Exogenity and Endogenity: the amplification of ideas in the News Cycle

Keywords: News Cycle, discourse, keyphrase extraction, collective attention, narrative

In this paper we explore the amplification of ideas in the news media through an analysis of front-page articles from August 2016 – 2018 in five major news publishers. First, we develop and operationalize the notion of a *storyline*, or a thread of coverage that, through many articles, elaborates a continuous story, scene, or situation. We operationalize the construct of the news storyline with a novel, unsupervised NLP approach. Second, we offer qualitative and quantitative evidence to suggest that some storylines are more endogenously driven, while others more exogenously driven.

An exogenously driven storyline corresponds to a news cycle that is driven by important, largely unexpected events that happen independently of the news media's apparatus of coverage. Exogenously driven news stories can be viewed as plucked from a conveyor belt of new events and information that can be observed but not influenced. As a result, we expect exogenously driven stories to resemble bursty responses to external "shocks" that end when the "action" ends.

In contrast, endogenously driven news coverage is self-propagating; it is in conversation with itself; it develops in response not only to random strings of external events but to its own cultural wake. American sports media, as an example, seems to adhere to one especially vulgar version of self-propagating logic. In the morning, a talking head will speculate some incendiary narrative or "hot take" about a team or player. In the afternoon, analysts will debate whether it could be true. Later at the game, reporters ask questions about the claim, seeing as everyone else is talking about it. In this manner ESPN generates a lot of content, but it is not only ESPN. The basic logic underlying the continuing coverage of the sports example may be present any time that the news media makes public opinion, broadly construed, the object of its coverage, since public opinion is not independent of the news media. As a result, we expect endogenously driven coverage to be more consistent over longer periods of time than exogenously driven coverage.

In this paper we make an initial analysis of the range of factors that drive news coverage through an operationalization of the notion of *storyline*. To find distinct storylines in the news article data (N=18,807 articles; from *NYTimes, Boston Globe, Chicago Tribune, Atlanta-Journal Constitution, LA Times*) we begin with an observation: journalists develop relatively stable ways of referencing—or names for—the things they talk about the most. Surfacing the most repeatedly mentioned names, or key phrases, thus indexes various storylines. "Hurricane Maria", "Hillary Clinton's private email server", and "the war in Ukraine" are a few examples. To identify the most-covered stories, we propose a three-step process, beginning with unsupervised detection of storyline mentions and ending with a supervised classification of articles belonging to storylines.

First, we use a simple heuristic to identify phrases that occur roughly verbatim in the data much more often than would be expected at random. After filtering out long name titles (e.g., "Senate Majority leader Mitch McConnel") and auxiliary jargon (e.g., "according to anonymous sources"), the resulting phrases appear to span the major events and issues repeatedly covered by the news

While step one produces a nice array of story-names, naturally not every reference to a given story uses the same exact wording. Thus, in the second step, to find variants of the phrases surfaced in the first step and improve the phrase-cluster recall, we make use of large language models which have been trained specifically on phrase-level paraphrase data generated by GPT-3 [1]. Embeddings from this model allow us to do a simple semantic

search of phrases in the corpus, using the phrases surfaced in step one as queries, to vastly improve recall with respect to all explicit mentions of the storyline in the corpus. Finally, the large sets of phrases generated from outputs of the first step are taken as mentions of each corresponding storyline. The various mentions of <story name> are used to heuristically label articles as being a part of the storyline about <story name>, yielding a ground truth with which to train a classifier.

Prior work by Lescovec et al [2] on the dynamics of the news cycle has similarly focused on short phrases ("memes") that index a particular idea and then on tracking the lifespan of the idea as measured by prevalence of the indexical phrases. However, Lescovec et. al. as well as other work on the lifespan of ideas in mass culture generally [3,4] assume that various ideas compete for attention according to an essentially homogeneous logic of amplification. In their study of the coverage of political quotes they model all threads of coverage (based on a particular quote) as sharp bursts followed by exponential decays. In figure 1, on the other hand, we suggest that the trajectories of different storylines may be qualitatively different.

Both panels in figure 1 have units of articles per sliding 14 day window. The first panel shows the time series of the coverage of five storylines, each of which is bursty and sees significant stretches of time with no activity. In contrast the story lines in the second panel do seem to be responding to random shocks, but these shocks appear superimposed over a baseline of steady coverage. The coverage baselinse in the second part of figure 1 is much larger than that of the first plot even when normalizing for total volume of coverage. This suggests that the storylines in the second panel are responding to a greater degree to something other than random exogenous events. In further work we run more formal tests on the significance of this difference.

To further explore what appear to be more regularly occurring swells of attention to certain storylines we also read the actual articles and find qualitative differences in the explicitly and implicitly expressed justifications for articles in different storylines. We find that endogenous storylines tend to justify coverage in headlines and opening sentences by appealing to perceptions, feelings ("analysts fear") and public narratives, while exogenous narratives tend to be more action driven.

We plan to expand the dataset to include more publishers over a longer period of time and expand the analysis to more formal quantitative comparisons of time series and between-publisher comparisons of the shape and volume of coverage devoted to different storylines.

References

[1] Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10837–10851, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[2] Leskovec, Jure et al. "Meme-tracking and the dynamics of the news cycle." Knowledge Discovery and Data Mining (2009).

[3] Robert J. Shiller, 2017. "Narrative Economics," American Economic Review, American Economic Association, vol. 107(4), pages 967-1004, April.

[4] Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. Nature Communications, 10, 1-9.

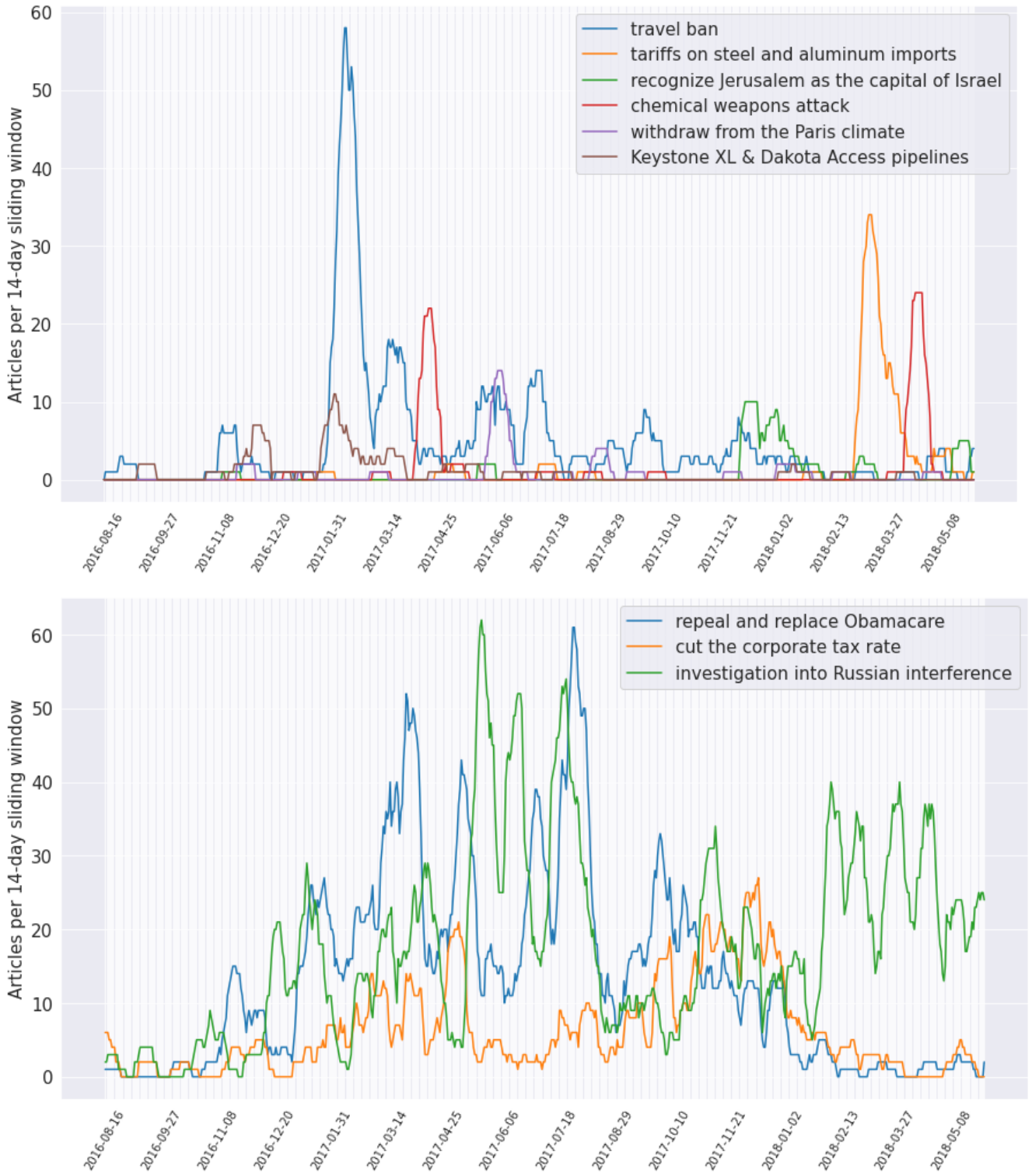

Figure 1 Front page articles written about 8 different storylines aggregated from five major US newspapers. The y-value corresponding to a given date equals the number of articles written in the past two weeks leading up to the day. The story lines in the top panel appear moreso to be responding to exogenous shocks than the bottom panel, whose storylines see a hum of regular coverage between pursty periods.