The Effects of Native Advertisement on the U.S. News Industry

by

Manon Revel
Diplôme d'Ingénieur, Ecole Centrale Paris (2017)

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of
Masters of Science in Technology and Policy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2019

Author .....
Institute for Data, Systems, and Society
May 10, 2019

Certified by .....
Munther Dahleh
William A. Coolidge Professor, Electrical Engineering and Computer Science
Director, Institute for Data, Systems, and Society
Thesis Supervisor

Certified by .....
Ali Jadbabaie
JR East Professor of Engineering
Associate Director, Institute for Data, Systems, and Society
Thesis Supervisor

Certified by .....
Dean Eckles
KDD Career Development Professor in Communications and Technology
Thesis Supervisor

Certified by .....
Adam Berinsky
Mitsui Professor of Political Science
Thesis Supervisor

Accepted by .....
Noelle Eckley Selin
Director, Technology and Policy Program
Associate Professor, Institute for Data, Systems, and Society and Department of Earth, Atmospheric and Planetary Sciences

The Effects of Native Advertisement on the U.S. News Industry

by

Manon Revel

Submitted to the Institute for Data, Systems, and Society
on May 10, 2019, in partial fulfillment of the
requirements for the degree of
Masters of Science in Technology and Policy

Abstract

The migration of news to the web has given advertisers new opportunities to target readers with ever more personal and engaging ads. This sponsored content, known as native advertising, is placed in news publications often camouflaged as legitimate news. Though native ads bring revenue to the struggling U.S. news industry, their ability to draw loyal readers off-site could hurt publishers in the long run. Herein, I measure the quality and the impact of the ads from Content Recommendation Networks (CRN) on the U.S. news industry, between March 2016 and February 2019. A CRN controls both the third-party ads and the house ads — recommendations for news articles from the host publisher — on a news publisher's website. During the 2016 presidential election, I found that 17% of ad headlines were political, and 67% of the stories were clickbait. Over the 2018 midterm elections, 15% of the ads were political, and 73% were clickbait. While third-party ads are more clickbait than house ads, the increase in clickbait between 2016 and 2018 is larger for the house ads. Further, I investigate the effect that a one-time exposure to these ads have on the perceived credibility on news articles. Four publishers were under study: CNN, Fox News, *The Atlantic* and *Sacramento Bee*. A one-time exposure to CRN ads was found to have no significant effect on the credibility of traditional publishers. Yet, the CRN ads impacted the credibility of less well-known publishers: ads increased the credibility of the news on *Sacramento Bee*, and decreased it on *The Atlantic*.

Thesis Supervisor: Munther Dahleh
Title: William A. Coolidge Professor, Electrical Engineering and Computer Science
Director, Institute for Data, Systems, and Society

Thesis Supervisor: Ali Jadbabaie
Title: JR East Professor of Engineering
Associate Director, Institute for Data, Systems, and Society

Thesis Supervisor: Dean Eckles
Title: KDD Career Development Professor in Communications and Technology

Thesis Supervisor: Adam Berinsky
Title: Mitsui Professor of Political Science

To Thomas.

I wish you a delightful journey until the roof of the World.

Mount Everest is waiting for you!

Acknowledgments

First and foremost,

To Professors Munzer Dahleh, Ali Jadbabaie, Dean Eckles and Adam Berinsky: thank you for your mentorship over the past two years. Thank you for your support and for allowing me to learn and grow from your rich perspectives. I have been thrilled to discover the research process through your complementary approaches.

I feel grateful to the Technology & Policy Program for offering me the opportunity to be part of the very special TPP family. To Noelle Selin, Frank Field, Barb DelaBarre and Ed Ballo: thank you for your help navigating at MIT.

Thank you to the MIT community, for always being prompt to discuss and to help. In particular, thanks to Amir for the collaboration and the great discussions. Thank you to Tesalia, Chris and Paige from MIT Perl for participating in our research, and to Betsy for the patient help.

Not to forget...

Grace A.: you show me the way! Thank you for your infallible friendship.

Guillaume C.: you are my role model!

Grisha K.: thank you for sharing your passion for the piano.

Louis C.: I will not forget the cooking skills and the laughs...

Matt B.: thank you for the endless stories and the guffaws in IDSS.

Shreyas G.M.: thank you for the relaxing lunches an dinners.

Julien B.: thanks for stopping at that Blue Bike station!

Jocelyn W.L.: thank you for the music!

Jean-Baptiste S.: thank you for philosophical discussions before the midterms...

Tuhin S.: thank you for scaring me about the PhD!

Tanguy M.: thank you for the Tatte breakfasts.

Jude S.: thank you for the wise pieces of advice and the rides in that crazy car.

Eleonore D., Anna C. and Kathy L.: thank you for stopping by Cambridge.

Julie P.: thank you for the butternut squash soups.

Dani R.L.: thank you for the Catalan nougat in New Orleans.

Maxime V.: thank you for your feedback.

Paula M.: thank you for the lazy swimming.

TSAPRR: *We are the champions!* MIT would not be MIT without all of you.

Ferran L.J.: thanks for the honesty, the endless walks, and the alluring singing.

Alex A.U.: let's conquer a Whole New World!

Paolo M.: the five-minute coffees are my favorites.

Aristeidis K.: thanks for wearing a suit on the finals.

Oriol R.B.: thanks for widening my personal library... and for making us dream with a basketball ball in your hands.

Last, but certainly not least,

Even from the other side of the ocean, thank you to my brothers Pierre-Hadrien and Côme: you are my sources of joy and inspiration. Thank you, dad, for your unconditional support! Mum, thank you for being that extraordinary woman who shows me everyday what being strong and brave means.

This research was funded by the Vannevar Bush Faculty Fellowship from the Department of Defense awarded to Professor Jadbabaie.

Contents

List of Figures

List of Tables

Chapter 1

Journalism Crises and Emergence of Content Recommendation Networks

In 2018, only 23% of the Americans said they have "quite a lot" to a "great deal" of trust in newspapers, making the media one of the less-trusted institution[1]. At the same time, traditional media have suffered important layoffs in the United States[2]. A credibility crisis, highly interlinked with a financial crisis, has been hitting journalism in the last decade. The following introductory chapter reflects the situation of journalism in the Web era and introduces a marketing network that financially supports publishers in the short-term.

1.1 Political challenges of the journalism crises

The Web era has dramatically reshaped the news industry. Information reaches a broader audience at a faster pace and journalists must adapt to that new audience and its appetite for immediacy. Consumers tend to get their news on the Web and social media in particular overtake traditional sources as news sources for some audience

[1]https://news.gallup.com/poll/236243/military-small-business-police-stir-confidence.aspx

[2]https://www.npr.org/2019/01/24/688372865/what-the-latest-layoffs-mean-for-digital-journalism

[17], and the variety of sources is incomparably richer than during the printed press era. The news industry struggles to position itself between social media and crowd-sourced approaches and has thus locked itself into an unstable situation and a poor business model that triggered both a financial crisis and a credibility crisis.

Hereafter, I look at the challenges journalism meets in the Web era. I will investigate the worrisome situation of journalism and reflect on the reasons why the Web changed the way journalism operates. Finally, I will focus on sensationalism, suspected to be flourishing on publishers' websites and to harm their credibility.

1.1.1 Worrisome situation of journalism

In the printed press era, the editorial market was centered around oligopolies that secured audience. However, the amount of free news' sources online has disrupted that status. "Nobody has figured out how to get people to pay for outstanding quality" said George Taber, editor at Time Magazine. Consumers are in fact not willing to pay for news online [7] — worldwide, 1 of 10 people pays for news online [25]. Further, while users remember if they searched on Google about the news or if they were on a social media when they were exposed to it, 37% and 47% of the users respectively do not remember the brand from which they read the news [25]. Traditional publishers have been losing their close connection to their audience: they reach the consumers indirectly and through bigger platforms, and they get indirectly paid by marketing.

The number of newsroom employees declined by 23% between 2008 and 2017, and by 45% if we look at the number of newsroom employees in newspapers [3] [14]. Furthermore, while newsroom employees are twice as likely than any other workers to have a college degree, they earn 20% less income than other college-educated workers [14]. The situation worsened in recent years, and "at least 36% of the largest newspapers across the United States as well as at least 23% of the highest-traffic digital-native news outlets experienced layoffs between January 2017 and April 2018." [15]

[3] Newspapers are defined as the "website of a legacy news brand" that was "born" offline.

Journalists create a service and struggle to sell the product of their work. Does that mean that the good itself is obsolete? Or that traditional journalism is unfitted for the Internet?

C.P. Scott (1921) theorized the mission of independent journalism as "gathering the news" making sure that the "supply is not tainted" [29]. Such control over the quality of information requires time to screen the validity and the news; this control has slowly become antinomic with the Web's immediacy. The relation between the media and the society also became more horizontal as one can easily enter the editorial market online, then reply and weight equally as traditional media. Our collective use of the "synchronised, delocalised and correlated" online information sphere (infosphere) [12] seems to render journalism obsolete by creating several challenges, among which the three key are:

- In a "synchronised (time)" infosphere, journalists are expected to deliver the news as things happen, which is incompatible with the mission to rigorously collect and screen information;
- In a "delocalised (space)" infosphere, field journalists often become an unnecessary investment, and the information is gathered from the newsroom on the Web, sometimes at the expense of the (ideal) journalistic objectivity and;
- In a "correlated (interactions)" infosphere, interactions between actors are disrupted as the entrance barrier of the editorial market is close to nonexistent. On the one hand, journalists risk to act as political activists at the expense of their duty to be objective and lose the audience's trust [16]; on the other hand, political activists, bloggers, social media influencers or marketing actors create their own editorial content that replaces journalistic sources and captivates an increasing audience [10].

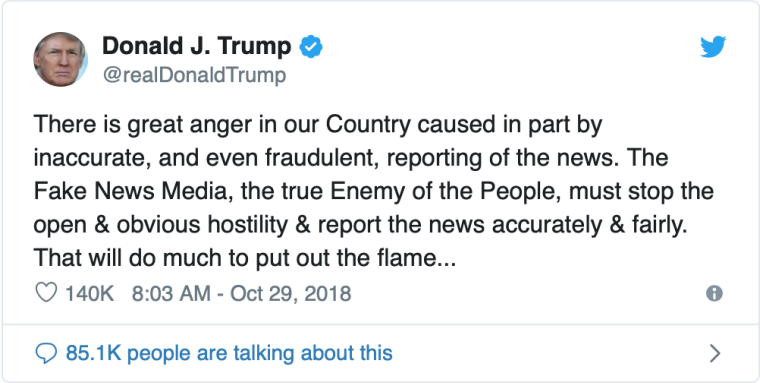


Figure 1-1: Tweet from Donald Trump in December 2018

In 2018, Donald Trump reported that there was "anger in [the] Country," blaming it on the "Fake News Media [that] must stop open & obvious hostility & report the news accurately & fairly." Based on this statement, one could wonder whether media became too partisan to accurately report the news, or if the society became too polarized to listen to cross-cutting voices. Partisanship in media is not new, nor is polarization in society. Yet, phenomena might have scaled up on the infosphere [12]: in this fast-paced Web filled with a myriad of sources, news immediacy overtakes news quality, and catchy titles overtake trustworthy headlines. Yet, legitimate sources that promotes a high-quality journalism fail to make their voices heard. In this thesis, I am interested in understanding what drives outlets' credibility. In particular, I will focus on ad networks that shape the environment of online news and investigate whether they impact the users' perception of news credibility on various sources. Two of these ad networks are prevalent on most of the news publishers and occupy the space under news articles with ads that appear, at first glance, sensationalist.

1.1.2 Sensationalism and clickbait

The possibility to access a wide variety of sources renders the audience volatile. Sensationalism [18] [30] has then been used as a strategy to attract eyeballs on a web-page, creating information of which value and interest are artificially inflated. In particular, clickbait headlines that aim at increasing click-through rates on links emerged as an especially cheap version of sensationalism.

Bandura's research in the *Social Cognitive Theory of Mass Communication* underlines that a common distorted use of mass media relies story-lines that "speak ardently to people's hopes and aspirations for better life" [2]. Bandura claims that sensationalism is motivated by more powerful story-lines that dramatize the realities of people's lives. Similarly, Graber found that sensationalism provoke emotions and empathy and that it is used in TV news: "framing is deliberately dramatic so that the policy-relevant aspects of news are often overshadowed by entertainment features." [13].

Clickbait is defined [4] as "content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page", especially when this content is of "dubious interest or value". Clickbait headlines aim at arousing the readers' curiosity [22], but they might deceive readers by distracting them from their main objective [24], [11]. Headlines such as "Man tries to hug a wild lion, You wont believe what happened next!" or "Mycha started drinking two glasses of bitter-guard juice everyday for seven days and the results are amazing" flourish online.

The reason for sensationalism and clickbait to be under surveillance is because it is suspected to change the way people interact with the media over the long-term. Clickbait is by nature unclear and designed to increase click-through rates. Even if it is not trying to meet a political agenda, clickbait content may play a role in the disinformation phenomenon [5]. Several researchers [28], [6], [5], [27], [1] developed methods to detect clickbait. Further, initiatives enhanced the understanding of the impact of clickbait headlines on news' credibility [20], [9]. However, little is known about the impact of clickbait ads (that surround news online) on the perception of news credibility.

[4]https://www.merriam-webster.com/dictionary/clickbait
https://en.oxforddictionaries.com/definition/clickbait

[5]Starbird et al. defined disinformation as the action to confuse — not to convince — and showed the relation between confusing information and of the dynamic of spreading disinformation [31] [19]

1.2 Advertising and online journalism: a freeze-frame on the Content Recommendation Network (CRN)

Different initiatives aim to financially support the media industry. High-quality outlets created paywalls, and they have required subscriptions to access their articles. While that solution secures a revenue, it creates a self-selected audience. High-quality news become a scarce resource that benefits a fraction of wealthy or aware people. Further, many outlets ask for contributions from their readers to help support a free and independent journalism. Most importantly, publishers earn revenue from advertisement. In particular, recommendation networks distinct from Google ads emerged online. They display third-party ads and share revenues only with publishers. This networks embed native ads — ads that "match the look, feel and function of the media format in which they appear." [6]. Thus, the ads may be mistaken for legitimate news. These networks gained an increasing power over the years, but they are under scrutiny as they are suspected to spread low-quality content through traditional publishers and to create a suspicious news ecosystem on the publishers' websites. These networks are the object of my Master's thesis. Let us introduce them and their characteristics.

1.2.1 Content Recommendation Networks

Content Recommendation Networks (CRNs) are marketing providers that recommend content on publishers' websites. CRNs display widgets filled with native ads at the bottom of news articles, and recommend content that comes comes from third-party providers or from the host publisher. Figure 1-2 is a typical example of CRNs' widget that appear at the bottom of an online article.

Previous research on the impact of native ads on news' credibility [8] [21] found that native ads could have a negative impact on news credibility, even when these

[6] https://www.outbrain.com/native-advertising/

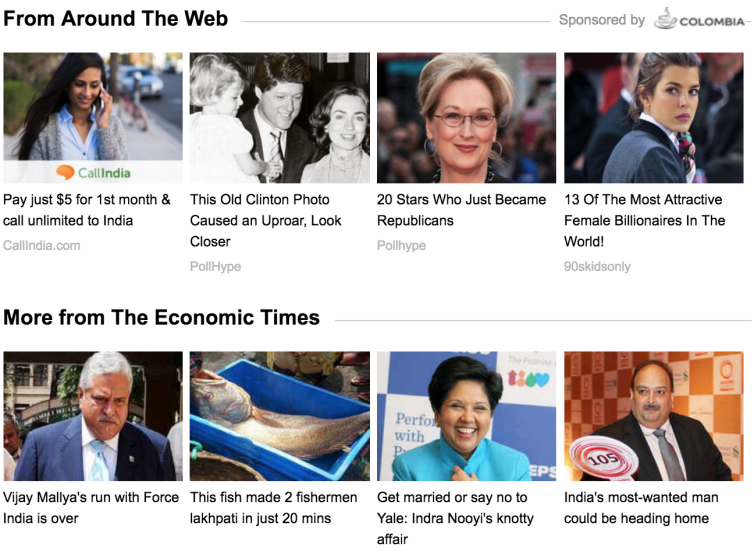


Figure 1-2: CRN's ads promoted by the Indian company Colombia in The Economic Times, on August 8th 2018.

ads were of good quality. Indeed, "content-related ads may cast user doubts over the distinction between content and ads." [8]

Further, previous research on the CRNs [3] found that CRNs conveyed low-quality ads that were not following the disclosure guidelines that applied to sponsored content.

1.2.2 What are CRNs?

CRNs are an intermediary between the publishers and the advertisers. Rather than looking for individual contracts with different advertisers, publishers rely on CRNs to aggregate ads from different sources. All the same, advertisers launch marketing campaigns through the CRNs' platform and thus could appear on multiple publishers' websites. CRNs aggregate ads on the one end, aggregate publishers on the other end, and place ads on publishers as a result. Outbrain and Taboola hold the biggest shares of the market. Other like Danoni, ZergNet or Colombia operate in other market segments. The information below comes from an 2017 interview with a former CRN employee.

Publisher-CRN: a financially-motivated partnership

CRNs became popular among online publishers because they have financially supported publishers. Each reader's click on an ad amounts to an evenly shared income between the publisher and the CRN. CRNs' officials claimed [7] that CRNs generate the greatest part of publishers' revenue. Additionally, a CRN is embedded on a variety of distinct publishers' websites: they can collect valuable information about the readers and provide recommendation for further reading to the user. To that extent, CRNs do not solely recommend ads; they also recirculate house ads (content from the publisher itself). Finally, CRNs also have a consulting activity, as they can track trendy topics elsewhere and provide editorial advice to their privileged partners among pubilshers.

1.2.3 Where to find CRNs?

One can find CRNs everywhere! As of August 2018, 36 out of the 50 top news publishers worldwide[8] embed one or more CRNs. If one does not count the weather and the news aggregators' websites, 35 out of the 40 remaining websites count one to three CRNs.

1.2.4 Why do CRNs matter?

A CRN is embedded on a variety of publishers' websites: as of February 2019, Outbrain's widgets are embedded on 90k websites and Taboola's on 177k [9]. Hence, a CRN collects information about the users' behaviors on different websites. It knows more than the publisher about the publisher's audience. Publishers have given away their recommendation network to the better-informed CRNs, relinquishing at the same time an important part of any online business — the one that allows targeting

[7] http://blogs.ft.com/tech-blog/2014/01/how-outbrain-hooked-publishers-on-content-marketing/

[8] https://www.alexa.com/topsites/category/News

[9] https://www.similartech.com/compare/outbrain-vs-taboola

and re-targeting the audience.

In 2016, the *New York Times*' journalists Sapna Maheshwari and John Herrman raised concerns about those "Around the Web" ads — these ubiquitous promoted stories one encounters at the bottom of articles[23]. Maheshwari and Herrman reported that readers were "starting to express discontent about these articles." The online publisher slate.com removed these ads in 2016 as their quality and ulterior motives were conflicting with the company's "No. 1 priority, readers' trust and loyalty". However, when the site faced the pragmatic reality of the online editorial market, it reversed its decision. slate.com did not reach its 2016 partner Outbrain though; in 2018, it partnered with Taboola — Outbrain's biggest rival.

Further, CRNs recirculate news articles in the ads widgets among the other third-party ads. Researchers found that "recognition of the article as advertising led to decreased perceptions of article quality" [32], and others uncovered that the unclear "separation of editorial functions from advertising" may harm the publishers' credibility [4]. This facts suggest that CRN's camouflage strategy may be a threat to news publishers' credibility.

While the ads are not the primary content users look at, they have become part of the online news environment. Their prevalence as side banners on publishers' websites justifies a deep analysis of their impact on the users' information diet. In the CRNs ecosystem, a tension seems to arise between the short-term economic sustainability of the publishers and the long-term social welfare of the readers. Hence, I ask herein the following research questions:

- **R1:** Are the ads highly politicized and what is the amount of clickbait ads in CRNs?
- **R2:** What is the impact of the CRNs on the readers' perception of the publishers?

1.3 Contributions

This work quantifies the amount of ads that are political and clickbait in the CRNs during the 2016 U.S. presidential campaign, and during the 2018 midterms elections. Chapter 2 presents the data and methodology through which I assessed the topic distribution and the amount of clickbait in CRN ads. Chapter 3 summarizes in 12 claims the key findings about CRNs. They provide, to my knowledge, the first extensive descriptive analysis of the CRN environment.

Further, this work investigates the causal effects that these ads have on the news credibility. A lab experiment was conducted to observe how the CRNs impact differently four publishers. This work builds on previous research [8] that studied native advertisement on news websites to specifically observe the CRNs. Chapter 4 presents the methodology used to detect the causal effects. Chapter 5 summarizes in 4 claims the effects of native ads on news credibility on four publishers: CNN, Fox News, *The Atlantic* and *Sacramento Bee*.

Chapter 2

Data and Methodology to Analyze CRNs

The first section introduces the datasets, the second section shows core statistics about these datasets, and the third one presents the models used to perform the descriptive analysis. The results of the analysis are displayed in Chapter 3.

2.1 Data collection

Two datasets, collected in 2016 and in 2018, in the context of the 2016 US presidential campaign and of the 2018 midterm elections respectively, were used in this research. Both datasets contain information scraped from CRNs' widgets on a variety of publishers' websites.

2.1.1 2016 data

In 2016, Bashir et al. [3] scraped 500 publishers' websites between February 26-March 4, 2016. Note that Super Tuesday (day during the electoral campaign when 14 U.S. states had their primary elections) occurred on March 1, 2016, hence in the middle of the scraping. This dataset contains information about 1 million ads' outline (including headline, host URL, destination URL, visit's time, CRN company), as well

as scraped content from 130,000 webpages that were linked to these ads.

2.1.2 2018 data

In 2018, we scraped 39 targeted publishers between November 1, 2018-February 2, 2019. Note that the midterm elections took place on November 6, and that a government shutdown happened between December 22-January 25. This dataset contains around 50,000 ads' outlines and gathered the headlines, host URLs, destination URLs, visits' time, and CRN companies as well as URLs of the ads' image, and IP addresses (from Boston, Seattle, Houston, San Jose, New York, Miami, Atlanta, and Cambridge).

Choice of publishers: The scope of the scraped publishers was narrowed down from 500 (in the 2016 scrapping) to 39 in 2018. Indeed, the goal was to research the impact of the CRNs on the traditional media, hence to focus on well-known publishers. We combined 23 mainstream publishers with 16 publishers chosen at random among the 2016 data. Potential differences in ad patterns between traditional publishers and random publishers can then be researched.

2.1.3 Data characteristics

Headlines often are redundant, as they could be scraped multiple times. Table 2.1 shows the proportion of unique headlines in each dataset. Unless if explicitly mentioned, all the statistics that follow are drawn from the set of unique headlines. Furthermore, as mentioned earlier, the CRNs' widgets contains both ads from third-party advertisers and house ads from the publishers. House ads are identified when the URL of the visited publisher matches with the URL of the ad. It is of utmost importance to differentiate both categories in the analysis. Table 2.2 shows the proportion of ads vs. house ads.

datasets	Total headlines	Unique headlines	% of unique headlines
2016	1480429	105317	7.1
2018	425942	50785	11.9

Table 2.1: Data Information

datasets	% of third-party ad headlines	% of house ad headlines
2016	57	43
2018	68	32

Table 2.2: Data Information

2.2 Models for descriptive analysis

This section presents the supervised learning methodology used to discover the topics and to uncover the quality of the ads. Further, it presents the Bayes classifiers used for the labeling task.

2.2.1 Building a training set for topic distribution

The goal here was to define a set of topics and to use Amazon-platform Mechanical Turk to build a training set.

Choice of Topics First of all, Natural Language Processing allowed to infer which topics could be spotted within the dataset. Leveraging the 2016 complementary dataset with the ads' content to improve the performance of the NLP algorithms, topics were identified through standard methods Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF).

The clusters that clearly emerge are a Financial topic (clustered around the words: pay, credit, card) and an Entertainment topic (clustered around the words: star, movie, celebrity, hidden, photo). A Political topic also emerges (clustered around the words: trump, cruz, donald). Along these lines, a Technology topic, a Healthcare topic and a Sports topic emerge.

Figures in Appendix C-1 are static representations of the LDA results for a random sample of 10,000 headlines. The Financial topic and the Entertainment topic emerge

as the primary topics in the dataset. Figure C-1b represents the clusters that contain the term "cash back." It shows that a Financial topic emerges including the clusters 1, 2, 3, 7 and 18. All the same, an Entertainment topic emerges on Figure C-1c, including 5, 9, 22, 24 and 28. We see from Figures C-1d that some words are trans-topical. In particular, the word "mortgage" seems to belong to several topics. In Figure C-1e, Cluster 23 appears to be formed around a word directory reflecting a political topic. The word "mortgage" is a salient term in cluster 23 (Political) and in clusters 1, 2 and 3 (Financial). Finally, "trump" is also trans-topical, and appears in the Entertainment and Financial clusters.

Hence, 6 topics emerged: Entertainment, Finance, Politics, Healthcare, Sports and Technology. A Mechanical Turk experiment [1] allowed to gather additional topics. The following topic dictionary was thus established: Entertainment, Finance, Politics, Technology, Healthcare, Sports, Culture/Religion, Romance/Dating, Local News, Retail and Education.

Training Set Two training sets (one for the 2016 data, one for the 2018 data), each containing 5,000 English headlines, were built. Leveraging Mechanical Turk, 1,000 people labeled 100 headlines (each person would be exposed to 100 random headlines from the 5,000 total). A total of 4,816 headlines was labeled in 2016, and 4,223 in 2018. For each headline, the workers were asked to choose up to 2 topics in the topic dictionary. If they chose two, they were asked to specify which was the primary topic and which was the secondary topic. Each headline received an average of 10 labels, and the topic that had the most vote was then chosen for the training set. Overall, we gathered a dataset that could be seen as follows:

[1] Thank you to MIT PERL for helping me setting up the MTurk task, and in particular to Tesalia!

Headlines	**Primary Topic**	**Secondary Topic**
h1	Finance	Healthcare
h2	Politics	NA
...	...	...

Table 2.3: Collected Set for Topics

We turn this data set into a collection of binary classifiers, where topic T is attributed to a headline if the top primary topic, or the top secondary topic is Topic T. The resulting training set is shown below:

Headlines	**Finance**	**...**	**Healthcare**	**Politics**
h1	1	...	1	0
h2	0	...	0	1
...	...	...	...	...

Table 2.4: Training Set for Topics

2.2.2 Building a training set for clickbait detection

We defined clickbait as follows: "something (such as a headline) designed to make readers want to click on a hyperlink especially when the link leads to content of dubious value or interest"[2]. In the survey design mentioned in Sec. 2.2.1, workers read the definition of clickbait and were asked to assess whether the headlines were clickbait or not. We turned clickbait into a binary category, where a headline is clickbait if the majority of the workers agreed and labeled that headline as clickbait. Overall, we gathered a dataset that could be seen as follows:

[2]from https://www.merriam-webster.com/dictionary/clickbait

Headlines	**Clickbait**
h1	0
h2	1
...	...

Table 2.5: Training Set for Clickbait

2.2.3 Bayes detector

A Bayes classifier was ran on the training sets to label the test sets.

We have 13 binary categories (12 topics and Clickbait). Each category is taken independently from the others.

Bayes detector We define a category C, a headline h which is a list of k words (w_1, w_2, ..., w_k), and a random variable H that takes the values 1 when h belongs to C and 0 otherwise. The decision rule to asses whether H is 1 or 0 is as follows:

$$H = \begin{cases} 1 & \text{if } P(H=1|w_1 \cap w_2 \cap ... \cap w_k) \geq \lambda * P(H=0|w_1 \cap w_2 \cap ... \cap w_k) \\ 0 & \text{if } P(H=1|w_1 \cap w_2 \cap ... \cap w_k) * \lambda < P(H=0|w_1 \cap w_2 \cap ... \cap w_k) \end{cases} \tag{2.1}$$

The factor λ added takes into account the fact that the classification may be noisy when the majority of the words in a headline from the test set where not present in the training set. Such headlines would have a ratio of probabilities close to 1 and the labeling decision would not be informative. To remove noisy classification, λ was set to 5.

Following from the Bayes rule:

$$\begin{aligned} P(H=1|w_1 \cap w_2 \cap ... \cap w_k) &= \frac{P(H=1 \cap w_1 \cap w_2 \cap ... \cap w_k)}{P(w_1 \cap w_2 \cap ... \cap w_k)} \\ &= \frac{P(w_1 \cap w_2 \cap ... \cap w_k|H=1)P(H=1)}{P(w_1 \cap w_2 \cap ... \cap w_k)} \end{aligned}$$

We assume that the w_i's are independent, then:

$$P(H=1|w_1 \cap w_2 \cap ... \cap w_k) = \frac{\prod_{i=1}^{k} P(w_i|H=1)P(H=1)}{\prod_{i=1}^{k} P(w_i)}$$

Similarly:

$$P(H=0|w_1 \cap w_2 \cap ... \cap w_k) = \frac{\prod_{i=1}^{k} P(w_i|H=0)P(H=0)}{\prod_{i=1}^{k} P(w_i)}$$

Hence,

$$\boxed{\frac{P(H=1|w_1 \cap w_2 \cap ... \cap w_k)}{P(H=0|w_1 \cap w_2 \cap ... \cap w_k)} = \frac{\prod_{i=1}^{k} P(w_i|H=1)P(H=1)}{\prod_{i=1}^{k} P(w_i|H=0)P(H=0)}} \tag{2.2}$$

The quantities $P(H=1)$, $P(H=0)$, $P(w_i|H=1)$, and $P(w_i|H=0)$ can be learned from the training set.

$$P(H=1) = \frac{\text{Number of headlines with label 1}}{\text{Total number of headlines}}$$

$$P(H=0) = \frac{\text{Number of headlines with label 0}}{\text{Total number of headlines}}$$

$$P(w_i|H=1) = \frac{\text{Number of occurrences of } w_i \text{ in headlines labeled as 1}}{\text{Number of words in headlines labeled as 1}}$$

$$P(w_i|H=0) = \frac{\text{Number of occurrences of } w_i \text{ in headlines labeled as 0}}{\text{Number of words in headlines labeled as 0}}$$

Finally, we regularize and we take into account the fact that some words that appear in the test set might not appear in the training set. As a result, we add a Laplacian smoothing parameter α in the conditional probability:

$$P(w_i|H=1) = \frac{\text{Number of occurrence of } w_i \text{ in headlines labeled as } 1 + \alpha}{\text{Number of words in headlines labeled as } 1 + \alpha * \text{Number of words in H}}$$

$$P(w_i|H=0) = \frac{\text{Number of occurrence of } w_i \text{ in headlines labeled as } + \alpha}{\text{Number of words in headlines labeled as } 0 + \alpha * \text{Number of words in H}}$$

Equation 2.2 defines a rule to label the headlines from the test set based on the empirical quantities derived from the training set.

In the training set, one headline would belong at most to two topics. As each "topic" category is treated independently, one headline in the test set may belong to one, two or more topics. This methodology choice is motivated first by the goal to study primarily the number of politicized headlines, and second by the goal to understand the overlap of different topics, which is better achieved with such a method compared to a multinomial regression. However, a multinomial regression gave similar trends in the topic distribution — with different orders of magnitudes though, as fewer headlines are labeled as 1.

2.3 Model validation

The performance of the detectors was evaluated through cross-validation on the training set. The most salient terms per category also allows to make sure that the topics captured made sense. The results for the 2016 and the 2018 training set can be found in Table 2.6.

While these results show that the detectors perform a cogent work, it is interesting to note that the salient words within the clickbait category evolve over time, and adapt whatever trendy topic there is at this point in time.

As mentioned earlier, the methodology used to label the topic categories may create more than two labels for one headline. Also, because the headlines may be formed of words not found in the training set, we may fail to label some headlines.

Detector	Accuracy	Salient words
Clickbait	92%	trump, us, donald, best, celebrity, credit, clinton, new, star, top, photos, history, hillary, president, love
Political	98%	trump, donald, clinton, us, gop, hillary, president, sanders, obama, bernie, republican, cruz, trumps, campaign, bill
Entertainment	97%	celebrity, trump, us, oscars, donald, star, tv, movies, best, movie, photos, hollywood, new, oscar, wars
Finance	99%	credit, card, us, cards, stocks, pay, paying, market, highest, dividend, security, stock, best, business, social
Healhcare	99%	health, cancer, weight, us, skin, signs, new, loss, zika, women, best, ways, care, lose, treatment
Technology	98%	apple, security, fbi, us, iphone, cars, english, globes, new, google, email, cyber, data, car, business
Local News	97%	shooting, trump, us, woman, school, police, man, donald, new, death, clinton, car, high, basketball, gop

Table 2.6: Model Validation for the 2016 training set

Detector	Accuracy	Salient words
Clickbait	95%	photos, new, trump, pics, best, gallery, one, says, home, people, us, know, top, life, heres
Political	98%	trump, us, says, house, trumps, president, new, white, china, election, democrats, war, trade, obama, gop
Entertainment	98%	photos, gallery, pics, new, see, like, world, best, life, wife, man, know, never, years, people
Finance	97%	home, best, credit, stocks, could, market, insurance, get, card, money, us, people, seniors, heres, new
Healhcare	99%	new, signs, cancer, health, top, know, never, foods, hearing, eat, one, says, arthritis, try, best
Technology	98%	new, theinquirer, best, solar, home, cars, drone, search, tech, get, us, security, free, time, could
Local News	98%	new, man, florida, woman, drivers, california, police, pics, photos, yearold, found, dead, trump, massachusetts, years

Table 2.7: Model Validation for the 2018 training set

In the 2016 test set , we successfully labeled 80% of the headlines and 79% of the headlines belong to one or two topics. In the 2018 test set , we successfully labeled 77% of the headlines, and 73% of the headlines belong to one or two topics.

Chapter 3

Descriptive Analysis of the CRNs

In this chapter, I answer the first research question: Are the ads highly politicized? What is the amount of clickbait ads, and what does the clickbait measure represent? Leveraging the methodology exposed in Chapter 2., I conclude that the proportion of clickbait ads increased over time, and that up to 1 ad out of 7 was politicized in 2018. To our knowledge, these ads are not trying to push a political agenda. They may be contributing to what Starbird [1] calls the "disinformation process" that spreads confusing and misleading information and increase the level of uncertainty and the appetite for sensationalism.

3.1 Topic distribution in CRNs

3.1.1 Claim 1: top topics are Entertainment, Politics and Personal Finance

Claim: CRN do not embed placement of product ads. Instead, the top topics in the ads are Entertainment, Politics and Personal Finance.

Evidence: In the list of the 50 most common words in 2016's third-party ads, one finds the words "best," "photos," "new," "celebrity," "believe," "amazing," "tips" as

[1] https://medium.com/hci-design-at-uw/information-wars-a-window-into-the-alternative-media-ecosystem-a1347f32fd8f

well as "credit," "stocks" and "trump." In the list of the 50 most common words in 2018's third-party ads, one finds the words "new," "photos," "best," "pics," "top" while the word "trump" become more popular than in 2016.

Figure 3-1 show the topic distributions for both datasets. Entertainment is the most common topic, followed by Politics and Personal Finance.

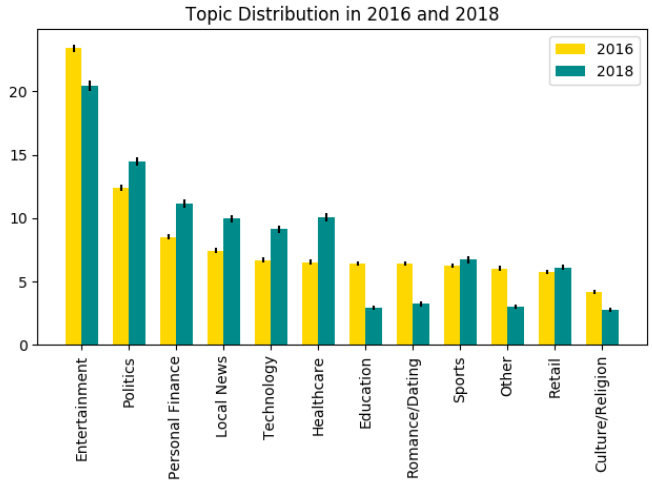


Figure 3-1: Topic distributions in 2016 and 2018.

3.1.2 Claim 2: different topics for different ad types

Claim: While third-party ads and house ads are mixed on CRN widgets, third-party ads and house ads have different topic distributions.

Evidence: Figures 3-2 and 3-3 show that the main topics are different in third-party ads and in house ads. Politics is the major topic in house ads, while it is the third one in third-party ads. Similarly, Local News is the third topic in house ads while it is the least represented in third-party ads, and Sports is more frequent in house ads than in third-party ads. On the other hand, Personal Finance is much rarer in house ads than in third-party ads. Entertainment is prevalent in third-party ads, but often appears in house ads too.

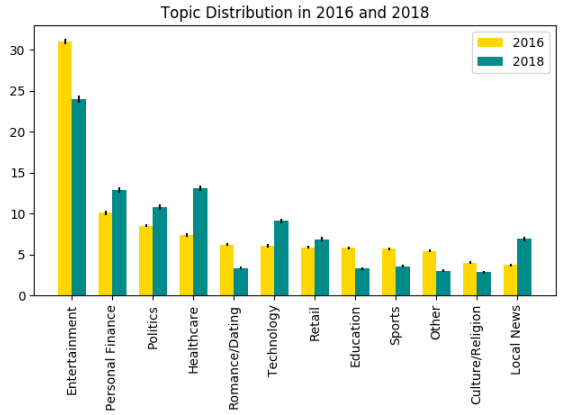


Figure 3-2: Topic distributions in 2016 and 2018 in third-party ads.

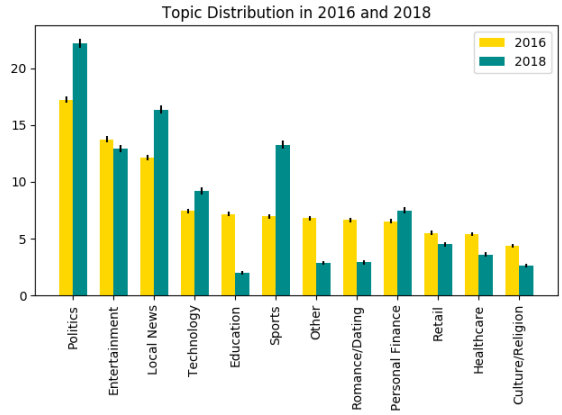


Figure 3-3: Topic distributions in 2016 and 2018 in house ads.

Similarly, Figures 3-4, 3-5, and 3-6 show the time series of the topics in the aggregated ads, third-party ads and house ads respectively, between November 2018 and February 2019. Different patterns emerge.

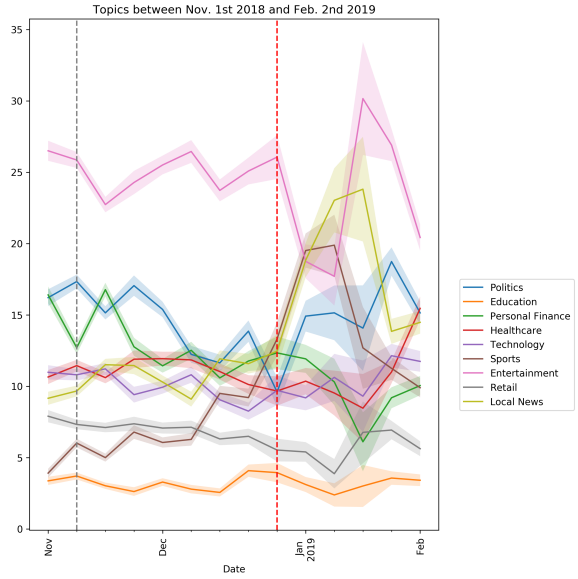


Figure 3-4: Time Series of topics in ads.

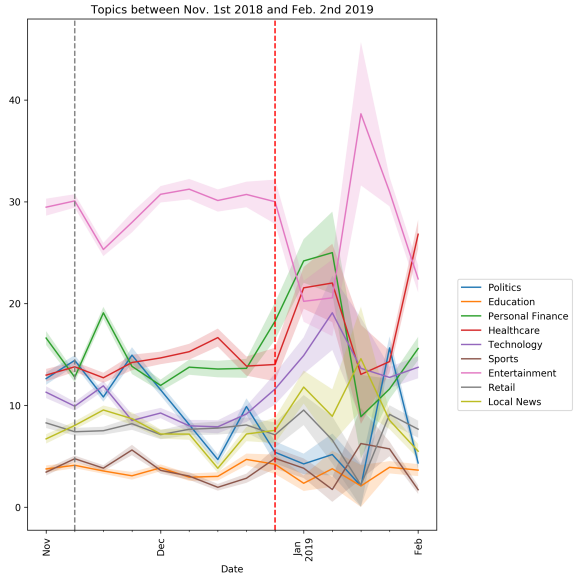


Figure 3-5: Time Series of topics in third-party ads.

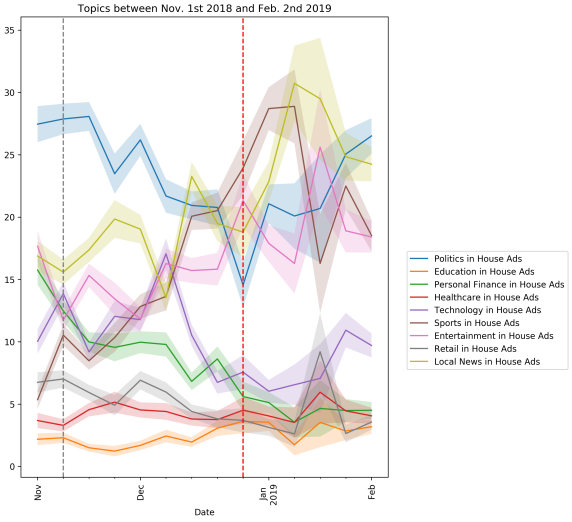


Figure 3-6: Time Series of topics in house ads.

3.2 Political Ads in CRNs

3.2.1 Claim 3: CRN ads are politicized

Claim: On average, 1 out of 7 ad was political either in both electoral periods — 2016 and 2018. Similarly, 1 out of 9 third-party ads was political.

Evidence: Political keywords appear in the datasets: 2.3% (resp. 4.1%) of the headlines in 2016 (resp. 2018) contain the words "donald" or "trump." In 2016, 35% of these headlines were third-party ads, while they were 50% of such third-party ad headlines in 2018. Further, headlines containing "trump" or "donald" in the third-party ads headlines were three times more frequent than headlines containing the words "hillary" or "clinton." This factor becomes 20 in the 2018 data, as Hillary Clinton was not a highly topical issue anymore.

Overall, we found that 17% of the ads were political in 2016, compared to 15% in 2018. In both years, 11% of the third-party ads were political. As expected, the house ads have higher concentration of political content than third-party ads (11% of the third-party ads were political vs. 23% of the house ads in 2016; and 11% of the third-party ads were political vs. 27% of the house ads in 2018).

3.2.2 Claim 4: CRN ads are more politicized during political events

Claim: The amount of political ads was higher during the midterms and during the government shutdown than around Christmas.

Evidence: Figure 3-7 shows the evolution of the political content from November 2018 until February 2019. A decline in the proportion of political ads is observed around Christmas, while the amount of political ads is higher around the midterms and the government shutdown.

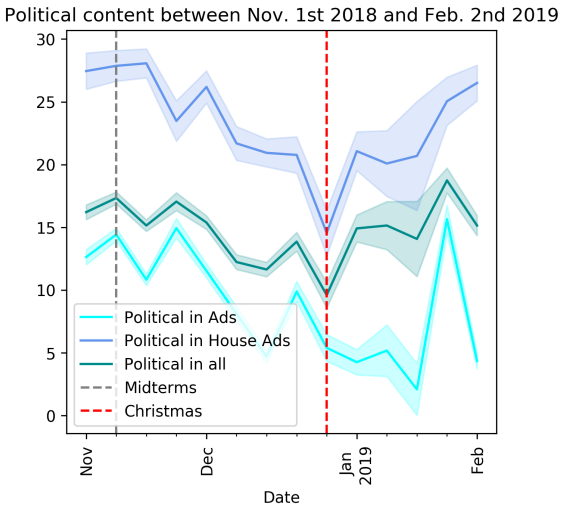


Figure 3-7: Time Series of political ads.

3.3 Clickbait ads in CRNs

3.3.1 Claim 5: "clickbait" describes the style and the tone of the headlines

Claim: In contrast with what was expected, clickbait is not a proxy for quality. What differentiates clickbait headlines from non-clickbait headlines is the style and the tone.

Evidence: The features that differentiate clickbait ads from not clickbait ads are the use of words and the use of punctuation. Clickbait ads call out the reader and in fact the word "you" is used 7.5 times more often in clickbait ads than in non-clickbait ads, such as in "[Gallery] Stars You May Not Know Passed Away". All the same, clickbait language uses nearly twice as much active verbs such as *see*, *can*, *look* or *do* than non-clickbait language, such as in "Trump successes are boosting GOP candidates in midterms Dont expect a blue wave". Finally, clickbait style contains nearly three times more punctuation (?, ! and ...) than non-clickbait style. Non-clickbait headlines tend to be more formal and factual, even if the facts they state

are not credible or irrelevant, such as in "Secret Service notified after Peter Fonda tweets about Barron Trump" or in "10 Signs of Bipolar Disorder".

Overall, these results are consistent with the general idea of clickbait that aims at filling the curiosity gap, engaging the readers to click on the links.

3.3.2 Claim 6: the majority of the CRN ads are clickbait

Claim: Between 67 and 73% of the headlines were clickbait either in 2016 or 2018. 81% and 83% of the third-party ads were clickbait in 2016 and 2018 respectively. This fact means that in the ads campaign, over 4 out of 5 headlines are clickbait. The increase in the proportion of clickbait ads between 2016 and 2018 was higher for house ads than for third-party ads.

Evidence: Figure 3-8 shows the proportion of clickbait ads per ad types. The proportion of clickbait ads increased by 2% for the third-party ads, and by 4% in the house ads. Figure 3-9 shows the evolution of the proportion of clickbait ads between November 2018 and February 2019. Third-party ads are more clickbait than house ads. Further, the amount of clickbait ads increased around Christmas in third-party ads while it decreased in house ads.

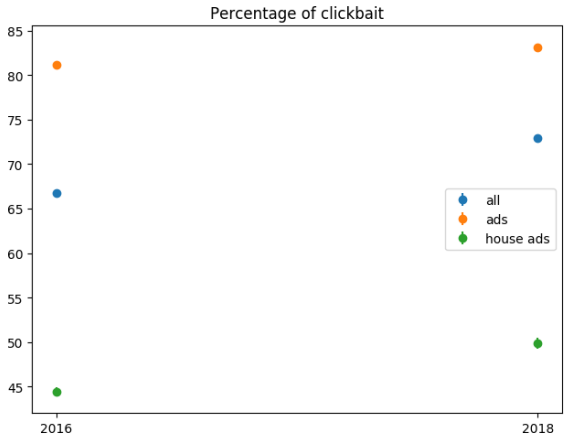


Figure 3-8: Clickbait percentage per ad type.

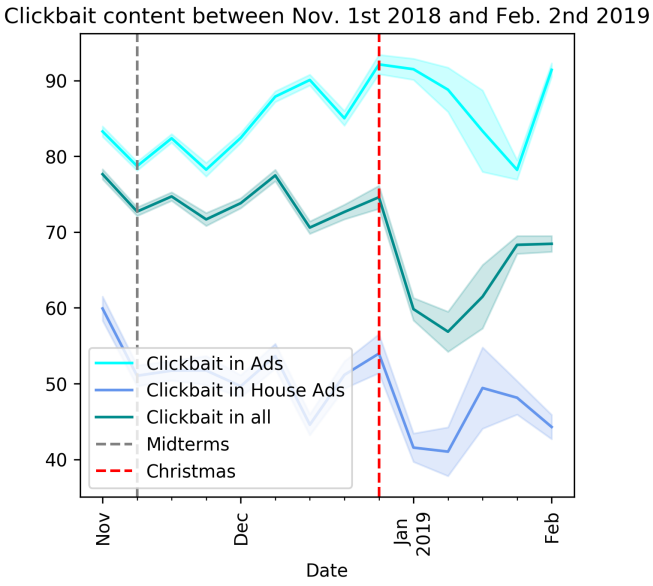


Figure 3-9: Time series of clickbait ads.

3.3.3 Claim 7: all topics are clickbait

Claim: Clickbait is shared among the topics.

Evidence Figure 3-10 shows that in each topic, at least 40% of the ads are clickbait. In 2016, Local News and Politics had less clickbait ads than the other topics, but clickbait still represents 40% of the political ads.

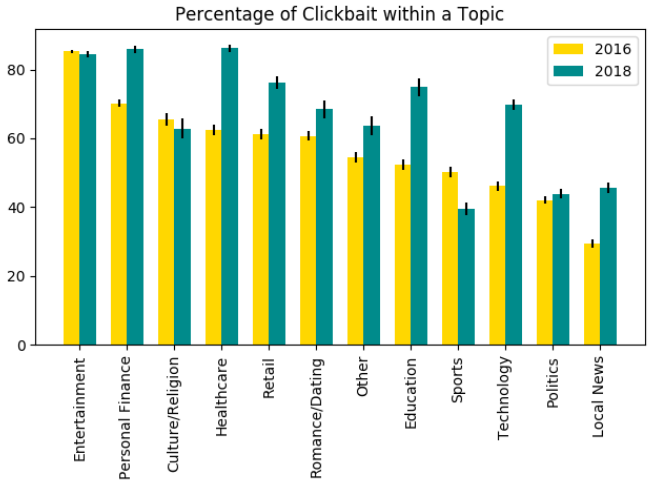


Figure 3-10: Clickbait headlines in house ads.

3.4 Ad campaigns target the users

3.4.1 Claim 8: the audience is targeted based on location

Claim: CRN ads differ based on the reader's location. In fact, an advertiser can target its audience based on an location.

Evidence The ads' patterns differ based on the IP addresses used. In 2018, data were scraped through IP addresses corresponding to Seattle, San Jose, New York, Miami, Houston, Boston and Atlanta. Figures 3-11 and 3-12 respectively show the proportion of political and clickbait headlines respectively per city with the 95% confidence intervals.

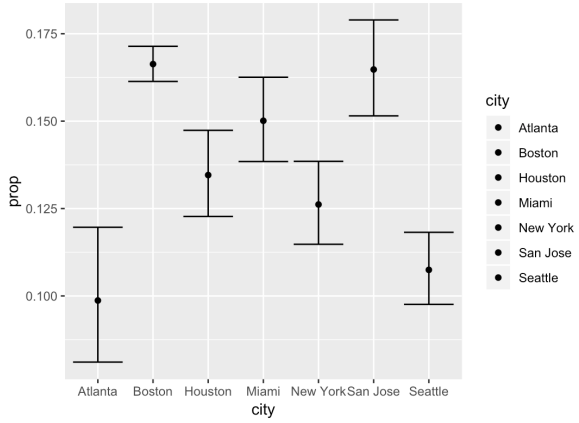


Figure 3-11: Political proportion per location.

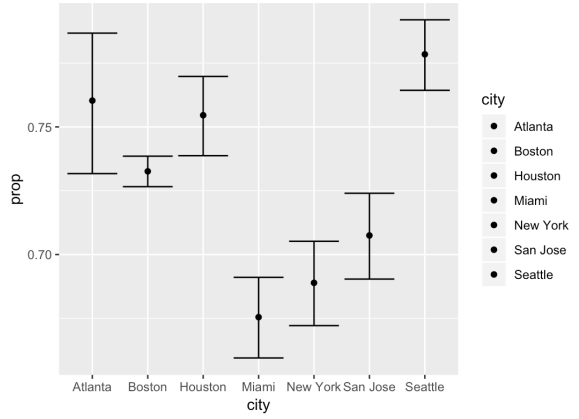


Figure 3-12: Clickbait proportion per location.

These results suggest that the ad campaigns significantly differ based on certain location. This could be a result of the optimization of the ad auctions, but it is interesting to notice that both Taboola and Outbrain's websites offer a feature that allows advertisers to target audiences (see 3-13 and 3-14).

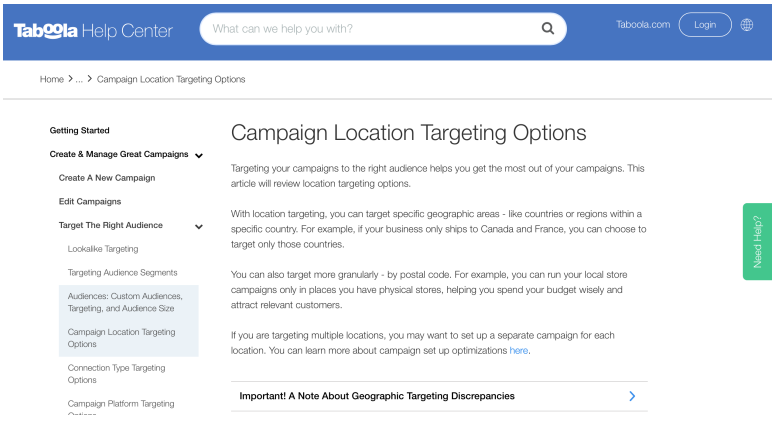


Figure 3-13: Taboola's policies.

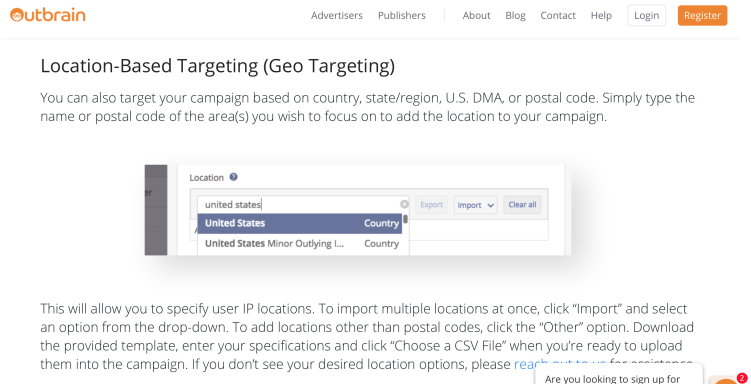


Figure 3-14: Outbrain's policies.

3.5 Different patterns in publishers

3.5.1 Claim 9: the ad campaigns differ per publisher

Claim: Different publishers have different patterns.

Evidence: The Huffington Post and the Washington Post had much more political third-party ads in 2016 than the other publishers, and CNN and Fox News had much more political third-party ads in 2018 than the other publishers. In 2018, Fox News shows an interesting pattern: it has less clickbait third-party ads than other

publishers, but it has more third-party political ads. To be noted that *The Atlantic* removed CRN's ads during our 2018 studies. Figures 3-15, 3-16, 3-17 and 3-18 show the different trends in the different publishers for proportion of political content in ads and third-party ads and proportion of clickbait content in ads and third-party ads respectively.

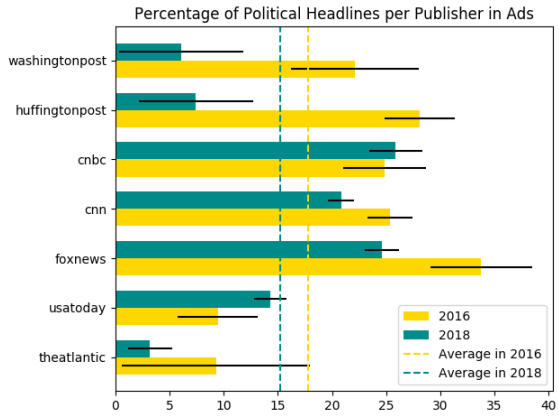


Figure 3-15: Political among all ads.

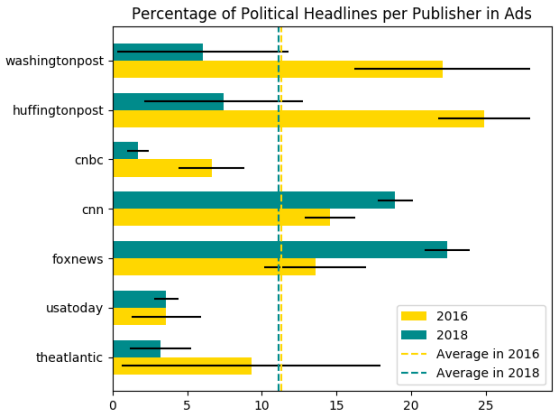


Figure 3-16: Political among third-party ads.

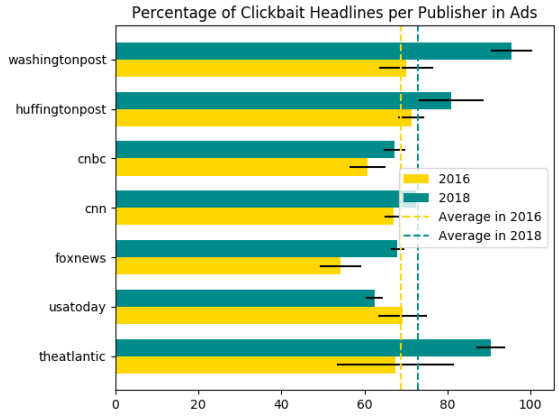


Figure 3-17: Clickbait among all ads.

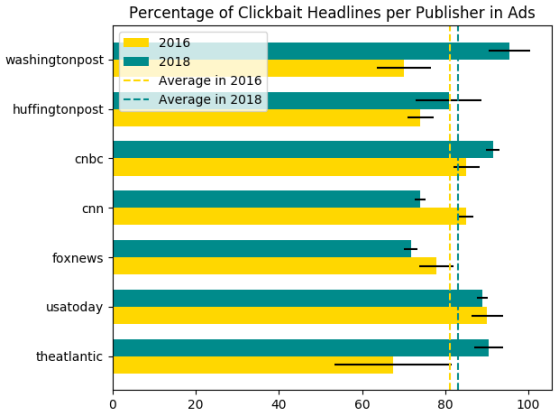


Figure 3-18: Clickbait among third-party ads.

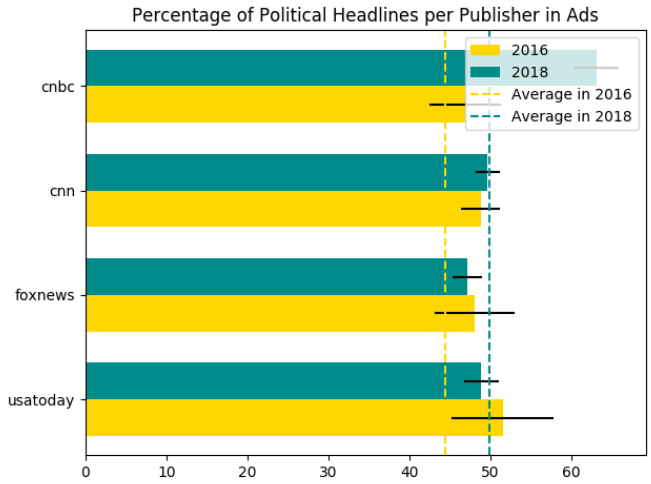


Figure 3-19: Clickbait among house ads.

3.5.2 Claim 10: tone of news articles vary per publisher

Claim: Further, the tendency to have more clickbait content does not apply exclusively to the third-party ads. House ads have also increasingly become clickbait, meaning that the news articles headlines are perceived as clickbait.

Evidence:

Further, news headlines from the house ads are also sometimes perceived clickbait. Figure 3-19 show that the percentage of clickbait house ads increased between 2016 and 2018, and that different publishers have different amount of clickbait house ads. It is important to keep in mind that the measure of clickbait is not a quality measure. It is also not defined systematically, but was learned through a labeling task performed on two sets of 1,000 subjects on Mechanical Turk. As mentioned earlier, clickbait relates to the style of the headlines, yet not-clickbait headlines can still be of low-quality.

3.5.3 Claim 11: ads are a shared resource

Claim: Sometimes, the same ad appear on different publishers.

Evidence: As the two biggest CRN companies share most of the publishers, one

could expect that the same ad campaign is ran on 2 publishers which partners with the same CRN. One could wonder if two publishers, very different in appearance, could share the same ads pattern as a result of their financial partnership with the same CRN.

Indeed, looking at the redundant headlines, 17% of the headlines were found to be shared by more than one publisher. Among these shared headlines, one headline was shared on average by 4 publishers. Surprising associations arose such as headlines shared by CNN and Breitbart. The same headlines usually came from the same advertiser.

On a randomized set of 1,000 headlines (reduced for computational ease), headlines with 60% or more of similarity were grouped together. Similar headlines were found to be shared among a wide variety of publishers. For instance, Breitbart shared similar headlines with (ordered as the number of shared headlines decreased): *NY Daily News, CNN, FoxNews, The Atlantic, Telegraph.* Similarly, *FoxNews* shared similar headlines with: *Breitbart, USA Today, The Guardian, The Atlantic, NY Daily News* and *CNN.*

3.6 Two big players in the CRNs' market

3.6.1 Claim 12: CRNs embed different ad campaigns

Claim: Taboola and Outbrain, the two biggest CRN companies embed different patterns of ad campaigns. Outbrain had more political third-party ads than Taboola and Taboola had more clickbait third-party ads.

Evidence: Outbrain constantly had more political headlines than Taboola. The proportion of political ads on Outbrain decreased between 2016 and 2018 (Figure 3-20), while the proportion of third-party political ads stayed constant (Figure 3-21). Hence, the decrease is due to the house ads, i.e., to the publisher's editorial choices. Similarly, while ads on CRNs were more clickbait on Outbrain than on Taboola in 2018 (Figure 3-22), Taboola had more third-party ads than Outbrain (Figure 3-23).

Hence, Outbrain's partners create much more clickbait news headlines than Taboola's partners.

Interestingly, Outbrain used to lead the market, but Taboola overtook the lion's share over the past years. If Taboola were to partner with high-quality publishers, Taboola may be more likely to embed house ads that are not clickbait but that are political. This is a possible interpretation for the facts explained in Claim 12.

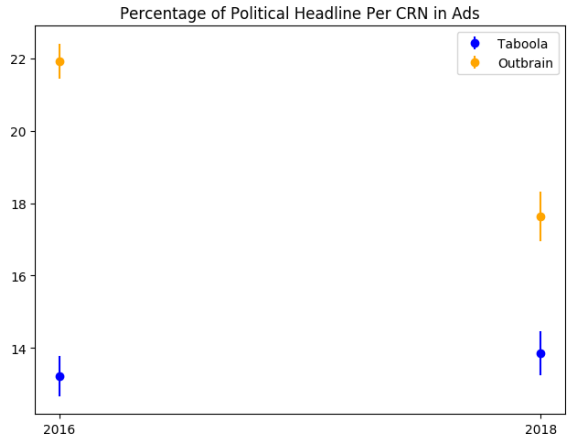


Figure 3-20: Political among all ads per CRN.

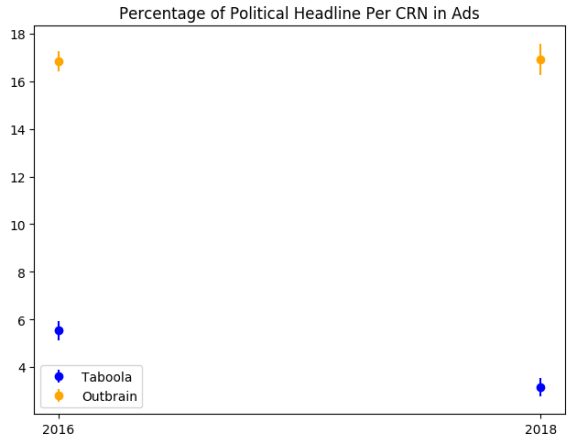


Figure 3-21: Political among third-party ads per CRN.

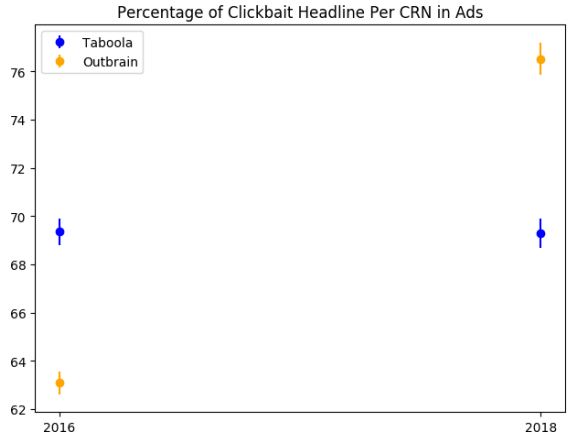


Figure 3-22: Clickbait among all ads per CRN.

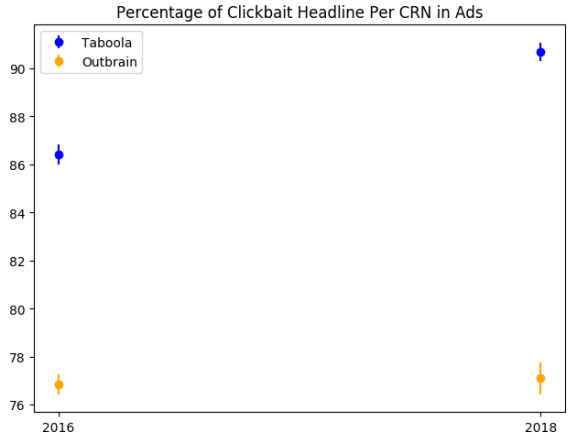


Figure 3-23: Clickbait among third-party ads per CRN.

3.7 Key takeaways

CRN third-party ads were found to convey political information during election periods. Further, evidence proved that CRN ads spread clickbait content, calling out the reader to click on the third-party ads. Such clickbait headlines could serve as a substitute to the news story, hence divert consumers away from the news website. It is in the interest of native advertisement to camouflage as a legitimate news, but it might have a negative impact on the publishers' traffic if it pull the audience away from the news.

Further, because the CRNs also control the recommendation of the publishers, these datasets allowed to look into the style of the house ads — news headlines created by the publishers. Interestingly, house ads were also perceived as clickbait. This might suggest that sensationalism is not solely present in marketing — it might also be in journalism.

While this chapter exposed a descriptive analysis of the CRNs, its impact the users' perception of the news is still not understood. Chapters 4 and 5 explore that question.

Chapter 4

Design of behavioral experiment

Here, I present the methodology used to measure the impact that CRNs have on the publishers' credibility. While I aim at understanding the long-term impact of the CRNs on the audience's loyalty, the scope of this thesis extends to measuring the impact of a one-time exposure to a set of CRN ads.

4.1 Design

Hereafter, I explain the design of the behavioral experiment measures the CRNs' impact on the publishers' credibility.

4.1.1 Articles' design

The goal is to create twelve pages, where each page comprises an article and a mix of CRN ads. Each article comes from a different publisher. There are a total of four articles (hence four publishers) and three ad mix, so a total of twelve pages.

Choice of publishers: The four publishers under study are: CNN, FoxNews, *The Atlantic* and *Sacramento Bee*. Two traditional publishers with different political affiliations (CNN and FowNews) and two less well-known publishers (*The Atlantic* and *Sacramento Bee*) were selected.

Choice of articles: One timely article per publisher was chosen the week before the launch of the pilot experiment. The four articles were soft news stories that most likely did not already reached our audience. Table 4.1 shows the four headlines of the four articles from the four publishers.

Publisher	Headline
CNN	Ex-nurse accused of impregnating a severely disabled Arizona woman pleads not guilty
FoxNews	Cuomo blames federal tax law for $2.3 billion New York budget deficit
The Atlantic	Now your groceries see you, too
SacBee	California's retail marijuana industry is struggling. Will tax breaks and bank helps?

Table 4.1: Headlines of articles per publisher.

Choice of ads: We selected eight CRN third-party ads: four are general ads and four are political ads. In Figure 4-1 one can see the three different ad regimes.

4.1.2 Lab experiment

Questionnaire : In order to collect answers, monitor the activity, and understand the characteristics of the surveyed people, we designed a questionnaire [1] on Qualtrics [2]. Table 4.2 shows the survey flow in blocks to which survey takers were exposed in chronological order. Table 4.2 also shows the purpose of each block.

[1] Thank you to MIT PERL for helping me setting up the MTurk task, and in particular to Chris and Paige!

[2] https://qualtrics.mit.edu

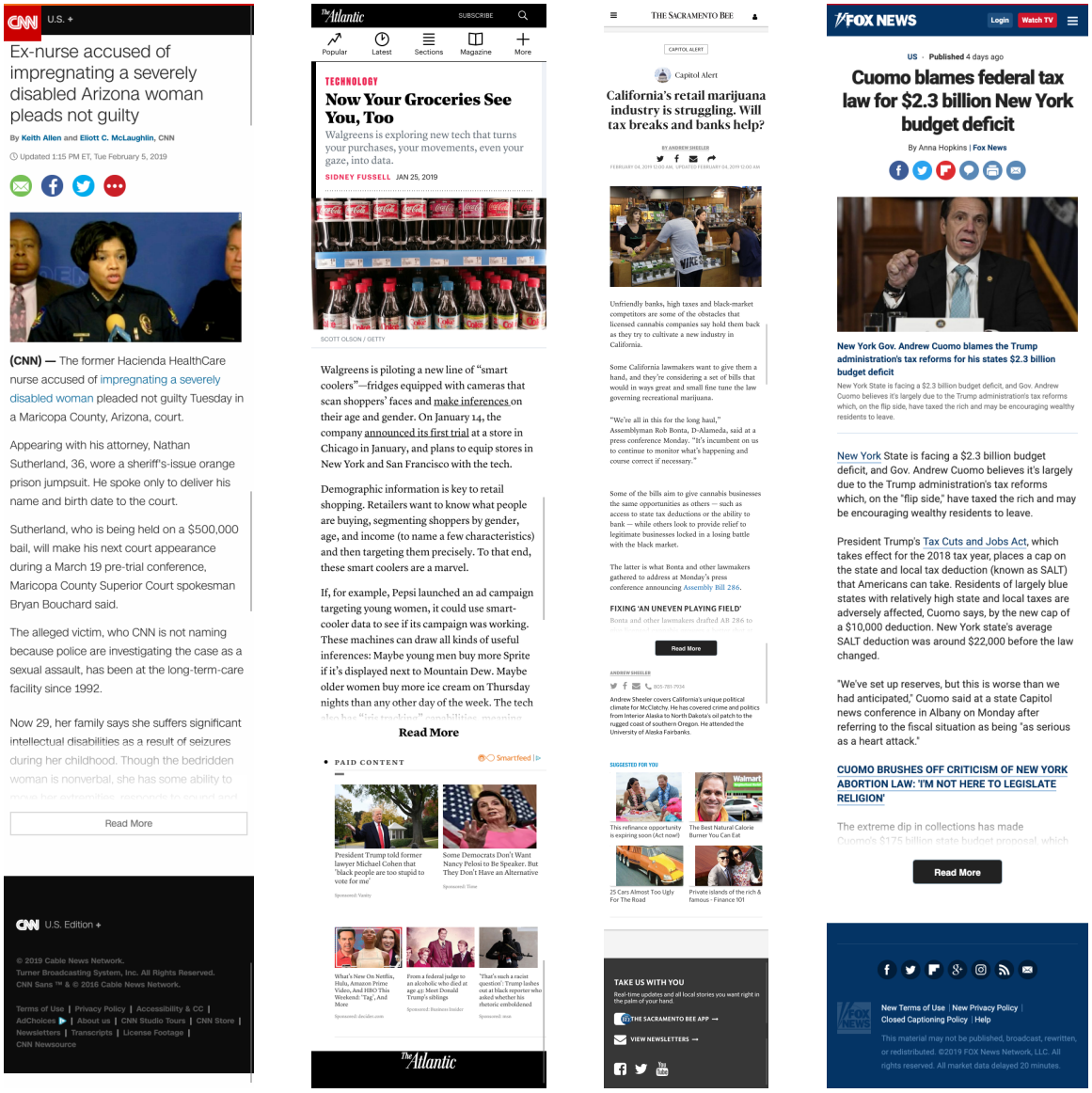


Figure 4-1: Sample publisher-ad combinations.

Survey Flow	Purpose	Number of Questions
Consent	Collect legal consent	1
Attention Screener	Check if person reads the prompt	1
Demographics	Make sure the sample is representative	19
Political Knowledge	Measure the knowledge about political facts	5
Cognitive Reflection Test	Measure the capacity to think logically	3
Media Engagement	Account for habits	2
Attention Screener	Check if person reads the prompt	1
Articles specific question	Blur the trust questions with general questions	Depends
Trust questions	Measure articles' credibility	5
Media Hostility	Account for the prior opinions	4
Attention Screener	Check if person reads the prompt	3

Table 4.2: Questions' headings, questions' use, and number of questions per block.

The **attention screener blocks** aim at monitoring whether the survey takers pay attention to the question or click without reading the prompt. They typically look as in Figure 4-2, expect at the end of the survey where they look as in Figure 4-3. Three attention screener were displayed in the survey to monitor a potential drop in the subjects' attention.

We would like to get a sense of your general preferences.

Most modern theories of decision making recognize that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. To demonstrate that you've read this much, just go ahead and select both red and green among the alternatives below, no matter what your favorite color is. Yes, ignore the question below and select both of those options.

What is your favorite color?

- ☐ White
- ☐ Pink
- ☐ Black
- ☐ Green
- ☐ Red
- ☐ Blue

Figure 4-2: Attention Question Type 1.

Please mark whether you agree or disagree with each of the statements below.

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly
World War I came after World War II	○	○	○	○	○
The federal government should guarantee health insurance for all citizens	○	○	○	○	○
In order to reduce the budget deficit, the federal government should raise taxes on people that make more than $250,000 per year	○	○	○	○	○
Gays and lesbians should have the right to legally marry	○	○	○	○	○
The country is headed in the right direction	○	○	○	○	○
Please click the "neither agree nor disagree" response	○	○	○	○	○

Figure 4-3: Attention Question Type 2.

Further, the **cognitive reflection test** is a common test used in behavioral studies to measure the level of cognitive reflection of test takers [26]. It allows to monitor the subject's capacity to think about logic problems. The **political knowledge questions** assess the political knowledge about basic facts. Typical questions for both cases are shown in 4-4 and 4-5 respectively.

If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

min

Figure 4-4: CRT question.

Whose responsibility is it to decide if a law is constitutional or not?

- ○ The President
- ○ Congress
- ○ The Supreme Court

Figure 4-5: Political knowledge question.

The **media engagement** and **media hostility questions** help understand the prior opinions of the survey takers. Indeed, we want to differentiate users who are very hostile vs. very pro CNN when we compare their answer about the credibility of the CNN article.

There are five **trust questions**. We ask the test takers if they strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree or strongly disagree with the following five statements:

- q_1 The article was trustworthy.
- q_2 The article was false.
- q_3 The article was credible.
- q_4 The article was biased.
- q_5 The article provides news information.

Go to the field: We partnered with Lucid [3] to survey a representative set of American people.

4.1.3 Randomized experiment

Finally, we randomized the experiment according to a mixed design approach: treatment varies within-subjects, but subjects do not read all 12 conditions (each four article exists under all three ad treatments). Indeed, each survey taker was exposed to all three CRNs conditions and all four publishers. Each persons saw two articles with no ads, one article with mixed ads, and one article with political ads. The publisher-ad combination was randomized for each person, and the order in which they appear was also randomized. Finally, the order in which the trust questions were asked was also randomized.

[3] https://luc.id

4.2 Analysis

Through the experiment, we collected people's answers to trust questions for different treatment conditions. Condition c_0 is the control treatment, i.e., subject is exposed to no ads, and c_m and c_p characterize subjects exposed to mixed ads and political ads, respectively. I present hereafter the stratification strategy as well as the Fisherian randomization and the mixed effects model used to measure the impact of the ad treatments on the publishers' credibility.

4.2.1 Stratification strategy

The study was conducted on all the surveyed people who agreed and finished the survey, or on a stratified set of attentive people. Based on people's answers to the attention screener, data were stratified between people who correctly answered the first attention question and people who did not. Further, studying the effects per publisher, we stratify the data based on subject's prior familiarity with the outlets.

4.2.2 Fisherian randomization

The idea of the Fisherian randomization is to reveal if the observed average treatment is likely to occur under the null hypothesis. Let us assume the case of a binary treatment assignment (treated vs. control). I_0 is defined at the set of individuals i's assigned to control, and I_1 the set of individuals i's assigned to treatment. The outcome $Y_i(0)$ denotes the answer about the article's credibility given by subject i in I_0, and the outcome $Y_i(1)$ denotes the answer about the article's credibility given by subject i in I_0. The table of the results looks like in Tab. 4.3.

Person	Publisher	Treatment	$Y_i(0)$	$Y_i(1)$
Id1	CNN	1	?	5
Id1	FoxNews	1	?	4
Id1	*The Atlantic*	0	4	?
Id1	SacBee	0	4	?
Id2	CNN	0	2	?
Id2	FoxNews	1	?	3
Id2	*The Atlantic*	0	4	?
Id2	SacBee	1	?	3
Id3	CNN	0	2	?
Id3	FoxNews	1	1	?
...	...	...	...	...

Table 4.3: Sample results for one trust question.

Each person read four articles, hence four sample were collected per person. Let define as a unit the unique pair subject-publisher (a unit is then a row in 4.3). We denote as U the total number of units. One computes the treatment effect, i.e., difference in the average scores between the treated units and the controlled units:

$$ATE^* = \frac{\sum_{i \in I_1} Y_i(1)}{|I_1|} - \frac{\sum_{i \in I_0} Y_i(0)}{|I_0|}$$

It is here (5+4+3+3)/4 - (4+4+2+4+2)/4. Here we have an observed average treatment effect $ATE^* = -0.8$.

Under the null hypothesis, we can fill the counterfactual outcomes (question mark in Table 4.3) with the observed outcome. Let define $Y_i'(0)$ and $Y_i'(1)$ as follow:

$$Y_i'(0) = \begin{cases} Y_i(0) & \text{if } i \in I_0 \\ Y_i(1) & \text{otherwise} \end{cases}$$

$$Y_i'(1) = \begin{cases} Y_i(1) & \text{if } i \in I_1 \\ Y_i(0) & \text{otherwise} \end{cases}$$

Then, the re-assignment of people in treatment can be simulated, and the average treatment effect ATE can be computed using the counterfactual outcomes (that are now filled with the observed value in each row). For each simulation s, we randomly assign the units to treatment and control, hence we define I_0^s and I_1^s that are two distinct partitions of a random permutation Σ of [1, ..., U]. Then, we compute ATE_s as follow:

$$ATE_s = \frac{\sum_{i \in I_1^s} Y_i(1)}{|I_1^s|} - \frac{\sum_{i \in I_0^s} Y_i(0)}{|I_0^s|}$$

Doing that simulation N times, we have a vector of N average treatment effects ATE_1 to ATE_N that gives us the distribution of the average treatment effects under the null hypothesis. We can also compute the probability for an average treatment effect to be more extreme than our observed average treatment effect ATE^*. This is the p-value based on which we may reject the null hypothesis.

We can use this exact method when for the average treatment effect per publisher. However, for the average treatment effect across all publishers, one wants to make sure to account for the fact that each person is exposed to four different publishers. Hence, the re-assignment of people in treatment is blocked so that each user is exposed to all four articles, two of which are with no ad, one with the ads mix, and one with the political ads. This is a Blocked Fisherian randomization.

4.2.3 Mixed effects model

Another approach to study the causal impact of the ads on the publishers' credibility is to use mixed effects model — a mix between fixed effect model and random effect model. The idea is to **linearly regress** the outcome vector Y^q (where q corresponds to one of the five trust questions) on the explanatory variables, adding random effects for effects that might be variables-specific. One subject might always rate articles higher than another one. Hence, a person-specific random effect is included. All

the same, publisher-specific random effects account for the fact that some publishers may consistently appear more credible than others. We have 5 Y^q, one for each trust question. Hence, if $Y_{i,j}$ is the observed outcome for person i reading article j, the random effect model gives (4.1):

$$Y_{i,j}^q = \mu + \alpha_p * P_i + \alpha_c * C_i + \beta_i * X_i + U_i + V_j \tag{4.1}$$

where μ is average effect on all units, P is a dummy vector to identify the publisher (so P_i is j), C is a dummy vector to identify the ad menu, X_i is a matrix of covariates (with the familiarity with the publishers prior to the study, and the political leaning of the subject takers), β_i is a vector of weight for each covariate, U_i is the random effect for user i, and V_i is the random effect for publisher j.

Because P and C are dummy variables, the regression coefficient α_p and α_c, from the regression of Y^q on the explanatory variables, gives us the average treatment effect. We can also compute the 95% confidence interval taking 1.96∗std_err.

Chapter 5

The impact of the CRNs on the publishers' credibility

This chapter[1], answers the second research question: What is the impact of the CRNs on the readers' perception of the publishers? The ads have no significant effect overall. However, CRN ads impact the credibility of the less-known publishers. This chapter presents the data characteristics, the measure used to asses the news credibility, and the ad impact on the news credibility overall and per publisher. The publishers under study are CNN, Fox News, *The Atlantic* and *Sacramento Bee*.

5.1 Representative set of respondents

This section presents the characteristics of the data collected. Over the 6,000 responses collected, 4,767 agreed to participate and finished the survey. The following statistics are drawn from that set of respondents.

Gender: 53.2% of the participants were women, 46.5% were men, and 0.3% identified neither as man nor as woman.

Age: The respondents were born between 1931 and 2000, as shown is Figure 5-1.

[1]This chapter is the result of a joint work with Amir Tohidi

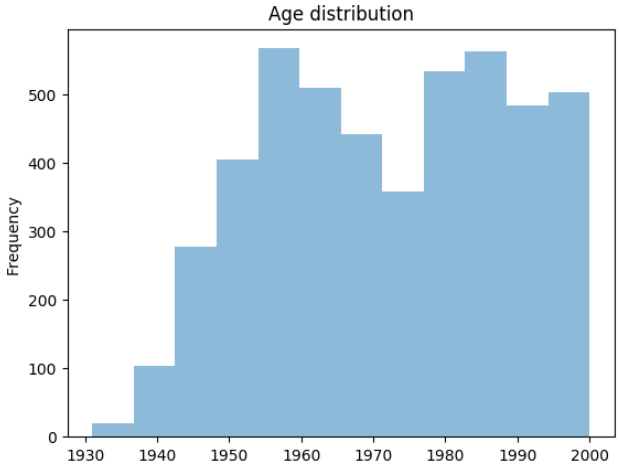


Figure 5-1: Respondents self-report their date of birth.

Ethnicity: The respondents represent different ethnicity, as shown is Figure 5-2.

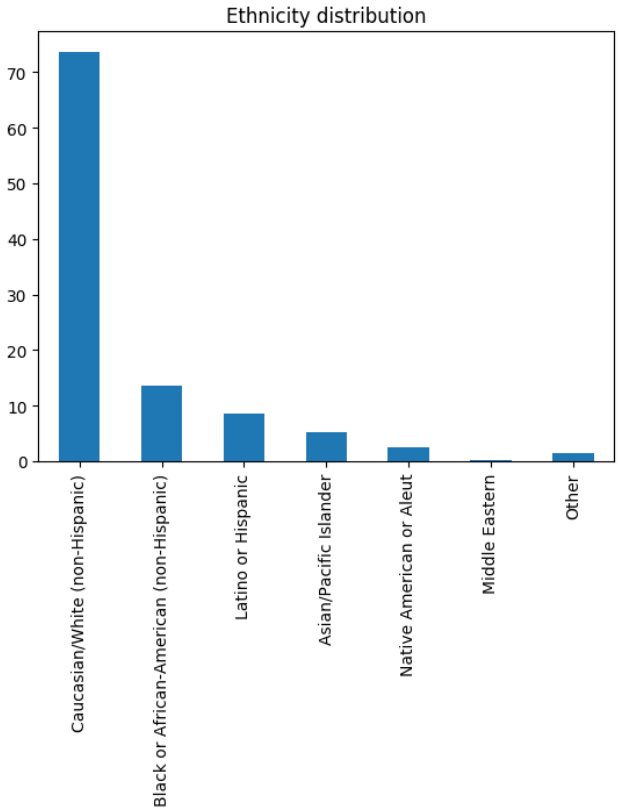


Figure 5-2: Respondents self-identify from different ethnicity.

States: All 50 states are represented in our dataset; as shown in Appendix B.2.

Political Affiliation: The political affiliation of the respondents follow the distribution shown in Figure 5-3.

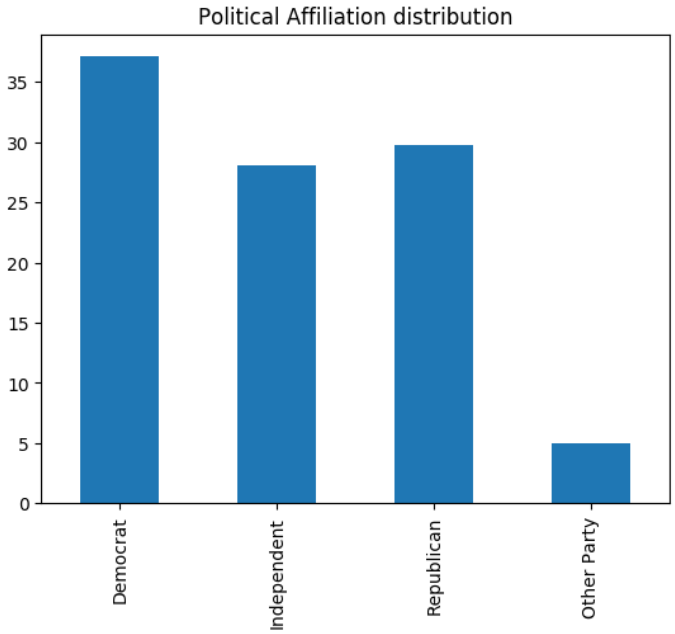


Figure 5-3: Respondents self-identify from different political parties.

Media knowledge and engagement prior to the study: Respondents were asked if they knew about the publishers before the survey, if they trusted the publishers, and the number of days per week they read with the publishers.

Fox News: Fox News was known by 88% of the respondents; 43% of the people trusted Fox News and 38% did not. Among the respondents, 13% said they interact everyday with Fox News and 45% never do. Fox News is more trusted among Republicans than among Democrats: 61% of the Republicans believed that Fox News conveys truthful content and 24% did not, while 34% of the Democrats believed it conveys truthful content and 50% did not.

CNN: CNN was also known by 88% of the respondents; 48% of the people trusted CNN and 34% did not. Among our respondents, 9% said they interact everyday with CNN and 47% never do. CNN is more trusted among Democrats than among Republicans: 67% of the Democrats believed that CNN conveys truthful content and 17% did not, while 31% of the Republicans believed it conveys truthful content and

54% did not.

***The Atlantic*:** *The Atlantic* was known by 40% of the respondents; 21% of the people trusted *The Atlantic* and 19% did not while most, 60%, were unsure. The results are different among those who were familiar with *The Atlantic* prior to the study as 41% of those who knew about *The Atlantic* believed it and 34% did not. Three percent of the respondents said that they interact everyday with *The Atlantic* and 78% never do.

***Sacramento Bee*:** *Sacramento Bee* was known by 22% of the respondents; 14% of the people trusted *Sacramento Bee* and 16% did not while most, 70%, were unsure. The results are different among those who knew *Sacramento Bee* prior to the study as 42% of those who knew about *The Atlantic* trusted it and 21% did not. Two percent of the respondents said that they interact everyday with *Sacramento Bee* and 83% never do.

As expected, *The Atlantic* and *Sacramento Bee* are far less known than Fox News and CNN.

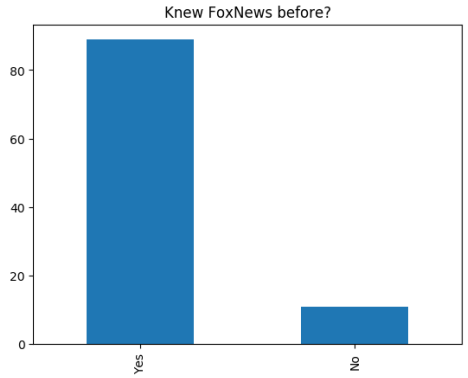


Figure 5-4: Percentage of people who knew Fox News prior to the study.

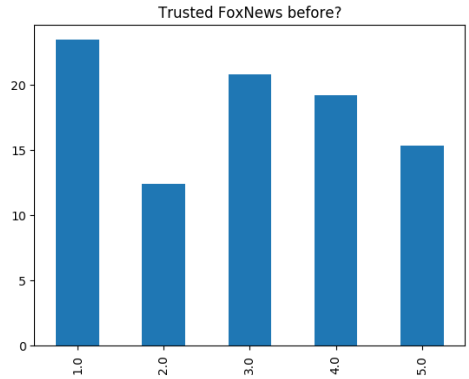


Figure 5-5: Percentage of people who trusted Fox News prior to the study on a 5-point scale.

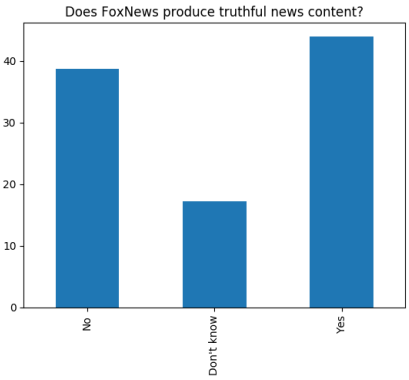


Figure 5-6: Percentage of people who believe Fox News conveys truthful information.

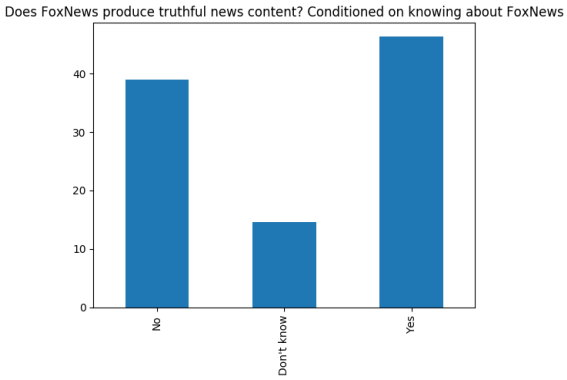


Figure 5-7: Percentage of people who believe Fox News conveys truthful information among those who knew about Fox News prior to the study.

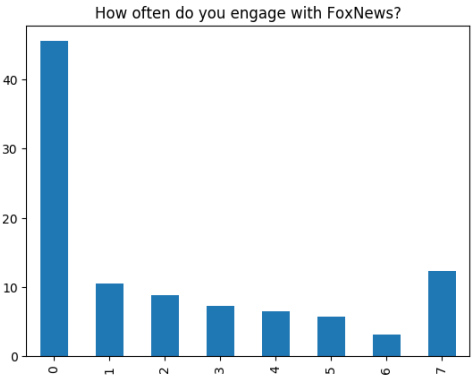


Figure 5-8: Percentage of people who engage x days per week with Fox News.

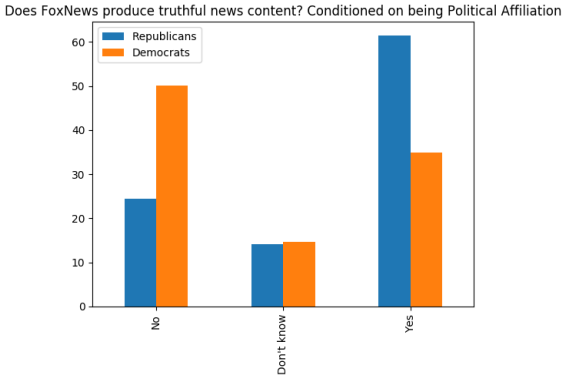


Figure 5-9: Trust in Fox News per political affiliation.

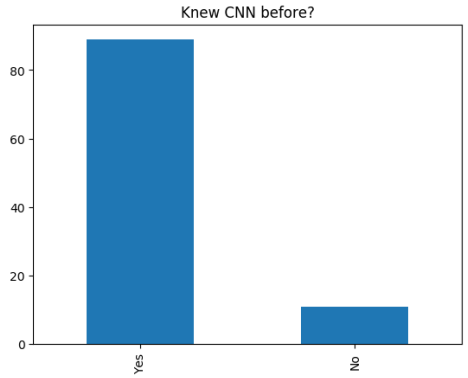


Figure 5-10: Percentage of people who knew CNN prior to the study.

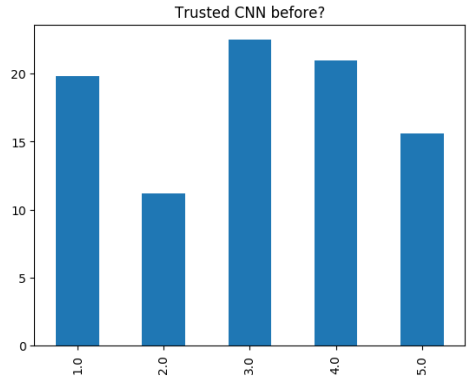


Figure 5-11: Percentage of people who trusted CNN prior to the study on a 5-point scale.

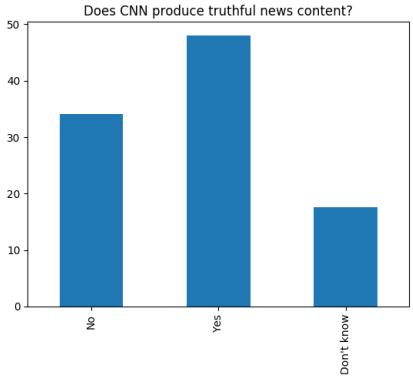


Figure 5-12: Percentage of people who believe CNN conveys truthful information.

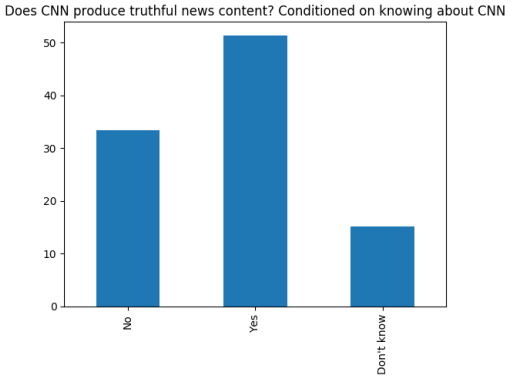


Figure 5-13: Percentage of people who believe CNN conveys truthful information among those who knew about CNN prior to the study.

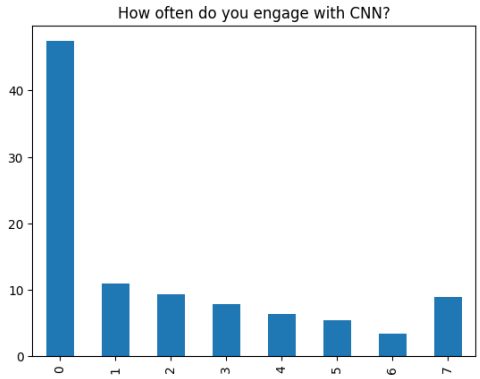


Figure 5-14: Percentage of people who engage x days per week with CNN.

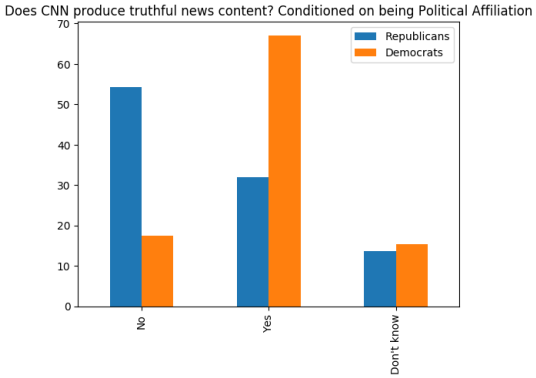


Figure 5-15: Trust in CNN per political affiliation.

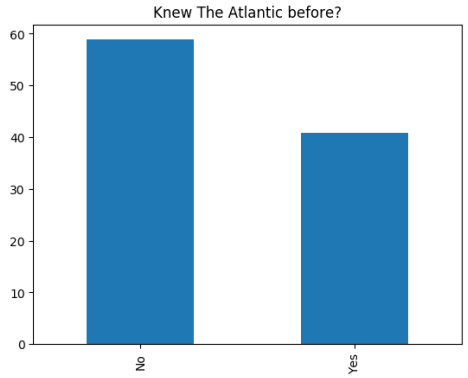


Figure 5-16: Percentage of people who knew *The Atlantic* prior to the study.

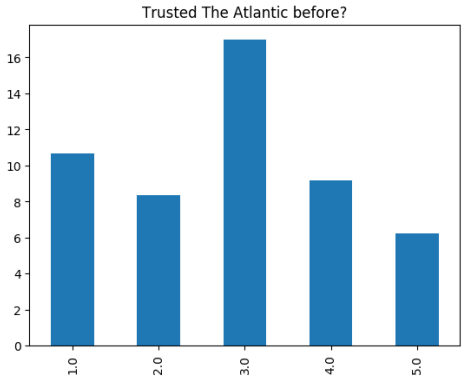


Figure 5-17: Percentage of people who trusted *The Atlantic* prior to the study on a 5-point scale.

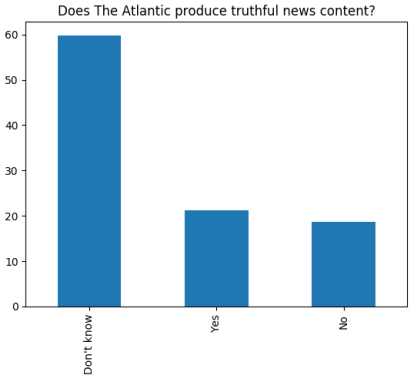


Figure 5-18: Percentage of people who believe *The Atlantic* conveys truthful information.

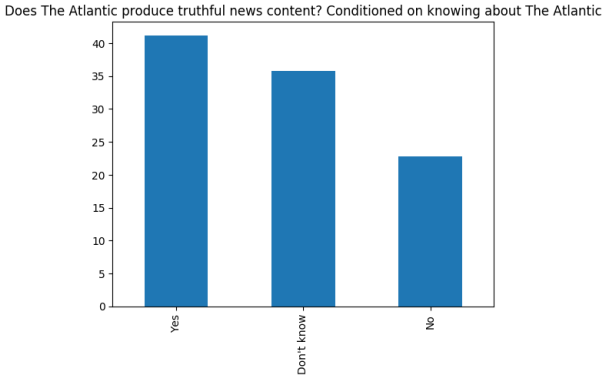


Figure 5-19: Percentage of people who believe *The Atlantic* conveys truthful information among those who knew about it prior to the study.

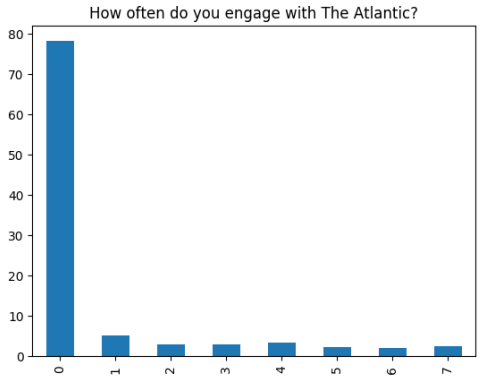


Figure 5-20: Percentage of people who engage x days per week with *The Atlantic*.

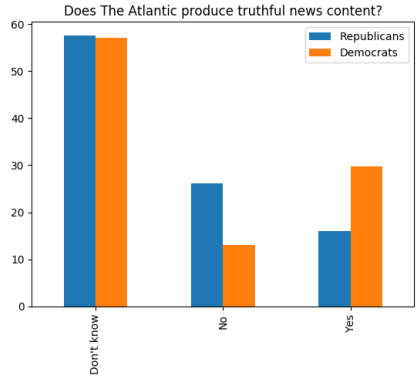


Figure 5-21: Trust in *The Atlantic* per political affiliation.

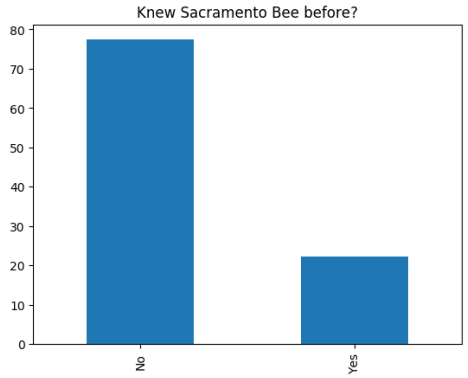


Figure 5-22: Percentage of people who knew *Sacramento Bee* prior to the study.

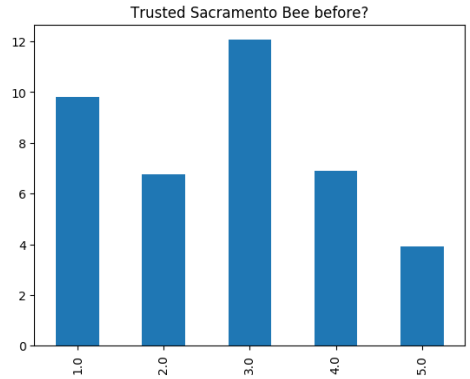


Figure 5-23: Percentage of people who trusted *Sacramento Bee* prior to the study on a 5-point scale.

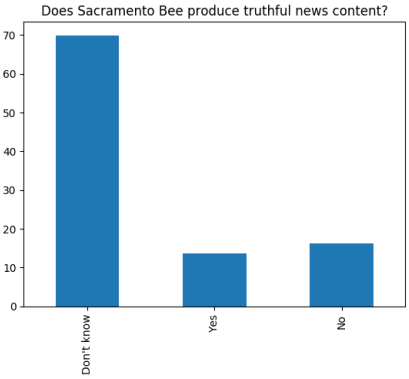


Figure 5-24: Percentage of people who believe *Sacramento Bee* conveys truthful information.

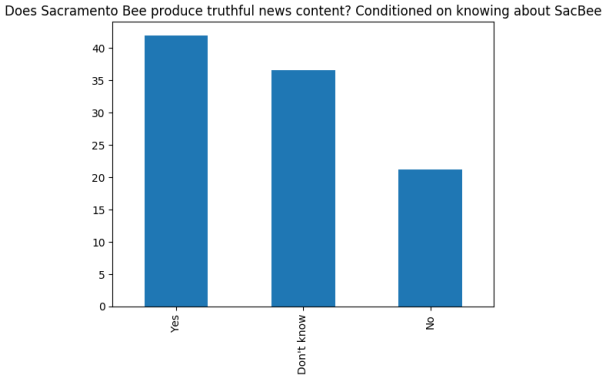


Figure 5-25: Percentage of people who believe *Sacramento Bee* conveys truthful information among those who knew about it prior to the study.

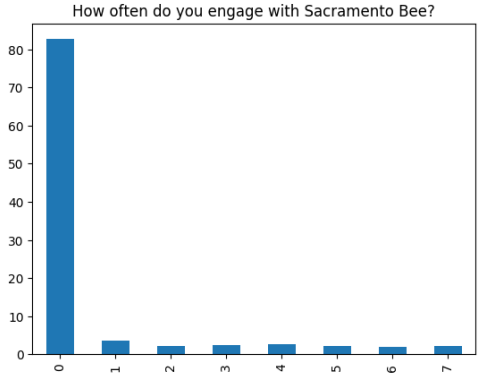


Figure 5-26: Percentage of people who engage x days per week with *Sacramento Bee*.

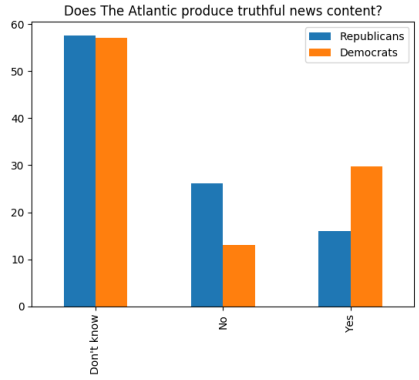


Figure 5-27: Trust in *Sacramento Bee* per political affiliation.

5.2 Correlation between questions

Second, we investigate the correlation between the five trust questions, in order to understand which one can be combined. As mentioned in Chapter 4, the five trust questions are:

- q_1 The article was trustworthy.
- q_2 The article was false.
- q_3 The article was credible.
- q_4 The article was biased.
- q_5 The article provides news information.

These five questions are answered for each article read, hence defined five different measures. Let define X, a matrix of size $U * 5$, one finds the U units studied and their 5 scores for each trust question:

$$X = \begin{bmatrix} q_1 & q_2 & q_3 & q_4 & q_5 \\ 3 & 4 & 5 & 4 & 1 \\ 2 & 5 & 3 & 1 & 4 \\ ... & ... & ... & ... & ... \\ 1 & 1 & 5 & 3 & 4 \end{bmatrix}$$

Then, the covariance matrix C is computed as follows:

$$C = \frac{1}{U-1} X^T X \tag{5.1}$$

The two principal components are computed by choosing the eigenvectors that correspond to the two largest eigenvalues of the covariance matrix C. Then, the vectors corresponding to each question can be expressed as a linear combination of the principal components. Figure 5-28 visualizes the Principal Component Analysis on all five questions.

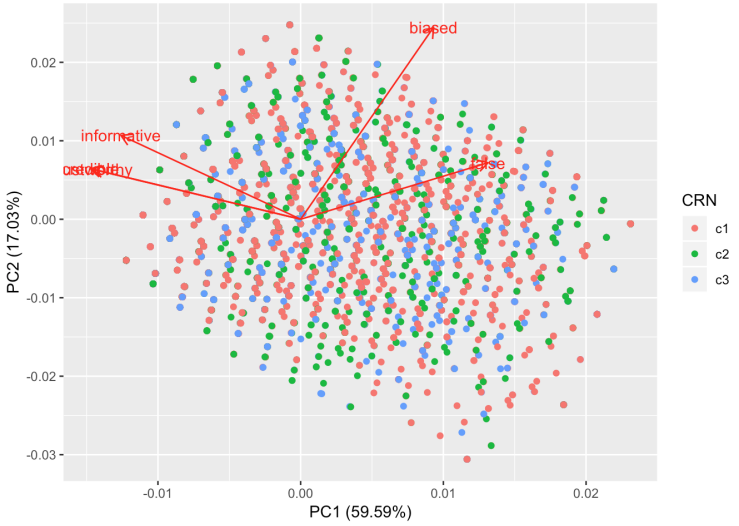


Figure 5-28: Principal Component Analysis of the Trust questions.

q_1 *trustworthy* and q_3 *credible* are colinear. q_5 *informative* is positively correlated with q_1 and q_3. In contrary, q_2 *false* is negatively correlated with q_1 and q_3 q_4 *biased* evolves in another direction and describes a part of the space significantly distinct from the others.

5.3 Ads' impact on the news credibility

Let define three ads regime: c_0 no ads, c_m mix ads, and c_p political ads. Two studies are conducted in parallel:

- The first study measures the impact of ads' presence on the news credibility. Hence, this study treats c_m and c_p as a unique condition. In this setting, the control is c_0 and the treatment is $c_m + c_p$.

- The second study measures the impact of the ads' style on the news credibility. Hence, this study focusess on c_m and c_p only. In this setting, the control is c_m and the treatment is c_p.

5.3.1 Claim 13: there is no overall impact

Claim: There is no significant effect on the aggregated data, neither in the study of ads vs. no ads (c_0 vs. $c_m + c_p$), nor in the study of mixed ads vs. political ads (c_m vs. c_p).

Evidence: Table 5.1 and Figure 5-29 show that no significant effect is witnessed.

The table bellow presents the effects and the p-values associated with the blocked Fisherian randomization.

Publisher		Trustworthy q_1	False q_2	Credible q_3	Biased q_4	Informative q_5
All	Effect	0.004	-0.004	0.003	0.008	-0.001
	p-value	0.72	0.70	0.78	0.46	0.84

Table 5.1: Effects and the p-values from blocked Fisherian randomization.

Figure 5-29 show that no effect appear through the mixed effects model in both settings: on top, c_0 vs. c_m & c_p; in the bottom c_m vs. c_p.

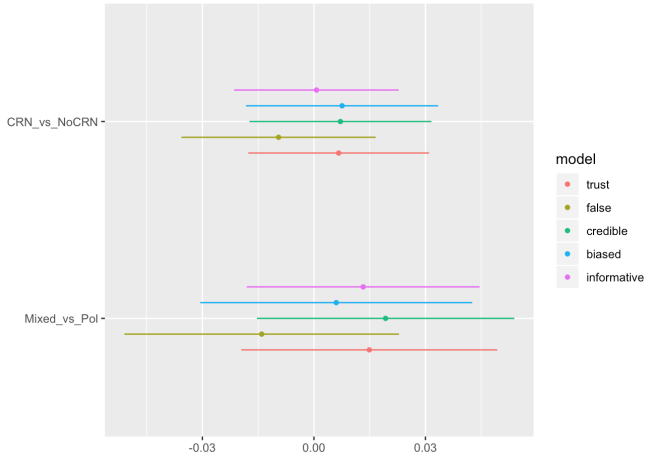


Figure 5-29: Random Effects on all five Trust questions.

5.3.2 Claim 14: a significant impact for less-known publishers is detected

Claim: Effects emerge for less-known publishers *Sacramento Bee* and, to a lesser extent, *The Atlantic*. Ads increase trust in *Sacramento Bee* and decrease it in *The Atlantic*.

Evidence: Table 5.2 and Figure 5-30 show that the two models agree on the conclusion: the effect of ads is positive for all trust questions for *Sacramento Bee*, and negative for *The Atlantic*'s trustworthiness. The result is surprising for *Sacramento Bee*. Perhaps, these CRN ads became the new standard as all traditional media have them. Then, an unknown source may earn credits from the CRN environment that looks alike a news environment in well-known source. This effect is particular to the one-time exposure to these pages. Long-term effects of the CRNs is beyond the scope of this thesis, but definitely of interest in future work.

Publisher		Trustworthy q_1	False q_2	Credible q_3	Biased q_4	Informative q_5
CNN	Effect	0,018	0,015	0,003	0,005	-0,021
	p-value	0,507	0,611	0,909	0,89	0,43
Fox News	Effect	0,005	-0,012	0,001	0,0421	0,0245
	p-value	0,856	0,708	0,963	0,193	0,363
The Atlantic	Effect	-0,069	0,050	-0,040	0,036	-0,043
	p-value	0,027	0,124	0,21	0,253	0,145
Sacramento Bee	Effect	0,075	-0,089	0,062	-0,054	0,041
	p-value	0,009	0,003	0,024	0,1	0,134

Table 5.2: Effects and the p-values from Fisherian randomization per publisher.

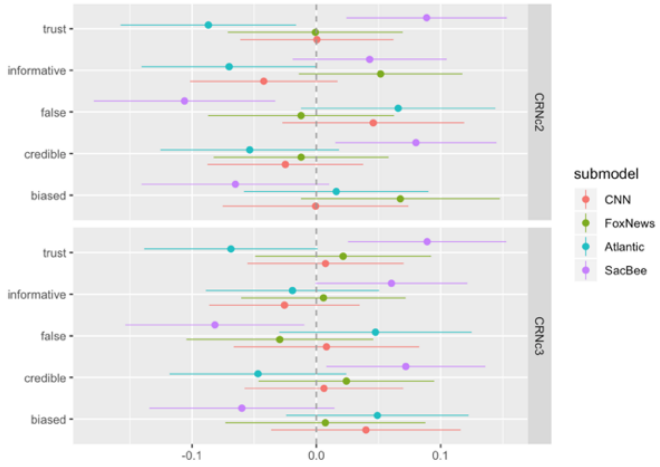


Figure 5-30: Effects per publisher from linear regression

5.3.3 Claim 15: the effects are stronger among the attentive persons

Claim: Stronger and more significant effects are witnessed on less-known publishers among the attentive persons.

Evidence: Table 5.3 and Figure 5-31 shows the causal effects of the ads on the news, for the subjects who answered correctly to the first attention screener. The trends are similar as the one observed among all the subjects, but the effects are stronger and more significant.

Publisher		Trustworthy q_1	False q_2	Credible q_3	Biased q_4	Informative q_5
CNN	Effect	-0,003	0,010	-0,013	0,012	-0,009
	p-value	0,891	0,786	0,704	0,782	0,783
Fox News	Effect	0,001	0,003	-0,026	0,085	0,021
	p-value	0,979	0,951	0,481	0,028	0,545
The Atlantic	Effect	-0,088	0,077	-0,041	0,021	-0,067
	p-value	0,015	0,053	0,251	0,566	0,057
Sacramento Bee	Effect	0,094	-0,142	0,084	-0,097	0,061
	p-value	0,008	0,001	0,011	0,017	0,045

Table 5.3: Effects and the p-values from Fisherian randomization per publisher for attentive subjects.

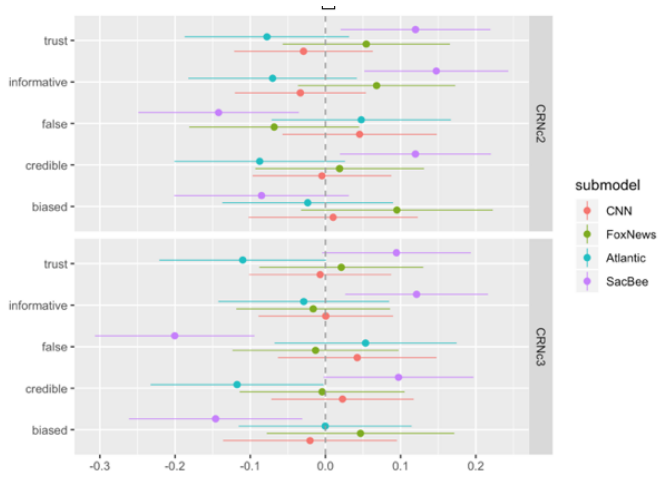


Figure 5-31: Effects per publisher from Linear Regression on people who passed the first distraction check

5.3.4 Claim 16: familiarity with publishers drives the results

Claim: The effects are driven by the audience's lack of familiarity with *Sacramento Bee*, and by the audience's familiarity with *The Atlantic*.

Evidence: Figure 5-32 shows the effect of the ads per publisher on data stratified

based on the audience's familiarity with the sources. Significant effects appear on Sacramento Bee, that come from the audience's unfamiliarity with the publisher. This is consistent with the argument according to which unknown outlets earn credits from the CRNs as they are the standardized default news environment online. In contrast, the ads further harm *The Atlantic*'s credibility among its audience. This fact is also consistent with our hypothesis that the audience loses trust after seeing native ads embedded by CRNs.

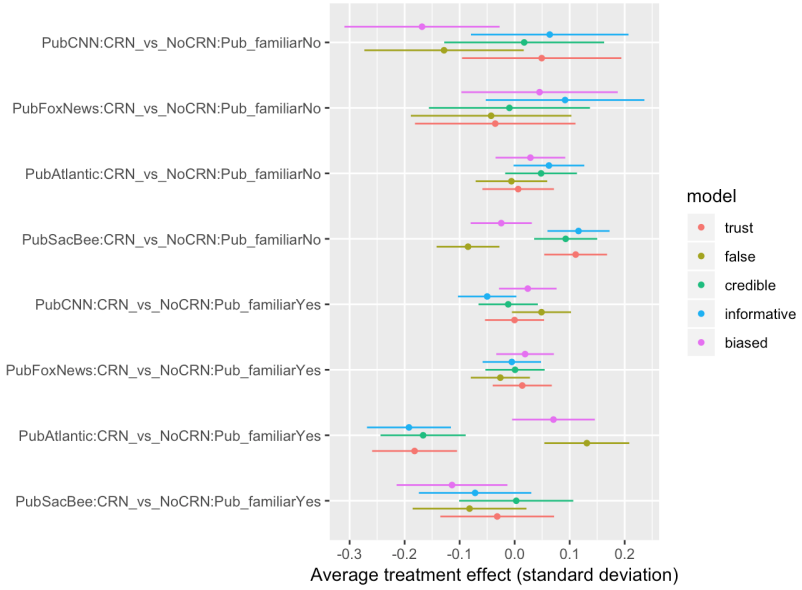


Figure 5-32: Effects per publisher from Linear Regression conditioned on subjects' familiarity with the source.

5.4 Key takeaways

No significant effects appear on the aggregated data, but ads impact less well-known publishers' credibility. Interestingly, the ads increase trust in *Sacramento Bee* (a local news publisher). The effect is even larger on subjects who were not familiar with *Sacramento Bee* prior to the study. This result suggests that an unknown source earns credits from CRNs — that are now the standardized default setting for online

news publishers. In contrast, the ads lower the trust in *The Atlantic*, particularly among the subjects who were familiar with the publisher prior to the study. This findings suggest that for high-quality news publishers, CRN ads increase scepticism toward the news content. *The Atlantic* actually removed the CRN ads during the study period. Further study will include testing the effect witnessed on less well-known publishers with a wider variety of articles and of publishers, as well as testing the long-term exposure to these ads through a partnership with a publisher and A/B testing.

Chapter 6

Conclusion

The Web audience is volatile and unwilling to pay for news. Hence, news publishers face an unprecedented challenge in attracting audiences and generating revenue. Even traditional sources create sensationalist headlines in an attempt to cope with the financial crisis. News publishers rely on marketing companies that bring revenue through clicks on ads. However, these marketing companies themselves proceed with clickbait in headlines to fill curiosity gaps. In particular, this study found that CRNs uniformly spread clickbait ads. Further, cilckbait increased in ads between 2016 and 2018, and it increased even more in the news' headlines than in the third-party advert headlines. In addition, CRNs were found to convey political ads within their widgets, and in particular during electoral periods. While no significant impact appeared for a one-time exposure to CRNs on publishers' credibility, less well-known publishers' credibility is impacted by the CRNs. Further research would involve measuring the repeated exposure to the CRN's ads and its impact on the audience's loyalty.

While CRNs are the default setting on news publishers, they are of low quality and they divert people from the news websites. While publishers' partnership with CRNs generate revenues in the short-term, it might harm the publishers' credibility in the long-term. A per-article cost instead of monthly or yearly subscriptions could allow to generate short-term revenues in place of the CRN ads. In the longer run, news could be aggregated on a unique platform, such as YouTube for videos, Netflix for movies and Spotify for music. Such an initiative could create an environment that

is user-friendly and could incentivize readers to follow and rely on identified sources. While this would have the downside of allowing users to read a brand as well as its competitors on any given topic, it would face a reality long overdue: alternative free sources — even if not journalistic — already exist, and are taking over the market.

Appendix A

Glossary

- **CRN ads:** CRN ads are the ads one finds in the CRN widgets. They are of two types: third-party ads and house ads.
- **CRN widget:** A CRN widget is the box that appear on a publisher's website with an aggregation of ads. A CRN — that is neither a publisher, nor an advertiser — controls the content of the widget.
- **House ads:** A house ad is a self-promotional ad for a publisher that is ran on the publisher's website.
- **Third-party ads:** A third-party ad is an ad ran on a publisher's website for another company that is not the publisher.

Appendix B

Tables

Reduced Publishers set
thehill, cnn, foxnews, huffingtonpost, theatlantic, bbc, cnbc, nypost, usatoday, euronews, cbsnews, breitbart, washingtonpost, wsj, time, heraldtribune, theguardian, forbes, latimes, chicagotribune, nydailynews, abcnews, nbcnews, dailymail, tmz, telegraph, independent, businessinsider, theonion, superherohype, democratandchronicle, brownsvilleherald, hollywoodlife, chicago.suntimes, seattletimes, livemint, greenbaypressgazette, theinquirer, power99, montgomeryadvertiser

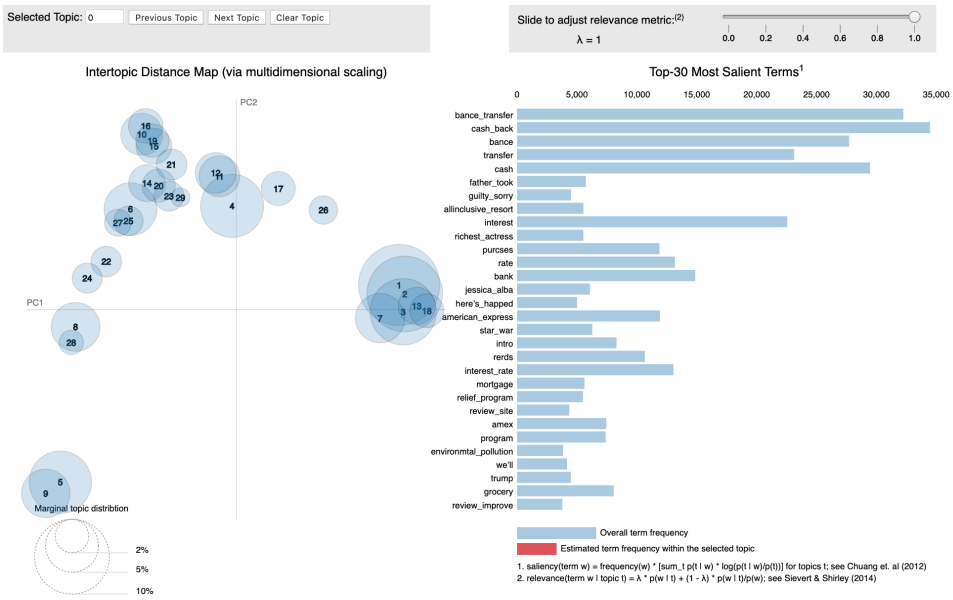
Table B.1: Reduced Publishers Set.

State	% (data)	% (U.S.)	State	% (data)	% (U.S.)
Wyoming	0.1	0.17	Oregon	1.7	1.27
Vermont	0.1	0.19	South Carolina	1.8	1.54
Delaware	0.2	0.29	Wisconsin	1.8	1.76
North Dakota	0.2	0.23	Alabama	1.8	1.48
District of Columbia	0.2	0.21	Maryland	1.8	1.83
New Hampshire	0.3	0.41	Kentucky	1.9	1.35
Rhode Island	0.3	0.32	Massachusetts	2.1	2.09
South Dakota	0.4	0.27	Washington	2.1	2.28
Nebraska	0.4	0.58	Indiana	2.2	2.02
Maine	0.4	0.40	Missouri	2.2	1.85
Montana	0.4	0.32	Arizona	2.3	2.17
New Mexico	0.5	0.63	Tennessee	2.4	2.05
Idaho	0.5	0.53	Virginia	2.5	2.58
Kansas	0.6	0.88	Michigan	2.9	3.02
Mississippi	0.7	0.90	New Jersey	3.1	2.69
Iowa	0.8	0.95	Illinois	3.1	3.85
West Virginia	0.8	0.55	Georgia	3.1	3.18
Utah	1.0	0.96	North Carolina	3.4	3.14
Arkansas	1.0	0.91	Ohio	4.2	3.53
Connecticut	1.1	1.08	Pennsylvania	5.3	3.87
Louisiana	1.1	1.41	Texas	6.6	8.68
Oklahoma	1.3	1.19	Florida	7.6	6.44
Minnesota	1.3	1.70	New York	7.7	5.91
Nevada	1.4	0.92	California	9.4	11.96
Colorado	1.4	1.72	Other	0.4	NA

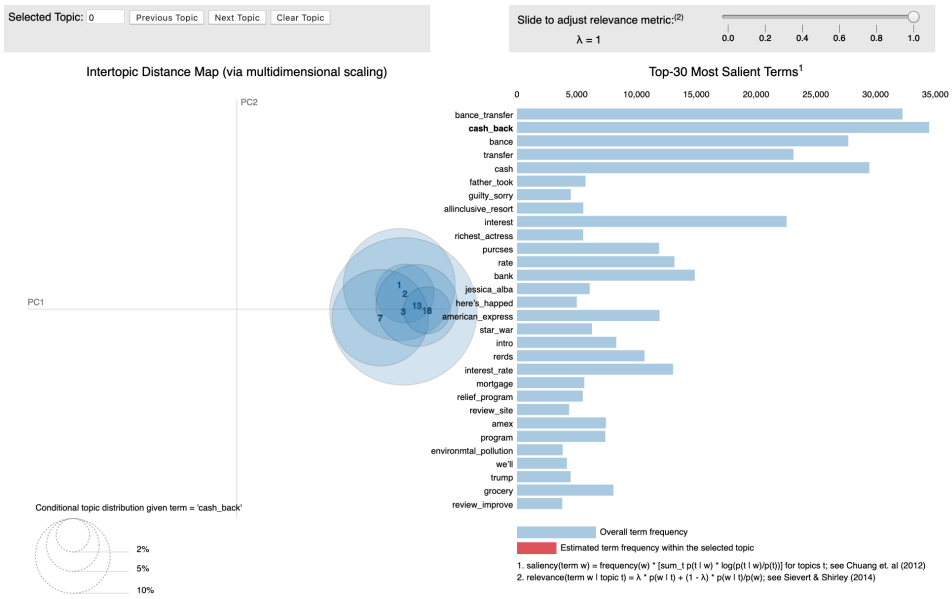
Table B.2: Percentage of respondents per U.S. State.

Appendix C

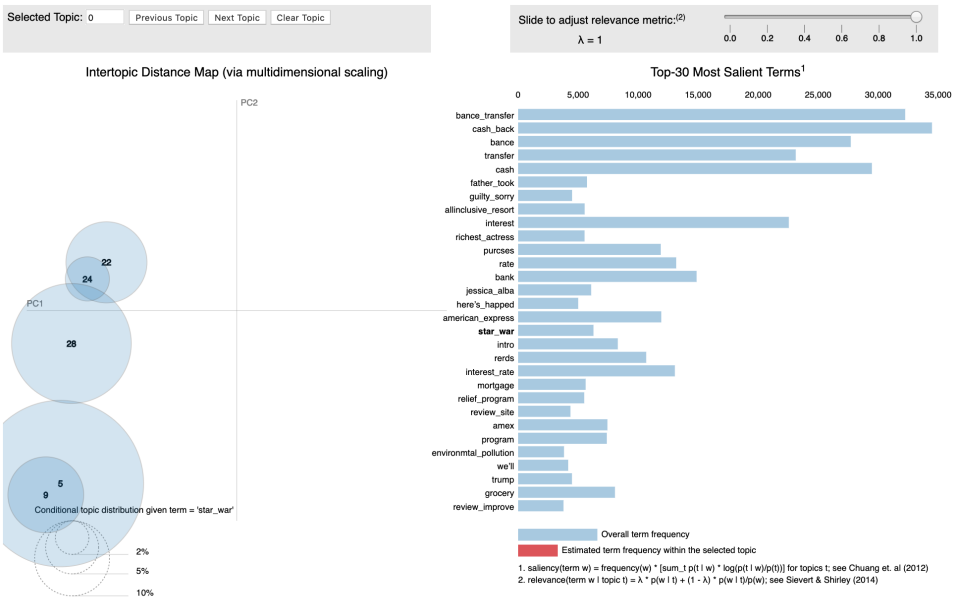
Figures



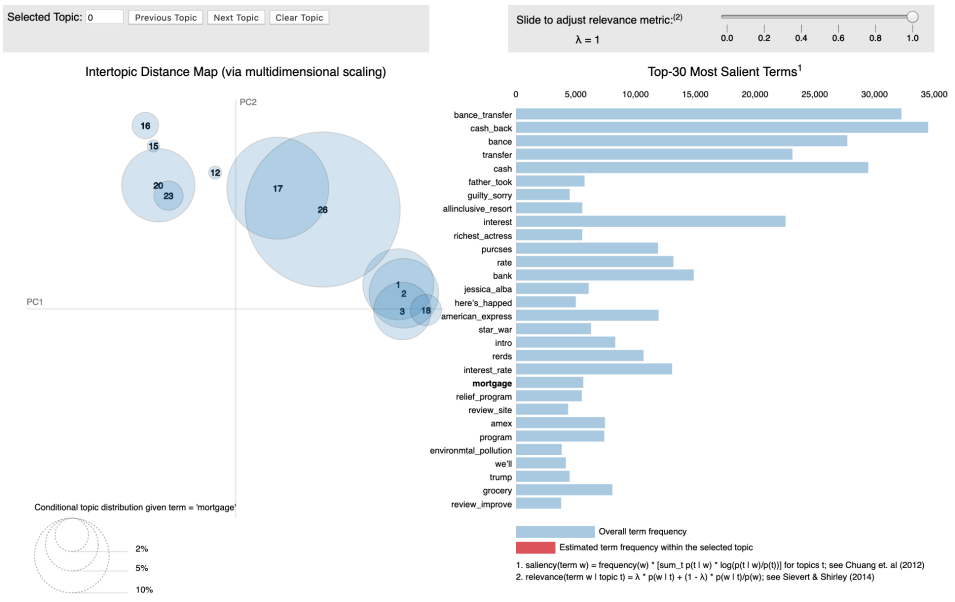
(a) LDA representation...



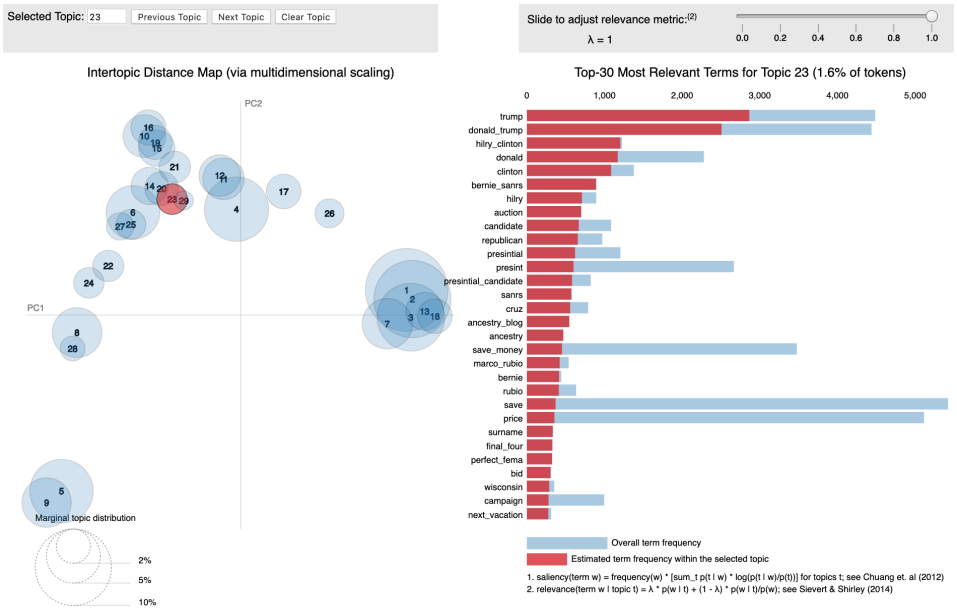
(b) ...conditioned on the words "cash back..."



(c) ...conditioned on the words "star wars..."



(d) ...conditioned on the words "mortgage."



(e) ... for topic #23

Figure C-1: LDA results on a random sample of 10,000 articles.

Bibliography

[1] Amol Agrawal. Clickbait detection using deep learning. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pages 268–272. IEEE, 2016.

[2] Albert Bandura. Social cognitive theory of mass communication. In *Media effects*, pages 110–140. Routledge, 2009.

[3] Muhammad Ahmad Bashir, Sajjad Arshad, and Christo Wilson. "recommended for you": A first look at content recommendation networks. In *Proceedings of the 2016 Internet Measurement Conference*, IMC '16, pages 17–24, New York, NY, USA, 2016. ACM.

[4] Matt Carlson. When news sites go native: Redefining the advertising–editorial divide in response to native advertising. *Journalism*, 16(7):849–865, 2015.

[5] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16. IEEE, 2016.

[6] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19. ACM, 2015.

[7] Hsiang Iris Chyi. Willingness to pay for online news: An empirical study on the viability of the subscription model. *Journal of Media Economics*, 18(2):131–142, 2005.

[8] Henriette Cramer. Effects of ad quality & content-relevance on perceived content quality. In *proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 2231–2234. ACM, 2015.

[9] Julio Cesar Soares dos Rieis, Fabrício Benevenuto de Souza, Pedro Olmo S Vaz de Melo, Raquel Oliveira Prates, Haewoon Kwak, and Jisun An. Breaking the news: First impressions matter on online news. In *Ninth International AAAI conference on web and social media*, 2015.

[10] Elizabeth Dubois and Gaffney Defin. The multiple facets of influence: Identifying political influentials and opinion leaders on twitter. *American Behavioral Scientist*, 2014.

[11] Jeffrey Dvorkin. Column: Why click-bait will be the death of journalism. *pbs.org/newshour/making-sense/what-you-dont-know-about-*, pages click–bait–journalism–could–kill–you/.

[12] Luciano Floridi. *Information: A Very Short Introduction*. Oxford University Press, 2010.

[13] Doris A Graber. The infotainment quotient in routine television news: A director's perspective. *Discourse & Society*, 5(4):483–508, 1994.

[14] Elizabeth Grieco. Newsroom employment dropped nearly a quarter in less than 10 years, with greatest decline at newspapers. *Pew Research Center*, 2018.

[15] Elizabeth Grieco, Nami Sumida, and Sophia Fedeli. About a third of large u.s. newspapers have suffered layoffs since 2017. *Pew Research Center*, 2018.

[16] John Herrman. Have the tech giants grown too powerful? that's an easy one. *The New York Times*, 2018.

[17] P. Hiebert. Social media overtakes television as news source for millennials. *YouGov*, 2016.

[18] C Richard Hofstetter and David M Dozier. Useful news, sensational news: Quality, sensationalism and local tv news. *Journalism Quarterly*, 63(4):815–853, 1986.

[19] Linlin Y. Huang, Kate Starbird, Mania Orand, Stephanie A. Stanek, , and Heather T. Pedersen. Connected through crisis emotional profit and the spread of misinformation online. *Computer-Supported Cooperative Work and Social Computing*, 15.

[20] Nathan Hurst. *To clickbait or not to clickbait? an examination of clickbait headline effects on source credibility*. PhD thesis, University of Missouri–Columbia, 2016.

[21] Magnus Hoem Iversen and Erik Knudsen. When politicians go native: The consequences of political native advertising for citizens trust in news. *Journalism*, page 1464884916688289, 2017.

[22] George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75, 1994.

[23] Sapna Maheshwari and John Herrman. Publishers are rethinking those around the web ads. *The New York Times*, 2016.

[24] Gloria Mark. Click bait is a distracting affront to our focus. nytimes.com/roomfordebate/2014/11/24/you-wont-believe-what- these-people-say-about-click-bait/click-bait-is-a-distracting-affront- to-our-focus.

[25] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David Levy, and Rasmus Kleis Nielsen. Digital news report. *Reuters Institute*, 2017.

[26] Gordon Pennycook and David G Rand. Who falls for fake news? the roles of analytic thinking, motivated reasoning, political ideology, and bullshit receptivity. *SSRN Electronic Journal*, 2017.

[27] Martin Potthast, Tim Gollub, Kristof Komlossy, Sebastian Schuster, Matti Wiegmann, Erika Patricia Garces Fernandez, Matthias Hagen, and Benno Stein. Crowdsourcing a large corpus of clickbait on twitter. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1498–1507, 2018.

[28] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. Clickbait detection. In *European Conference on Information Retrieval*, pages 810–817. Springer, 2016.

[29] Charles Prestwich Scott. A hundred years. 1921.

[30] Jung S Ryu. Public affairs and sensationalism in local tv news programs. *Journalism Quarterly*, 59(1):74–137, 1982.

[31] Kate Starbird, Emma Spiro, Isabelle Edwards, Kaitlyn Zhou, Jim Maddock, and Sindu Narasimhanm. Could this be true? it think so! expressed uncertainty in online rumoring. *ACM CHI*, 7:–.

[32] Bartosz W Wojdynski. The deceptiveness of sponsored news articles: How readers recognize and perceive native advertising. *American Behavioral Scientist*, 60(12):1475–1491, 2016.