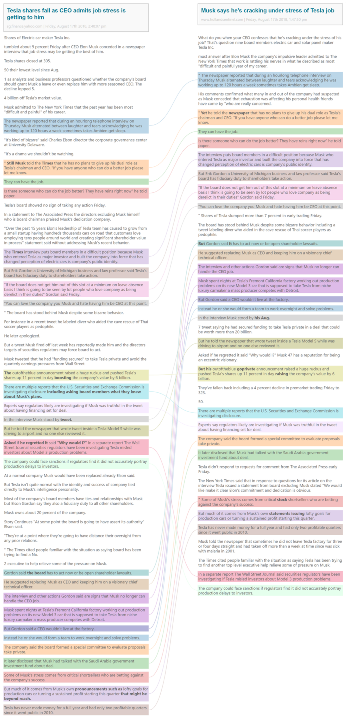
News Provenance: Revealing News Text Reuse at Web-Scale in an Augmented News Search Experience

Nathan Evans
Microsoft Research
Silverdale, WA, USA
naevans@microsoft.com

Jonathan Larson
Microsoft Research
Silverdale, WA, USA
jolarso@microsoft.com

Darren Edge
Microsoft Research
Cambridge, UK
daedge@microsoft.com

Christopher White
Microsoft Research
Redmond, WA, USA
chwh@microsoft.com



Tesla shares fall as CEO admits job stress is getting to him

sg.finance.yahoo.com | Friday, August 17th 2018, 2:48:07 pm

Shares of Electric car maker Tesla Inc.

tumbled about 9 percent Friday after CEO Elon Musk conceded in a newspaper interview that job stress may be getting the best of him.

Tesla shares closed at 305.

50 their lowest level since Aug.

1 as analysts and business professors questioned whether the company's board should grant Musk a leave or even replace him with more seasoned CEO. The decline lopped 5.

4 billion off Tesla's market value.

Musk admitted to The New York Times that the past year has been most "difficult and painful" of his career.

The newspaper reported that during an hourlong telephone interview on Thursday Musk alternated between laughter and tears acknowledging he was working up to 120 hours a week sometimes takes Ambien get sleep.

"It's kind of bizarre" said Charles Elson director the corporate governance center at University Delaware.

"It's a drama we shouldn't be watching.

" **Still Musk** told the **Times** that he has no plans to give up his dual role as Tesla's chairman and CEO. "If you have anyone who can do a better job please let me know.

They can have the job.

Is there someone who can do the job better? They have reins right now" he told paper.

Tesla's board showed no sign of taking any action Friday.

In a statement to The Associated Press the directors excluding Musk himself who is board chairman praised Musk's dedication company.

"Over the past 15 years Elon's leadership of Tesla team has caused to grow from a small startup having hundreds thousands cars on road that customers love employing tens people around world and creating significant shareholder value in process" statement said without addressing Musk's recent behavior.

The **Times** interview puts board members in a difficult position because Musk who entered Tesla as major investor and built the company into force that has changed perception of electric cars is company's public identity.

But Erik Gordon a University of Michigan business and law professor said Tesla's board has fiduciary duty to shareholders take action.

"If the board does not get him out of this slot at a minimum on leave absence basis I think is going to be seen by lot people who love company as being derelict in their duties" Gordon said Friday.

"You can love the company you Musk and hate having him be CEO at this point.

" The board has stood behind Musk despite some bizarre behavior.

For instance in a recent tweet he labeled diver who aided the cave rescue of Thai soccer players as pedophile.

He later apologized.

But a tweet Musk fired off last week has reportedly made him and the directors targets of securities regulators may force board to act.

Musk tweeted that he had "funding secured" to take Tesla private and avoid the quarterly earnings pressures from Wall Street.

The outoftheblue announcement raised a huge ruckus and pushed Tesla's shares up 11 percent in day **boosting** the company's value by 6 billion.

There are multiple reports that the U.S. Securities and Exchange Commission is investigating disclosure **including asking board members what they knew about Musk's plans.**

Experts say regulators likely are investigating if Musk was truthful in the tweet about having financing set for deal.

In the interview Musk stood by **tweet.**

But he told the newspaper that wrote tweet inside a Tesla Model S while was driving to airport and no one else reviewed it.

Asked if **he regretted it** said **"Why would I?"** In a separate report The Wall Street Journal securities regulators have been investigating Tesla misled investors about Model 3 production problems.

The company could face sanctions if regulators find it did not accurately portray production delays to investors.

At a normal company Musk would have been replaced already Elson said.

But Tesla isn't quite normal with the identity and success of company tied directly to Musk's intelligence personality.

Most of the company's board members have ties and relationships with Musk but Elson Gordon say they also a fiduciary duty to all other shareholders.

Musk owns about 20 percent of the company.

Story Continues "At some point the board is going to have assert its authority" Elson said.

"They're at a point where they're going to have distance their oversight from any prior relations.

" The Times cited people familiar with the situation as saying board has been trying to find a No.

2 executive to help relieve some of the pressure on Musk.

Gordon said **the board** has to act now or be open shareholder lawsuits.

He suggested replacing Musk as CEO and keeping him on a visionary chief technical officer.

The interview and other actions Gordon said are signs that Musk no longer can handle the CEO job.

Musk spent nights at Tesla's Fremont California factory working out production problems on its new Model 3 car that is supposed to take Tesla from niche luxury carmaker a mass producer competes with Detroit.

But Gordon said a CEO wouldn't live at the factory.

Instead he or she would form a team to work overnight and solve problems.

The company said the board formed a special committee to evaluate proposals take private.

It later disclosed that Musk had talked with the Saudi Arabia government investment fund about deal.

Some of Musk's stress comes from critical shortsellers who are betting against the company's success.

But much of it comes from Musk's own **pronouncements such as** lofty goals for production cars or turning a sustained profit starting this quarter **that might be beyond reach.**

Tesla has never made money for a full year and had only two profitable quarters since it went public in 2010.

Musk says he's cracking under stress of Tesla job

www.hollandsentinel.com | Friday, August 17th 2018, 1:47:50 pm

What do you when your CEO confesses that he's cracking under the stress of his job? That's question nine board members electric car and solar panel maker Tesla Inc.

must answer after Elon Musk the company's impulsive leader admitted to The New York Times that work is rattling his nerves in what he described as most "difficult and painful year of my career.

" The newspaper reported that during an hourlong telephone interview on Thursday Musk alternated between laughter and tears acknowledging he was working up to 120 hours a week sometimes takes Ambien get sleep.

His comments confirmed what many in and out of the company had suspected as Musk conceded that exhaustion was affecting his personal health friends have come by "who are really concerned.

" **Yet** he told the **newspaper** that has no plans to give up his dual role as Tesla's chairman and CEO. "If you have anyone who can do a better job please let me know.

They can have the job.

Is there someone who can do the job better? They have reins right now" he told paper.

The interview puts board members in a difficult position because Musk who entered Tesla as major investor and built the company into force that has changed perception of electric cars is company's public identity.

But Erik Gordon a University of Michigan business and law professor said Tesla's board has fiduciary duty to shareholders take action.

"If the board does not get him out of this slot at a minimum on leave absence basis I think is going to be seen by lot people who love company as being derelict in their duties" Gordon said Friday.

"You can love the company you Musk and hate having him be CEO at this point.

" Shares of Tesla slumped more than 7 percent in early trading Friday.

The board has stood behind Musk despite some bizarre behavior including a tweet labeling diver who aided in the cave rescue of Thai soccer players as pedophile.

But Gordon said **it** has to act now or be open shareholder lawsuits.

He suggested replacing Musk as CEO and keeping him on a visionary chief technical officer.

The interview and other actions Gordon said are signs that Musk no longer can handle the CEO job.

Musk spent nights at Tesla's Fremont California factory working out production problems on its new Model 3 car that is supposed to take Tesla from niche luxury carmaker a mass producer competes with Detroit.

But Gordon said a CEO wouldn't live at the factory.

Instead he or she would form a team to work overnight and solve problems.

In the interview Musk stood by **his Aug.**

7 tweet saying he had secured funding to take Tesla private in a deal that could be worth more than 20 billion.

But he told the newspaper that wrote tweet inside a Tesla Model S while was driving to airport and no one else reviewed it.

Asked if he regretted it said "Why would I?" Musk 47 has a reputation for being an eccentric visionary.

But his outoftheblue **goprivate** announcement raised a huge ruckus and pushed Tesla's shares up 11 percent in day **raising** the company's value by 6 billion.

They've fallen back including a 4 percent decline in premarket trading Friday to 323.

50.

There are multiple reports that the U.S. Securities and Exchange Commission is investigating disclosure.

Experts say regulators likely are investigating if Musk was truthful in the tweet about having financing set for deal.

The company said the board formed a special committee to evaluate proposals take private.

It later disclosed that Musk had talked with the Saudi Arabia government investment fund about deal.

Tesla didn't respond to requests for comment from The Associated Press early Friday.

The New York Times said that in response to questions for its article on the interview Tesla issued a statement from board excluding Musk stated "We would like make it clear Elon's commitment and dedication is obvious.

" Some of Musk's stress comes from critical **stock** shortsellers who are betting against the company's success.

But much of it comes from Musk's own **statements issuing** lofty goals for production cars or turning a sustained profit starting this quarter.

Tesla has never made money for a full year and had only two profitable quarters since it went public in 2010.

Musk told the newspaper that sometimes he did not leave Tesla factory for three or four days straight and had taken off more than a week at time since was sick with malaria in 2001.

The Times cited people familiar with the situation as saying Tesla has been trying to find another top level executive help relieve some of pressure on Musk.

In a separate report The Wall Street Journal said securities regulators have been investigating if Tesla misled investors about Model 3 production problems.

The company could face sanctions if regulators find it did not accurately portray production delays to investors.

Figure 1: Sentence reuse across two articles on "self-driving cars". Our pipeline detected reuse even across edits, e.g.: *"...Musk's own **statements issuing** lofty goals for production cars or turning a sustained profit starting this quarter"* vs *"...Musk's own **pronouncements such as** lofty goals for production cars or turning a sustained profit starting this quarter **that might be beyond reach"*** detected as variants.

Abstract

The media industry has a practice of reusing news content, which may be a surprise to news consumers. Whether by agreement or plagiarism, a lack of explicit citations makes it difficult to understand where news comes from and how it spreads. We reveal *news provenance* by reconstructing the history of near-duplicate news in the web index – identifying the origins of republished content and the impact of original content. By aggregating provenance information and presenting it as part of news search results, users may be able to make more informed decisions about which articles to read and which publishers to trust. We report on early analysis and user feedback, highlighting the critical tension between the desire for media transparency and the risks of disrupting an already fragile ecosystem.

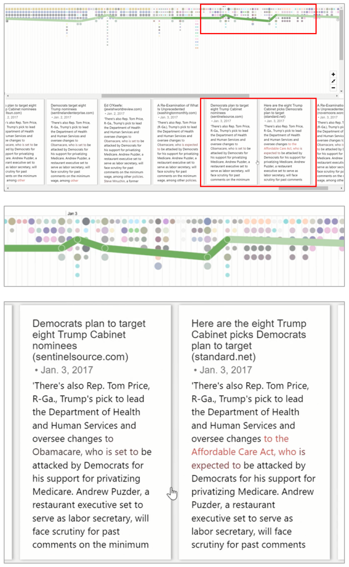
Author Keywords

news media; news search; media transparency; provenance analysis; provenance visualization

CCS Concepts

•**Human-centered computing** → *Activity centered design; Visual analytics;* •**Information systems** → *Near-duplicate and plagiarism detection; Data provenance; Presentation of retrieval results;*

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.
© 2020 ACM ISBN 978-1-4503-6819-3/20/04 ...$15.00.
http://dx.doi.org/10.1145//3334480.3375225



Democrats plan to target eight Trump Cabinet nominees (sentinelsource.com)
• Jan. 3, 2017

'There's also Rep. Tom Price, R-Ga., Trump's pick to lead the Department of Health and Human Services and oversee changes to Obamacare, who is set to be attacked by Democrats for his support for privatizing Medicare. Andrew Puzder, a restaurant executive set to serve as labor secretary, will face scrutiny for past comments on the minimum

Here are the eight Trump Cabinet picks Democrats plan to target (standard.net)
• Jan. 3, 2017

'There's also Rep. Tom Price, R-Ga., Trump's pick to lead the Department of Health and Human Services and oversee changes to the Affordable Care Act, who is expected to be attacked by Democrats for his support for privatizing Medicare. Andrew Puzder, a restaurant executive set to serve as labor secretary, will face scrutiny for past comments

Figure 2: News Threads visual for Power BI (not released). Articles are shown as vertical dot plots of individual paragraphs and ordered by publication time. Paragraphs are clustered by similarity, with dots colored and sized based on cluster identity and republication count. Selecting a paragraph's dot reveals its provenance "thread" and lists clustered paragraphs for comparison. TF-IDF is applied at the trigram level to highlight (in red) text that is distinctive within a cluster, e.g., *"...changes to **Obamacare** who is **set** to be attacked..."* vs *"...changes to **the Affordable Care Act** who is **expected** to be attacked..."*.

Introduction

Since the first newspapers were printed in the seventeenth century, following the news has become a central activity of everyday life. While the shift to digital distribution has increased both the production and consumption of news, the pressures of the 24-hour news cycle combined with shrinking revenues and resources has led to a "crisis in journalism" and a "publish first, verify later" culture [17].

The consequences of this crisis are many, from news plagiarism [16] to state-sponsored disinformation campaigns [1]. We are interested in a third phenomenon largely ignored in the public debate, however: the practice of journalistic text reuse. Earlier work contrasted the "benign and acceptable" reuse of news agency text ("copy") from the more harmful and unacceptable practice of plagiarism [5]. Today, however, content can easily be copied from anywhere by anyone, without agreement and with little chance of repercussions. The resulting potential for mass republication creates a multitude of threats – an economic threat to news publishers, a political threat to democratic processes, and an epistemological threat to news consumers who no longer know what to believe. For publishers and consumers alike, it has never been more important to understand *news provenance* – where news content has come from and how it has been modified and republished over time.

This case study reports on our early efforts to reveal such news provenance based on automated analysis of news article text – identifying original and duplicated sentences in ways that allow us to create a rich graph of content reuse across both articles and publishers. Our work takes advantage of the fact that Microsoft is one of only several organizations with a broad and near real-time index of online news, combined with a search engine and news portal through which many millions of users find and consume news every day. It also builds on our prior work developing text analytics capabilities for Microsoft Power BI [7], where we observed first-hand the difficulty of making sense of a news ecosystem distorted by republication (and designed a Power BI visual to explore patterns of reuse – Figure 2).

Related Work

News text reuse has been common practice since the era of print publication. The 1999 METER project [5] presents the most in-depth study of the phenomenon, contributing a corpus of 772 Press Association articles and 944 articles from 9 UK daily newspapers, with both article and sentence level reuse annotated by domain experts. The project reported that 78% of news articles were derived from news agency sources (46% partially, 32% wholly) and introduced a variety of techniques (n-gram overlap, greedy string tiling, and sentence alignment) for classifying the extent of derivation.

The concept of data provenance is similarly well known, gaining prominence with the development of the semantic web and motivated by the observed reuse of news text [10]. The resulting data model, PROV-DM, has since been used to describe the provenance of news stories as inferred from the named entities mentioned within them (evaluated on a week of 410 publications by a single publisher) [6].

For the problem of identifying shared text rather than shared entities, Locality-Sensitive Hashing (LSH) of individual sentences has been proposed as a logical but expensive approach [15], with document fingerprinting based on the Discrete Cosine Transform (DCT) preferred for detecting 'local' text reuse across a newswire collection (Table 1).

Analyzing news provenance on the web presents additional challenges in terms of document quantity, quality, and variety, as well as the need to retrieve the appropriate document set for analysis. One approach to web text

Reuse type	Prevalence
Most–Most	8%
Most–Much	10%
Most–Some	10%
Much–Much	17%
Much–Some	27%
Some–Some	28%

Table 1: Text reuse types and their relative prevalence in a TREC newswire collection of 758,224 articles by the Associated Press, Wall Street Journal, Financial Times, and Los Angeles Times, 1987–1994 [15]. Reuse is defined in terms of mutual containment over article pairs, with *Most* as 80-100% containment, *Much* as 50-80%, and *Some* as 10-50%. 51% of articles overlapped with at least one other article.

provenance [3] is to systematically construct, evaluate, and rank the results of multiple search queries comprising structured subsets of the target text based on its noun phrases and named entities. While this approach can leverage the comprehensive web index of the underlying search engine, such indices are not designed with text provenance in mind and cannot be queried at the scale required to reconstruct provenance graphs for large document collections.

The challenge of news provenance on the web is exacerbated by the growing number of publishers arising from the low barriers to entry. Recent work [11] describes how mass news republication by different sources can distort perceptions of story significance, or develop a sense of credibility for an otherwise unreliable source which may combine legitimate, unattributed articles with alternative, fake, or conspiracy-oriented news. The authors study this phenomenon by analyzing verbatim text reuse across 54k political news articles in 2017. Of 92 publishers, 67 were found to republish at least one article, with reuse occurring more frequently among publishers with similar audiences (e.g., for mainstream vs alternative vs hyperpartisan news).

A study of user evaluation strategies for news articles appearing in the Facebook news feed [8] has shown that the source of an article is a critical component for evaluating its trustworthiness, yet readers rarely question whether the listed source is accurate and legitimate. One successful approach to representing credibility in the context of search results [14] is to augment results with visual badges, ratings, or rankings of attributes that would be difficult or impossible for users to assess directly.

In summary, prior work has studied news provenance and search result augmentation, but never in combination as a means of supporting the real-world activities of both news publishers and consumers.

Implementation of News Provenance

Our prototyping of news provenance capabilities has focused on the iterative and incremental development of data pipelines supporting differentiated user experiences.

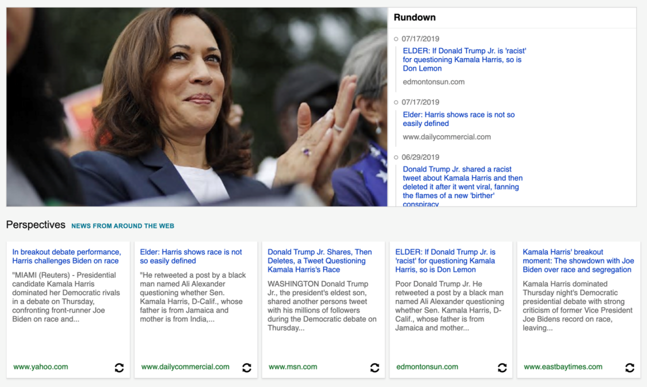
Common Data Pipeline

Our technical approach is based on the NLP technique of *w-shingling* for computing text similarity – extracting fixed-length n-grams from each comparable text unit and taking the Jaccard similarity (intersection divided by union) of these token sets. We use a Locality-Sensitive Hashing (LSH) index of n-gram token sets represented as MinHash values, which provides sub-linear estimation of Jaccard similarity for any given text query. For a given similarity threshold, our pipeline will cluster similar text units and sort them by the publication time of the containing article, thus reconstructing the provenance and republication of all observed text units. This provenance information can then be aggregated to the article and publisher levels to give a complete picture of text reuse within the news ecosystem.

By first applying our pipeline at the document level, we can identify independent clusters of documents within which sentence clustering is computationally feasible (since clustering requires pairwise comparisons that grow much more rapidly for sentences than documents). Generalizing this idea, our pipeline supports efficient hierarchical clustering of arbitrary text units. It first populates a sparse matrix with pairwise text distances when the Jaccard similarity exceeds a certain threshold, before applying DBSCAN clustering in a hierarchical fashion. At each iteration, the espilon parameter is reduced to progressivly decrease the variation within each identified cluster, and clustering only takes place within previously identified clusters. The result is a hierarchical clustering moving from all text units at the root to duplicate (or near-duplicate) text units at the leaves.

Reuse type	Prevalence
Most–Most	46%
Most–Much	7%
Most–Some	4%
Much–Much	7%
Much–Some	5%
Some–Some	31%

Table 2: Text reuse types and their relative prevalence as analyzed by our news provenance pipeline over a corpus of 45,300 news articles matching "Kamala Harris" from May–July 2019. 97% of articles were derived ($\geq$10% sentence similarity with other articles), as were 97% of sentences ($\geq$50% trigram similarity with other sentences). Reuse types were classified using the scheme from [15], aggregated over all pairwise article comparisons, and normalized to 100%. In contrast to prior results from the era of print publication (Table 1), we can see that the newly dominant form of reuse is one in which entire articles are copied directly with only minor edits and additions.



Rundown

07/17/2019
ELDER: If Donald Trump Jr. is 'racist' for questioning Kamala Harris, so is Don Lemon
edmontonsun.com

07/17/2019
Elder: Harris shows race is not so easily defined
www.dailycommercial.com

06/29/2019
Donald Trump Jr. shared a racist tweet about Kamala Harris and then deleted it after it went viral, fanning the flames of a new 'birther' conspiracy

Perspectives NEWS FROM AROUND THE WEB

In breakout debate performance, Harris challenges Biden on race
"MIAMI (Reuters) - Presidential candidate Kamala Harris dominated her Democratic rivals in a debate on Thursday, confronting front-runner Joe Biden on race and...
www.yahoo.com

Elder: Harris shows race is not so easily defined
"He retweeted a post by a black man named Ali Alexander questioning whether Sen. Kamala Harris, D-Calif., whose father is from Jamaica and mother is from India,...
www.dailycommercial.com

Donald Trump Jr. Shares, Then Deletes, a Tweet Questioning Kamala Harris's Race
WASHINGTON Donald Trump Jr., the president's eldest son, shared another persons tweet with his millions of followers during the Democratic debate on Thursday...
www.msn.com

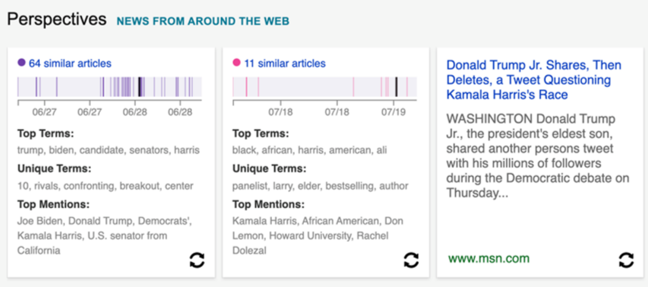
ELDER: If Donald Trump Jr. is 'racist' for questioning Kamala Harris, so is Don Lemon
Poor Donald Trump Jr. He retweeted a post by a black man named Ali Alexander questioning whether Sen. Kamala Harris, D-Calif., whose father is from Jamaica and mother...
edmontonsun.com

Kamala Harris' breakout moment: The showdown with Joe Biden over race and segregation
Kamala Harris dominated Thursday night's Democratic presidential debate with strong criticism of former Vice President Joe Bidens record on race, leaving...
www.eastbaytimes.com

Figure 3: Consumer experience for the news query "Kamala Harris", showing a timeline of published articles (right) and a card carousel for clusters of similar articles (bottom).

Consumer Experience
The Bing News Spotlight feature [18] provides a model for giving news consumers an overview of major developing news stories, including a topic timeline and presentation of different "perspectives". We adopted the same model for the design of our consumer experience (Figure 3). Clicking the icon in the lower-right corner of a perspective card reveals more information about the article cluster (Figure 4), while opening the article reveals sentence-level provenance (Figure 5). Selecting two articles shows a side-by-side view of fuzzy sentence matches and their differences (Figure 1).

To recreate the "perspectives" view based on clusters of similar articles at the text level, we configure our pipeline to operate on document unigrams (i.e., focusing on the presence but not order of words) and use our adaptive clustering method to identify clusters of related articles. Sentences are then clustered as near-duplicates when their trigram sets have a Jaccard similarity of 0.5 or greater.



Perspectives NEWS FROM AROUND THE WEB

● 64 similar articles

Top Terms:
trump, biden, candidate, senators, harris
Unique Terms:
10, rivals, confronting, breakout, center
Top Mentions:
Joe Biden, Donald Trump, Democrats', Kamala Harris, U.S. senator from California

● 11 similar articles

Top Terms:
black, african, harris, american, ali
Unique Terms:
panelist, larry, elder, bestselling, author
Top Mentions:
Kamala Harris, African American, Don Lemon, Howard University, Rachel Dolezal

Donald Trump Jr. Shares, Then Deletes, a Tweet Questioning Kamala Harris's Race
WASHINGTON Donald Trump Jr., the president's eldest son, shared another persons tweet with his millions of followers during the Democratic debate on Thursday...
www.msn.com

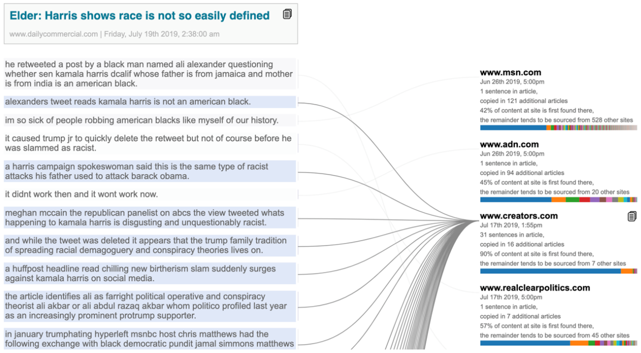
Figure 4: Cards can be flipped to reveal more details about the article cluster. Selecting "N similar articles" drills down to the individual articles of that cluster.

Publisher Experience
Publishers deeply understand the ecosystem of news wires, syndication, and editorial decision-making in a way that most consumers do not. Our publisher experience therefore includes support for verifying sources, understanding content propagation, and revealing both gaps in reporting coverage and patterns in publication timing. The interface shows publication timelines for content variants, navigable document clusters, and a treemap of news domains encoding both popularity and share of original content (Figure 6).

Analyst Experience
The Power BI visual we created to explore "news threads" (Figure 2) confirmed what we had informally observed in our prior development of news analytics applications [7] – that near-verbatim duplication of entire paragraphs was widespread across the results of news search queries. We observed likely cases of automated syndication (sequences of identical dot patterns), rolling updates to breaking news (dot patterns shifted downwards over time as new paragraphs were inserted at the top), remixing of content from multiple distinct sources, and edits both benign and slanted.

Quote variations
"you know" harris said turning to former vice president joe biden *"there was..."*
...harris informed him that *"there was..."*
"...every day" harris 54 told biden.
"...every day" harris whose father is black and mother is of indian descent said.
"...every day" then the killer line *"that little girl was me."*
"...was me" harris a california democrat said.
"...was me" harris said at times looking directly at biden as he gazed elsewhere.
"...was me" the moment apparently planned ... led to harris being unofficially declared the debate winner.

Table 3: Example sentence variants reporting the Kamala Harris quote *"and you know there was a little girl in california who was part of the second class to integrate her public schools and she was bused to school every day and that little girl was me"*. Different spellings, insertions/deletions, and paraphrasing also cause variation.



Elder: Harris shows race is not so easily defined

www.dailycommercial.com | Friday, July 19th 2019, 2:38:00 am

he retweeted a post by a black man named ali alexander questioning whether sen kamala harris dcalif whose father is from jamaica and mother is from india is an american black.

alexanders tweet reads kamala harris is not an american black.

im so sick of people robbing american blacks like myself of our history.

it caused trump jr to quickly delete the retweet but not of course before he was slammed as racist.

a harris campaign spokeswoman said this is the same type of racist attacks his father used to attack barack obama.

it didnt work then and it wont work now.

meghan mccain the republican panelist on abcs the view tweeted whats happening to kamala harris is disgusting and unquestionably racist.

and while the tweet was deleted it appears that the trump family tradition of spreading racial demagoguery and conspiracy theories lives on.

a huffpost headline read chilling new birtherism slam suddenly surges against kamala harris on social media.

the article identifies ali as farright political operative and conspiracy theorist ali akbar or ali abdul razaq akbar whom politico profiled last year as an increasingly prominent protrump supporter.

in january trumphating hyperleft msnbc host chris matthews had the following exchange with black democratic pundit jamal simmons matthews

www.msn.com
Jun 26th 2019, 5:00pm
1 sentence in article,
copied in 121 additional articles
42% of content at site is first found there,
the remainder tends to be sourced from 528 other sites

www.adn.com
Jun 26th 2019, 5:00pm
1 sentence in article,
copied in 94 additional articles
45% of content at site is first found there,
the remainder tends to be sourced from 20 other sites

www.creators.com
Jul 17th 2019, 1:55pm
31 sentences in article,
copied in 16 additional articles
90% of content at site is first found there,
the remainder tends to be sourced from 7 other sites

www.realclearpolitics.com
Jul 17th 2019, 5:00pm
1 sentence in article,
copied in 7 additional articles
57% of content at site is first found there,
the remainder tends to be sourced from 45 other sites

Figure 5: Selecting the article title shows each sentence of the article annotated with the article where it was first observed. Selecting a publisher highlights all associated derivation links.

To support this analyst experience, we configure our data pipeline to operate on paragraph-level text units and output the data tables needed to drive this visual within Power BI.

Example Analysis

We have analyzed a wide range of news topics including healthcare, self-driving cars, and eSports, as well as more controversial topics such as gun control. Our goal was to understand how news is copied and modified across syndication outlets, local affiliates, and content farms over time, as well as to evaluate and refine our provenance pipeline and the experiences built upon it. Here we describe one such analysis of a controversial news cycle about Kamala Harris – a US Senator whose race and eligibility as a presidential candidate were called into question following the launch of her 2020 US election campaign.

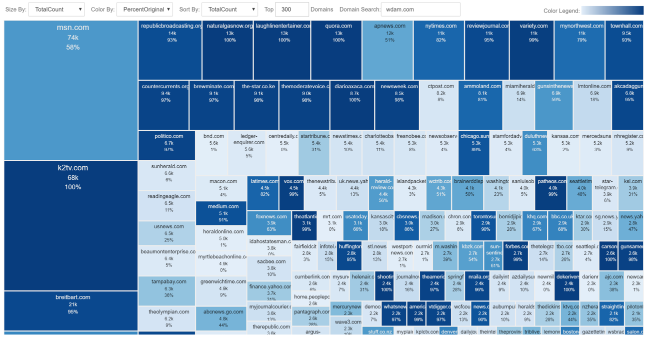
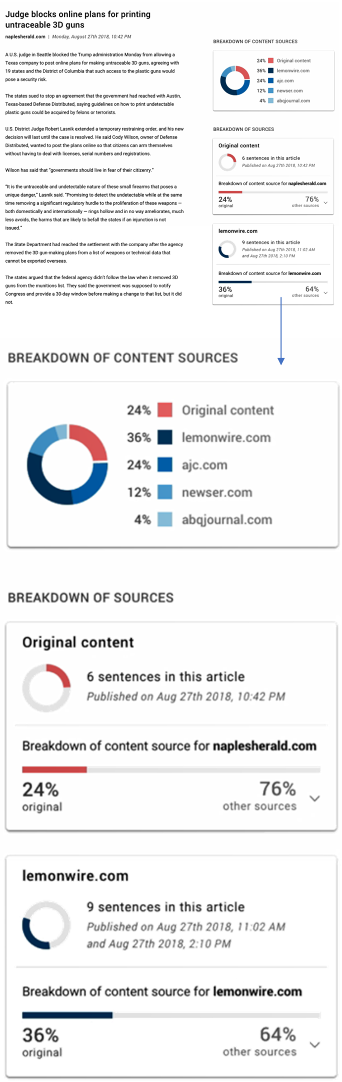


Figure 6: Domain treemap from publisher experience. Cell size and color represent article count and original content respectively for the topic "gun control".

We indexed a corpus of 45,300 news articles containing "Kamala Harris" from the period May–July 2019, reliably tracing content provenance back to a widely-copied opinion piece by Larry Elder (Figure 5). Analysis of reuse types using the method of [15] is shown in Table 2. We observe dominance of the *Most–Most* reuse type across the majority of our analyses, confirming a substantial change in news text reuse since the shift to online distribution.

Overall, we found 254,094 unique sentences from 1,522,340 sentence-to-document mappings, giving an average of 34 sentences per document. Clustering these sentences using a trigram similarity threshold of 0.5 resulted in 212,714 sentence clusters, with a mean of 7.2 *sentence publications* per cluster (median 1, maximum 3003). The mean number of *sentence variants* within each cluster was 1.2 (median 1, maximum 105 – see Table 3 for analysis of this cluster).



Judge blocks online plans for printing untraceable 3D guns

naplesherald.com | *Monday, August 27th 2018, 10:42 PM*

A U.S. judge in Seattle blocked the Trump administration Monday from allowing a Texas company to post online plans for making untraceable 3D guns, agreeing with 19 states and the District of Columbia that such access to the plastic guns would pose a security risk.

The states sued to stop an agreement that the government had reached with Austin, Texas-based Defense Distributed, saying guidelines on how to print undetectable plastic guns could be acquired by felons or terrorists.

U.S. District Judge Robert Lasnik extended a temporary restraining order, and his new decision will last until the case is resolved. He said Cody Wilson, owner of Defense Distributed, wanted to post the plans online so that citizens can arm themselves without having to deal with licenses, serial numbers and registrations.

Wilson has said that "governments should live in fear of their citizenry."

"It is the untraceable and undetectable nature of these small firearms that poses a unique danger," Lasnik said. "Promising to detect the undetectable while at the same time removing a significant regulatory hurdle to the proliferation of these weapons — both domestically and internationally — rings hollow and in no way ameliorates, much less avoids, the harms that are likely to befall the states if an injunction is not issued."

The State Department had reached the settlement with the company after the agency removed the 3D gun-making plans from a list of weapons or technical data that cannot be exported overseas.

The states argued that the federal agency didn't follow the law when it removed 3D guns from the munitions list. They said the government was supposed to notify Congress and provide a 30-day window before making a change to that list, but it did not.

BREAKDOWN OF CONTENT SOURCES

24% 🟥	Original content
36% ■	lemonwire.com
24% ■	ajc.com
12% ■	newser.com
4% ■	abqjournal.com

BREAKDOWN OF SOURCES

Original content

6 sentences in this article

Published on Aug 27th 2018, 10:42 PM

Breakdown of content source for **naplesherald.com**

24% original — 76% other sources

lemonwire.com

9 sentences in this article

Published on Aug 27th 2018, 11:02 AM and Aug 27th 2018, 2:10 PM

Breakdown of content source for **lemonwire.com**

36% original — 64% other sources

BREAKDOWN OF CONTENT SOURCES

24% Original content
36% lemonwire.com
24% ajc.com
12% newser.com
4% abqjournal.com

BREAKDOWN OF SOURCES

Original content
6 sentences in this article
Published on Aug 27th 2018, 10:42 PM
Breakdown of content source for naplesherald.com
24% original 76% other sources

lemonwire.com
9 sentences in this article
Published on Aug 27th 2018, 11:02 AM and Aug 27th 2018, 2:10 PM
Breakdown of content source for lemonwire.com
36% original 64% other sources

Figure 7: Concept video image.

Concept Testing

We performed both consumer surveys and publisher interviews to understand attitudes towards news provenance. Both used a concept video walk-through of a representative interface showing a news reading experience (Figure 7).

Consumer Survey

We recruited 40 participants to complete an online survey (20 male/female, average age 33). All were online news consumers with 78% checking the news at least daily. Participants were assigned to four conditions in a between-subjects design comparing publisher type (local vs national) and degree of originality (mostly original vs mostly sourced). A 7-point scale from extremely low to extremely high was used to understand how these factors may influence user confidence in article content.

We found that while *original local* content increased confidence (9/10 high with 5/10 extremely high), *sourced national* content decreased confidence (4/10 low with 4/10 slightly low). The majority of participants rated the interface as easy to use (35/40 easy with 21/40 extremely easy).

Participants were also asked to rank the concepts of authority, credibility, trustworthiness, and originality in order of importance. 19/40 each ranked trustworthiness and credibility as most important, with 26/40 participants ranking originality as least important. Multiple participants reported that they were indifferent to originality so long as other conditions were met, e.g., "As long as the source is accurate and truthful and well-recognized as credible" and "News is news. If it is accurate, I am less concerned with originality". Some noted how content reuse could increase credibility, e.g., "It makes me think they are more credible if they are willing to pull from other sources. It seems more balanced", while others saw reuse as beneficial only if it came from a trusted source or was accompanied by commentary on the reused text. Concerns about representing news provenance included accuracy (e.g., "how do you confirm the actual original source of the information?"), bias (e.g., "it might be possible ... to take a side and make people think that some articles are actually more original than others"), misinterpretation (e.g., "just because something is 100% unoriginal, that isn't bad"), and misuse (e.g., "I'd be afraid that this would be used to discredit legitimate websites").

Publisher Interviews

We interviewed representatives of eight major news publishers in the United States for critical feedback on the concept of news provenance. All supported the use of provenance interfaces as newsroom tools for research and reporting (e.g., "News rooms would love to see it"), noting how they could be used to identify original sources, showcase information from trusted sources, support proper attribution, detect plagiarism, track stories across networks, observe how different publishers use different sources, and find sources that are cited both widely and rarely.

As a consumer-facing capability, one broad view was that "publishers should have nothing to fear" because "more transparency and [content] tracking is a good thing" that would "incentivize original content" in ways both "good for user and original news org" . Reported benefits to consumers included access to a greater diversity of sources in search results that could also reduce redundancy.

At the same time, publishers raised concerns about possible unintended consequences of such capabilities (e.g., source devaluation and effects on search result rankings) and the difficulties of both calculating and interpreting aggregate measures of original content. "Publishers could get upset" if they felt they were unfairly represented, while "users would be shocked at [the level of] copying". Four publishers also stated that they would like to see an evaluation of source quality included as part of news provenance, although the question of who gets to judge quality and how are open questions that are significantly more contentious than any measures of original versus sourced content.

Discussion and Conclusion

This case study brings together diverse strands of work and applies them in ways that enrich our understanding of modern day news. It also suggests multiple directions for future work. For example, the analysis of text reuse is not limited to news content – it can be applied to other texts just as similar methods have been applied in the diverse areas of ancient literature, [13], state legislature [4], and Wikipedia articles [2]. The concept of news provenance is also not limited to news text, e.g., the New York Times has announced a blockchain project to track multimedia provenance [12]. Finally, Microsoft is not the only company analyzing originality – Google has also said it will be elevating the ranking of original content in its search results [9].

The results of our news provenance concept testing suggest that while news consumers may prioritize credibility over originality in abstract terms, the explicit representation of original versus sourced content has a concrete impact on consumers' evaluation of the same underlying news text. Great care must therefore be taken when deciding if and how to communicate such high-level provenance statistics. Media transparency might ultimately be best served by using inferred provenance to encourage publishers to provide more complete attribution for their articles, as a "source checking" counterpart to fact checking. Interfaces might also simply refer to the "first observations" of individual text units, rather than report the proportion of "original content" in ways that might be misinterpreted or manipulated. On this final point, while full disclosure of provenance algorithms could reduce misinterpretation, it could also create confusion and increase the risk of adversarial attacks. All of these factors need to be considered when designing future mechanisms for news provenance.

Acknowledgements

We thank our study participants, our team at MSR and Uncharted Software, our partners in the Search & Ads Studio, Microsoft AI + Research, and our collaborator Carey Priebe and his team at the Johns Hopkins University Applied Mathematics and Statistics Department.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[2] Milad Alshomary, Michael Völske, Tristan Licht, Henning Wachsmuth, Benno Stein, Matthias Hagen, and Martin Potthast. 2019. Wikipedia Text Reuse: Within and Without. In *European Conference on Information Retrieval*. Springer, 747–754.

[3] Michael Bendersky and W Bruce Croft. 2009. Finding text reuse on the web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM, 262–271.

[4] Matthew Burgess, Eugenia Giraudy, Julian Katz-Samuels, Joe Walsh, Derek Willis, Lauren Haynes, and Rayid Ghani. 2016. The Legislative Influence Detector: Finding Text Reuse in State Legislation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 57–66.

[5] Paul Clough, Robert Gaizauskas, Scott SL Piao, and Yorick Wilks. 2002. Meter: Measuring text reuse. In *Proceedings of the 40th Annual Meeting on*

Association for Computational Linguistics. Association for Computational Linguistics, 152–159.

[6] Tom De Nies, Sam Coppens, Davy Van Deursen, Erik Mannens, and Rik Van de Walle. 2012. Automatic discovery of high-level provenance using semantic similarity. In *International Provenance and Annotation Workshop*. Springer, 97–110.

[7] Darren Edge, Jonathan Larson, and Christopher White. 2018. Bringing AI to BI: Enabling Visual Analytics of Unstructured Data in a Modern Business Intelligence Platform. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article CS02, 9 pages. DOI: http://dx.doi.org/10.1145/3170427.3174367

[8] Martin Flintham, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. 2018. Falling for Fake News: Investigating the Consumption of News via Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 376, 10 pages. DOI: http://dx.doi.org/10.1145/3173574.3173950

[9] Richard Gingras. 2019. Elevating original reporting in Search. https://www.blog.google/products/search/original-reporting/. (2019). Published Sep. 12, 2019. Accessed Oct. 1, 2019.

[10] Paul Groth, Yolanda Gil, James Cheney, and Simon Miles. 2012. Requirements for provenance on the web. *International Journal of Digital Curation* 7, 1 (2012), 39–56.

[11] Benjamin D Horne and Sibel Adali. 2018. An Exploration of Verbatim Content Republishing by News Producers. *arXiv preprint arXiv:1805.05939* (2018).

[12] Sasha Koren. 2019. Introducing the News Provenance Project. https://open.nytimes.com/introducing-the-news-provenance-project-723dbaf07c44. (2019). Published Jul. 23, 2019. Accessed Oct. 1, 2019.

[13] John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 472–479.

[14] Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1245–1254.

[15] Jangwon Seo and W Bruce Croft. 2008. Local text reuse detection. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 571–578.

[16] Rui Sousa-Silva. 2015. Reporter fired for plagiarism: a forensic linguistic analysis of news plagiarism. (2015).

[17] Kate Starbird, Dharma Dailey, Owla Mohamed, Gina Lee, and Emma S. Spiro. 2018. Engage Early, Correct More: How Journalists Participate in False Rumors Online During Crisis Events. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 105:1–105:12. DOI: http://dx.doi.org/10.1145/3173574.3173679

[18] Bing Team. 2018. Bing helps you learn more about the news in less time. https://blogs.bing.com/search/2018-08/Bing-helps-you-learn-more-about-the-news-in-less-time. (2018). Published Aug. 27, 2018. Accessed Oct. 1, 2019.