The Modern Gatekeepers in Mass Media

Patrick Kraft
PhD. Student
Stony Brook University

Yanna Krupnikov
Assistant Professor
Stony Brook University

Kerri Milita
Assistant Professor
Illinois State University

John Barry Ryan
Assistant Professor
Stony Brook University

March 25, 2016

Abstract

The abstract goes here.

Prepared for presentation at the annual meeting of the Midwest Political Science Association, Chicago, April 8, 2016. The manuscript and code are available on GitHub: https://github.com/pwkraft/nyt.

1 Introduction

I will write an introduction.

2 The More Things Change...

With technology, it is very tempting to herald the era following every advancement as fundamentally different from the era that proceeded it. The reality, however, is more nuanced than that. From the invention of the printing press to television to streaming services, each new technology has allowed creative people to tell their stories to a wider audience. Still, while each medium has its subtle differences in how it tells stories, there is something fundamental about drama and comedy that remains unchanged. That is what allows modern people to enjoy stories dating all the way back to ancient Greece and even before.

There is a similar tendency among some to say that social media has transformed how we consume political information. While their clearly some differences, we would argue it is foolish to assume the old theories no longer apply. Social media allows communication with larger numbers of people in dispersed locations–including people who have never been in the same room. At the same time, despite these differences, communication via social media is still, at its heart, one person sending messages to another person. In this fundamental way, it is no different from talking at a very large table.

To examine how much has changed in the era of Facebook and Twitter, we can examine how the "two-step flow" of information is different now. In the most basic form, the two-step flow would suggest that most people receive information from their friends, family, and co-workers who pay attention to the news media (Katz, 1957). If

this were different, then we would expect that more people got their news from the primary sources–e.g., newspapers, television news, or internet news sites. If anything has changed, fewer people are receiving information directly from these news outlets necessitating the two-step flow of information even more (**?**).

With off-line social networks, there is a tendency towards political homophily (**?**). While it is unlikely that many people choose their discussion partners solely on the basis of politics, people do associate with individuals who are similar in terms of race, religion, education, and other factors which often correlate with political affiliations (**?**). Despite the tendency toward homophily, the complex patterns of social relations allows people to hold opinions different from some of their acquaintances (Huckfeldt, Johnson, and Sprague, 2004). In fact, the more one talks about politics, the more likely they are to encounter disagreement (**?**). This means that 21st century on-line social networks, which tend towards homophily but allow for some cross-cutting exposure, are not that different from the pre-internet social networks.

On-line social networks are larger than off-line social networks, but it is unclear how this difference matters. To describe some of the "friends" on social media as "weak" ties would stretch the definition of the concept. Granovetter (**?**) noted that weak ties can provide information unavailable among an individual's intimate associates. Only a bold individual would ask for a job based on a Facebook post from someone he or she met a few times five years ago. Further, the most influential ties are those with some intimacy (Kenny, 1998). That is, the most important ties are those with whom one has a feeling of closeness. This feeling is more important than frequency of discussion or the duration of the connection (**?**). Hence, researchers suggest individuals should not look at the entire on-line social network if one wants to understand the truly influential connections. Rather, the meaningful social network is made of the individuals with whom one has direct communication–e.g., they comment on each other's photos (**?**).

Finally, the most politically engaged are the most likely to share political information on social media (**?**). The most politically engaged are also the most likely to discuss politics (Huckfeldt, 2001). The motivations in both instances are quite similar. As Ahn, Huckfeldt, and Ryan write (2014), "We cannot expect opinion leaders to provide information that is balance, objective, and value free" (255). Opinion leaders discuss politics and post information about it on social media because they care about politics and they want to influence people. This is the most important way that the social communication is unchanged even with the internet. The modern two-step flow, like the two-step flow of previous eras, operates because individuals want to influence how people think about politics.

3 Endogenous Agenda Setting

To the extent that social media changes communication, it is not about interactions among peers. The major change involves how the news media interacts with the politically engaged members of the public. Whether referring to the news media as agenda setters or gatekeepers, the major theories of information flow in the 20th Century assumed that the media were the first movers in deciding what issues were important (Iyengar and Kinder, 1987). Obviously, this view was always a bit oversimplified–as the "two-step flow" implies, the politically engaged members of the public acted as a second gatekeeper in regards to what stories most people heard. Further, 20th century editors would choose stories in part based on what stories they thought would sell newspapers (**?**). For this reason, Thorson and Wells (**?**) argue that we should think of modern gatekeeping in terms of "curated flows": journalists select certain stories to post on a website, engaged members select a subset of those to read, and they share an even smaller subset.

Social media adds another wrinkle to this agenda setting story. While 20th Century journalists had to anticipate which stories attract public attention and then receive noisy signals about the stories did capture the public, 21st Century journalists are able to quantify just how much attention a story is receiving. They then act on this information as they decide which stories to write about and which stories to feature (**?**). In this way, modern agenda setting is endogenous. Media gatekeepers set the agenda for the engaged members of the public. Engaged members of the public act as gatekeepers for everybody else sharing certain stories and not other. The media then writes more stories about the topics the engaged members of the public are sharing.

When engaged members of the public having greater say over the agenda, this has the potential to fundamentally change the news that makes the agenda. The most basic difference comes from the beliefs of the different agenda setting actors. In journalism, there are norms that encourage journalists to be neutral and, with some notable exceptions, many view their main role as informing the public (**?**). The modern engaged public is marked by extreme positions (**?**) and their main goal is to persuade (Ahn, Huckfeldt, and Ryan, 2014). This would suggest that there will be biases in terms of what stories are passed along. Typically, we think about biases in social communication as selective sharing of information in order to make one's own party look better (Ahn and Ryan, 2015; Pietryka, Forthcoming). The more important biases in this case, however, may result from simply which types of stories receive attention.

Hence, while this new form of agenda setting has many potential implications, this paper's goal is to explore a simple question: are the types of stories the news media values different from the types of stories the political engaged members of the public think are important? Is this the first question that any investigation in endogenous agenda setting must address. If the news media and the engaged public believe the same stories are important, then the feedback journalists receive now provides no new

information to them. On the other hand, if the engaged public differs from journalists in its beliefs about what is important, then stories that would have previously been buried may begin to receive front page attention now that the journalists are more likely to follow the public.

4 Data

The data for this study was collected from the website of *The New York Times* between February 18 and April 28, 2015. We created a macro using Visual Basic that performed an exhaustive scrape of key areas of the website.[1] The macro consisted of three web scrapers that extracted article title, author, date, time, and URL for (1) the front page of *The New York Times*, which mirrors the lead stories on the hard copy of the newspaper, (2) the digital front page, which contains more stories than the hard copy of the front page, and (3) a subset of the front page that lists the top ten most viewed, most emailed, most shared articles.

This data collection allows us to analyze every story that appeared on the hardcopy and digital front pages of the *New York Times*. It also allows us to see which types of stories the journalists see as more important–those that make the front page or receive a prominent placement on the digital front page. It also allows us to separate the agenda of the engaged public into separate categories in a way that is not capable with off-line

[1] Standard web scraping platforms and programs (e.g., Outwit, rvest) were unable to extract all of the requested information, as the three subsections intermittently experienced html changes, such as an alteration in the font size or heading classification for a particular article. Changes such as this will cause traditional web scraping programs to skip over these articles. Thus, it was necessary to write a macro that was customized to these three subsections of *The New York Times* website. Using html parsing, we documented the source code for each subsection and noted common changes to the html over a course of three days. For instance, if a text-only article is replaced on the following day by one with embedded video, the source code will change to reflect that. The macro was programmed to run every 12 hours and to export the scraped material to an excel file with three tabs (one for each subsection). Every time the macro ran, it created an additional excel file (e.g., NYT01, NYT02, etc.).

social network research. We can see the stories they believe are important–those that make the most viewed list–and separate those stories from the stories they believe are the most important for other people to know–the most share on Facebook, Twitter, or via email.

5 Description of Dataset

In order to analyze the content of the NYT articles using the structural topic model approach presented by **?**, I transformed the scraped articles to a reduced dataset where each unique article is included as a *single* observation. Articles that appeared several times in the original raw collection (e.g. most tweeted article for several days) or through different channels (e.g. most tweeted and most viewed) were combined in a single observation. Overall, the reduced dataset contains 5592 *unique* articles for subsequent analyses.

For each observation, I created a vector of dichotomous variables indicating whether the respective article was included in each of the categories (emailed, facebook, etc.) at least once. Here is a sample of observations from this reduced dataset (article body and keywords are omitted).

The dataset contains a unique id for each article which can be used in subsequent analyses to link the observations back to the full (raw) dataset including multiple instances for each article. The variables `emailed` through `digital` represent the matrix of covariates that will be used in order to model differences in topical prevalence in the collection of documents.

6 Initial Results for 20 Topics

In order to provide a first validation of the method, I estimated a structural topic model with 20 topics using the `stm` package in `R` (**??**). Depending on the specific research question, the number of topics can be increased in subsequent iterations in order to capture more fine-grained differences in article contents. The following output presents an overview of the extracted topics by displaying words that are highly associated with the respective topic (using highest probability, FREX, Lift, Score, see **?**). It should be noted that I used the spectral initialization implemented in the `stm` package in order to specify starting values for the subsequent model estimation. **?**, 12-13 discuss different alternatives to specify starting values and argue that in practice, the spectral initialization can be utilized successfully with vocabularies smaller than 10,000 entries. However, the vocabulary for the analyses of NYT articles has a size of 10268, so it might make sense to look at alternative strategies to specify starting values as well. However, I'll leave this issue for future iterations of the analyses.

Overall, the extracted topics have high face validity in the context of newspaper articles. Some of them clearly focus on political issues, such as the US presidential race, whereas others represent other themes common in newspapers, such as art, media, sports, or health issues. Based on the high probability words that are strongly associated with each topic, we can assign descriptive labels for each of the extracted topics. I decided to label the topics in the following way:

1. Presidential race
2. Sports
3. Books
4. Health

5. Art
6. Iran/Israel
7. Banking/Financial
8. New York/Food
9. Real Estate
10. Technology
11. Fashion
12. Terrorism/Iraq
13. Education
14. Media
15. Police
16. Plane Crash
17. Gender
18. Film/Shows
19. Legal/Court
20. Business

We can also investigate how frequently each topic was mentioned in the articles. The following plot displays the proportions of each individual topic in the overall text body.

Interestingly, “Books” appears to be the most prevalent topic in the set of articles analyzed here. However, it should be kept in mind, that each observation in the document matrix can represent multiple instances of an article in the raw collection. The proportions presented here only describe the proportions in unique articles but does not take into account how often each of the articles was included originally (e.g. as most tweeted, or most viewed multiple times).

We can also investigate the correlations between topics. Due to the fact that the structural topic model does not assume that each document has to be ascribed to a single topic but rather to a collection of topics, we can investigate the extent to which topics co-occur in documents. The following plot visualizes correlations between topics by linking topics that are correlated above a certain threshold (set at 0.01).

The connections between topics shows some interesting and plausible patterns. For example, it can be seen that the topics “Fashion”, “Art”, “Film”, “Books”, and “Media” cluster together or that “Police” is related to “Legal/Court”. Overall, the results provide some additional validity for the topics extracted using the structural topic model.

7 Differences in Topic Proportions between Categories

As described in **?**, the structural topic model not only extracts topics from a collection of documents but also allows us to directly model the prevalence of topics in specific documents based on a matrix of meta-covariates. While it is also possible to use covariates in order to model differences in words used to describe certain topics, we only focus on differences with regard to *how much* a topic is discussed in specific articles.

The following figures display the change in the proportion to discuss each of the 20 topics for articles that were included in each of the categories or not.

For articles that were most emailed, we see for example that the proportion to

discuss the presidential race is significantly lower (as compared to articles that were not most emailed). On the other hand, articles that were most emailed included higher proportions of the “Health” topic.

For articles that were most shared on facebook, there are hardly any significant differences with regard to the proportions of either of the 20 topics. Only the proportion of “Business” topics appears to be significantly lower in articles that were shared on facebook as compared to articles that were not shared on facebook.

Looking at the topic proportions for front page articles, we can again observe several significant differences. For example, front page articles included higher proportions of political issues (presidential race, Iran/Israel, Terrorism/Iraq) and lower proportions of fashion, media, literature, and related topics. This result provides some additional validity for the model results.

Articles that were most tweeted in the period under consideration encompass higher proportions of topics described as “technology”, “terrorism”, “financial”, and “health”. Interestingly, the proportion of “sports” is lower in tweeted articles as compared to the remaining articles.

Looking at articles that were most viewed, we can see that the proportion of the topic “presidential race” is significantly larger. On the other hand, the proportion of “health”, “art”, or “technology” is lower in articles that were most viewed.

For articles that were included in the digital first page, no clear significant differences are observed. This could be due to the fact that most of the articles included in the analyses were part of the digital edition at some point (Only 4.99% of the articles included in the analyses were not part of the digital edition.)

8 Conclusion / Next Steps

Overall, the structural topic model appears to recover plausible topics from the set of articles analyzed here. Subsequent iterations could investigate results for larger numbers of topics. Furthermore, additional analyses should be conducted in order to check robustness with regard to varying starting values. Other possible subsequent steps include:

- investigate influence of other characteristics of articles (e.g. length of article)
- link topic model results back to full data, include information about persistence in different categories
- set up hazard model, explain how long a topic stays in the cycle

References

Ahn, T.K., and John Barry Ryan. 2015. "The Overvaluing of Expertise in Discussion Partner Choice." *Journal of Theoretical Politics* 27 (3): 380-400.

Ahn, T.K., Robert Huckfeldt, and John Barry Ryan. 2014. *Experts, Activists, and Interdependent Citizens: Are Electorates Self-Educating?* New York: Cambridge University Press.

Huckfeldt, Robert. 2001. "The Social Communication of Political Expertise." *American Journal of Political Science* 45 (2): 425-438.

Huckfeldt, Robert, Paul E. Johnson, and John Sprague. 2004. *Political Disagreement: The Survival of Diverse Opinions within Communication Networks*. New York: Cambridge University Press.

Iyengar, Shanto, and Donald R. Kinder. 1987. *News That Matters: Television and American Opinion*. Chicago: University of Chicago Press.

Katz, Elihu. 1957. "The Two Step Flow of Communication: An Up-to-Date Report on an Hypothesis." *Public Opinion Quarterly* 21 (1): 67-81.

Kenny, Christopher. 1998. "The Behavioral Consequences of Political Discussion: Another Look at Discussant Effects on Vote Choice." *The Journal of Politics* 60 (1): 231-244.

Pietryka, Matthew T. Forthcoming. "Accuracy Motivations, Predispositions, and Social Information in Political Discussion Networks." *Political Psychology*. DOI: 10.1111/pops.12255.