9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Measuring Diversity in Online News

Keywords: online news, media consolidation, network analysis, natural language processing

Extended Abstract

News publishers have had a difficult time adapting from print to digital distribution. Advertising revenue from digital distribution channels has not kept pace with the massive decline of advertising and subscription revenue from print editions, and news aggregators and social media platforms enabled the consumption of news without reading the newspapers directly. Consequently, newsroom employment has dropped dramatically and is now only half of what it was as recently as 2008 [1]. However, much of this decline is not obvious to the average reader because struggling newspapers in the digital world rarely disappear: a publisher can reduce original reporting and fill a large share of its news sections with ready-made stories from wire services like Reuters and AP at low cost. While this was also an option in the pre-digital age, this pattern has likely accelerated with declining revenues in recent years. Alternatively, struggling publishers are often acquired by media conglomerates that centrally produce news stories that are then republished by their affiliates and subsidiaries [2]. This trend is antithetical to the idea of a diverse news diet and raises questions such as: What share of news reporting is original versus copied? Where is original news most often produced? And which factors influence the amount and distribution of original reporting across U.S. news publications?

In this paper, we quantify the amount of news overlap in the U.S. media landscape by analyzing the amount of duplication in news stories due to media consolidation. We identify duplicate news stories republished by different subsidiaries of the same media company using artefacts in article URLs, which also unveil ownership networks in the media landscape. We then expand this net by performing text-matching between articles to find similar content across different media companies.

To study this phenomenon, we collect a large dataset of over 26 million news articles over 2 years from more than 3,000 publishers covering 99% of U.S. news consumption. For creating the networks that link news websites which share content with each other, we present a novel technique that exploits common patterns in URLs of articles published by subsidiaries of a media company to find duplicate articles and automatically extract ownership structures in an unsupervised manner. We find that local news organizations owned by the same parent organization often have a shared back-end infrastructure for content management and merely republish a significant proportion of stories from their sibling websites. We are able to use these common patterns to build a graph of completely identical, shared news stories that connect otherwise distinct news publications, which allows us to not only visualize the ownership structure of media in the U.S., but also quantify with high precision the 'originality' of news publications and the amount of content they share with each other. Figure 1 shows such a network generated by our method from data that includes international websites as well.

We extend this analysis to identify content reuse that may exist beyond media consolidation and network affiliations by comparing the full text of articles published in the U.S. in the span of one month. Using GPT-3 [3], we obtain vector embeddings for the documents which represent their semantic meaning. We group all article published within a three-day window as candidates for duplicated content, and perform cosine similarity between all pairs in that group. Using a sufficiently high threshold of similarity (in this case, at least 0.95), we map these embeddings

as a network of articles connected by an edge of they are duplicates. Figure 2 shows such a network. This lets us quantify not only the lexical similarity between news articles, but also lets us study the semantic distribution of news content by clustering the document embeddings using a lower similarity threshold.

This work represents the first investigation of this scale into the originality of news in the U.S. by looking at it through the lens of both content overlap and ownership and affiliation structures. We hope that it encourages further research on the diversity of news, not only in the U.S., but also across the rest of world.

References

[1] E. Grieco, "U.S. Newspapers Have Shed Half of Their Newsroom Employees Since 2008," *Pew Research*, 2020.

[2] P. M. Abernathy, *News Deserts and Ghost Newspapers: Will Local News Survive?* Center for Innovation and Sustainability in Local Media, School of Media and Journalism, University of North Carolina at Chapel Hill, 2020.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

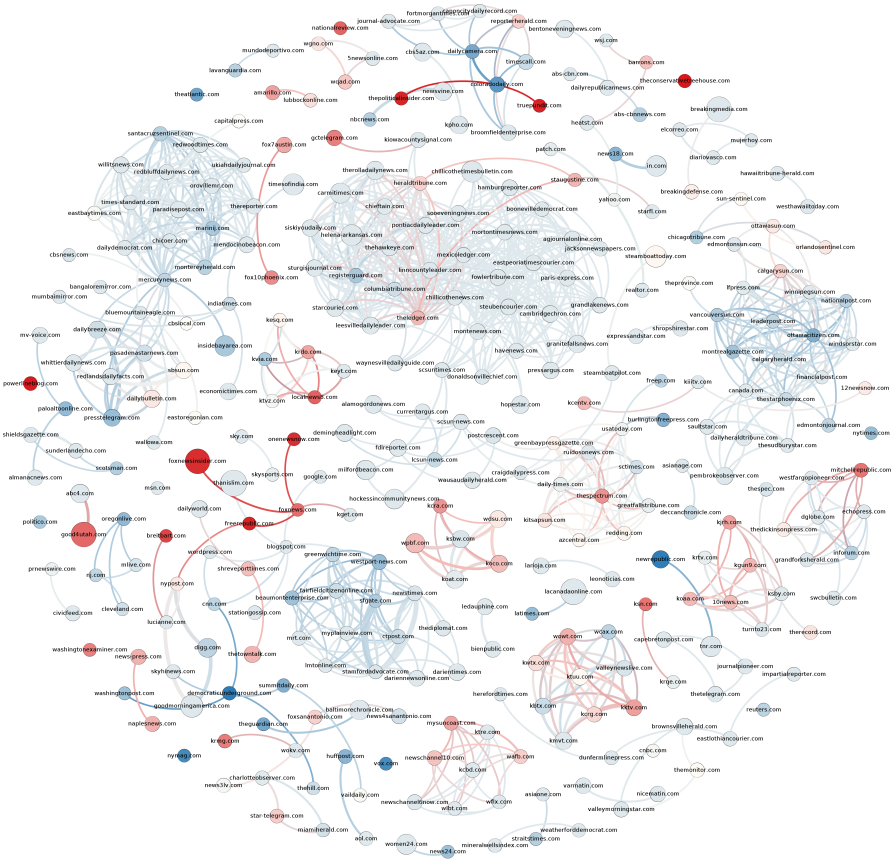

Figure 1: A network of shared URL structure between articles of different news publications, capturing both corporate ownership and network affiliation relationships. Nodes represent news publications: size denotes proportion of shared versus unique content and color denotes partisan bias. Edge weights represent the amount of shared content between a pair of publications. Observe the cliques of strongly interconnected news publications that are either owned by the same media conglomerate or affiliated with the same network.

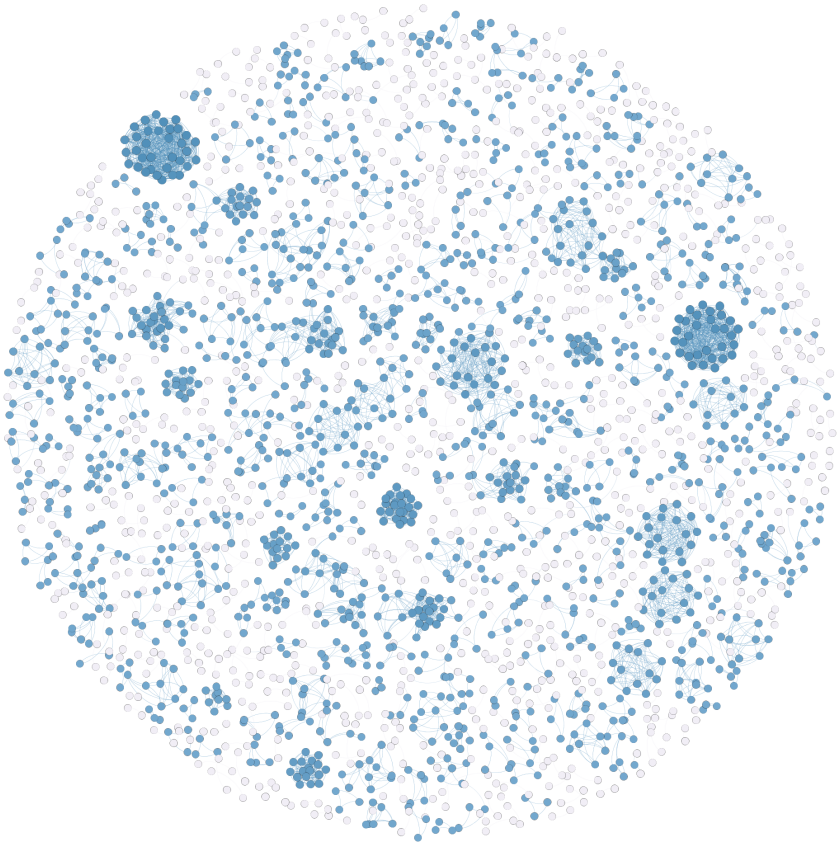

Figure 2: A network of news articles sampled from one day having a cosine similarity of at least 0.95 between their document embeddings. Nodes represent news articles and edge weights represent the cosine similarity between them. We find several cliques of strongly interconnected articles that contain text reused very frequently.