On the Bubble: An Internet-Scale Investigation into Location-Based Algorithmic Filtering of Political Content

Author 1 abc@tbd

Author 2 xyz@tbd

1 Introduction

The news media has been described as a market for information, in which news consumers seek news producers in order to acquire more knowledge about relevant matters (Hamilton, 2004). In fact, for certain types of information such as politics, the news is the only source of information for many individuals. As with any other market, consumer choice is restricted to that which is offered by producers. The market model has long been used to explain the substantial variance in the quality and focus on various issues amongst news producers. Creating news has fixed and variable costs associated with production and distribution; as a result, news organizations are forced to cater to the interests of the news-consuming public in order to remain viable.

Online media has largely been assumed to obey the same market principles as previous technologies such as television or print. Consumers are believed to interact with online media for the same information-seeking reasons as with traditional media. But in the Internet age, consumers have much more choice than compared to previous markets. Subsequently, consumers are forced to develop strategies that allow them to select relevant news producers more quickly. The diversity of online media has positive effects, namely that news consumers are more capable to gather knowledge that is used to form opinions on other matters, which should lead to a more informed population. However, there are also potentially negative effects, in that the strategies that news consumers develop are presumed to be based upon preference for ideologically similar information, a behavior referred to as "selective exposure."

1.1 Selective Exposure

Previous research on selective exposure has traditionally demonstrated that individuals largely prefer to consume information that confirms their own worldview, particularly in the case of political identity (Sears and Freedman, 1967), (Chaiken, 1980), (Arceneaux et al., 2012), (Iyengar and Hahn, 2009) . This preference is reinforced by any number of existing factors, such as the strength of their partisanship (Stroud, 2008) or situational enhancement (Bryan et al., 2009).

Generally, the model is described as follows: people form their political identities early in their lives, strengthen their understanding of their political identity with updated information via the news, which in turn restricts their willingness to select sources of information that provide counter-ideological views and further restrict their willingness to expose themselves to opposing perspectives. As a result, selective exposure is often suspected to be a major cause of partisan polarization. As individuals become less likely to "hear the other side," their animus towards the perceived opposition grows

(Levendusky, 2013). There are few barriers that prevent interested consumers from restricting their information consumption only to those specific sources that largely enhance the strengths of one's political identity while diminishing those of others (Sunstein, 2009). This common explanation of selective exposure primarily places the onus of responsibility for information choice solely upon the consumers.

1.2 The Rise of Recommendation Systems

In its current state, the literature on online media consumption continues to view news exposure from the assumption that selective exposure is largely the result of individuals making product selections under reasonably fair conditions, meaning that there is a chance that they will exposed to counter-attitudinal information (i.e., a strong Democrat will occasionally be exposed to a presumed conservative-leading piece of information). In other words, it assumes that a person has an equal chance of being presented with ideologically favorable or unfavorable information.

However, some literature has pointed out a flaw in this worldview, namely that search engines and recommendation systems possess the ability to notably skew product offerings towards one's existing preferences already. Eli Pariser coined the term "filter bubble" to describe the phenomenon in which personalized recommendations make it significantly less likely for an individual to be presented with counter-attitudinal information (Pariser, 2011). Earlier concerns about filter bubbles focused solely on personalized recommendations based on previous behavior of the individual news consumer. However, with increasingly more sophisticated algorithms, recommendation systems are more capable of inferring preferences given seemingly innocuous attributes of a news consumer readily available through one's browser, such as their location or the search terms that lead to a user landing on a given site (otherwise known as the referrer path). This suggests that filter bubbles can arise even with completely pristine browsing history.

1.3 Fear of Filter Bubbles

In light of the increasing popularity of recommendation systems, some researchers have moved their focus away from understanding and identifying causes of selective exposure within consumers. Instead, they focus more on the influence that algorithms can have on the market available to an individual consumer. Much of this research consists of empirical observations about the state of recommended news.

While Facebook is not a traditional recommendation engine per se, their news feed does have some algorithmic complexity in determining what stories you should see. Bakshy, Messing, and Adamic observed that a Facebook user is likely to see counter-attitudinal news recommendations in their news feed, suggesting that Facebook does not encourage filter bubbles (Bakshy et al., 2015). However, an earlier study demonstrated that news aggregation follows a power law distribution, where only a few sites get the lion's share of visitor traffic (Hindman, 2008). An additional study found that the distribution of content on the first page of search engine results can result in attitude and opinion changes that can lead to different political outcomes (Epstein and Robertson, 2015). The combination of these two studies suggest that while the world of online news is diverse, only a few sites make it to the attention of most people, and most people only pay attention to the top ranked content generated by algorithms.

These observations suggest that selective exposure is not solely responsible for the media's contribution towards partisan animus, since news consumers may actually be more likely to be presented

with ideologically appealing content even without having built up a personalized set of recommendations. For example, simply sharing the same physical location as extreme partisans may lead to increased exposure to extreme news content.

1.4 Hypotheses

It is clear from the existing literature that to some extent, there is evidence of both selective exposure and filter bubbles in the context of consuming news in the market for information. However, it remains an active question as to whether selective exposure and filter bubbles are universal phenomena, or if some individuals are more prone to one phenomena over the other. Specifically, if a news consumer from a specific location seeks information from a news aggregator such as a search engine, are they more or less likely to be recommended ideologically aligned news sources or not? We developed two hypotheses to describe either outcome:

Selective exposure hypothesis: The likelihood of getting a specific news website in search results given a query term is independent of your location, indicating that algorithmic news recommendations are not likely to encourage polarization.

Filter bubble hypothesis: The likelihood of getting a specific news website in search results given a query term is **not** independent of your location, which would suggest that news aggregators like Facebook and Google are unintentionally encouraging the partisan divide.

2 Methods

2.1 SearchLight

SearchLight is an application designed to gather Google search and advertising data on a large scale by automating location-based searching by virtual users. Given a set of locations and search terms, SearchLight launches a browser session, sets the Google user's location in a profile, then performs the search, scrapes the results and ads, and stores this in a database. Each search begins on a fresh browser with no cookies or search history. The database stores the query, location, first page of search results, URLs for the results, advertisements, and the URLs of the ads. SearchLight also comes with several scripts that offer preliminary analysis of discrepancies that emerge between different locations. The code is free and open source, hosted on GitHub[1], and has been designed and run on the GNU/Linux distribution Ubuntu 14.04.

The application was programmed in Ruby and employs the Selenium, Capybara, and PhantomJS libraries. Selenium provides a 'headless' browser, which is designed to look and behave like a normal browser with a human user so that websites treat it like a real human user. Capybara automates clicking and typing within the application so that the user's profile can be populated with the desired location, then types the given search terms into the Google search bar. PhantomJS enables the headless browser to load and interact with client-side JavaScript content. The search and advertising content for each page of search results is then sent to a Mongo database on a central server.

[1] Link will be included upon acceptance

3 Data

3.1 Data collection design

In order to use SearchLight, a list of query terms and locations as either zip codes or city-state tuples is required. We selected 15 query terms, 9 of which were considered to be political in nature and 3 which were unrelated to political issues. A list of "swing states" were chosen according to their vote outcomes in the 2014 election, and we collected all cities within each state using Census data, resulting in a total of 5315 city-state tuples. SearchLight was instructed to run once a day and collect the results from the first page of the Google search results for each city-state tuple and query term combination (a total of about 80,000 queries given all city-state tuples). Refer to Table 1 for the full list of queries and states selected for data collection.

SearchLight collected data from the dates October 14, 2015 to November 1, 2015 for the purpose of writing this extended abstract. Additionally, SearchLight was run from nine separate virtual servers, one for each state, to minimize the risk of cross-contaminating data as the application simultaneously gathered data on different locations. Our experiment will continue to run for the duration of the 2016 election cycle, which will allow us to continually update our model with results as they are collected.

3.2 Preliminary results

The data collection design results in a matrix of search results, where each row of the matrix contains a tuple consisting of a single observation of (query, state, url). We have collected a total of 2,039,439 tuples as of time of writing. Among these search results, we have observed 1974 unique total URLs over the list of queries for all states. Table 2 highlights the top 10 most frequently occurring search results for each of the two query term groups by both raw count and proportional frequency. These tables demonstrate that there are substantial differences in the top domains returned by Google for collections of search terms. For example, a news consumer is about 3 times more likely to be provided a Wikipedia article when searching for politics than they are for our control queries.

4 Next steps

Given our study design, we intend to fit a model that estimates the likelihood of receiving a search result suggesting a particular news domain given a specific location and a query term. By modeling $p(url|state, query)$, we will be able to determine if a potential information-seeking voter in a given state is more or less likely to be directed towards a particular news website over another. If the selective exposure hypothesis is correct, then the probabilities for a given url should be identical regardless of location. But if the filter bubble hypothesis is correct, we should observe different probabilities of a url's recommendation given a specific location and a query.

5 Acknowledgments

Thanks to the ABCDEF[2] for funding the development of Searchlight and the servers needed to run this study.

[2]To be formally included upon acceptance

Query Terms	States
“Jeb Bush”	California
“Donald Trump”	Iowa
“Ted Cruz”	Ohio
“Ben Carson”	South Carolina
“Carly Fiorina”	Virginia
“Marco Rubio”	Florida
“Hillary Clinton”	New York
“Bernie Sanders”	New Hampshire
“republican”	Nevada
“democrat”	Colorado
“election”	
“candidates”	
“basketball”*	
“flu symptoms”*	
“island getaway”*	

Table 1: List of query terms and states. * indicates control query term.

Search result URL	Count	Freq
www.islandgetaway.com	95471	16.33
www.webmd.com	43952	7.52
www.nba.com	18291	3.13
m.christianpost.com	17979	3.07
en.wikipedia.org	17861	3.05
www.theatlantic.com	17857	3.05
www.wired.com	17857	3.05
www.newrepublic.com	17857	3.05
www.flu.gov	17854	3.05
www.cdc.gov	17853	3.05

(a) Control

Search result URL	Count	Freq
twitter.com	165746	11.39
en.wikipedia.org	142627	9.80
www.theatlantic.com	121476	8.35
www.facebook.com	106195	7.30
www.washingtonpost.com	80195	5.51
www.nytimes.com	79971	5.50
www.newyorker.com	52335	3.60
www.cnn.com	46114	3.17
www.biography.com	31250	2.15
www.ontheissues.org	23674	1.63

(b) Politics

Table 2: Top 10 search result counts and frequencies for control versus political query terms

References

Arceneaux, K., Johnson, M., and Murphy, C. (2012). Polarized political communication, oppositional media hostility, and selective exposure. *The Journal of Politics*, 74(01):174–186.

Bakshy, E., Messing, S., and Adamic, L. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, page aaa1160.

Bryan, C. J., Dweck, C. S., Ross, L., Kay, A. C., and Mislavsky, N. O. (2009). Political mindset:

Effects of schema priming on liberal-conservative political positions. *Journal of Experimental Social Psychology*, 45(4):890–895.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5):752.

Epstein, R. and Robertson, R. E. (2015). The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521.

Hamilton, J. (2004). *All the news that's fit to sell: How the market transforms information into news*. Princeton University Press.

Hindman, M. (2008). *The myth of digital democracy*. Princeton University Press.

Iyengar, S. and Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59(1):19–39.

Levendusky, M. (2013). Partisan media exposure and attitudes toward the opposition. *Political Communication*, 30(4):565–581.

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Sears, D. O. and Freedman, J. L. (1967). Selective exposure to information: A critical review. *Public Opinion Quarterly*, pages 194–213.

Stroud, N. J. (2008). Media use and political predispositions: Revisiting the concept of selective exposure. *Political Behavior*, 30(3):341–366.

Sunstein, C. R. (2009). *Republic. com 2.0*. Princeton University Press.