major threats to developed countries by the World Economic Forum [1].

The vast majority of fake news research has targeted social networks, with Twitter being by far the most studied network due to its relative accessibility. Such work studies topics such as the spreading patterns and velocity in social networks [2], [3]. In online social networks, this spread, i.e. information cascade, is comparatively easy to study since the channels along which information flows are given by the graph relation within the social network.

News websites, including those that belong to printed newspapers and independent ones play an important role in the spread of information, and with that potentially also misinformation. However, they are much harder to study quantitatively. While Twitter offers an API for extracting information from the network, news websites operate independently and use different technologies for content delivery. Thus, an existing project offers immense help for the task of news website analysis.

The Global Dataset of Events, Location, and Tone (GDELT) [4] uses a series of sophisticated tools to scrape articles from the global news landscape in essentially real time, extract the events that are being reported, as well as the location of these events along with a wealth of other information. The automatic coding allows a truly global view on current events and thus represents a massive step ahead compared to earlier methods [5]. It also allows an in-depth, data-driven analysis of global news reporting.

The main obstacle in using GDELT to perform a large scale analysis is the sheer volume of data which can limit what can be analyzed. To overcome this limitation we design and implement a lightweight, efficient system capable of reading the entire GDELT database and extracting information in real time.

We use this system to perform an in-depth analysis of the global English language news agencies for the years 2015 to 2019. We present statistics on the most active publishers, as well as their topic overlap measured by the co-publishing factor. We measure the coverage of each selected country in the press of other countries. Finally, we measure the publishing delay, i.e. the average amount of time that different news agencies take to publish an article about an event, thereby studying the fundamental question whether the news business is accelerating.

II. Related Work

Since its inception, GDELT has been the topic of more than 100 scientific articles. Some of them deal with the properties of GDELT [4] and analyze the news coverage [6], while others compare it to similar systems such as ICEWS [7] or EventRegistry [8]. However, the majority of articles use GDELT to either study current news coverage and its effects [9]–[12], attempt to track real-world trends [13], [14], or attempt to predict future events based on current news [15]–[19].

Similar to our approach, Lu and Szymanski [20] use high performance computing techniques to analyze large amounts of GDELT data quickly. However, while they perform a streaming analysis for the prediction of viral news events, we aim to make all GDELT data available in memory for rapid processing. Other works that use GDELT data to observe news media include Rappaz et al. [21], who propose a dynamic embedding model of the media landscape.

The size and the properties of the GDELT dataset have prompted the development of specialized solutions. Al-Naami et al. [22], [23] proposed a spatial query processing system for GDELT and observed it to be faster than a standard Hadoop based solution by orders of magnitude. Other spatial query systems [24], [25] have been benchmarked with GDELT data. The results emphasize the importance of in-memory processing. GDELT data can also be accessed using Google BigQuery[1]. However, BigQuery reports October 2018 as the time of the last update. A similar option is available on Amazon S3[2] although this is also not currently (February 2020) being updated. In any case, such systems are limited to SQL queries, and they do not allow running network analysis algorithms efficiently. Furthermore, using BigQuery has a cost based on the amount of data processed. Thus, performing truly massive investigations in the collected GDELT data is not encouraged when using such a system. A simple test query looking for mentions of a politician in a short span of time required processing of more than one TB of data. For these reasons, we opt to build a standalone system.

III. The GDELT System

The current version 2.0 of GDELT monitors both English and non-English news sources, with archives going back to 2015. Non-English articles in 65 languages are translated to English for further analysis using what is believed to be the largest realtime streaming news machine translation deployment in the world. It has the capacity to monitor news of the entire world. 98.4% of the monitored content is translated in real time. Thus, it is most likely the system with the widest reach w.r.t. media in the non-western world, although its reach in these areas is limited compared to the western world.

Every 15 minutes, the GDELT system uploads two files. The *Events* and the *Mentions* table of the last 15 minutes. Whenever a news article is scraped, its text is analyzed and the event it reports on is determined. If the event is new, it will be issued a unique event ID. If it has been reported on before, the news item will be linked to the event ID. In either case, it is added to the mentions table, along with the event ID. The mentions table thus contains the URLs of the articles along with supplemental information, as well as the ID of the event they are reporting on.

While the amount of data generated in this way is substantial, following current events only poses a moderate challenge for modern computers. One week worth of data, which is 672 sets of files covering 15 minutes each, amounts to approximately one GB. However, when gathering data over years, the analysis becomes more challenging. Thus, in this paper we present a system that can deal with such data efficiently on modern servers with sufficient memory. By creating native, in memory tables and using parallel processing, we can analyze several years worth of news data within seconds. Even complex operations such as tracking co-mentions among thousands of news agencies become feasible within a few minutes.

GDELT 2.0 also contains a large number of additional features such as realtime measurement of emotions and themes via a system called GDELT Global Content Analysis Measures (GCAM). For 15 selected languages, the measurement of emotions is performed natively, i.e. without prior translation to English. Moreover, GDELT is performing analysis of images embedded in the news articles, and take note of real world knowledge such as prices, amounts, names of organizations, dates, and legislations, and quotes. It also features the GDELT Global Knowledge Graph that connects such information.

However, the advanced features contained in GDELT 2.0 have so far not found wide adoption in the scientific community. By the same token, in this work, we focus on the monitoring of English language news itself, rather than the knowledge derived from it.

IV. High Performance GDELT Analysis

Unlike a standard database, our system works in a read-only manner after the initial data tables are set up. As a result, we can query large amounts of data much faster. The system is written in C++ using OpenMP for parallelization. It is designed to run on large memory nodes. The overall system structure is depicted in Figure 1. Before working with the data, we once convert GDELT database files with our preprocessing tool in order to build indexed version of the database which contains data fields in machine-readable binary format. User-defined queries to the database are processed via a query execution engine optimized for in-memory handling of previously converted GDELT data. We implemented parallel version of the most intensive aggregated queries (see subsection VI-G).

We run the system on a compute node equipped with dual socket AMD EPYC 7601 processors having 32 cores and 64 threads each. It is equipped with 2 terabytes of DRAM and has a STREAM [26] memory bandwidth of about 240 GB/s. The node is part of the Experimental Infrastructure for Exploration of Exascale Computing (eX3)[3]. Due to the large memory of the system, it is possible to load and query the data inside a single shared memory system. This obviates

[1] https://www.gdeltproject.org/data.htmlgooglebigquery

[2] https://registry.opendata.aws/gdelt/

[3] https://www.ex3.simula.no/

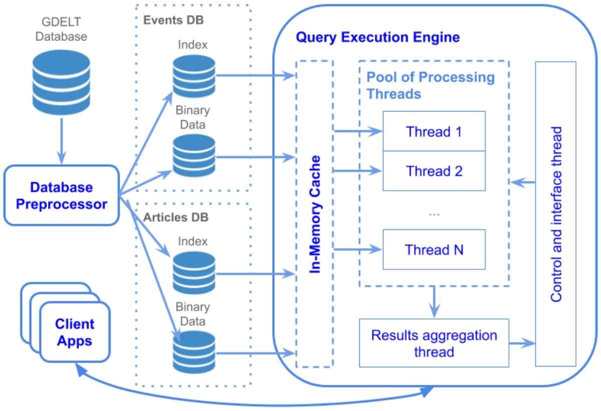


Figure 1: The figure shows the internal structure of the developed high performance GDELT analysis system.

the need for inter-node communication, which constitutes a potential performance bottleneck.

However, the chiplet based structure of the AMD EPYC 7601 processors implies a complicated non-uniform memory access (NUMA) architecture. The system has eight separate NUMA nodes, and bandwith between these is limited[4]. Consequently, care must be taken to correctly place the compute threads and distribute memory allocations among the cores and NUMA nodes in order to obtain the full performance of the machine.

V. Data Collection and Dataset

To construct our dataset we downloaded the GDELT 2.0 Event Database[5] for the period from the 18th of February 2015 to the 31st of December 2019. The start date is defined by the first date when the GDELT project started to collect the Event Database in the new detailed data format 2.0. Basic statistics are listed in Table I.

Table I: General dataset statistics

Number of	Value
Sources	20,996
Events	324,564,472
Capture intervals	168,266
Articles	1,090,310,118
Minimum number of articles per event	1
Maximum number of articles per event	5234
Articles per event (weighted average)	3.36

Table II: Problems found during the dataset analysis

Number of	Value
Missformatted dataset master list entries	53
Missing archives for dataset chunks	8
Missing event source URL	1
Recorder event date is in future compared to the recorded first article publication date	4

Interestingly, the number of news sources is far lower than the roughly 50,000 sources ingested by Google News [27]. Thus, we can conclude that despite its impressive reach, GDELT is still missing a large part of the global news landscape that is accessible online. Nevertheless, our data gathering extends to more than one billion articles covering more than 300 million events, thus constituting the largest such analysis so far. We are predominantly interested in the timing and location of world events, as well as the relative time, i.e. publishing delay, source, and location of the mentions.

Naturally, the number of articles reporting on a given event vary widely. The most reported event, which is the 2016 Orlando nightclub shooting, was picked up by more than a quarter of all tracked news sources, while the typical event is covered only by one to five sites. When considering that not all sources are active over time, the event was reported on by about 85% of the sources that were active at that time, as shown in Figure 3. The low number of active sites shows that many of the sources tracked by GDELT are periodical publications rather than daily newspapers.

As expected, the frequency of highly reported news follows a power law distribution [28], as shown in Figure 2. A similar observation was made previously by Lu et al. [20]. However, unlike Lu et al., our data shows a slight but noticeable deviation from the power law around the center of the graph. Note that unlike Lu et al., we take all sources and all articles into account.

A crucial component of our efficient handling of GDELT data is its conversion to a binary format. While doing so is straightforward, it requires cleaning and checking the data. Doing so, we found a small number of problems with the GDELT source data which are listed in Table II.

We present the development of key statistics over time. The number of sources is shown in Figure 3, events in Figure 4, and articles in Figure 5. For readability reasons we aggregated time into quarters. Note that the first entry begins on the 18th of February 2015, and thus does not represent a full quarter. The numbers are relatively stable over time, with a slight decrease in the years 2018 and 2019. Interestingly, while the number of sources is relatively stable, only about one third of the sources are active in any given quarter. 0 per page at the time they register.

VI. Experimental Results

In this section we present a series of experimental results which were obtained using our system. Some of these are grouped by world region, while others are grouped by time.

A. Articles over Time

Our first experiment is a simple count of articles per source. Based on the result, we determined the 10 most productive news websites. In Figure 6, we present their

[4] http://www.prace-ri.eu/best-practice-guide-amd-epyc

[5] https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/

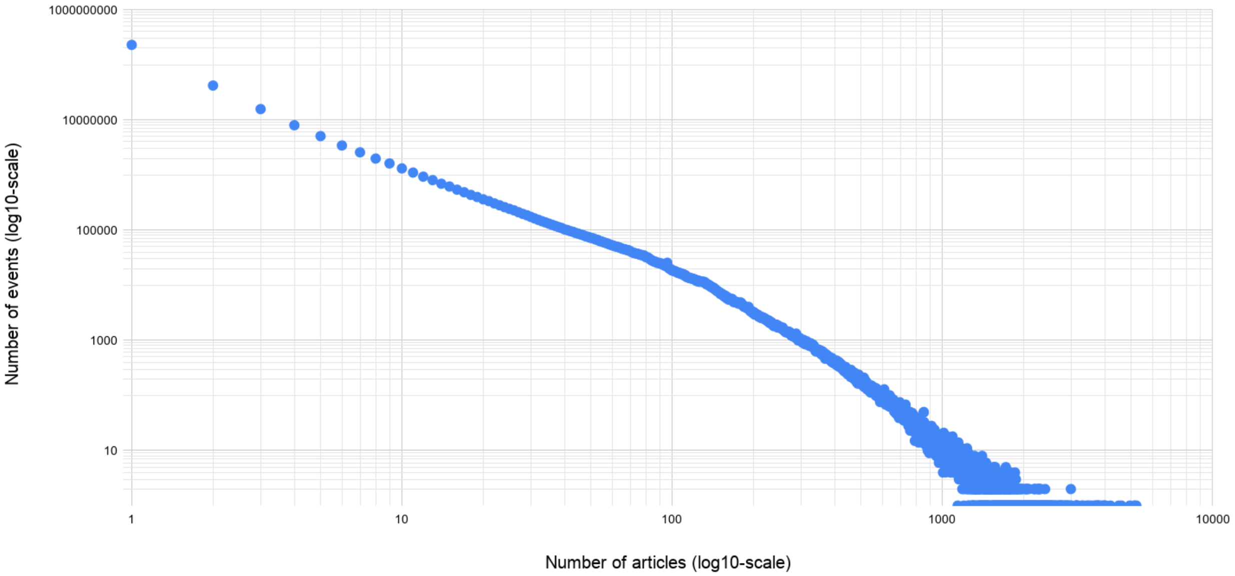


Figure 2: The diagram shows the number of events with certain number of articles.

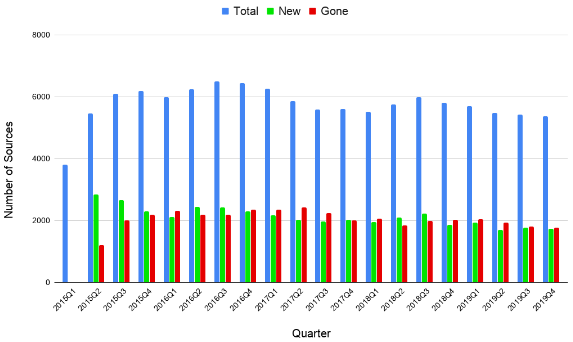


Figure 3: The diagram shows the number of sources that are active during each quarter.

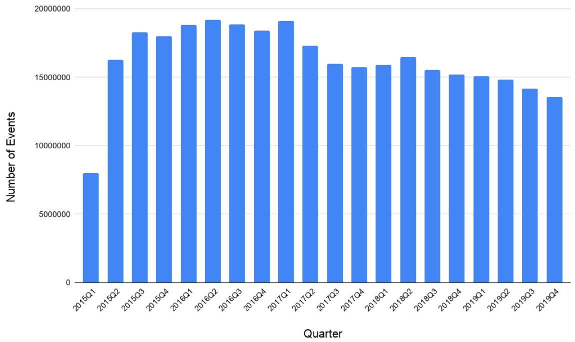


Figure 4: The diagram shows the number of events observed by quarter.

Table III: The ten most reported events

Mentions	Event source URL
5234	Orlando nightclub shooting, 2016
5147	Las Vegas shooting, 2017
5131	Shooting of Dallas police officers, 2016
4944	Shooting of Alton Sterling, 2016
4606	Donald Trump announces running for a second term, 2019
4501	Reactions to shooting of Dallas police officers, 2016
4196	Reactions to Orlando nightclub shooting, 2016
4037	El Paso shooting, 2019
3989	NRA activity, 2019
3984	Russian reaction to Donald Trump election, 2017

number of articles over time. In the following, we refer to this number as n_i for a source i.

While one would expect to find internationally known newspapers here, we observe that 8 out of the 10 of these websites are regional British newspapers, and most of them are owned by the Newsquest Media Group[6]. Cursory inspection of the news website suggests that most of these articles are comparatively short. The graphs also indicate a certain correlation over time, which we will investigate in the next experiment.

B. Common Reporting of the Top Publishers

We created the co-reporting matrix of all 20996 sources present in the data. For each pair of sources i and j, we measure their co-reporting c_{ij} by counting the number of events e_{ij} that both report on, divided by the total number of events the pair reported on, i.e. the Jaccard Index (see e.g. [29]). Note that this formulation heavily depends on using e_i i.e. *events reported on* by source i as a base statistic

[6] https://www.newsquest.co.uk/

Table IV: The follow-reporting matrix for ten most productive news websites lists the f_{ij} index for each pair of publishing websites. Websites are the same as in Figure 6.

		Follow-up Publishers									
		A	B	C	D	E	F	G	H	I	J
First Publisher	A	0.054	0.093	0.091	0.079	0.072	0.079	0.067	0.052	0.046	0.052
	B	0.093	0.053	0.089	0.077	0.071	0.076	0.066	0.048	0.050	0.049
	C	0.087	0.088	0.052	0.076	0.072	0.067	0.067	0.048	0.043	0.047
	D	0.077	0.073	0.082	0.044	0.077	0.073	0.064	0.049	0.050	0.046
	E	0.072	0.068	0.082	0.076	0.044	0.067	0.066	0.046	0.049	0.044
	F	0.071	0.072	0.088	0.076	0.072	0.039	0.066	0.047	0.041	0.048
	G	0.074	0.073	0.087	0.076	0.073	0.073	0.075	0.051	0.060	0.048
	H	0.051	0.049	0.081	0.067	0.070	0.071	0.062	0.028	0.042	0.047
	I	0.039	0.044	0.075	0.068	0.073	0.059	0.071	0.042	0.046	0.039
	J	0.048	0.048	0.086	0.069	0.064	0.065	0.064	0.047	0.041	0.029
	Sum	0.667	0.661	0.813	0.707	0.687	0.668	0.667	0.460	0.467	0.450

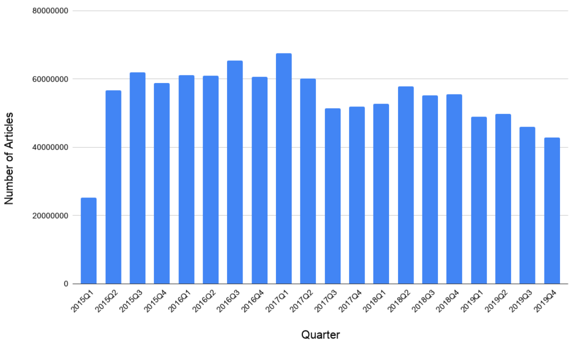


Figure 5: The diagram shows the number of articles observed by quarter.

rather than *number of articles*. Co-reporting is thus defined as:

$$c_{ij} = \frac{e_{ij}}{e_i + e_j - e_{ij}}$$

Since a dense representation requires only about 1.8 GB, this is the most efficient way of of computing the matrix due to the large number of updates. Even if more news sources were tracked by GDELT, the resources of modern hardware are likely sufficient for following this approach.

Furthermore, not all sources are reporting actively over time, as shown in Figure 3. Therefore, even if the number of sources is large, a global co-reporting matrix can be assembled from smaller matrices that cover only a limited time span. These matrices can then be compressed into a sparse format and assembled into a larger sparse matrix. Only in social networks where the number of sources, i.e. participants of discussion, is much larger than in the online news sphere would it be necessary to build a co-reporting matrix using sparse data structures.

The disadvantage of co-reporting is that it does not take time into account, i.e. it does not distinguish between who published first and who followed. Thus, we define follow-reporting f_{ij} as the *number of articles* published by site j on an event that site i published on before n_{ij}, divided by the number of articles that j published, i.e.:

$$f_{ij} = \frac{n_{ij}}{n_j}$$

Table IV gives the f_{ij} values among the 10 top publishers. Their sum shows the fraction of articles that are follow-reporting among the Top 10. As indicated by the results in Figure 6, this ratio is quite high, as one would expect for for the websites with the most articles. It also includes articles that follow up on reporting by the same news source. The corresponding rate is listed on the diagonal. Interestingly, the f_{ij} values for the Top 5 are also relatively balanced, indicating that there is no particular direction in the reporting, i.e. each site is roughly as often leader as it is follower.

We also visualize the follow-reporting matrix of the 50 most productive news websites identified in Section VI-A. Results are shown in Figure 7. We observe heavy follow-reporting among the top publishers from Table IV, some co-reporting between those and the rest, and low co-reporting among the rest. Considering that most of the top publishers are owned by the same media group, it is not surprising that they frequently report on the same topic. When tracking or predicting the spread of news, one should consider that such clusters will behave quite differently from independent groups of newspapers. More clusters of heavily co-reporting and likely co-owned news websites can be found by applying clustering algorithms (e.g. Markov clustering [30]) to the co-reporting matrix. Being symmetric, the co-reporting matrix is better suited for finding such clusters than the follow-reporting matrix.

C. Co-Reporting between Countries

The co-reporting matrix can also be used to analyze the connectedness of the news spheres of different countries. The results show which regions have a large overlap in the events that the main news sites report on, and which regions are more separate. GDELT itself does not provide information about the location of the news sources. Thus, we assign each news website a country based on its top-level

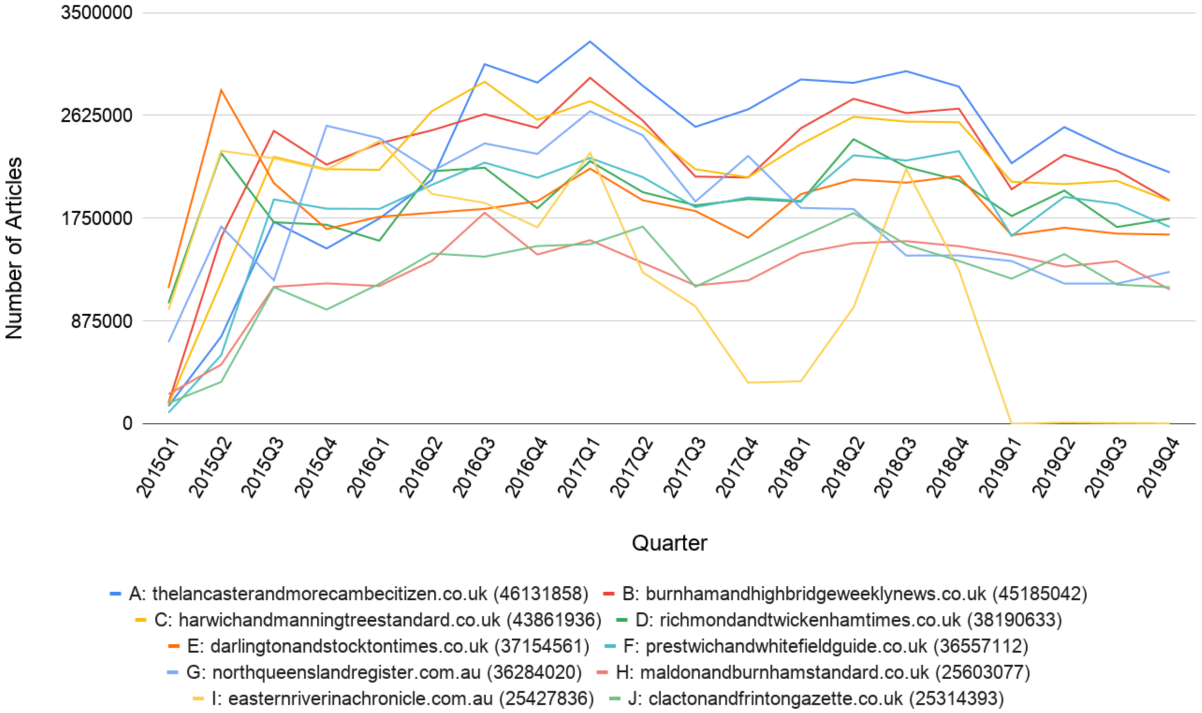


Figure 6: The diagram shows the number of articles for ten top publishers observed by the quarters. The numbers in brackets correspond to the total number of articles published by a publisher during all observation period.

Table V: Common Reporting between World Regions. Note that values are rounded.

	UK	USA	Australia	India	Italy	Canada	South Africa	Nigeria	Bangladesh	Philippines
UK		0.113	0.091	0.016	0.003	0.003	0.002	0.001	0.001	0
USA	0.113		0.103	0.02	0.004	0.004	0.002	0.001	0.001	0
Australia	0.091	0.103		0.028	0.006	0.006	0.004	0.001	0.001	0.001
India	0.016	0.02	0.028		0.02	0.02	0.009	0.006	0.006	0.002
Italy	0.003	0.004	0.006	0.02		0.014	0.005	0.003	0.004	0.003
Canada	0.003	0.004	0.006	0.02	0.014		0.01	0.006	0.006	0.001
South Africa	0.002	0.002	0.004	0.009	0.005	0.01		0.002	0.006	0.002
Nigeria	0.001	0.001	0.001	0.006	0.003	0.006	0.002		0.003	0.002
Bangladesh	0.001	0.001	0.001	0.006	0.004	0.006	0.006	0.003		0.002
Philippines	0	0	0.001	0.002	0.003	0.001	0.002	0.002	0.002	

domain. While this method is not entirely accurate (e.g. the British newspaper Guardian has *www.theguardian.com* as its base URL.), it was used due to a high number of news sources and a limited amount of man-power available for manual labeling of news sources' countries of origin.

Results are shown in Table V. As expected, we find a strong cluster between UK, USA, and Australia, with India having a somewhat weaker connection to the three. Interesting observing that Canada is not part of that cluster. The other countries seem to have much weaker co-reporting between them.

D. Connection between Reporting Country and Event Location

In addition to co-reporting, we analyze the reporting of the regions on events happening in different regions. Results are shown in Table VI. Unlike Section VI-C, we show the number of articles from region i reporting on events that happened in region j. As a consequence, the matrix is asymmetric. The Top 10 reporting countries are the same as in Section VI-C, while the countries reported on are the Top 10 by the total number of events recorded in that category, which includes non English-speaking countries.

Figure 8 gives an expanded view on the same data by showing cross-reporting for 50 countries using a log scale. Clearly, countries outside the Top 10 contribute little to the global English-speaking news. However, the bright first row

Table VI: The country-cross-reporting matrix for the ten most reported on (the country in which the event has happened) and most publishing (the country in which an article on the event was published) countries lists the number of articles for each pair of reported-publisher countries. Reported countries are arranged by the total number of events recorded. Publisher countries are arranged by the total number of articles recorded. Numbers represent the number of articles.

		Publishing Country									
		UK	USA	Australia	India	Italy	Canada	South Africa	Nigeria	Bangladesh	Philippines
Reported on Country	USA	188162540	142232473	63996675	4311031	798066	898807	393876	222035	128250	54194
	UK	24920353	16115992	7097182	634075	121076	99963	57203	17553	19274	6119
	India	12875274	8927574	4530107	369358	59451	49874	37845	8062	13758	4664
	China	11618761	8732422	4834084	331005	54289	55868	35289	9261	13349	5829
	Australia	13358384	10142138	8798879	290641	58434	58895	44547	8247	13086	15016
	Canada	10656237	8689060	4181162	261563	58742	80827	21567	9442	8898	2907
	Nigeria	6621603	4663538	2318310	163855	29339	25725	18608	7735	5934	2401
	Russia	14517877	10362300	4403051	366765	62420	61249	34305	18068	11068	3019
	Israel	12211243	8384841	3725686	329520	51075	47105	31785	10701	8960	3247
	Pakistan	6432568	4467034	2239956	169439	28122	23395	17276	5335	5902	2877

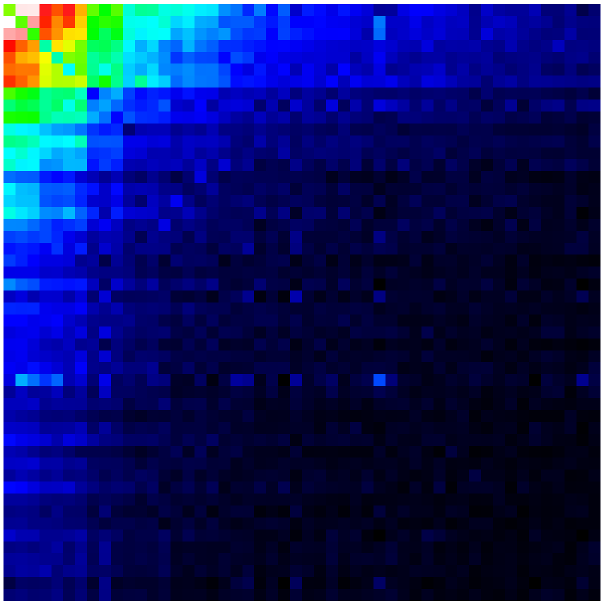


Figure 7: The diagram shows follow-reporting matrix for the fifty most productive news websites (represented in the same order in rows and cols).

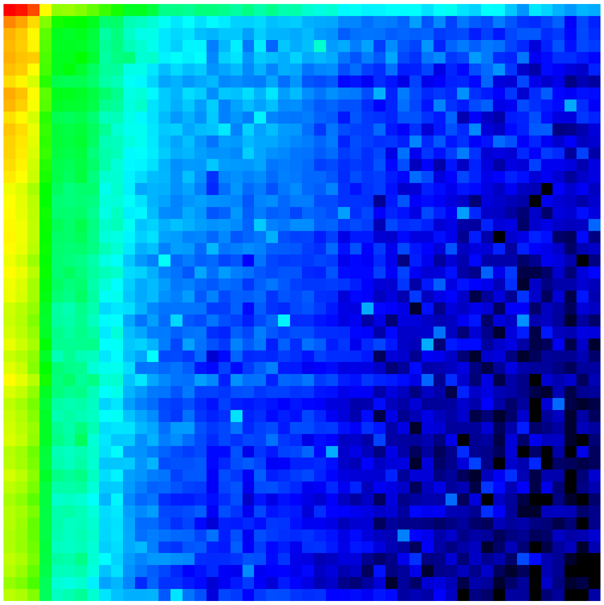


Figure 8: The diagram shows the countries-cross-reporting matrix for fifty most reported on and most publishing countries (represented in rows and cols respectively). Reported countries are represented as rows and arranged by the total number of events recorded. Publisher countries are represented as cols and arranged by the total number of articles recorded.

indicates that almost all of the 50 countries report heavily on the US.

In addition, in Table VII we show the Top 10 numbers as percentages. The numbers confirm that the US has a disproportionate presence in the news of the entire English speaking world. Similar to SectionVI-A, the UK is highly active as a source, but less so as a target of reporting. Interestingly, the percentages are relatively close for the different publishing countries, i.e. there is a large consensus on which countries' events are newsworthy. Note however that not all events have geotagging, and a large number of local news is not tagged in this way since it is assumed that the reader of a local newspaper knows the context. Thus, the numbers can provide relative, but not absolute information.

E. *Publishing Delay*

Our primary question about today's online news world is the speed with which articles are published, i.e. the delay between an event happening and the news reporting it. A common hypothesis is that the pressure to publish quickly favours the spread of misinformation. Bago et al. [31] describe evidence for the fact that at least for news consumers, deliberation, and thus time, reduces trust in misinformation. If this connection also holds for journalists, then the news becoming faster would be a likely cause of at least some of

Table VII: The fractional country-cross-reporting matrix for the ten most reported on (the country in which the event is happened) and most publishing (the country in which an article on the event was published) countries lists the number of articles for each pair of reported-publisher countries. Reported countries are arranged by the total number of events recorded. Publisher countries are arranged by the total number of articles recorded. Numbers represent the percentage of all articles from the publisher country reporting on the other countries.

		Publishing Country									
		UK	USA	Australia	India	Italy	Canada	South Africa	Nigeria	Bangladesh	Philippines
Reported on Country	USA	39.67	40.99	38.78	37.59	37.3	42.78	34.36	47.4	34.53	33.34
	UK	5.25	4.64	4.3	5.53	5.66	4.76	4.99	3.75	5.19	3.76
	India	2.71	2.57	2.75	3.22	2.78	2.37	3.3	1.72	3.7	2.87
	China	2.45	2.52	2.93	2.89	2.54	2.66	3.08	1.98	3.59	3.59
	Australia	2.82	2.92	5.33	2.53	2.73	2.8	3.89	1.76	3.52	9.24
	Canada	2.25	2.5	2.53	2.28	2.75	3.85	1.88	2.02	2.4	1.79
	Nigeria	1.4	1.34	1.4	1.43	1.37	1.22	1.62	1.65	1.6	1.48
	Russia	3.06	2.99	2.67	3.2	2.92	2.92	2.99	3.86	2.98	1.86
	Israel	2.57	2.42	2.26	2.87	2.39	2.24	2.77	2.28	2.41	2
	Pakistan	1.36	1.29	1.36	1.48	1.31	1.11	1.51	1.14	1.59	1.77

the increase in misinformation that was perceived in recent years. While establishing such a connection is beyond the scope of this paper, measuring the reporting delay over time is an important step for doing so.

The GDELT database denotes the time of all events it records. However, it does not record the publishing times of the mentions it gathers. While many news websites report publishing times to the minute, some prominent sites such as the New York Times (as of February 2020) do not. And even among those that do, reporting publishing times does not always follow a common standard. Therefore, setting up an automatic system for querying this information for more than 20,000 news sources would require immense effort.

However, the publishing time can be gauged from the 15 minute increment during which a news article was scraped and added to the database. This provides information with the best accuracy one can obtain at this time for the such a wide news sources coverage.

We record the average, minimum, median, and maximum publishing delay for each source. Average and median publishing delays represent, respectively, the average and median numbers of 15 minute intervals that pass between the event and the article mentioning it, over all articles published by that news source. Results are shown in Figure 9. With respect to the minimum delay, we see that about half the news sites have reported on at least one event within 15 minutes. The other half seems to roughly follow a power law, although very few seem to take longer than 96 intervals, i.e. 24 hours. There is however a group of outliers with a delay of more than 30000, which amounts to roughly one year.

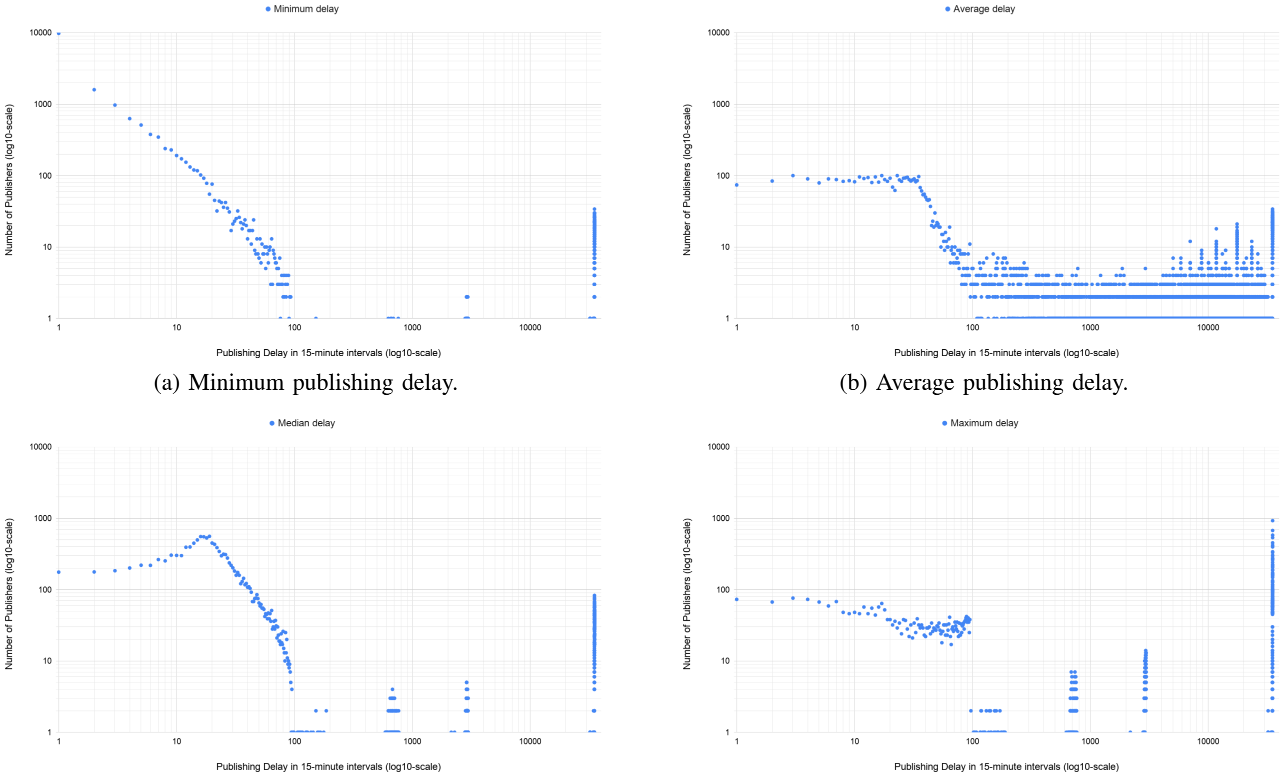
The average values are much more spread out, with a significant group reporting on events that are three month in the past on average. However, most fall within the window of 2 to 8 hours after the event. On the other hand, the maximum delays offer a relatively clear picture. The majority of publishers follows a 24 hour news cycle, which means that their maximum delay is close to 96 intervals, i.e. 24 hours. In addition to that, we clearly see three groups which correspond to a week, month, and year, which implies that many online publication still follow the formats established for print media. Finally, the median values reinforce the picture presented by the averages, which are somewhat distorted by articles that are published weeks after the event. A clear peak of publishers with a median delay of 4 to 5 hours emerges, with a rapid decay towards the 24 hour limit.

The data shows that while quickly written or copied articles are common, they do not seem to make up the majority of online news. We can roughly group online news sources into three categories. A relatively large slow group that reports on topics that are days or months in the past, a large average group that roughly follows the 24 hour news cycle with a median delay of about 5 hours, and a fast group, that typically reports in less than 2 hours. While smaller, there are several hundred publishers in that last group, and when studying the spread of news, especially digital wildfires, these represent a most important pool of core news sources that are as close to real time reporting as possible.

Table VIII: The publication delay statistic for ten most productive news websites. Publishers are the same as in Figure 5.

	Delay			
Publisher	Min	Max	Average	Median
A	1	35135	39	**16**
B	1	35135	39	**16**
C	1	35135	40	**16**
D	1	35135	39	**15**
E	1	35135	39	**15**
F	1	35135	41	**16**
G	1	35135	41	**13**
H	1	35135	46	**16**
I	1	35135	37	**14**
J	1	35135	48	**16**

In Table VIII, we give the minimum, maximum, average, and median publishing delay for the Top 10 news sites we found in Section VI-A. All of them belong to the average



(a) Minimum publishing delay.

(b) Average publishing delay.

(c) Median publishing delay.

(d) Maximum publishing delay.

Figure 9: The figures show the minimum, average, and maximum article publication delay.

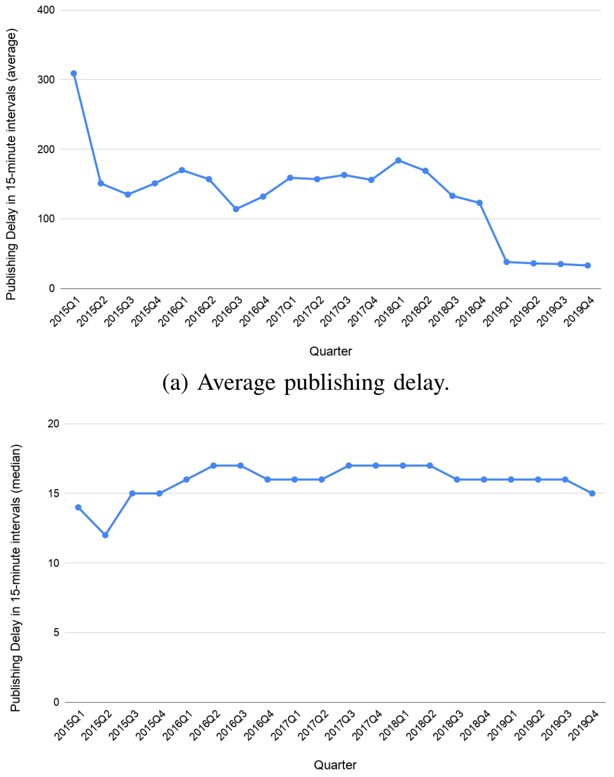
speed group, with a median delay of about 4 hours. Each of them has at least one article that appeared exactly one year after the event being reported, the average is highly skewed. This reveal an important follow-up GDELT data research aimed to both wildfires and fake news detection topics. Observed delay for the very first article from any source on a particular topic might be relevant to reporting speediness and potential news wildfires. Repeated articles on an event by a single source might very well be an indicator of thorough and responsible reporting. However, it could also be an indication of intentional spreading of misinformation.

F. Relation between Article Delay and Publishing Frequency

Finally we present the development of publishing delay over time. For each quarter since the beginning of 2015, we compute both the average and the median publishing delay over all articles published during that quarter. Results are shown in Figure 10. While Subfigure 10a shows a clear decline in average delay, especially in 2019. On the other hand, the median values in Subfigure 10b seem to be quite stable. Based on our analysis in Section VI-E, it is likely that the decrease in average value is due to a decrease in the number of high delay articles. As mentioned earlier, in-depth analysis of publishing delays will be performed in the follow-up research. At the moment, we check this by simply counting the number of articles with a publishing delay of more than one day, i.e. those that fall outside the 24 hour news cycle. The results can be found in Figure 11. We can clearly see a significant decrease in the number of these articles which does at least partially explain the reduction. We note however that the reduction does not entirely match up with the decay of the average, and even the median has decreased in the last quarter. Thus, as of now, we cannot establish a global increase in publishing speed from that data, although this does not preclude the existence of such a trend or the possibility of such a trend is currently starting.

G. System performance

A single aggregated query was used to obtain all data presented in Tables V, VI and VII. In single-threaded processing, the query took 344 seconds. The use of OpenMP-enabled implementation improved this to 43 seconds (see Figure 12). However, as one can see from the plot, the parallel performance improvement was hampered due to the need for I/O operations in single-node mode, and has some opportunities for subsequent improvements using distributed MPI-based computations.



(a) Average publishing delay.

(b) Median publishing delay.

Figure 10: Aggregated quarterly publishing delay in 15-minute intervals.

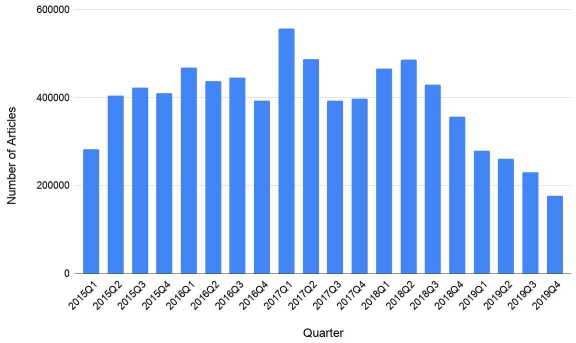


Figure 11: Number of articles with publishing delay greater than 24 hours.

VII. Conclusions and Future Work

We have presented a system capable of rapidly analyzing the entire contents of the GDELT 2.0 Database. We focused on the Events and Mentions table to uncover trends in the world of online news during the last four years. Such a system is extremely helpful in running large scale analyses on more than a billion news articles. While there are several alternatives for accessing this data, they have significant limitations which we aim to avoid. The massive size of the GDELT data collection does create the additional problem that the objects it tracks are very heterogeneous. Different news sources operate by completely different standards, and articles range from a few sentences to several pages. As a result, the Top 10 publishers by volume are local newspapers that push a large number of short articles which are hardly comparable to e.g. those of the *New York Times*. In the future we will work on identifying the different types of sources and articles.

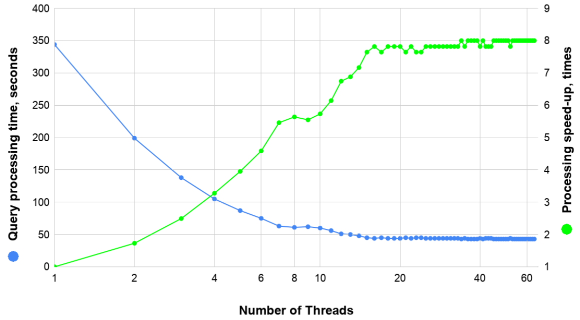


Figure 12: The plot shows the performance evaluation of the OpenMP-based parallel implementation of query execution engine.

Furthermore, we aim to add a Python interface for ease of use. The platform will allow more in-depth investigation of online news using clustering and graph discovery methods in order to discover and understand the spread of news, and in that manner learn more about the spread of fake news. Finally, we will extend the analysis to the non English-speaking world. It is expected that this will require adding distributed memory capabilities using MPI to handle the substantial amount of additional data.

Acknowledgment

This work is founded by the Norwegian SAMRISK-2 project ”UMOD” (#272019). The research presented in this paper has benefited from the Experimental Infrastructure for Exploration of Exascale Computing (eX3), which is financially supported by the Research Council of Norway under contract 270053.

References

[1] W. E. Forum, http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/, 2013.

[2] L. Wu and H. Liu, “Tracing fake-news footprints: Characterizing social media messages by how they propagate,” in *Proceedings of the eleventh ACM international conference on Web Search and Data Mining*, 2018, pp. 637–645.

[3] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[4] K. Leetaru and P. A. Schrodt, "Gdelt: Global data on events, location, and tone, 1979–2012," in *ISA annual convention*, vol. 2, no. 4. Citeseer, 2013, pp. 1–49.

[5] J. E. Yonamine, "Predicting future levels of violence in afghanistan districts using gdelt," *Unpublished manuscript*, 2013.

[6] H. Kwak and J. An, "A first look at global news coverage of disasters by using the gdelt dataset," in *International Conference on Social Informatics*. Springer, 2014, pp. 300–308.

[7] M. D. Ward, A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford, "Comparing gdelt and icews event data," *Analysis*, vol. 21, no. 1, pp. 267–297, 2013.

[8] H. Kwak and J. An, "Two tales of the world: Comparison of widely used world news datasets gdelt and eventregistry," in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[9] E. Boudemagh and I. Moise, "News media coverage of refugees in 2016: A gdelt case study," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[10] J. E. Yonamine, "A nuanced study of political conflict using the global datasets of events location and tone (gdelt) dataset," 2013.

[11] F. Peng, J.-b. GAO *et al.*, "Dynamic evolution of multinational relation's network in the south china sea arbitration based on massive media data analysis," *DEStech Transactions on Computer Science and Engineering*, no. pcmm, 2018.

[12] K. Chen, "Big data analysis: Trump effect on trade narratives," 2017.

[13] M. Bertl, "News analysis for the detection of cyber security issues in digital healthcare," *Young Information Scientist*, vol. 4, pp. 1–15, 2019.

[14] S. Brazys and A. Dukalskis, "Rising powers and grassroots image management: Confucius institutes and china in the media," *The Chinese Journal of International Politics*, vol. 12, no. 4, pp. 557–584, 2019.

[15] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, and H. Wang, "Predicting social unrest events with hidden markov models using gdelt," *Discrete Dynamics in Nature and Society*, vol. 2017, 2017.

[16] D. Galla and J. Burke, "Predicting social unrest using gdelt," in *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2018, pp. 103–116.

[17] R. Alamro, A. McCarren, and A. Al-Rasheed, "Predicting saudi stock market index by incorporating gdelt using multivariate time series modelling," in *International Conference on Computing*. Springer, 2019, pp. 317–328.

[18] Y. Keneshloo, J. Cadena, G. Korkmaz, and N. Ramakrishnan, "Detecting and forecasting domestic political crises: A graph-based approach," in *Proceedings of the 2014 ACM conference on Web science*, 2014, pp. 192–196.

[19] H. Roos, A. N. Usanov, N. Farnham, and T. Sweijs, "Improving the early warning function of civil war onset models using automated event data," *Available at SSRN 2948293*, 2018.

[20] X. Lu and B. Szymanski, "Predicting viral news events in online media," in *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2017, pp. 1447–1456.

[21] J. Rappaz, D. Bourgeois, and K. Aberer, "A dynamic embedding model of the media landscape," in *The World Wide Web Conference*, 2019, pp. 1544–1554.

[22] K. M. Al-Naami, S. Seker, and L. Khan, "Gisqf: An efficient spatial query processing system," in *2014 IEEE 7th International Conference on Cloud Computing*. IEEE, 2014, pp. 681–688.

[23] ——, "Gisqaf: Mapreduce guided spatial query processing and analytics system," *Software: Practice and Experience*, vol. 46, no. 10, pp. 1329–1349, 2016.

[24] D. Xie, F. Li, B. Yao, G. Li, L. Zhou, and M. Guo, "Simba: Efficient in-memory spatial analytics," in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 1071–1085.

[25] M. Shukla, R. Dos Santos, F. Chen, and C.-T. Lu, "Discrn: A distributed storytelling framework for intelligence analysis," *Big data*, vol. 5, no. 3, pp. 225–245, 2017.

[26] J. McAlpin, "The stream2 benchmark reference information," *Available from: University ofVirginia, Department of Computer Science Web site: http://www. cs. virginia. edu/stream/ref. html [Accessed: February 8, 2009]*.

[27] F. Filloux, "Google news: the secret sauce," https://www.theguardian.com/technology/2013/feb/25/1, February 2013, retrieved February 19, 2020.

[28] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[29] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.

[30] S. Dongen, "A cluster algorithm for graphs," 2000.

[31] B. Bago, D. G. Rand, and G. Pennycook, "Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines." *Journal of experimental psychology: general*, 2020.