A System for High Performance Mining on GDELT Data

Johannes Langguth

Simula Research Laboratory, Oslo, Norway

joint work with
Konstantin Pogorelov, Daniel Thilo Schroeder, and Petra Filkukova
(Simula Research Laboratory)

Studying Fake News

- Following the 2016 US presidential elections, the topic of Fake News has been studied intensively
- Highly interdisciplinary field with many different sub-topics
- Crucial contribution of computer science: large scale quantitative analysis
- Mostly focused on social media

Studying Fake News on Social Media

- Many social networks not open for research
- The majority of research focusses on Twitter
- Allows the analysis of structural information (Followers, retweets, etc.)
- Textual information limited due to 140/280 character limit

Studying Fake News on Websites

- A lot of fake news doe not come from Twitter, but from news websites
- News websites typically have more text, but no structural information
- Unlike Twitter etc., there is no single API to collect information from news websites
- However, an existing project

Studying Fake News on Websites

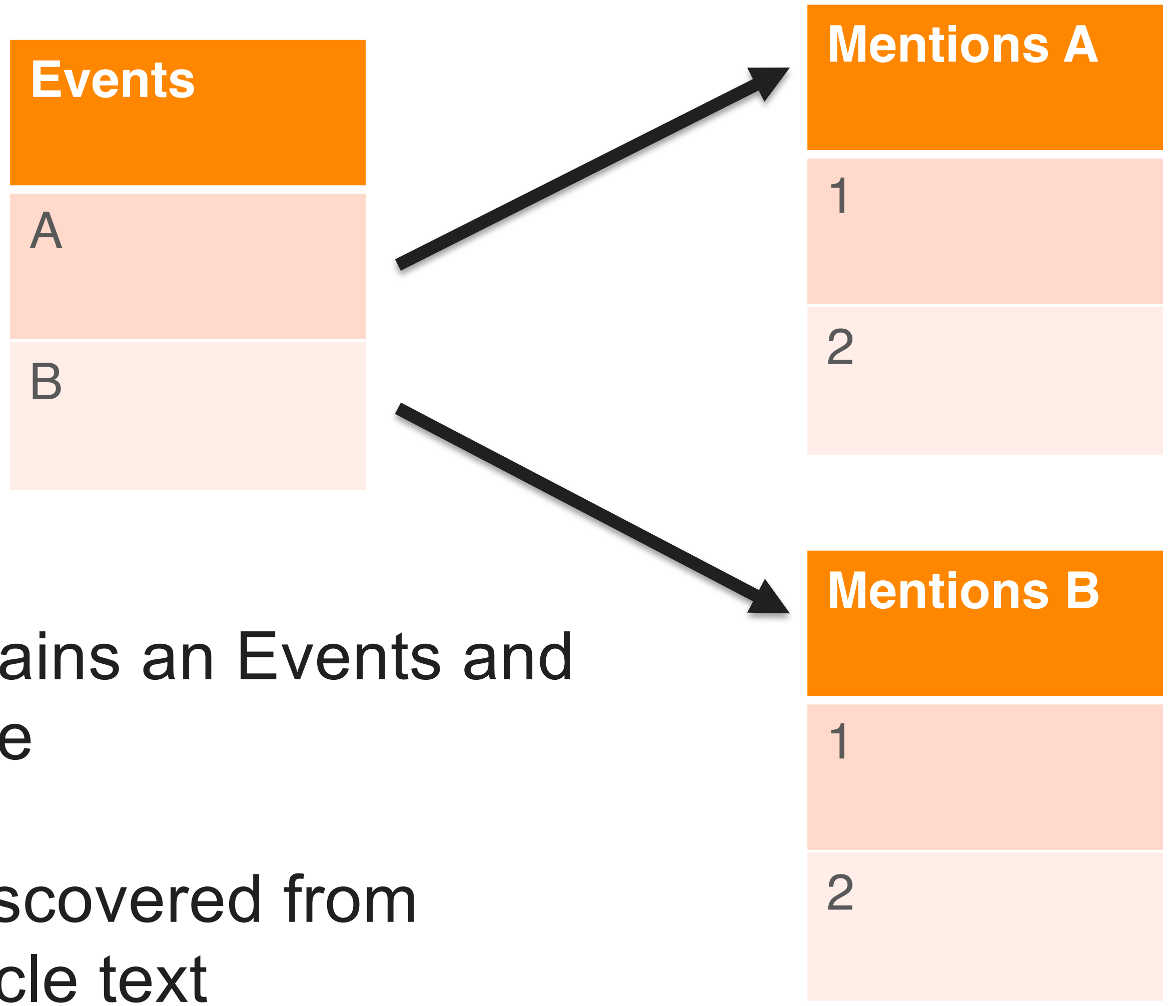
- A lot of fake news doe not come from Twitter, but from news websites
- News websites typically have more text, but no structural information
- Unlike Twitter etc., there is no single API to collect data from news websites
- However, an existing project offers a wealth of information

The GDELT Project

The Global Database on Events, Language and Tone collects news articles worldwide

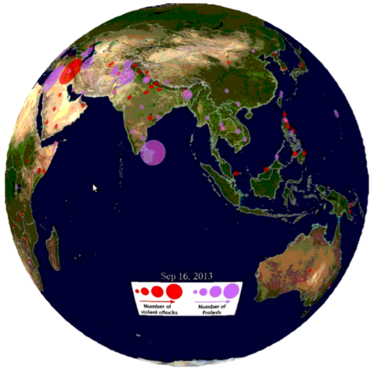
The current 2.0 version has collected data since 2015

Events and Mentions in GDELT



- GDELT maintains an Events and Mentions table
- Events are discovered from analyzing article text
- Each event typically has multiple mentions

GDELT Data Collection



- Every 15 minutes, GDELT 2.0 publishes a new events and mentions file
- Since 2015, more than a billion articles registered
- Data is freely available
- Main problem: how to analyze large amounts of data efficiently ?

Analyzing GDELT Data

- GDELT data is available in Google BigQuery and Amazon S3

- Neither is currently being updated

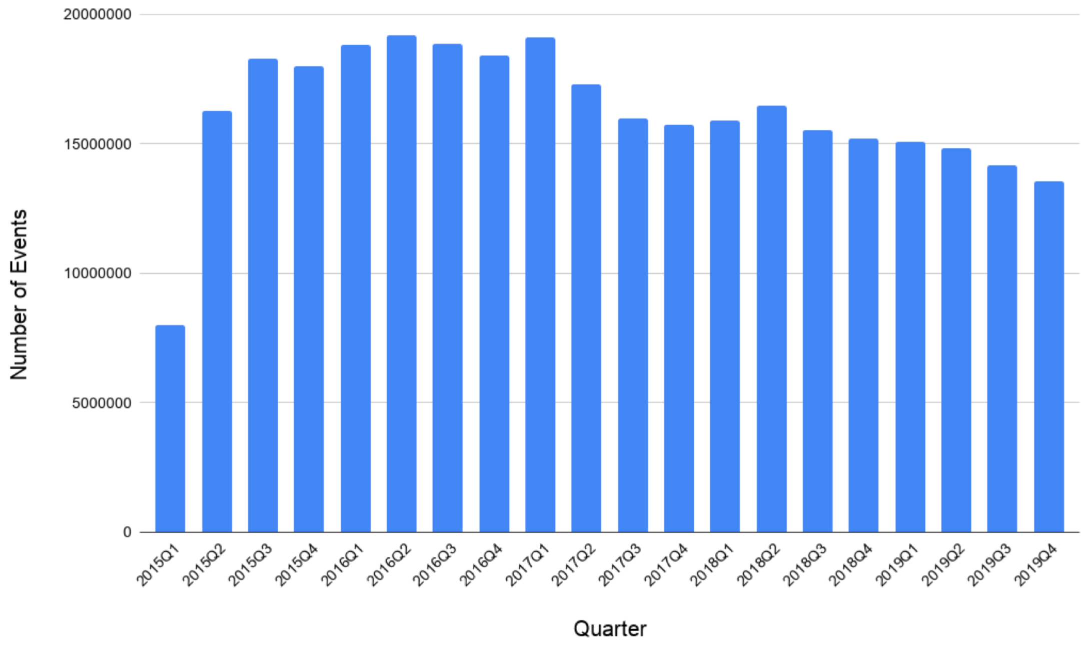
- Business model of Google BigQuery makes it very hard to estimate cost. Not suited for continuous trend detection.

- Our goal: develop a lightweight, efficient system for in-memory processing on large memory nodes

Parallel GDELT Data Analysis

- Tool is written in C++ / OpenMP, tuned for performance
- Can select which data fields are stored in memory. Allows control of memory consumption
- Most analyses require only limited information per article
- Relies on scripts that clean and preprocess the data
- Use system to detect trend in the news landscape

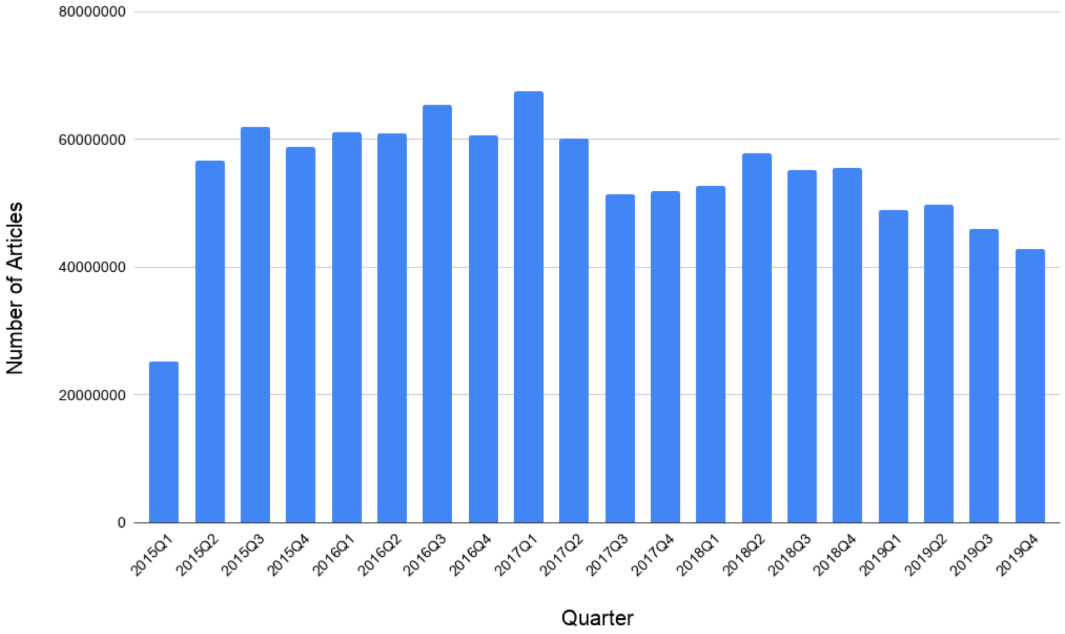
Detecting Trends in Global News

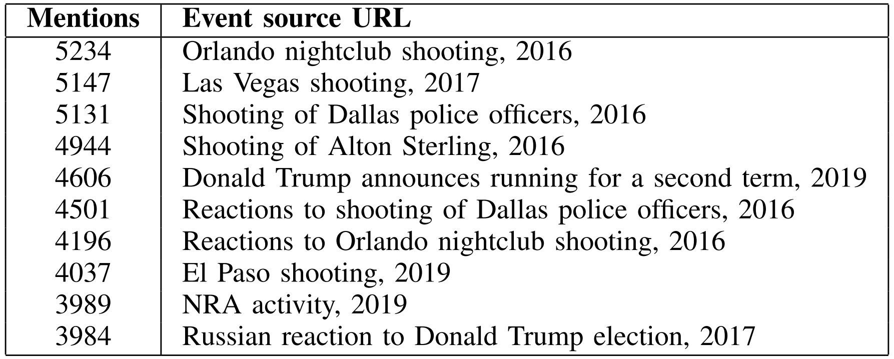


Number of	**Value**
Sources	20,996
Events	324,564,472
Capture intervals	168,266
Articles	1,090,310,118
Minimum number of articles per event	1
Maximum number of articles per event	5234
Articles per event (weighted average)	3.36

- Divide time since 2015 into quarters
- First quarter is shorter. System went live in February 2015
- Number of events tracked relatively constant

Detecting Trends in Global News



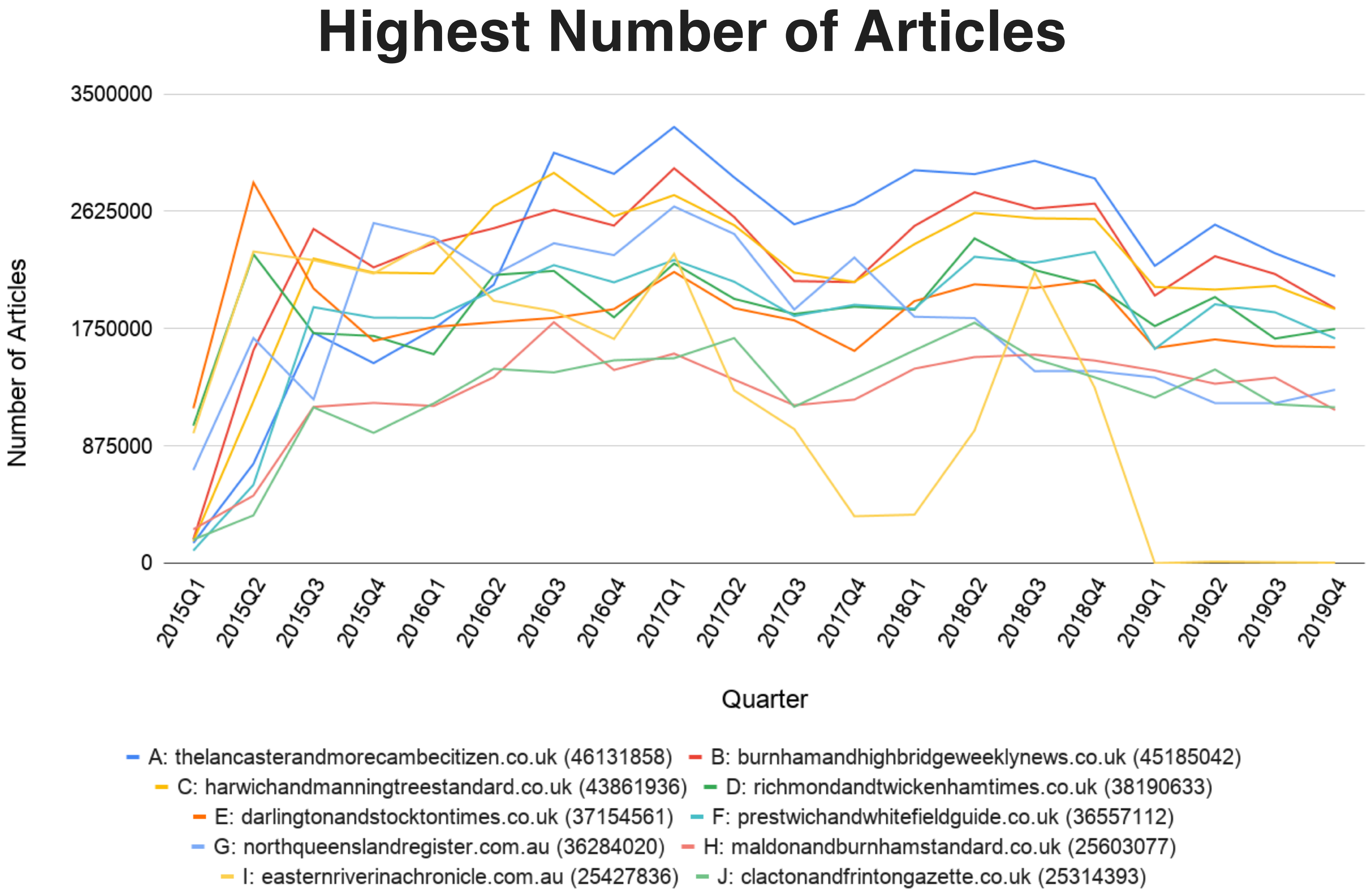


Mentions	Event source URL
5234	Orlando nightclub shooting, 2016
5147	Las Vegas shooting, 2017
5131	Shooting of Dallas police officers, 2016
4944	Shooting of Alton Sterling, 2016
4606	Donald Trump announces running for a second term, 2019
4501	Reactions to shooting of Dallas police officers, 2016
4196	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 tracked relatively constant

- Highest number of articles per event typically in US mass shootings (85% of active sources report on)

- Analyze English language only



- 7 out of 10 top publishers by number of articles belong to Newsquest Media Group (UK)
- Regional newspapers, typically short articlesster

Co-reporting

- Define co-reporting similar to Jaccard index:

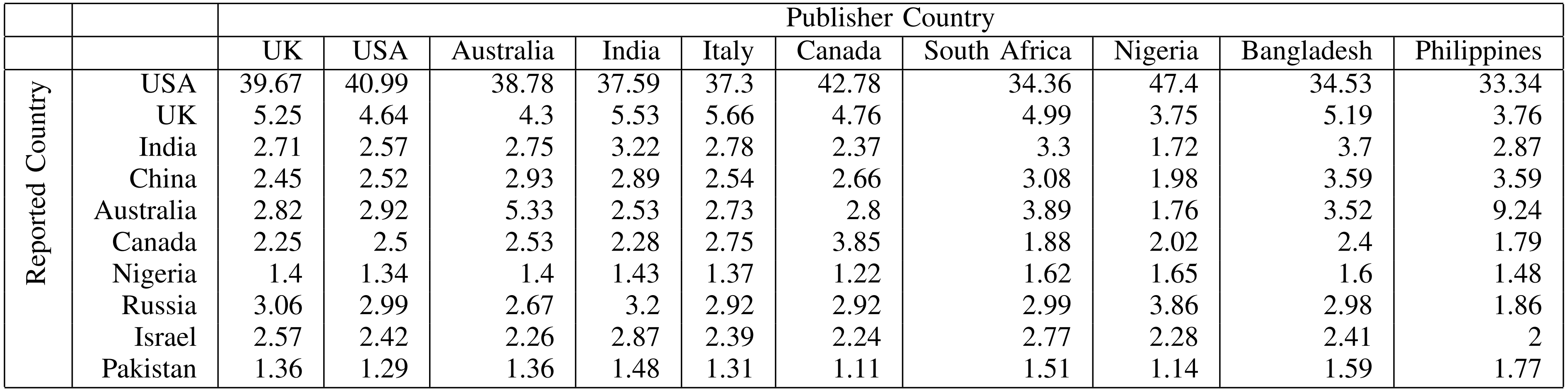
$$c_{ij} = \frac{e_{ij}}{e_i + e_j - e_{ij}}$$

- co-reporting shows strong connection between top 10 sites

- Likely vector for the spread of fake news

Reporting between countries

- Similar statistic: fraction of news about other countries:



		Publisher Country									
		UK	USA	Australia	India	Italy	Canada	South Africa	Nigeria	Bangladesh	Philippines
Reported 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

- News about USA dominate in all countries

- News topics relatively consistent internationally

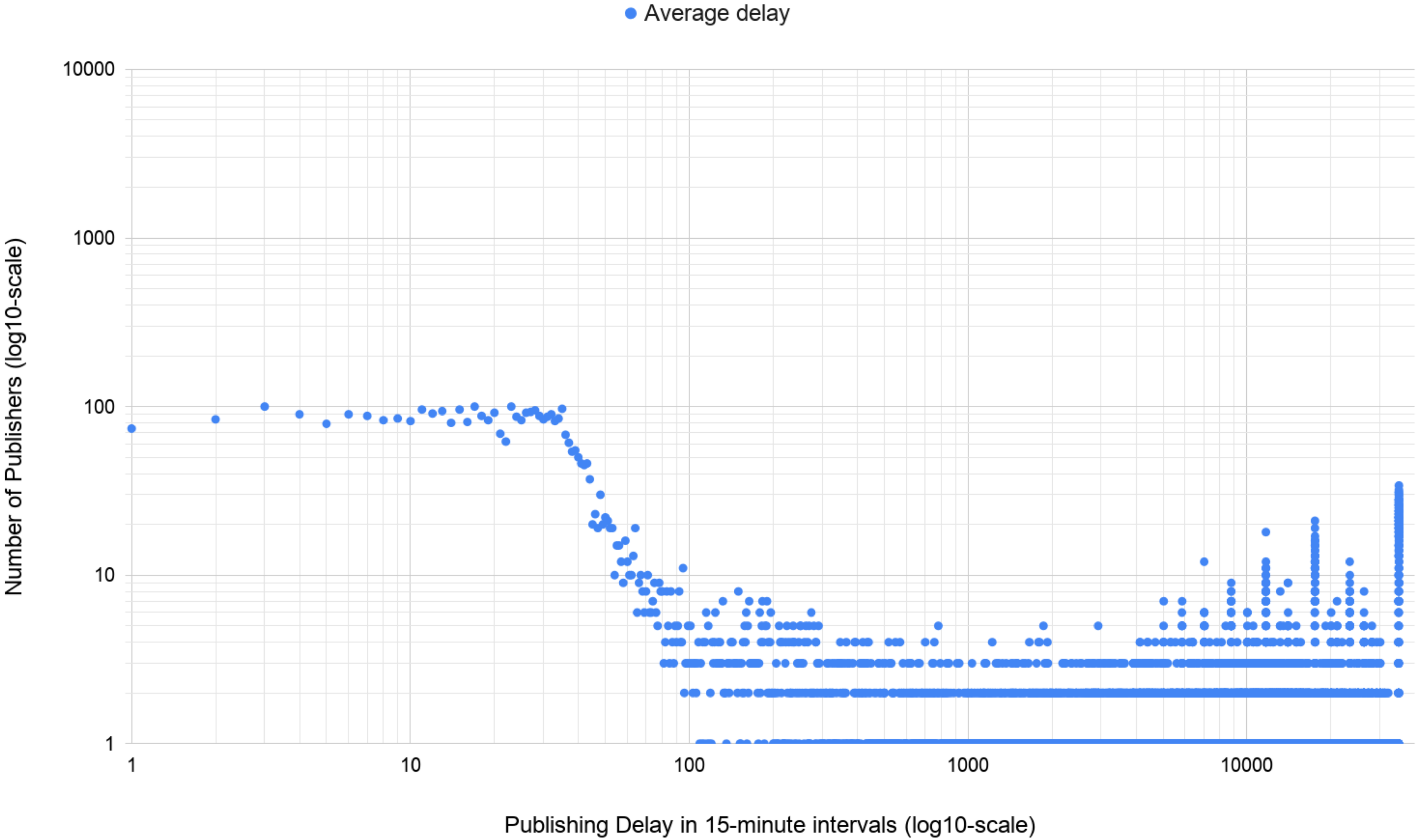
- Data is based on GDELT NLP. May contain a substantial error rate.

Is the news getting faster ?

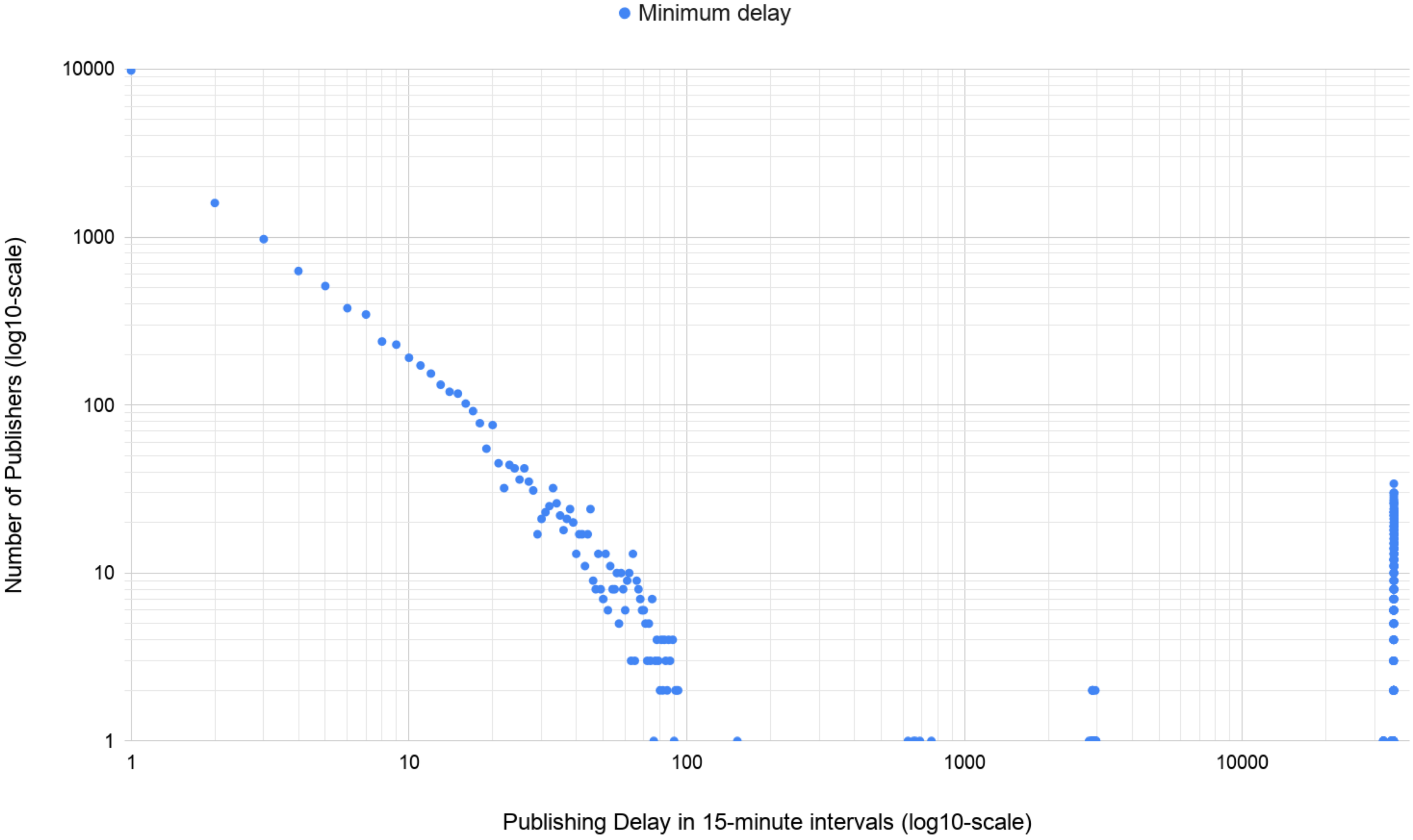
- Traditional news followed a 24 hour news cycle
- No longer true for online news
- Pressure on journalists to produce articles faster
- This pressure is often cited as a reason for lack of fact-checking/propagation of fake news
- We analyze mine GDELT to answer the question:

Is the time between an event and news that mentions it getting smaller ?

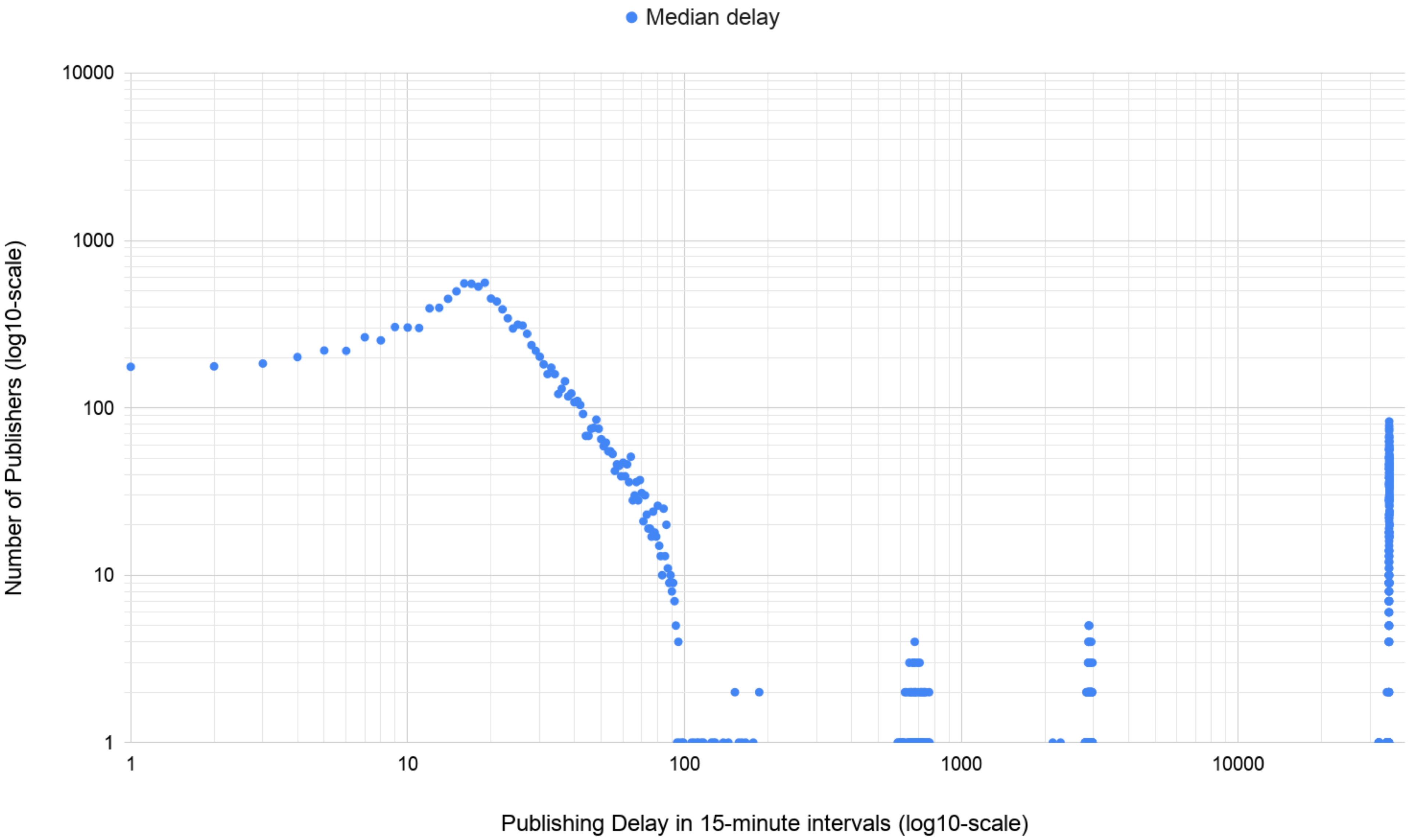
Number of Publishers by Average Delay



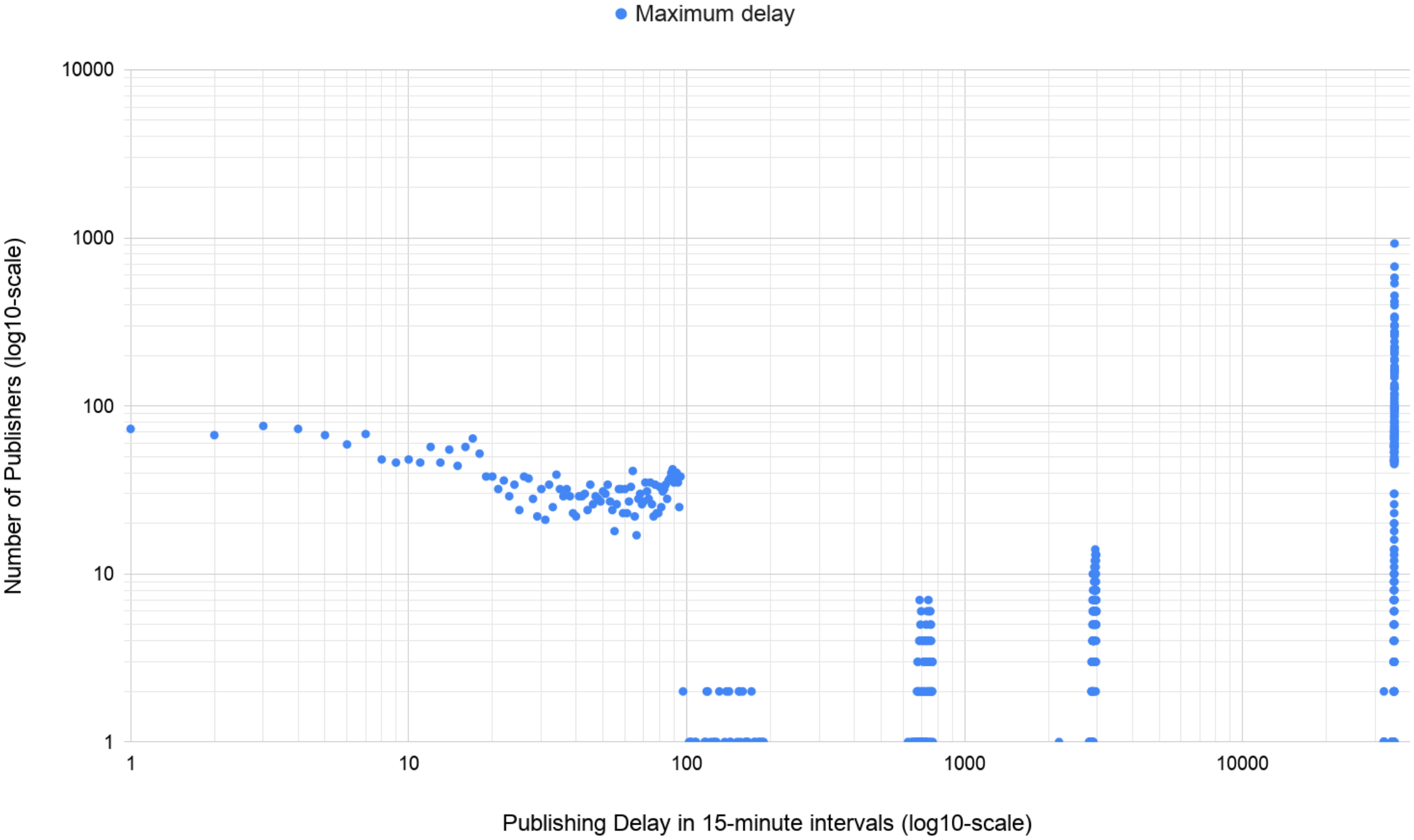
Number of Publishers by Minimum Delay



Number of Publishers by Median Delay



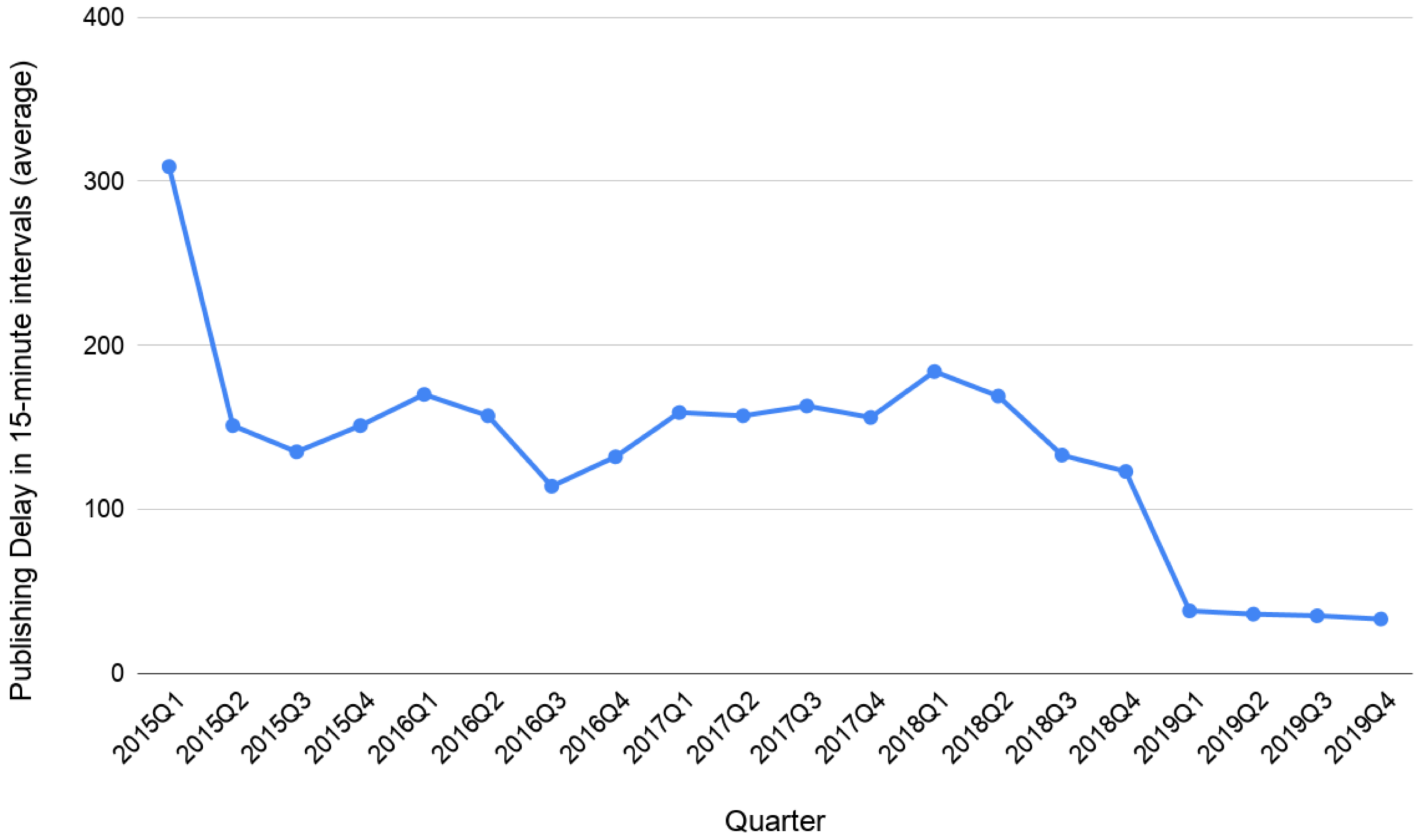
Number of Publishers by Maximum Delay



Publishing Delays

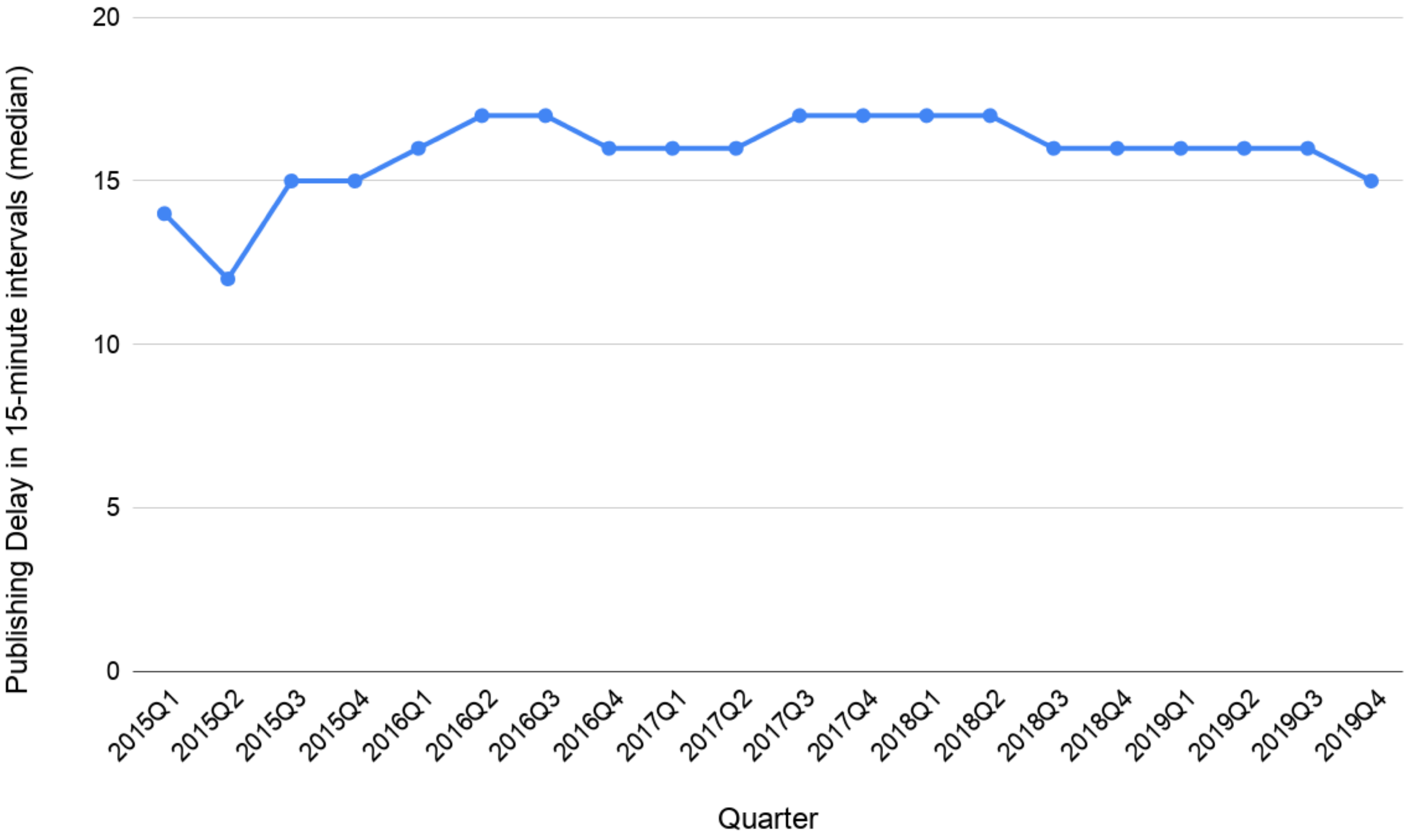
- Event time is not available. Delay is counted from the first mentioning article.
- Maximum delay clearly separates news sources into daily/weekly/monthly/yearly group
- Median peaks at about 4 hours
- Need to aggregate data by quarter

Is the news getting faster ?



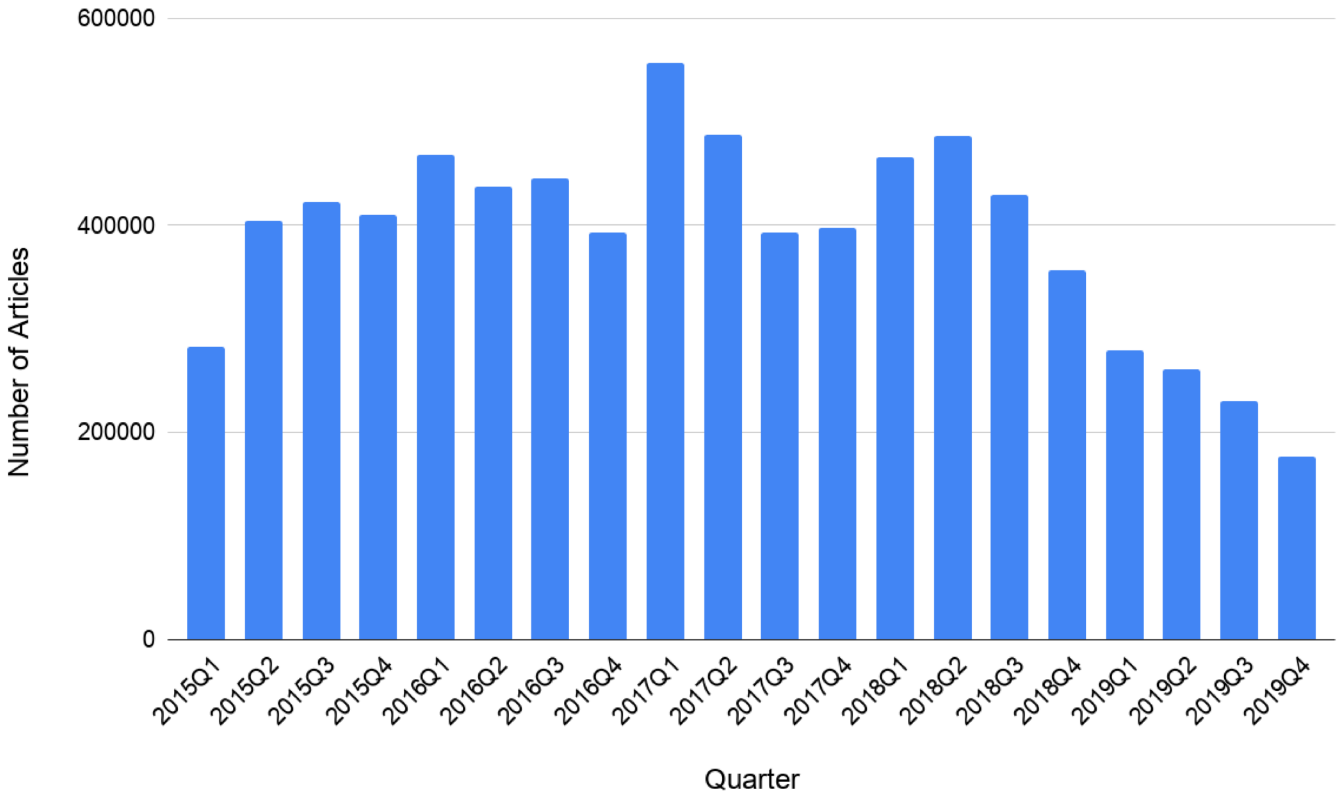
- Average publishing delay went down significantly in 2019

Is the news getting faster ?



- On the other hand, Median is quite stable

Number of Articles with more than 1 Day Delay



- Reduction of the average is due to reduction in “slow news”

- Thus, news is getting faster, although not by speeding up an already fast cycle