Universidad de Concepción
Dirección de Postgrado
Facultad de Ingeniería - Programa de Doctorado en Ciencias de la Computación

INVESTIGACIÓN COMPUTACIONAL DEL "PROPAGANDA MODEL" (WHAT THE MEDIA DO IN THE SHADOWS: A COMPUTATIONAL INVESTIGATION OF THE PROPAGANDA MODEL)

Tesis para optar al grado de
DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

POR
Erick Elejalde Sierra
CONCEPCIÓN, CHILE
Marzo, 2018

Profesor guía: Leo Ferres & Johan Bollen
Departamento de Ingenería Informática y Ciencias de la Computación
Facultad de Ingeniería
Universidad de Concepción

To my family ...

ACKNOWLEDGMENTS

I want to thank professor Leo Ferres for all the hours dedicated to me and this project. Also, thanks to professor Johan Bollen for his contribution and supervision of this thesis.

I especially want to thank professor Eelco Herder, from the L3S Research Center in Hannover, Germany (now in Radboud Universiteit). Not only he received me during my internship, but he helped to guide a major part of this work. Most definitely he deserves to be on the list of supervisor.

I want to recognize the invaluable help of all the incredible people that I had the opportunity to collaborate with for my publications. They are: Prof. Leo Ferres from Universidad del Desarrollo, Prof. Johan Bollen from Indiana University, Prof. Eelco Herder from Radboud Universiteit, Prof. Barbara Poblete from Universidad de Chile, and Prof. Rossano Schifanella from Universita degli Studi di Torino. Their work is also reflected in this thesis.

I also want to thank my fellow PhD students from the program for always being willing to discuss new ideas and give me their opinion.

Thanks to all other professors in the Computers Science department which in a way or another offered me their support during all these years. With an especial distinction and all my gratitude to Professor Andrea Rodriguez.

I also would like to thank the members of my thesis committee: Diego Seco, Giancarlo Ruffo, Bernardo Riffo, and Julio Godoy. Thank you for the detailed revision of this work and all the helpful advice and comments.

Finally, I want to thank CONICYT for giving me their financial support to accomplish this goal.

Abstract

The Propaganda Model (PM) discussed in Manufacturing Consent is a theory in political economy that states that the mass media are channels through which governments and major power groups pass down certain ideologies and mold a general consent according to their own interests. According to the authors, every piece of news have gone through a set of filters that has ultimately yielded the source event as newsworthy. Current developments in communications, the digital availability of a huge amount of news on-line streaming from every corner of the world, together with our increasing capability to process all this information in a lot of different ways, give us the perfect environment to validate social theories using quantitative methods. In our work we take advantage of all these data to prove, empirically, the theory laid out in the PM. Previous work has had used machine learning and natural language processing techniques, but they have focused only on showing some leaning to a political party by a sample of the major news outlets. Here make a first attempt in the formalization of the model and the filters, and we help to provide and explanation on how the media works taking a computational approach. Results illustrate a measurable media bias, showing a marked favoritism of Chilean media for the ruling political parties. This favoritism becomes clearer as we empirically observe a shift in the position of the mass media when there is a change in government. Furthermore, results support the PM by showing that high levels of concentration characterizes the Chilean media landscape in terms of ownership and topical coverage. Our methods reveal which groups of outlets and ownership exert the greatest influence on news coverage and can be generalized to any nation's news system. Our studies on the geographic news coverage also give indications of the presence of the second filter (advertising). Experiments on predicting the communes with the biggest share of readership show this to be highly correlated with those regions with the greatest population, better socioeconomic status and a distinct political preference. As far as we know, this is the first time that there has been an attempt to empirically prove this (or any other) political economy theory using data science.

Table of Contents

List of Tables

Note: All Tables in this document were created by the author.

List of Figures

Note: All Figures in this document were created by the author.

Chapter 1

Introduction

The Propaganda Model discussed in Manufacturing Consent (henceforth MC, [72]) is a theory in political economy that states that mass media are channels through which governments and major power groups pass down certain ideologies and mold a general consent according to their own interests. The authors hypothesize that every piece of *published* news has gone through a "filtering process" that has ultimately yielded the source event as newsworthy. If they in fact exist, the filters mentioned above apply in many different ways. They are not usually presented as explicit censorship, but in subtle ways. For example, by favoring the hiring of reporters with a certain discourse.

Until a few years ago, the quantitative studies that we were able to do in political economy were limited to analyzing no more than a few dozens of sources (e.g., newspapers), and then again focusing on very specific issues, almost as anecdotal evidence. For example, in MC, the authors use a few recent events to point out how the press applies the word "genocide" to cases of victimization in non-allied states, but almost never to similar or worse cases committed by the home state or allied regimes. In the latter case, they could use the word "repression of insurgency", for example.

Thanks to new developments in communication and network technologies, nearly all the news generated by the mass media is available in digital form, together with endless discussion by other journalists and the general public in the form of on-line comments, blogs, tweets, etc. This makes data on the behavior of mass media more accessible than ever before. Moreover, new algorithms for the statistical analysis of complex systems also allow for the automatic processing of these large volumes of information to, for the first time, validate social theories using quantitative methods.

Relatedly, social networks have attracted a big part of the Eath's population, contributing to their exponential growth. For example, Twitter reported 328 million monthly active users in the first quarter of 2017 [160], and Facebook has almost 1.5 billion. On these social networks people share, in real time, almost every aspect of their social life, even their current state of mind or political inclinations. Since these interactions are digitally stored, they can be persisted and later recovered. This constitutes a great source of data on social behavior. The anonymity and the unawareness of being under observation, allow for a set of "ecologically-valid" experiments that were impossible at this scale just a few years ago.

In our work we take advantage of all these data to prove, empirically, the theory laid out in Herman and Chomsky's Manufacturing Consent. We carry out experiments using traces of human and corporate behavior preserved in digital form. These traces effectively constitute the largest repository of ecologically-valid (not elicited in a lab

or through a questionnaire) and complex human behavior data in the history of the species. Such analysis can only be tackled by making use of large computational systems. At the same time, this computational systems require the use of novelty techniques in the fields of natural language processing, set theory, graph theory, information retrieval, to name just a few. We are able, as never before, to analyze every news by searching for sentiments, intentions, topics, vocabulary and put it all together to see the bigger picture in the global context. This allow us to empirically show systemic tendencies where ever they exist [48].

It is important to notice, however, that we only have access to the corpus of published news (and the associated commentaries/blogs/tweets/Facebook posts, etc.). Thus, we only have access to the *final product* of the system of news production; that is, it has already gone through the filtering process. Given this, we are limited to testing if a piece of news (to repeat, *published* news) is accepted by the predicates of the filters and look for statically relevant tendencies in the media.

As far as we know, this is the first time that there has been an attempt to empirically prove a political economy theory using data science. This work provides some evidence on how the media works (with or without the existence of filters), it may help to create a less biased view of the world by suggesting alternative/complementary pieces of news, and it validates a methodology of computational approaches to social science (computational social science, from here onwards).

Chapter 2

Reviewing the Propaganda Model

2.1 The Propaganda Model

2.1.1 What is it?

The **Propaganda model** (henceforth, PM) is a conceptual model presented by Edward S. Herman and Noam Chomsky in their book *Manufacturing Consent: The Political Economy of the Mass Media* [72]. In it, the authors use a political economy approach to study the mass media: they study the rules dominating production and distribution of the news and how these link to political and economic power relationship (such as government and big corporate interests). According to Herman and Chomsky, everyday editorial decisions, like long term news outlets policies (e.g. hiring personnel that share their personal and political values with the company), will define the outlet's economic ideology and, as such, they should be explained and discussed. Like in many political economy analysis, the PM also include elements of game theory. The model tries to explain how political and economic forces control the mass media and exploit this dominance to manipulate the population and create a global view that is biased to satisfy their own interest through what the authors call *propaganda*. The classes in power try to build a global and homogeneous consent in topics such as economics, external affairs, politics and social policies.

2.1.2 What are the filters?

The PM attempts to explain how groups dominating the media *filter* what is finally published so they can marginalize dissent, allowing their own private interests to get their messages across to the public. The PM is based on a set of five of these filters: (1) Concentrated **ownership** - profit orientation of the dominant media firms; (2) **advertising** - the primary income source of the mass media; (3) **sourcing** - the reliance of the media on information provided by "experts" funded by these primary agents of power; (4) **"flak"** - a mean to penalize unaligned behavior of the media; and (5) **fear** - a vehicle to rally people against a common enemy. These elements interact with and reinforce one another giving rise to sophisticated behaviors. They are not usually presented as explicit censorship, but in subtle ways. For example, by creating an internal culture that incorporates preferences and definitions of newsworthiness that accommodate to institution's policy. Each linguistic account of an event, however, must pass through *all* these filters. The filters then define what is newsworthy and mold the discourse and

interpretation of events. According to Herman and Chomsky,

The first filter (ownership) The high cost of starting and/or maintaining a media outlet, at least one with a sizable outreach, makes it prohibitive for most people[1]. Thus, the ownership and control of the media is limited to a very small number of very big companies. Initially it was the government who tried to impose this price limitation, through the application of high taxes and the implementation of libel laws. Ultimately, it was the free market forces themselves what led to the high concentration of ownership that we can see today [128]. The integration of the media industry into the market have been accelerated by favorable policies from governments by eliminating or relaxing rules against concentration of ownership, cross-ownership and allowing non-media companies to gain control of media outlets. In turn, the open market leads to an increased pressure to attract investors and the temptation to have a profit-driven organization. The problem with an increasingly high percentage of the stocks trade in the open market is that media outlets are in close relation with the mainstream corporate community by sharing boards of directors and social links. The media companies also have links to bankers through loans, credits and financial advisers. So, banks and other investors, even if they don't have majority control, can make themselves heard. If the manager of the outlets fail to comply, investors may take actions that could affect the company (e.g. sell their stocks and depress its price, favoring potential takeovers). Investors will try to push the media towards objective strictly design to maximize profit. Finally, the media also depend heavily on the government for the acquisition of licences and franchises, policy support, interest rates, etc. This dependency makes them vulnerable to government interference or harassment. As a result, media protect themselves by lobbying and cultivating political relationships. The media companies also depend on the government to maintain good diplomatic relations with other countries where they or their associated companies may have businesses.

Herman and Chomsky give some statistics (Tables 1-1 and 1-2 in [72]) showing the huge benefits (a median of $183 million after taxes) earned by the 24 larger US media corporations and the percentage of their business (i.e. influence) that is still retained by the control group (in most cases the founding family). For example, General Electric and Westinghouse own NBC, hence its board of director is predominantly form by corporate executive rather than media executive. The data presented by the authors give an idea of how exposed these businesses are to external pressures, which should be a matter of concern for the public interest. However, they do not give any empirical evidence showing that the content produced by the media have been actually affected by such power structure (see chapter 4).

In summary, the first filter states that large companies dominate the mainstream media. These companies, in turn, are controlled by very wealthy people or managers that answer to owners and other profit-oriented forces. Ultimately, the media is ruled

[1] At least at the time when the PM was first proposed. See section 2.1.4 for an updated analysis of the PM in the internet era.

by the common interest of major corporations, banks, and governments.

The second filter (advertising) Advertising, being a fundamental source of income for news outlets, also plays an important role to maintain the leadership of the top companies in the free market. Outlets that can secure good advertising contracts may afford lower sell prices and become more competitive. This business model breaks the natural market rules that give the final buyer's choice the power to decide. In this case, the advertisers' settlements have a significant impact on the media growth or even their survival. Thus, outlets are forced to comply and demonstrate to advertisers how the content serves their private needs. This agreement with the advertisers has led the media to shift to lighter and less controversial content (e.g., lifestyle, fashion, sports, etc.) [67, 124, 125].

As a general rule, advertisers will prefer to take their businesses to outlets with target audiences of high purchasing power (some others may want to reach audience base on ideologies - e.g., political parties, church, etc). Working class and radical newspapers will be marginalized and eventually driven out of the market. For this reason, the media will strive to expand their audience but mainly focusing on acquiring "quality" readers[2]. Outlets discrimination can be also influenced by political reasons, with advertisers declining to do business with media that are perceived as ideological enemies or generally unfavorable to their interests. According to Chomsky and Herman, all these characteristics make the news media system comparable to a political scheme where votes are weighted by income.

The Third filter (sourcing) Mutual interests and economic reasons have created a deep connection between the media and the government as a reliable source of news. This link also applies to powerful business corporations. With limited resources, the media have to select where to allocate its assets carefully. Having a steady stream of news from the government and other well known and "reliable" sources spare them a significant amount of workforce. With the extra advantage that the content produced by these sources can be taken as newsworthy and factual, minimizing investigative overheads. They even make every effort to smooth the transit of their content into the news (e.g. provide advanced copies of speeches and reports, schedule press conferences at hours that satisfy news media deadlines). In exchange, these sources are granted exclusive access to the media to spread their message. The budget allocated to public relations and communications of any of these sources overshadow all dissenting voices combined (sometimes by orders of magnitude). So, governments can use their influence to coerce the media. They can manipulate stories by flooding the media with information that adjust to the desired viewpoint. The "official" sources also dominate the discussion by adopting "experts" (e.g., through paying them as consultants

[2]Understand by reader any person who, in general, gets information from the media, and *quality readers* those that are of some value for advertising purposes

or funding their research) or just by creating them (e.g., giving them more coverage or supporting their perspective).

The Fourth filter (flak) The authors give several examples of mechanisms through which governments and big corporations punish and try to correct any behavior of the media that deviates from the line. The most common instrument is the sponsorship of institutions whose supposed objective is to hold the media responsible for the biased content (e.g., think-tanks), but which in reality only respond to the political and economic interests of the power groups.

The counterattack or flak to the media for unwanted behavior may come in various forms (e.g. public letters, lawsuits, speeches, bills before Congress, withdraw of advertisement contracts). Flak may be an organized effort or separate actions. Flak has the potential to become inconvenient and expensive to media. Of course, the more powerful and influential the person or institution, the more damaging the flak it can produce. So, media will avoid content that might create flak.

The Fifth filter (fear) Originally postulated as "anticommunism", this filter has evolved since. After the fall of the Soviet Union and the socialist bloc, anti-terrorism became an adequate replacement. Broadly, the fifth filter characterizes the use of an ideologically deep-seated common enemy as a control mechanism. It represents the general idea of creating fear to activate a consistent response of the masses against some "evil". A similar idea have been presented by other authors [4]. Since this "enemy" is ill-defined, it can be directed to any subject perceived as a menace to the interests of the elite or their way of life. Any, otherwise unthinkable, action presented as necessary to defeat the communism/terrorism (or any of its substitutes) will be accepted as the lesser of two evils (e.g., NSA massive surveillance on US citizens). The mass media plays a crucial role in the propagation of this ideologies by relaxing the need for evidence when referring to this matters and making "rooting for our side" a genuine form of news.

2.1.3 How do they support the filters?

H&C claim that the PM predicts a "[...] *systematic and highly political dichotomization in news coverage based on serviceability to important domestic power interests*". The authors support the theory postulated in the filters by discussing a few, almost anecdotal, paired examples of the different treatment of the media to similar events that occurred in close time proximity. For example, the authors discuss a few recent developments (at the time of the book) to point out how the press applies the word "genocide". The media widely use this defamatory term in cases of victimization in non-allied states (e.g., Yugoslavia/Kosovo, Iraq(Saddam Hussein)/Kurds). But they almost never use it in similar or worse cases committed by the home state or allied regimes (e.g., Turkey/Kurds, Indonesia/East-Timor). In the latter case, they could use the word "repression of insurgency", for example. In general, throughout the book, the authors use these differential

treatments in coverage of what they call “worthy” vs. “unworthy” victims to show that the media bias, as predicted by the PM, most of the time aligns with the government political interests or can only be explained by economic concerns.

In the PM, the authors use the paired example methodological approach to show variations in the volume and quality of the coverage, but the analysis is limited to just a small sample of some of the most influential newspapers. Their analysis indicates that the output produced by the news media system is consistent with that predicted by the PM. We think it is worth noticing that the methodology used by the authors is aimed to show bias in the media. However, as mention by some of it critics, it is sometimes hard to tell from their methodology which filter is acting in each case. This may work to demonstrate the performance of the model, but it makes it more difficult when applying the PM in a different event/context or if we try to operationalize the model.

2.1.4 Relevance of the Propaganda Model after 30 years

In an interview with Mullen [110], Herman and Chomsky express their confidence in the applicability of the model to forms of media other than newspapers, especially the Internet, where traditional news outlets compete with new digital media and advertising is more relevant than ever before. If anything, the “old” news industry has evolved and has adapted to the new environment. Rampton [131] discuss how some characteristics of the Internet have the potential to challenge the system modeled by the PM, and go into details for each filter. For example, he recognize that now we can find a ‘blogosphere’ fed by individuals talking about events with a discourse not always favoring the groups in power, and showing content (e.g., photos, leaks, etc.) otherwise captured by the filters. This is some times called “citizen journalism”, and it has been used as a measure to detect bias in the traditional media [178, 91]. But even these alternative media (such as blogs, podcast, etc.), most of the time, limit themselves to discuss news that they have read or seen in a major news outlet. In other words, they practice critical analysis rather than news-making. Rampton arrives at the conclusion that the filters are still relevant due to the mainstream media hegemony as news consumers’ preferred source. Other authors have analyzed the model and share the same ideas on the relevance of the model more than 20 years after it was first introduced [112, 60].

Pedro [123] also makes a filter-by-filter verification of the continued relevance of the PM in the 21st century by analyzing a series of more recent events that fit the predictions of the PM (e.g., the purchase of the Tribune Company, coverage of wars in Iraq and Afghanistan, the growing influence of bloggers and social media users as “flak agents”, etc.). The author points out that the media market has shown a tendency towards a bigger concentration (from 50 dominant major corporations dominating the media in U.S. in 1884 to only 5 in 2004 [9]). The competition to gain the attention of advertisers has intensified as well (about 20 media companies hoard three-quarters of the global investment in advertising [99]) which in turn give advertisers more influence over the content that is preferred in the regular programming [37].

In [167], the author describes the evolution of the PM in the Internet era. He, as

Rampton [131], emphasizes the potential of the new medium, if used properly, to overthrow the filters. But the author also calls for our attention to possible risks, and he noticed the emergence of new filters, such as the over-exposition to unstructured (in a lot of cases contradictory) information. The frustration caused by this information overload creates **self-censorship** and leads to users only looking at sources that reinforce their own believes. This behavior ultimately creates what is known as echo chambers and filter bubbles [120], and makes groups to become more extreme.

More recently, Robinson [135] revisited the model to evaluate if the introduction of new communication technologies and the Internet have affected the influence of the economic structure over the news media system. He argues that even though there has been a shift from the printed news to the digital media (newspaper website audiences grew by 7.4% in 2012 [43]), the news cycle is still controlled by the big news corporations. The three most significant outlets in U.S. (i.e., Wall Street Journal, USA Today and New York Times) had a circulation of over 5 million users (which include digital subscribers); any of them counting at least one order of magnitude more readers than the next closest competitor [43]. According to a 2012 Pew "State of the Media" study, social media such as Facebook and Twitter act more as a path to the news than as a replacement source [135]. With regards to advertising on Internet, the author concludes that news outlets had to rely even harder on this source of income: online subscription revenue does not cover the previous earnings made from selling print newspapers. So, the content has became more profit driven with a shift to soft-news and corporate-friendly reports [43].

On the same note, Pew Research Center reported that nowadays up to 93% of U.S. adults consume some news online [153], and to do so they use official news organizations websites and social media in equal shares (36% vs. 35% respectively) [104]. However, regardless of the path they use to get to news, readers overwhelmingly favor professional news organizations to get their news (76%) [103] With the increasingly important roll of social media in the news system, news outlets have again adapted by embracing the new platform and now they are able to create original content for them (i.e., Facebooks instant articles[3] allows any publisher to create a piece of news directly in the Facebook platform). This move is also deeply motivated by the market and the task to increase advertising revenue. In 2016 digital advertising constituted aprox. 37% of all advertising revenue (not just news outlets), and Facebook along comprises for 35% of digital display advertising revenue [153].

More specifically, in [95], the authors analyze the dynamics of news and journalism on the Twitter platform. They found that 0.8% of the tweets are news media related. This gives an idea of how significant are news media to Twitter. They also report, confirming the results of Robinson [135], that the traditional notion of gatekeeping and news production have not changed and large news organizations still control what is newsworthy. Moreover, they show that news entities do not use the social media to engage with their audience but rather as a way for content dissemination (mostly by redirecting

[3] https://instantarticles.fb.com/

users to their own websites). Also, they find difference in interest (topic-wise) between the general Twitter users and the news outlets. This difference of focus gives some evidence on the agenda-setting behavior of the news industry, which reinforces the hypothesis of a profit-driven system, instead of an informative one.

The applicability of the PM in the Internet era and the transferability of the filters to other forms of media mean that we should be able to verify the existence of these mechanisms acting in the content published by the press on places like Twitter. This is one of the primary goals of the study that we present in these pages.

2.2 How other authors received the model

2.2.1 Criticisms to the Model

Chomsky explains that the PM makes predictions at various levels [31]. Chomsky argues that the PM makes what he terms "second-order predictions" about how the news media performance will be discussed and evaluated. Due to the nature of the Propaganda Model (itself drastically violating the filters), Chomsky expects it to be ignored or heavily criticized. Indeed, the text has been labelled as a conspiracy theory and anti-American [45, 65]. Some critics accuse the model of underestimating the news consumers (the media's audience) [90]. Herman and Chomsky have explained that the propaganda model is a decentralized model and it is not the result of one overseeing person or entity, but the independent actions of many individuals and organizations [71, 110]. They also have clarified that the PM only aims to explain the behavior of the media and not the impact that this behavior has on the audience, or whether or not the propaganda dissemination is effective [71]. However, the authors also identify what they term "first-order' predictions, analyzing the media's *behavior* as dictated by the filters. These are, specifically, the kinds of behaviors we are going to analyze in this thesis, leaving aside any opinion-base discussion and concentrating on the PM *qua* a formal system, measuring the predictive power of (most) of the filters themselves.

Mullen [113, 111] shows that the PM has been also systematically marginalized in the fields of media and communication studies. In a sample of articles (from both media theory and media practice) extracted from ten media and communication journals published between 1988 to 2007 in Europe and North America, only $2,6\%$ make reference to the PM. Even within this $2,6\%$, Mullen shows that there is little engagement with the PM (most of the articles only include a reference to *Manufacturing Consent* in the bibliographies). He found a similar dismissal of the PM in a sample of media and communication textbooks from the same period. Mullen and Klaehn [85] identify two phases in the reception of the PM within the academic community. According to them, since the book's publication in 1988 and into the 1990s, the PM was received with hostility and criticism. From the early 2000s, the criticism became more constructive and there was a greater engagement with the PM. During this second phase the authors highlight the works of Boyd-Barrett [17] and Sparks [151].

In [151], Sparks agrees with the general idea of the PM and the analytic framework it provides, but at the same time notices of some its flaws and try to address some of the criticism received by the PM by complementing it. According to the author, one of the most significant issues with the model is its oversimplification in the predicted tendency to media uniformity. Instead, he proposes a more heterogeneous media motivated by the divided nature of the capital class. The author identifies six ways in which the PM can be extended. In the first argument, Sparks concur that the owners, and the elite class they represent, steer the media according to their capitalist interests. But he notes that, besides this common general class concern, they all have their individual prize in sight and most of the time they compete and conflict with each other. Herman and Chomsky do have referred to situations of elite dissent and how these impact the media behavior [110]. The second way Sparks proposed to improve the PM would be by reflecting the unusual but not nonexistent opposing forces withing the political elite (such as socialist parties in most European countries). The third way regards the political economy of the mass media: the author agrees with the premise that media adjust their editorial strategy according to the intended public, but he criticizes that the PM describes the content of the media as mainly directed to an elite audience as a mean to increase theirs followers 'quality'. Sparks uses as an example how media dedicate rather long sections to sports, art, and other forms of entertainment; articles that according to him are not intended to convey any policy and are only purposed to engage their audience. Even though, Herman and Chomsky clearly stated (second filter [72]) that the PM predict a shift in the media to content that gets the audience into "buying mood" (materials that avoid serious complexities and disturbing controversies). This tendency is more attractive to sponsors and creates political apathy in the general public (contemporary equivalent of the Roman "games of the circus"). In his fourth observation, Sparks stresses that to be applicable in places other than the USA, the PM should also represent capitalist democracies (such as most European countries) with public/state service broadcasting and a more politically diverse media. This argument is also used by Corner in his criticism of the model [36]. For the fifth way in which the PM could be enhanced, the author points out that, although journalists heavily rely on sources, they usually may count with more than one authority. Different sources may contradict each other depending on the faction of the elite they are representing. But, in favor of the PM, bringing opposing views is sometimes used by the media as a marketing strategy. The last criticism is directed against the failure of the PM to recognized journalists (those in lower positions that make the majority of the press) as part of the working force, and thus caught in the same class struggle that ultimately leads to labor organization and clashes with the elite powers. Other critics of the model also point out this last observation [65]. Some responses to the criticisms made by Sparks are reviewed in [134]. Across this thesis we show empirical evidence that support the postulates of the PM against most of these criticism. We also complement the analysis with alternative models that back up the theoretical framework offered by the PM.

Boyd-Barrett also engage with the model and proposed a sixth filter: "buying out"

of journalist [17]. This extra filter implies the direct but hidden control of the media by the government and even intelligence agencies, this in addition to the more subtle mechanism proposed in the PM (e.g., source and flak). The author criticizes the focus of the PM in the structural factor of the news selection that leaves aside intentionality. He also points out the lack of precision in the definition of the filters in the PM (to which this thesis will hopefully contribute), in turn makes it difficult in some cases to distinguish when to apply one filter or the other. The methodology used by the author (similar to that of Herman and Chomsky) is based on case studies (i.e. the US invasion of Iraq in 2003) and the analysis of the coverage (and manipulation of information) that these events received from some of the major newspapers in the USA. Boyd-Barrett discusses some formalization of the filters and provide some evidence to support the presence of at least the last three filters (namely: relaying on official sources, fear of flak and ideological convergence).

Another critical review of the PM is presented by Corner [36]. Its main criticism is directed at the way in which the model categorically defines the media system and conclusively asserts the effectiveness of the filters. Chomsky [31] has addressed this critique by explaining that the PM does give space to alternative voices but, even in most of these cases, they benefit from the illusion of diversity that they convey. In agreement with Boyd-Barrett [17], Corner also questions the intentionality predicted by the model. The vocabulary used by Herman and Chomsky (*e.g.*, filters, manufacturing, propaganda, etc.) might imply a deliberate manipulation of the content. The author suggests that the essence of the message is not modified but generated in that way within the system. Corner also respond to Klaehn's claims that the PM have been widely ignore in the academic community [81] by noticing a similar avoidance by the PM advocates of the previous literature debating the topic of the media and its relation to the state and the free market.

For a thorough review of criticism received by the PM and some of the responses given by Herman, Chomsky and some other followers of the model we refer the reader to [112].

2.2.2 Endorsements of the Model

Despite its limited adoption, some advocates of the PM have tried to advance its analytical and ideological viewpoints. For example, Mullen, a strong supporter of the model, co-edited an issue of the *Westminster Papers in Communication and Culture* journal that was entirely dedicated to the PM [109].

In the 30 years since it publication, researchers have used the PM to describe the media in countries other than United States, such as Canada [80], UK [37, 60], Netherlands [13] and Spain [122]. These studies invalidate the criticism claiming that the PM is only applicable to the U.S media system [36]. Moreover, the model have been successfully used to analyze the media in the transitioning system of the People's Republic of China [68, 69]. This growing body of work shows that the PM is a valid theoretical framework that can be used to explain the news production of mainstream

media across different socio-political and economic systems.

Apart of Herman and Chomsky, Klaehn has been one of the most active supporters of the model. Besides responding to criticism against the model [82, 83, 85, 112], he has analyzed the methodological approach to the PM [81, 84] and has shown empirical evidence to support its five filters [80]. However, Klaehn's analysis uses the same methodology of paired examples employed by Chomsky and Herman and hence, it has the same limitations.

Mullen [111] identifies seven points that need to be addressed to ensure that the PM stays relevant in the 21st century: (1) Reevaluate the role of the state in today's capitalist societies to keep updating our understanding of corporate-state and corporate-corporate power struggle that lead to elite consensus/dissensus; (2) Empirically test the hypotheses of the PM in countries with a different economic and political system; (3) Operationalize the filters of the PM; (4) Revisit the five filters and modify them if needed; (5) Test if the PM can predict the behavior of modern media forms such as social media, blogging, etc.; (6) Investigate if the PM fit with other models of media performance and (7) Evaluate the effect of the operation of the filter on the media audience to have a complete view of the propaganda system. Our present work tries to understand some of these issues by means of big datasets and automatic analysis. For example, we address the first point in chapter 3. Moreover, our entire thesis is intended to contribute to the second and third point, at least for some of the filters (see section 2.4.4). By testing our hypothesis using data extracted from social media and other online sources we directly apply ourselves to the fifth point. Finally, concerning the sixth point, in the next section we introduce alternative models on the political economy of the mass media that are also used to support our analysis in the following chapters.

2.3 Alternative models

Although we focus on it for the purposes of this thesis, the PM is by no means the only model in the literature that tries to explain the behavior of the media and its relation with the political and economic forces that interact in modern society. In this section, we discuss two alternative models that fundamentally address the same concerns raised by the PM. The first one, presented by Hallin and Mancini [66, 65], proposes three variations of the media system that can be found in most of the mutations of the western democracy found across Europe and North America. The second model is a clear example of the higher order predictions made by Chomsky [31, 113, 111]. Prat and Strömberg [129] present a theoretical framework to study the Media Capture and Political coverage. Without any reference to Herman or Chomsky, this work agrees with most, if not all, of the PM's operative principles. We will show that these "alternative" models do not fall far from the description of the media system proposed by the PM. In the next few chapters, we adopt some of the arguments that are used by these authors either to complement the theory of the PM or to support its hypotheses further.

2.3.1 Three models of media and politics

Hallin and Mancini offer a seminal work [66] in which they approach the study of the press by means of comparative analysis. Comparing two or more systems allows to keep the similarities as constants and highlight the difference as variables that can then be used to explain alternatives behaviors of the media. They propose three models that should represent three different systems. These models presented by Hallin and Mancini are intended as a replacement/modernization of the models introduced in the influential work of Siebert et al. *"Four theories of the press"* [145]. The authors agree with their predecessors in that the media have to be analyzed following the social and political structures within which it operates. Hence, the three proposed models try to capture the main features of three competing media landscape identified by the author as representative of most western democracies. The first model, the *Liberal Model*, can be applied where the free market and private commercial media are the ruling forces of the system. This one represents countries like the UK, Ireland and North America. The *Democratic Corporatist Model*, which dominate across northern continental Europe, describes a system where commercial media and media tied to organized social and political groups coexist. Finally, the *Polarized Pluralist Model* characterizes a system dominated by public/state media and an integration of the press with party politics. According to the authors, this last model would apply in most of the Mediterranean countries of southern Europe.

For their comparative analysis, Hallin and Mancini [66] propose a framework based on four primary dimensions: (1) the development of media markets, (2) political parallelism, (3) the development of journalistic professionalism, and (4) the degree and nature of state intervention in the media system.

The first dimension discusses whether the media distribution is intended for a selected elite audience or to the general public. This will also determine the role of the media: to open debates and negotiations among different sides of the elite (horizontal process) or to intercede between the political elite and the ordinary citizen (vertical process). The authors note that "quality" papers, designated for the elite, have a more sophisticated and politicized content. Meanwhile, the newspapers addressed to a mass public generally do not engage with the political world. As we mention before, Herman and Chomsky describe this tendency as a mechanism of the system to create political apathy in the general public (second filter [72]).

The second dimension, or political parallelism, refers to the alignment of the outlets in the media system to political trends. In other words, the media outlets are popularly perceived as leaned to one side or the other in the political spectrum (not necessarily linked to a political party but rather to a political range). The idea of political parallelism is closely related with the two ways in which it is possible to achieve diversity/pluralism in the media system: External Pluralism (EP) and Internal Pluralism (IP) [128]. EP requires that all political opinions have room and are represented in at least some of the suppliers of content in the media market. On the other hand, IP is achieved when every media company covers all sides of the main political issues in society. This

means that a system with high IP will have low levels of political parallelism.

This notion of the media acting as an instrument to create consensus among the elite or as a conveyor of their political message in exchange of economic support or preferential access (referred by the authors as “organizational connections”) is in alignment with the predictions of the Propaganda Model. We deal with these topics throughout the chapters of this thesis.

The third dimension proposed by Hallin and Mancini is the development of journalistic professionalism. Professionalism and Professionalization is the notion of individuals in the exercise of the profession based on a set of ethics rules and principles. The authors measure the degree of professionalization in the media according to specific parameters such as autonomy, distinct professional norms and public service orientation. The authors recognize that professionalization can be jeopardized by political instrumentalization, commercialization or both at once. Instrumentalization is defined here as the control over the media exercised by external actors (e.g., a political party, advertisers, etc.). As we mentioned before, Hallin and some other critics [65, 151] have pointed out that the Propaganda Model does not take into account the role of journalists and their professionalism in the process of news production. However, Herman and Chomsky explain that achieving journalists conformity with corporate and state interests can be obtained in more subtle ways. For example, they may hire a staff that shares a social and political background with the news organization. Likewise, Hallin and Mancini [66] recognize that journalist are usually politically active figure and they tend to build their career in news institutions with a matching political leaning.

The last dimension defined by the author considers the role of the state in the media. They grant that the state is a dominant force in molding the media system. In the different countries included in the study, how the government influence the system varies from ownership to laws and regulations of the market to flak (for example through libel, defamation, and right-of-reply laws). The authors also point out that the state and state-owned enterprises are also significant advertisers; which in turn, as discussed in the PM, give them a great deal of influence over the media. Even more, Hallin and Mancini concede that “the state always plays an important role as a source of information and primary definer of news [64], with enormous influence on the agenda and framing of public issues”. Herman and Chomsky share this same view of the role of the state in the media, primarily in their third filter of the PM. Our work here shows some media behaviors that support the idea of an implicit influence/intervention of the state in the system (see section 3.3.3). In general, the same arguments and cites to mechanisms of control are critical elements in the discussion of the five filters of the Propaganda Model.

2.3.2 The media capture and political coverage

In [129], Prat and Strömberg make an in-depth survey of the political economy of the mass media. They divide their study in four sections: transparency, media capture, informative coverage and political bias. For each of the dimension in their study, the

authors first present a theoretical framework and then a review of supporting empirical evidence.

The authors use for their analysis of the media capture a model first presented by Besley and Prat in [15]. Their model predicts that the government benefits the most from a media system with a low number of independent outlets (low pluralism), in which case the media can be fully capture (Proposition 1 [129]). A reduced number of owners makes it easier for the incumbent to bribe the media. They assign to the bribing process a transaction cost that represents the necessity of the incumbent to appeal to alternative instruments to reward compliant outlets. In extreme cases this can be a direct payment in cash or threatening journalist with prosecution. But it can also be more indirect, such as offering preferential news access to friendly outlets or by introducing regulations that favor the owners of the media companies. Besley and Prat [15] recognize that the media capture model can be also applied to other actors different from the government such as corporations. Corporate media capture can be achieve, for example, through advertising [52].

In general, the media capture model highlights the role of ownership concentration as a facilitator for capture (equivalent to the first filter of the PM). Moreover, the authors give evidence showing that even an ideological leaning in the news outlets can be economically motivated [57]. The model presents the benefits of the media in indulging the incumbent (cover by the third filter of the PM) and the mechanism of pressure available to the government to manipulate the media (referred as flak in the PM).

Prat and Strömberg also make a rich analysis on the media coverage and its effect on policymaking. In general, their model predicts that a more informed audience will have a higher impact on the election of politician and hence the politician will pay more attention to these sectors of the population. This creates a loop that ends up neglecting (policy-wise) the most vulnerable segments of society just for having limited access to the media. The model also predicts that to increase revenues, newspapers will cover more issues that are interesting to people who are valuable to advertisers and to which it is cheap to deliver the news either because it is easier to send reporters or because they are geographically closer to the source (Proposition 4 [129]). This model of coverage, in fact, offers a more formal definition of the same concepts presented in the second filter of the PM.

Surprisingly (or maybe not [113]), even though it is possible to find a direct match between most of the filters of the PM and the main arguments and discoveries presented in this work, there is no reference to any article in this subject authored by Herman and/or Chomsky.

2.4 Our contribution

Our contribution to the existing literature is three-fold. First, we investigate the Propaganda Model using a computational approach and massive datasets of digital records. This allows us to show empirical evidence that supports the existence of the filters but

at a scale that becomes impractical when using the traditional methods. Second, we explore the real impact of the filters in the new channels of communication. Social networks went from convenient to unavoidable for the media. With most of the social and political discussion taking place on the social media, every news outlet with a steady readership has its presence in sites like Facebook or Twitter [153]. Finally, our analysis relies on information about Chilean news outlets. The Chilean media have established a significant social media presence with a high number of Chilean users [106]. The Chilean media landscape is furthermore well documented due to the availability of detailed, publicly available data with respect to its ownership structure, compiled by *Poderopedia* [127], a journalist NGO that aims to understand power relationships between people, companies, and organizations. The case of Chilean news thus provides an interesting addition to previous media studies that are mostly focused mostly on Northern hemisphere countries, with a strong inclination to the United States and Western Europe. Still, we believe the methods can be readily generalized to other media systems of the world, and so can some of the findings.

2.4.1 Hypothesis

The Propaganda Model predicts the behavior of on-line media.

2.4.2 Goals

Main Goal

- Demonstrate the applicability of the Propaganda Model to on-line media using computational approaches with, as of now, the most representative sample of sociological data as evidenced by Twitter, both in terms of ecological validity and sample size.

Specific Goals

- Formalize and operationalize the filters of the Propaganda Model.
- Gather a representative database with information from different sources.
- Analyze the nature of the bias in the media to look for signs of consensus/dissent.
- Study the influence of the Chilean media's power structure over news editorial policies to support the first filter.
- Study the effects of consolidation and ownership on news diversity to further validate the existence and implications of the first filter.
- Quantify how much of the news outlets coverage can be explained by geographic bias and the politic/socioeconomic profile of the areas they serve. The evidence should be consistent with the postulates of the second filter.

2.4.3 Definitions

In this section we give the definition (a modest attempt of formalization) for the first two filters as these are the primary focus of our thesis. These definitions will guide our research and allow us to measure in some form the extend of the applicability of each filter in the media system under study.

First Filter (Ownership)

The strongest and most explicit claim proposed in the first filter of the PM is that free markets forces operating over the media system have led to the concentration of ownership and the opening of news companies to external investment. In turn, the authors postulate, this changes in the news media power structure have morphed the news schema into a profit-driven mechanism used by the political and economic elite (which own the media) to transmit a message that guarantees the *status quo* in a socioeconomic system from which they are the prime beneficiaries.

The tendency of the media system to an increasing concentration of ownership is a proven fact [9, 41]. As of December 2017, even with the reduced number of major corporation remaining, this relentless process keeps converging to a more and more concentrated system [61]. So, in order for the first filter to apply, we need to verified the predicted influence of the ownership over the editorial policies of their news outlets and show that this influence is motivated by their own private interest.

The political part of the prediction can be supported by showing a political media bias that systematically favors the government. Evidence of media capture (as defined by Prat and Strömberg [129]) will also reinforce the existence of this filter. On the other hand, the promotion of economic or self-interests promise a bit harder to prove at a system level. First, we have to show that there is an actual influence of the owners over their outlets' selection of content. Then, we should be able to see patterns in the choice and framing of the news that consistently protect/endorse a group of entities (e.g., members of the board of directors, high executives, parent and sibling companies of the news outlets, investors, etc.). The existence of any of these factors will uphold the hypothesis of the first filter.

Second Filter (Advertising)

As we mention earlier, the model used by Prat and Strömberg [129] to describe the media coverage agrees in most aspects with the predictions of the PM (see Section 2.3.2). In their *Proposition 4*, the authors laid the conditions that need to be matched by an issue to be published under a strictly profit-driven media system. These are: (1) to be of interest to a large group of people; (2) or to be potentially attractive to advertisers; (3) or to be journalistically newsworthy (e.g. a volcanic eruption or the Olympic Games) and (4) to have a low distribution cost to the group concerned with the issue.

The definition given by Prat and Strömberg overlaps with more than one filter of the PM. However, it mainly describes the produced content of the media as a mean to reach an audience that can be later commercialized and turn into profit. Since it covers the deliberate targeting of a preferred audience, the desired to appeal to advertisers and the lack of interest for maximizing social welfare (as opposed to economic profit), we consider that *Proposition 4* completely covers the basis of the second filter of the PM. So, to take advantage of the formal theoretical framework provided by Prat and Strömberg, we will assume the model and predictions of their *Proposition 4* as a superset of the second filter of the PM.

2.4.4 Methodology

In this thesis we will focus only on the first two filters (i.e., **ownership** and **advertising**) as they can be explained by studying the behavior of the media by itself. According to Herman and Chomsky [110], the first two filters belong to institutional analysis since they aim to explain how media firms operate given a set of both empirical and theoretical rules. On the other hand, "**sourcing**" and "**flak**" act as control mechanisms in any form of government. If the media is in fact subjected to the interest of the elite classes, then they would have to adapt to these processes. The last filter is an excellent source of media content and brings people together for a common cause, overcoming class differences. So, it is also a control mechanism that diverts the animosity of the public to a provisioned evil. For these last filters, we would have to study also the actions and behavior of other players (e.g., governments) in more complex processes that ultimately also shape the media but indirectly. Nonetheless, as mentioned before, the filters enter into sophisticated interactions and we will try to untangle them in these pages, always focusing in the ones that are of particular interest here.

To create our database of outlets, we used different sources, with Poderopedia's "influence" database [127] as our baseline, manually adding other news outlets in Chile. Our database contains 403 *active* accounts. An account is considered *active* if it tweets at least once a month. We enriched the information of each outlet by adding relevant information such as geographic location, scope, Twitter account, number of followers, etc. We use the Twitter's API to collect the tweets generated by these outlets automatically. As of December 2017, our dataset counts with almost 7 million tweets that expand over a period of two years. Depending on the experiment, we work with different subsets of our data to be able to compare our results and improve our analysis.

The authors of the PM have explained that they do not have rules, but rather observations on how and on what cases the filters will apply [110]. They point out that the elite consensus, which leads to an uncompromised servilism of the media, is more noticeable when it involve matters of external affairs that enhance the perception of power (such as wars). Another type of issues that bring consensus is when these controversies threaten fundamental interests of the ruling classes (e.g., military industry expansion, free trade, labour organization). Herman explains that the media also may be relatively more open to debate in cases where other groups in society show more

interest in the subject, are better inform or present a more organized front to the elite interests [71].

In our study of the media, we look for signs of this consensus/dissent by analyzing the nature of the (largely implicit) socio-economic bias of news outlets in the Chilean context. We propose to automatically categorize news outlets by analyzing what they "think" about certain relevant, controversial topics using their tweet content and then map these worldviews onto a well-known political quiz. By comparing the position of the outlets in two different periods we can identify the influence of government orientation on the media landscape.

To find evidence on the existence of the first filter, we research the influence that ownership relations have on news media content and coverage by quantifying the strength of the relation between news media ownership and the selection of topics they report in Twitter. We analyze the user accounts of news media outlets to study how their content overlaps, and whether or not these observations are linked to their known ownership structure. As explained in the first filter of the PM, the market-driven consolidation of the news media industry may lead to concentration of ownership. This concentration of ownership may have direct and indirect effects on editorial policies. For example, the editorial board of a newspaper owned by a group that also invests in agriculture may perceive a pressure to report more favorably about agricultural initiatives. Unlike other factors such as the (frequently explicitly publicized) political and historical mission of the outlet and its readership, ownership bias may thus exert a more insidious effect on editorial policies that is difficult to operationalize and quantify.

Furthermore, to assess the impact of the ownership concentration process over the news diversity, we uses a set of ecological indicators to analyze the "news ecosystem" as viewed from Twitter. We apply well-known ecology indices to quantitatively measure how "healthy" – diverse – our system is. We assume that diversity of content is a desired property of any news system. Both the *Media Capture Model* and *The Propaganda Model* warn against the negative consequences of the concentration of ownership in the mass media. Having a large share of the media industry in the hands of just a few mega-conglomerates poses the risk of the system not necessarily representing the interest of the common good, the media's original primary purpose.

For the second filter, we use statistical models to test how much of the distribution of Twitter followers can be explained based on the geographic and socioeconomic features of the different areas. We try to find empirical evidence on each of these factors at a scale that it is only feasible thanks to new communication technologies and the massive adoption of social networks. Although these correlations do not directly proof causality in the behavior of the media, it gives supportive evidence on the predictive power of the Propaganda Model.

Chapter 3

Nature of Real and Perceived Bias in the Mainstream Media

News consumers expect news outlets to be objective and balanced in their reports of events and opinions. However, the political and economic interests of news outlets and the people who control it have its impact on the news that the population of a territory gets served. As we saw in the discussion of the Propaganda Model (see Chapter 2), Herman and Chomsky [72] argue that political and doctrinal interests have penetrated the press at different stages of the news generation process, deliberately or accidental (e.g., through homophily effects).

People usually have some intuition of media bias. For average readers, though, it is very difficult and time-consuming to be aware or even find the bias of all media outlets, let alone quantify these biases and give them a total order in terms of the magnitude of the leaning. If, for example, a newspaper claims to be objective, but is in fact "right-wing, conservative" (as is the case with El Mercurio in Chile [169]), people should be able to recognize this and take this bias into account when reading its content. The case of El Mercurio is quite clear, and being a very old, traditional newspaper, the bias is actually known and arguably accepted. It is important to emphasize here that "bias" is not categorical, but comes embedded in a geopolitical news context determined by other outlets in the region [162]. In other, bolder words, some bias is inherent to all media, but how biased they are, depends, to an extent, upon a comparison to other media.

Bias in the media is a global phenomenon, not exclusive to one kind of economy or particular political system. As such, there is now a quickly growing body of empirical evidence on its existence [102, 38]. It have been showed that there are several types of bias in media coverage [159, 137]. What has not been studied as deeply, however, at least not quantitatively, is *how* outlets could be positioned in a socio-economic space. Knowing *the nature* of media bias will give is a better understanding on *why* certain news outlets occupied a specific position in the socioeconomic/political spectrum. According to Herman and Chomsky [110], there are some topics that bring elite consensus, which leads to an uncompromised servilism of the media. However, Herman explains that the media also may be relatively more open to debate in cases where other groups in society show more interest in the subject, are better inform or present a more organized front to the elite interests [71].

In this chapter, we automatically identify the (largely implicit) socio-economic "relative bias" of news outlets in the context of Chilean media. A socio-economic study at this scale help uncover patterns of editorial policies that show a systematic bias

that favors governments' propaganda or private economic interests over social welfare. Operationalizing bias is a difficult task. It relies not only on linguistic information, but also on the actual geo-socio-economic, and even historical, context of the newspaper. We propose to automatically categorize news outlets by analyzing what they "think" about certain relevant, controversial topics using their tweet content and then map these worldviews onto a well-known political quiz: "The World's Smallest Political Quiz" (henceforth *PolQuiz*) [155].

The *PolQuiz* has ten question, and it was originally intended for an American audience. Although we believe this does not imply a loss of generality wrt Latin American culture, at least in the topics chosen. It does, obviously, impact the polarity of attitudes towards those topics, but that is what we explore in these pages. It was designed by the Libertarian Advocates for Self Government [157], created by Marshall Fritz in 1985. The quiz is based on the one proposed by David Nolan Chart in 1971 [116], which in turn can be traced back to a 2D chart proposed in 1968 [23], representing variations in political and socio-economic orientation.

In short, we use what the media say on Twitter to position them in a Cartesian plane that tells us more about their orientation based on Fritz's quiz[1]. Then, we conduct a deeper investigation into the nature of the found bias, which we study through the vocabulary used and entities covered by the news outlets. Finally, we show the results of a survey that confirm that media bias has a noticeable impact on how news related to controversial topics are presented.

3.1 What we know about media bias detection

There are several works related to the topic of media bias [121, 57, 180, 59, 38]. Some works do not try to identify bias directly, but instead try to identify and track events in order to present different points of view of the same affair to the readers in order to counteract these possible bias [121]. These are complemented by works like J. An *et.al.* [6], which create a so-called landscape of newspapers based on the similarity of their communities. They measure the exposure of Twitter users to politically diverse news. Other authors assume a certain leaning by contacts association [34]. In [137] the authors go deeper and try to identify different kinds of bias, what they term gatekeeping, coverage and statement bias, according to the stage at which the news acquire the alleged bias.

Most outlets identify themselves as unbiased free press, which makes the discussion on the direction and degree of media bias very controversial. To be fair, it is true that "bias" in journalism may arise naturally out of the interaction of reporters, rather than *a prior* [66] (see Section 2.3.1). Media bias is usually found in the editorial policies that ultimately decide which stories are worth publishing and which amount and angle of coverage they get [137, 121, 154].

[1] We presented an earlier version of this work in HT'17 [44].

This bias reflects the political and socio-economic views of the institution, rather than the point of view of a particular reporter. For example, in the PM [72] the authors use a few recent events to point out the usage of different vocabulary applied to similar issues depending on their own convenience (see Section 2.1.3).

In [93], the authors defined a model to predict political preference among Twitter users. Through this model they calculate, for each user, a ranking of the likelihood that they prefer a political party over another. This model is based on the usage of *weighted words*. The words and their weights are extracted from tweets of candidates of certain political parties. Using these weights, in combination with Twitter specific features (retweets, following, etc.) the authors train classifiers that achieve a performance similar to that of human annotators. Similarly, in [57], the authors estimate the bias in newspapers according to how similar the language is to that used by congressmen for which a right/left stand is known. One interesting result is that bias in the news is found to be correlated to political inclinations of readers, showing a tendency in these news outlets to maximize profit by "catering" to a certain audience. We deep in the analysis of this specific behavior in another Chapter (see Chapter 6).

The topology of the social network on its own has also been shown to give enough information to create classifiers concerning a user's preference, even when the choices are very similar [54](e.g. Pepsi vs. Coca Cola, Hertz vs. Avis or McDonalds vs. BurgerKing). Although we carefully select the dataset to use in our experiments to achieve extensible results [33], we notice here that in our dataset, news outlets (which may be considered the participants of our study), regularly talk about these controversial topics, and thus, it is possible to use more traditional methods (such as NLP) to find a political stand.

Combining topological characteristics of the social networks with language features has also been tested [34], showing that users tend to interact more frequently with like-minded people. This phenomenon is known as *homophily*. As we mentioned before, our dataset is derived from a special type of users (news outlets Twitter accounts), and this method may not apply directly.

As an alternative approach, in [180] the authors propose a semi-supervised classifier for detecting political preference. They design a propagation method that, starting with a few labeled items and users, creates a graph representing the connections between users and items or even users with other users. Based on the same phenomena of homophily, they assume that users interacting with the same item, or with each other, most likely have the same political leaning. This way, they can propagate the labels from tagged users and tagged items to the rest of the graph. They report that the system achieves over 95% precision in cross-validation. In [59, 62], the authors also follow a propagation strategy to compute the political preference of Twitter users, but using Congress members as the initially tagged users.

In [92], the authors describe a framework to discover and track controversial topics that involve opposing views. They first use tags that represent each side (e.g. "#prolife" - "#prochoice") as seeds to find an expanded set of labels to represent each side. This

may also help in cases where labels may change over time as the result of new arguments for either side. With these sets of labels they identify strong partisans (anchors) that have a clear lean to one side. Having these anchors and a graph representing relationships between users (based on similarity of tweet content or based on re-tweets), they propagate the classification through the graph inferring the opinion bias of "regular" users.

Yet another approach to quantifying political leaning is presented in [102]. They based their analysis on the number of tweets and re-tweets generated about different political events associated with some predefined topics. The authors developed a model that takes into account both the sentiment analysis of the tweets and the number of time they are re-tweeted to calculate the political leaning score of each outlet.

In [162], the authors propose an unsupervised model based on how news outlets quoted president Barack Obama's speeches. The findings suggest that quotation patterns do reveal some underlying structure in the media, and that these may be evidence of bias. They found that one of the identified dimensions roughly aligns with the traditional left(liberal)-right(conservative) political classification and the other with a mainstream/independent one. This is a strong finding. Still, we believe this is to be somewhat expected, given the selected corpus; namely, presidential speeches in the strongly bipartisan system that dominates U.S. politics. Although this model helps classify and quantify bias in the media, it does not explain the causes and nature of this bias.

Here, we present a new methodology that quantifies the political leaning of news outlets based on the automation of a well known political quiz. The prediction of the answers for each question for each outlet is generated based on the polarity of their tweets on subjects related to the issues addressed in the quiz. The automation of a quiz has been used before to automatically classify mood [16] but, as far as we know, this is the first attempt to quantify media bias using this approach.

3.2 Methodology

As we mention in previous chapters, our database contains 403 *active* accounts. An account is considered *active* if it tweets at least once a month. We selected a subset of our data that contains 1,916,709 tweets, spanning a period of 8 months - from October 6, 2015 to June 4, 2016. The accounts vary dramatically in tweet publication behavior, with some having published more than a hundred thousand tweets to others with less than a hundred in this timeframe. Out of the 403 active accounts, only 269 outlets published at least one document about the topics of interest.

We treat every tweet as an independent document from which we can extract a statement. We assume that these reflect the ideology of the news outlet as an entity. As many others, we use Twitter as our source documents to study news [95, 179, 102]. Given that Twitter and other social media have become hubs of news for an

increasing number of users [103], we consider a tweet from a media outlet as a man-made summary of the news (usually in the form of a *headline*). It conveys the main idea, and hence arguably the main editorial point of view. Headlines of online news articles have shown to be slightly more reliable than full text for adequately providing a high-level overview of the news events [3, 40, 158]. These summaries are expected to be representative of the newspaper's bias [168, 114], but with the advantage that bias is easier to detect than in a full articles (shorter, to the point). Because of this, tweets, in all their simplicity, seem to be not only enough, but a better fit, for our profiling purposes.

3.2.1 PolQuiz

The *PolQuiz* has ten questions, divided into two groups: economic and personal issues, of five questions each. The answers to the questions may be *Agree*, *Maybe (or Don't Know)* or *Disagree*.

Personal issue questions:

1. Government should not censor speech, press, media or Internet.
2. Military service should be voluntary. There should be no draft.
3. There should be no laws regarding sex between consenting adults.
4. Repeal laws prohibiting adult possession and use of drugs.
5. There should be no National ID card.

Economic issue questions:

6. End corporate welfare. No government handouts to business.
7. End government barriers to international free trade.
8. Let people control their own retirement: privatize Social Security.
9. Replace government welfare with private charity.
10. Cut taxes and government spending by 50% or more.

Based on the answers to these questions, the quiz-taker is classified into one of the five categories: left-liberal, libertarian, centrist, right-conservative, or statist. *Left-liberalism* is a political ideology that supports governments that take care of the welfare of vulnerable people and keeps a centralized economy, but at the same time, allows a great deal of liberties in personal matters. *Libertarians* seek freedom in both economic and personal issues, minimizing the role of the state in all matters. An extreme position in this direction would be anarchism. On the other side, *statists* - or supporters of a big government - want the state to regulate both personal and economic issues. Examples of this position would be totalitarian regimes, such as Kim Jong-Un in North Korea. *Right-wing conservatives* are more reluctant to accept changes in personal issues and want official standards on these matters (i.e. morality and family traditions), but demand economic freedom and a free market. Finally, *centrists* accept or even support a balance between the government reach and personal/economic freedom. They favor selective government interventions to current problems while avoiding drastic measures that may shift society to either side of the spectrum.

For each *Agree* answer, we increase the score of the quiz-taker in the corresponding dimension by 20 points. If the answer is *Maybe (or don't know)*, we only add 10 points. Finally, if the answer is *Disagree*, no points are added. This way, if the quiz taker agrees with all the issues in one dimension, it will be in one end of that axis. In the other extreme of the axis, we will have a quiz-taker who disagrees with all issues in that dimension. In our study, we assume that news outlets are (or strive to be) unbiased, so in an ideal world, most of their comments should have no polarity toward any side of the issue and, as such, they should score as a *Maybe*. Another expected behavior would be that news outlets report on both sides of the issue to cover different points of view. Both approaches would result in the news outlet being in the center of the graph.

There is a long tradition of surveys to profile individuals and position them on a socio-economic landscape (see, for example, [46]). The instrument we use here, "The World's Smallest Political Quiz" (PolQuiz), follows this tradition. The theoretical foundations of the *PolQuiz* can be traced back to the works of David Nolan, Maurice Bryson and other political scientists in the late 60s and early 70s [116, 23]. Although there are other more "complete" quizzes online (see, for example, *The Political Compass* [119]), an advantage of the *PolQuiz* is that it is "open source", in the sense that the scoring system is known, unlike for example the Political Compass one mentioned above. It is also short (only five questions per dimension), very popular (the Advocates for Self-Government (founded by the creator of the quiz: Marshall Fritz) claims that the quiz has been taken online more than 23 million times [156], and it has been used, evaluated and cited scientifically [39, 32].

3.2.2 Operationalizing the Quiz

We filtered the tweets to get only those with information regarding the issues referred to in the *PolQuiz*. For this, we created a seed query for each question, containing a set of preselected keywords (see Table 3.1).

With the subset of documents returned by the seed queries, we then analyzed the hash-tags to find an expanded set of labels that may represent related aspects of the same issues [92]. We removed hash-tags that contain the name of a news outlet, as it is common practice in newspapers accounts to use hash-tags to refer to themselves or the original source of the news (regardless of the subject). We also remove hash-tags with names of politicians: even when these politicians could potentially provide some relevant documents, they also introduce a lot of noise, mostly due to the salience of politicians who appear regularly in the news for a wide variety of issues not necessarily related to the query in question. The new labels are added as keywords to the original query. Our enriched queries give us the final set of tweets used to evaluate any possible bias of each news outlet, see Table 3.2.

Having the set of tweets for each question, we classified their polarity *with respect to the corresponding question*. For example, for *question 7* (**q7**), a tweet classified as *Agree* is *"TPP abrira puertas a ms de 1.600 productos chilenos no incluidos en*

Question	Keywords
q1	(censura — libertad) & (prensa — discurso — expresion)
q2	(servicio — reclutamiento — entrenamiento — reserva) & (militar — ejercito — armada)
q3	(ley — legal — legislacion — regulacion — penalizacion) & (sexual — prostitucion — sexo — sodomia — gay) & ¬(infantil — menor — niño — acoso — abuso — agresion)
q4	(ley — legal — legislacion — regulacion — penalizacion) & (droga — marihuana — cannabis — psicotropico — cocaina)
q5	inmigracion — inmigrante — refugiado — xenofobia
q6	(subsidio — bienestar — ayuda) & (corporativa — empresa)
q7	(trato — tratado — convenio — negociacion — relacion) & (comercial — economica) & (internacional — bilateral — gobierno — libre — liberal — barrera — proteccion — bloque)
q8	(”seguridad social” — afp — pension — jubilado — prevision) & (privada — gobierno — estatal)
q9	(”beneficio sociale” — bono — ”ayuda sociale” — ”programa social”) & (gobierno)
q10	(reducion — recorte — aumento — incremento) & (impuesto — gasto) & (gobierno — gubernamental)

Our actual queries are designed so they can also find variations of the keywords (such as variations in gender and number). AFP stands for Administradoras de Fondos de Pensiones (Chilean pension system)

Table 3.1: Initial set of keywords for each query.

acuerdos vigentes.” (tr. TPP will open doors for more than 1.600 Chilean products not included in existing agreements). For that same question, the following tweet disagrees with it: *“El TPP: un misil contra la soberana” (tr. TPP: a missile against our sovereignty)*. In other words, we classify the polarity of the tweet with respect to the corresponding issue. As the number of tweets is too large to label manually, we created and trained a supervised model for each question. This approach also allows us to scale in the presence of an even larger number of resulting documents.

To create a representative sample for the training set, we randomly select, where possible, two tweets from each question from each news outlet. We took care to not include duplicate tweets (tweets with the exact same text) published by the same outlet. The training set consisted of 1916 documents (an average of about 190 documents per question). We manually classified this training set in four groups: *Agree*, *Maybe*, *Disagree* and *Out of topic (Not relevant)*. The distribution of each training set is shown in Table 3.2.

For the automatic classification task, we used a “Randomized Trees” model [58] (Implemented in the python library `scikit-learn` in the module `ExtraTreeClassifier`).

Qs	tweets	Training set (TS)	% Agr (TS)	% Mb (TS)	% Dis (TS)	% Not rel (TS)	Prc ($\pm 2 * stdev$)
q1	374	179	48.6	16.7	17.8	16.7	0.76 ($\pm$ 0.14)
q2	194	132	18.1	29.5	20.4	31.8	0.87 ($\pm$ 0.11)
q3	144	78	57.6	05.1	24.3	12.8	0.83 ($\pm$ 0.17)
q4	597	203	61.0	08.3	14.2	16.2	0.80 ($\pm$ 0.10)
q5	746	219	35.1	12.7	15.9	36.0	0.73 ($\pm$ 0.16)
q6	636	117	26.4	25.6	23.9	23.9	0.53 ($\pm$ 0.20)
q7	1162	238	29.8	24.7	39.9	05.4	0.76 ($\pm$ 0.09)
q8	251	117	21.3	09.4	41.8	27.3	0.76 ($\pm$ 0.13)
q9	298	167	05.9	13.1	69.4	11.3	0.87 ($\pm$ 0.09)
q10	8573	466	16.7	13.3	66.0	03.8	0.71 ($\pm$ 0.06)

The last column indicates the average precision obtained by the model in cross-validation (See Section 3.2.2)

Table 3.2: Tweets extracted from our corpus after applying the enriched queries.

Decision trees are less susceptible to overfitting, considering that we have relatively small training sets. Given that the classes in our training set are not evenly populated, we decided to evaluate the model using a 10 iterations stratified shuffle-split cross validation. Each fold leaves out 20% for validation. The other 80% is selected while preserving the percentage of samples for each class. The accuracy values for each model is presented in Table 3.2.

After the classification stage, we scored each news outlet on each question. We removed those documents classified as *off-topic (Not relevant)*. We scored the remaining documents' polarity according to the *PolQuiz* scoring system and we found the average for each question/news-outlet pair. For simplicity, in the question/news-outlet pairs for which we have no associated documents, we assume a *Maybe (or don't know)* answer. This assumption is the least disruptive towards the default supposition of an unbiased media.

In order to find out how sensitive the observed bias is to noise, we repeated the scoring steps 20 times. Each time we leave out 5% of the tweets, selected at random while maintaining the original distribution of documents per question. Each time we measure the average score of the news outlets for which we were able to answer at least one question in the corresponding dimension. We did not go over 5%, because the smallest news outlets already have a small set of documents: removing too many entries would have resulted in the elimination of an entire outlet, affecting the results.

Finally, we tested how the entire system adapts to the local environment. For a proof of concept, we introduced the subject of abortion in the personal dimension. This topic appears among the personal issues in other political quizzes (e.g., *The Political Compass* [119]) In addition, abortion was a very relevant and controversial topic in the Chilean media during this period because of a new bill presented by the president and

approved by the Chamber of Deputies to legalize the abortion on three grounds: pregnancy resulting from rape, lethal fetal infeasibility or danger to the life of the pregnant women.
We formulated this new question as follows:

0. All women should be free to choose whether she wants to terminate her pregnancy or not.

Notice that the question is formulated in the same "direction" as the rest of the questions. This is, agreeing with the statement will be an indicator of a more liberal tendency by the quiz taker.

Question	Keywords
q0	(ley — legal — legislacion — regulacion — penalizacion — despenalizacion) & (aborto — interrupcion — embarazo)

Our actual queries are designed so they can also find variation of the keywords (such as variations in gender and number)

Table 3.3: Initial set of keywords for q0 query.

We apply the same methodology described before to the original *PolQuiz*. We named this question **q0** and the query we applied (before injecting the hash-tags) is shown in Table 3.3. The enriched query returned 4891 documents from our corpus. We selected two random documents for each news outlets to create a training set containing 409 tweets. We had an average precision of 0.70 (± 0.08) in the 10 iterations stratified shuffle-split cross validation.

3.2.3 Rank difference

Using the *PolQuiz*, we aim to show empirically that the news media in Chile have some socio-economic leaning. This means that news outlets tend to have a stand in at least some of the controversial topics that dominate the political landscape of the country. However, remember that we are also interested in the *nature* of the bias regarding such controversial topics.

To do this, we use the *rank difference* method proposed in [79]. Rank difference is used to identify terms that characterize a specific domain. For example, the word *court* will be probably identified as a term if we are analyzing a corpus of legal documents. The method creates a ranking of words based on their frequency in a domain and a generic corpus. By comparing their relative position in both corpora, the algorithm identifies words that are significantly more used in a given domain. These unusual word frequencies are used as an indication of the importance of these words in the given domain. The formula for calculating rank difference is shown in Equation 3.1,

$$\tau(w) = \frac{r_D(w)}{\sum_{w' \in V_D} r_D(w')} - \frac{r_B(w)}{\sum_{w' \in V_B} r_B(w')} \tag{3.1}$$

where $r_D(w)$ and $r_B(w)$ are the ranks of word w in the domain and background corpus respectively. Rank normalization is done against the summation of all word rankings in the corresponding vocabulary (V_D and V_B).

3.2.4 Survey

To investigate to what extent the bias – as measured with the *PolQuiz* and investigated using the rank difference method – is perceived by the general audience, we conducted an online survey. We chose abortion as the topic of this survey, as this is (as explained in Section 3.2.2) a current and controversial item in Chile. In such cases the authors of the PM predict that the media will be open to debate that has received an important amount of coverage in the local media. This means that most people in Chile are aware of the discussion and probably have their own opinions. We also restricted our survey to the subset of news outlets who had relevant tweets for at least four questions per dimension (see Section 3.3.1) since these are the ones that we were able to position in the chart with the highest confidence.

We calculated the bi-grams' rank difference (see Section 3.2.3) for each news outlet. We decided to present bi-grams to users in the survey instead of words, because bi-grams offer more context, so it was easier for people to assess the connotation of a word or set of words within the selected topic. We also decided to use bi-grams over named entities because people not always recognize all the names involved in the discussion, although they do have an intuition in the discourse and the arguments used on both sides.

For each survey we presented a randomly selected and anonymised list (each list represents a news outlet) with the top-20 ranked bi-grams in one column and the bottom-20 bi-grams in another column. The top-20 list was presented as the words used with a relatively high frequency by one outlet. The bottom-20 list was presented as words the outlet tried to avoid or used with a relatively low frequency. The user had to answer if, based on these lists, he or she considered the outlet to be "in favor" or "against" abortion. The user could also respond with an "I can't tell" option. A user could answer the survey more than once, but the random selection was always made from the remaining lists.

We scored the "perceived bias" for each news outlet based on the answers we received in the survey. For each outlet, we calculated the percentage of users that answered "in favor" and subtracted the percentage of "against" answers. These percentage include the users that answered "I can't tell". So, we consider outlets with a negative score to have a conservative "perceived bias". Equivalently, outlets with a positive score are considered as liberals in our "perceived bias". It is worth noticing that an unbiased news outlet should be expected to score close to zero (because it should have mixed signals and, either a proportional number of user labeled in each direction or most users were unable to classify it).

3.3 Results

We found that the media landscape in Chile in terms of absolute positions is highly in line with the political orientation of the government. Within this landscape, relative differences in biases can be observed, which are in line with public perception – as captured by tendencies reported in Wikipedia. Further, we show that the nature of the bias can be explained and shown by the entities and sentiment related to the news outlet. Finally, we found empirical evidence on how the media landscape, in terms of absolute positions, shifted along with a shift of the government, which further support the existence of media capture.

3.3.1 Measuring bias using the PolQuiz

Our list of news media covers outlets of very different sizes (as measured by number of followers of that outlets's Twitter account). This difference in size is also reflected in the number of documents related to the issues in the *PolQuiz* that we are able to retrieve for each news entity. As we can see in Fig 3.1 there is a trend where the larger the outlet (measure in number of followers), the more they talk about (at least) the socio-economic and political issues related to the *PolQuiz*. Mid-size outlets are well represented and show an active participation. In this work we are interested in the behaviour and bias displayed by each different outlet in the media landscape, regardless of the size. This notwithstanding, we wanted to make sure we were covering the entire spectrum of the Chilean media. Our methodology is able to find the outlet's position even with just a few tweets. The evident extension of a more general analysis that takes into account the weight that each news outlet contributes to the global media bias is left for future work.

Absolute bias: the media landscape

To understand the results, a few preliminaries about the Chilean context is in order. First, the president during the observed period, Michelle Bachelet, was affiliated with the socialist party. In general, the ruling coalition was "Nueva Mayora", which mainly consisted of center-left to left-wing parties. Second, one of the strongest component in this coalition was the Christian Democratic Party. Christian democracy is still a center-left political ideology, but probably the most conservative within the government, especially in personal issues. With this in mind, Fig 3.2 shows the absolute positions of 269 news outlets that published at least one tweet related to the issues of at least one of the questions of the *PolQuiz*. Notice that for outlets with no information on a given question we assumed, conservatively, that they were "unbiased", in the sense that they did not explicitly pronounce their stance one way or the other. We can see that outlets tend to tweet more content pertaining to the economic axis than to the personal axis. This may suggest that communicating economic issues is more important to the news system in terms of reaching or influencing their audience. A drastic shift in economic

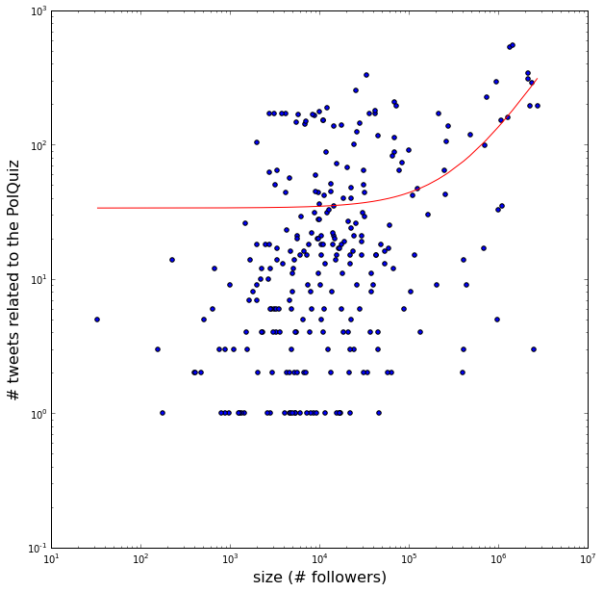


Figure 3.1: Distribution of tweets related to the *PolQuiz* per news outlet. The red line represents the least squares polynomial fit.

issues may invoke fear of losing your job or livelihood. Meanwhile, personal issues like freedom of speech are more ideological but of less immediate effect. The figure also shows that there is a clear preference in the Chilean media for the *left-liberal* end of the spectrum. This is explained, at least in part, by the political context of Chile during the observed period, discussed above. So, in this case, the found leaning of the media has a similar tendency to the political alignment of the ruling coalition. This result already lends some evidence to the *Propaganda Model* [72]. We explore more in how a change in government may influence the media position in Section *3.3.3*. Notice that this tendency also coincides with the "liberal media" label that is frequently used, implying a popularly perceived Left-liberal leaning in the majority of the outlets [55].

Of the 269 news outlets, our method yielded 26 that answered at least four questions on each dimension (we will call this subset of news outlets the ***26ers***). This represents 10% of our database and 13% of those that regularly report on economics and politics. The ***26ers*** account for 45% of the tweets relevant to the subjects in the *PolQuiz*. In Fig 3.3 we explicitly labeled some of the most prominent ones to help understand the general landscape.

Measuring perceived bias

We wanted to investigate if the outlets' bias obtained using the *PolQuiz* corresponded with the popular perception of their political orientation. We annotated the ***26ers*** political alignment using information extracted from Wikipedia, the official web site of the news outlet or the political alignment known for their owners. Since Wikipedia pages

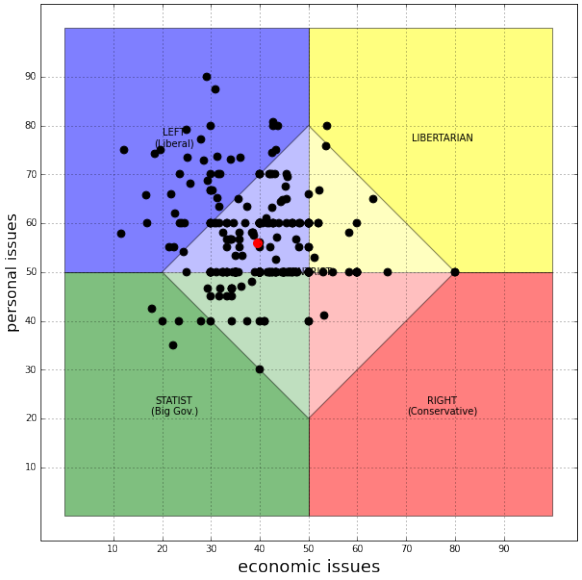


Figure 3.2: Absolute positions of the Chilean media. The chart shows the position of the outlets for which we were able to answer at least one question. The red dot shows the average position.

are crowdsourced content, we consider the political alignment extracted from there as either self-declared or a popular perception. Remember that the Christian Democratic Party is part of the center-left coalition that was ruling in Chile during the observed period, and was generally in favor of the social changes promoted by the government. We used the label “Christian democracy” to group the outlets associated to this party, and assigned them a left-leaning position in our analysis. This classification is taken as ground-truth to evaluate our model.

There are three of the ***26ers*** for which we could not find reliable information on their political leaning. Two of them (tele13_radio and t13) are owned by Chiles Grupo Luksic (one of the richest families in Chile and Latin America). The third one (pinguinodiario) is a mid-size daily local newspaper headquartered in Punta Arenas, in the south of Chile. In the ***26ers*** there are also two outlets that belong to international groups originating outside of Chile: publimetrochile and adnradiochile. The first one is owned by Metro International, a Swedish global media company based in Luxembourg that publishes the Metro newspapers in many big cities around the world. The other one is controlled by a subsidiary of the Spanish group PRISA. Although these international companies may also have a political stand, it is probably less influential in theirs outlets editorial policy. We did not make a bias inference either for these cases. Nevertheless, very much as for any other outlets outside the ***26ers***, some valuable information can be learn from their automatic classification.

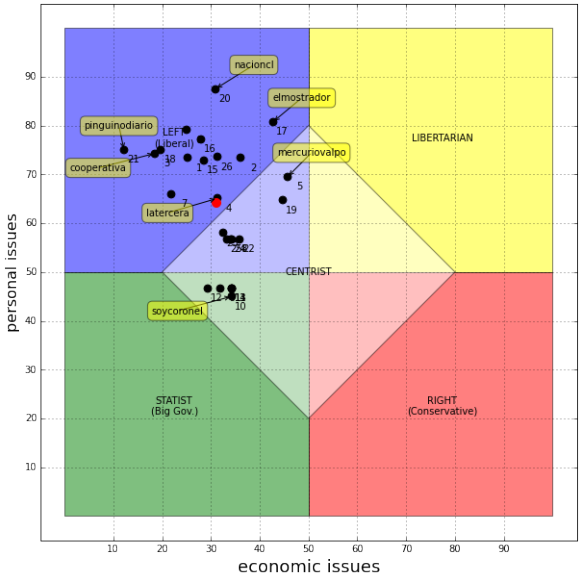


Figure 3.3: Absolute positions of the 26 news outlets who had relevant tweets for at least four questions per dimension (the *26ers*). The red dot shows the average position. 1. adnradiochile, 2. biobio, 3. cooperativa, 4. latercera, 5. mercuriovalpo, 6. publimetrochile, 7. emol, 8. soyarauco, 9 soyconcepcion, 10. soycoronel, 11. soyquillota, 12. soysanantonio, 13. soytalcahuano, 14 soytome, 15. dfinanciero, 16. el_ciudadano, 17. elmostrador, 18. tele13_radio, 19. el_dinamo, 20. nacioncl, 21. pinguinodiario, 22. soychillan, 23. soycopiapo, 24. soyvaldiviacl, 25. soyvalparaiso, 26. t13

In the rightmost of the group, we have `mercuriovalpo` (Tags in Fig 3.3 are the corresponding Twitter accounts (e.g., `https://twitter.com/mercuriovalpo`)) that represents *El Mercurio de Valparaso*, one of the oldest newspapers in Chile currently in circulation. This newspaper is part of a big conglomerate (*El Mercurio S.A.P*) that owns more than 20 news papers and several radio stations, among other broadcast media (such as magazines, TV cable, etc.). The regional newspaper *Soy Coronel* (`soycoronel`), on the bottom, is also part of this group. In fact, 11 regional newspapers of *El Mercurio S.A.P* are within these 26 and are all clustered bottom-right. As we mentioned earlier, the *El Mercurio*'s newspapers are popularly perceived as right-wing conservative.

La Tercera (`latercera`), is owned by *Copesa S.A.*, which is *El Mercurio*'s closest competitor. These two companies have a so-called news media duopoly. *La Tercera*, also in the lower-right, but closer to the center of the group, is thought to be moderate-conservative [171]. *El Mostrador* (`elmostrador`) is an on-line newspaper with a perceived orientation to progressivism [170].

Finally, we want to mention *La Nacin* (`nacioncl`) since it is a newspaper that currently only publishes its online edition and is partially controlled by the government.

This newspaper appears in the top region of the *personal* dimension. Compared to the other 25 news outlets, this one appears as the most progressive on personal issues. This score probably can be attributed to a series of populist reforms promoted by the government during the observed period (i.e. therapeutic marihuana legalization, decriminalization of abortion, anti-xenophobic campaigns, promote voluntary enlistment of women to the military service, etc.)

The perceived bias assigned to the rest of the ***26ers*** is shown in Table 3.5. Notice that many outlets in that list are perceived as *Right-wing, conservative*. There is also a group labeled as *Libertarian*. None of those outlets' *PolQuiz* automatic classification correspond with their popular recognized leaning. Our hypothesis is that this mismatch is produced by a lack of contextualization: if we look at the range of scores obtained by the automatic method, we notice that they are confined to a fraction of the entire space. In order to investigate our proposition, we normalized the original scores by making the range of observed values our entire universe. We discuss these results in the next section.

3.3.2 Relative positioning

We applied a normalization to contextualize the political leaning of the outlets to the reality of Chile. We normalized the scores on each axis in the range $[0, 100]$. Now our entire positioning universe is determined by the scope defined by the Chilean media. Fig 3.4 shows the relative position of the ***26ers***. Being the quantification for each outlet relative to the others, it means that we are now aiming for a comparative analysis between members of the media. These new position are much more inline with how people think of the leaning of these outlets.

To measure the improvement in the positioning of the outlets, we calculated the Mean Square Error (MSE) of the euclidean distance between the automatic calculated scores and the corresponding point estimated from the assigned perceived bias. Since the perceived bias is a qualitative measure, we estimated their position to be in the center of their corresponding quadrant (or in the border with the Centrist region for those outlets with a more moderate leaning (e.g., cooperativa – Left-Centrist border – associated with Christian democracy or latercera - Libertarian-Centrist border – associated with Classical liberalism)). Remember that we have the perceived position of 21 of the ***26ers***. The original position (without normalization) for these 21 outlets gives us a $MSE = 2104.93$, while with the relative localization it comes down to $MSE = 934.94$. When comparing against the perceived bias shown in Fig 3.5, we can notice by visual inspection that the normalized positions give a more accurate representation of the Chilean media landscape than its original counterpart.

This result shows that the bias should not be a categorical measure. Media bias comes embedded in a geopolitical news context determined by other outlets in the region. In other words, some bias is inherent to the media, but how biased they are (and on what direction they lean) will depend upon a comparison to other media in the same context. But, more importantly, *our own perception of the bias seem to be*

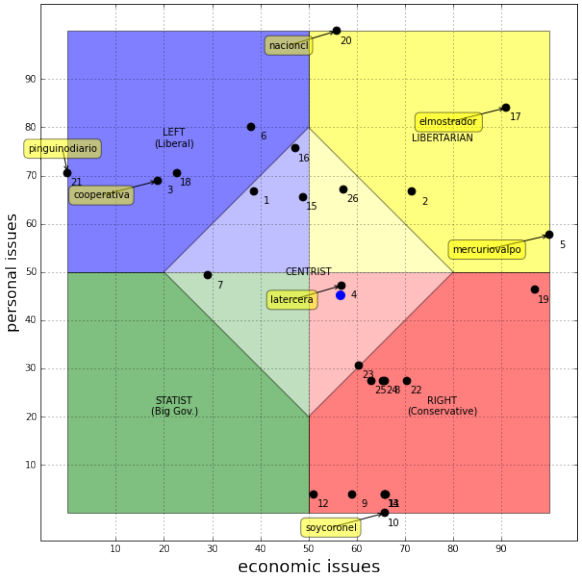


Figure 3.4: Relative position of the *26ers*. The blue dot shows the average position.

adjusted and limited by the political space defined by the news that we receive as a population. This phenomenon has important socio-politic implications (such as the possibility of artificial displacement of the center for political purposes), but we leave to social scientist further considerations on these matters. Our work, like the Propaganda Model, will not concern with the effect of the bias in the population, but without a doubt, this strengthens the importance of studying the media behavior.

We noticed that even with the normalized scores, the Chilean media is not balanced. For our statistical analysis we treated each axis independently, so we could work with values in only one dimension. We conducted a one-sample Student t-test (the QQplot and the histogram suggested normality was a reasonable assumption) for each dimension (economic and personal) to test if the mean score was significantly different from 50 (the assumed unbiased score). We used, for each dimension, the scores of those news outlets for which we were able to answer at least one question on that dimension. For the economic dimension, there is a significant bias, $t(254) = -10.93, p < .001$, with a leaning to the left-wing $(M = 40.28, SD = 14.21)$. In the personal issues the bias is lower, but still is statistically significant, $t(190) = -2.10, p < .05$, with a leaning to the conservative side $(M = 47.42, SD = 16.98)$.

Once again, we think the slight left-wing bias in the economic issues might be explained, at least in part, by the political context of Chile during the observed period (see Section *3.3.3*). On the personal issues dimension, we can also see some bias, although less prominent, tending to the conservative end of the spectrum. Other possible factors that might contribute to the observed tendency are the unavoidable bias introduced by the quiz itself and the presented methodology. The *PolQuiz* has been criticized as being biased by using leading questions, favoring libertarian results and

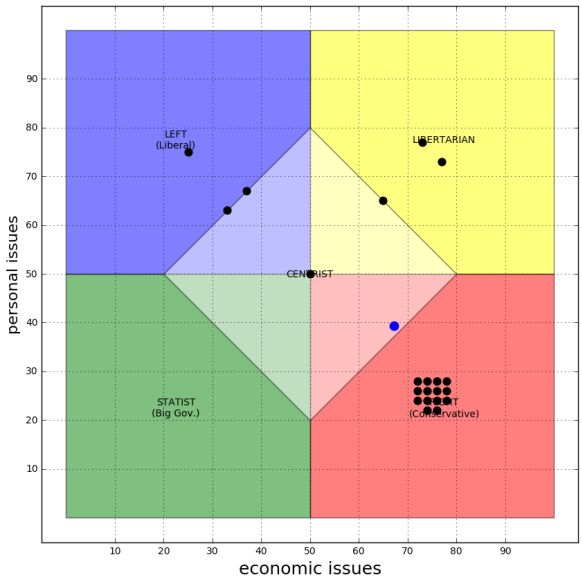


Figure 3.5: Popularly perceived position of the *26ers*. The blue dot shows the average position.

imposing the libertarian definition of freedom [105]. For our part, we tried to minimize the bias in the methodology by making a conscious collective selection of the initial keywords for each query, but there is always room for interpretation of the questions. Despite the alleged bias, we have shown that the quiz can differentiate outlets with opposite points of view in both dimensions and also that the automatic classification is in accordance with the widespread perception of the tendency displayed by the main outlets. This means that either the bias introduced by the appliances in the methodology is not significant or it is representative of the predisposition showed by the population that we are considering in our study.

Stability of the results

In order to find out the stability of the observed bias with respect to changes in the obtained evidence (i.e. the collected tweets), we repeated the scoring steps 20 times. Each time we leave out 5% of the tweets selected at random, while maintaining the original distribution of documents per question. Each time, we measure the average score of the news outlets for which we were able to answer at least one question in the corresponding dimension. In the *economic* issues, we could observe a consistent bias to the left ($M = 40.45,\ 95\%\ CI[36.91,\ 43.99]$). On the other hand, the *personal* dimension, although it is also leaning to one side, is much closer to the center of the spectrum ($M = 46.89,\ 95\%\ CI[43.99,\ 49.79]$). Figure 3.6 shows a similar analysis, but at an individual level in the ***26ers***. The mean for each individual score stays close to its original position, and each newspaper can be located in a relatively small neighborhood

with high confidence, meaning that there are not any drastic changes in the previous classification.

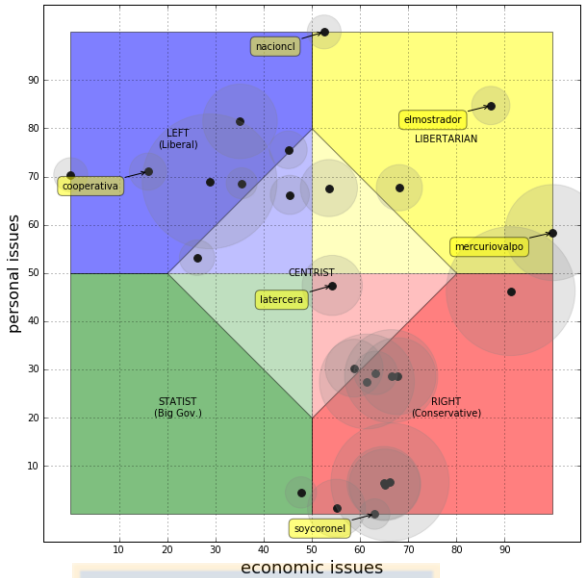


Figure 3.6: Relative position of the ***26ers***. *The score on each dimension is the average over 20 repetitions, leaving out each time a random 5% of the documents. Gray shade around outlets is its 95% confidence interval.*

The relatively low impact of leaving out data in the positioning process indicates that the results are not very sensitive to change and not influenced by only a small number of tweets.

Contextualizing the PolQuiz

We noticed that some of our queries, particularly in the personal issues dimension, returned only a small number of documents (e.i **q2** and **q3**). This is because of lack of interest or too few relevant events related to the corresponding topics during the observed period. We think that a way to counteract this environmental/circumstantial effect is to substitute the respective questions or to increase the number of questions. As a proof of concept, we repeated our analysis using **q0** as a replacement for question **q3** (related to laws concerning sex between consenting adult, see Section 3.2.1). We replaced **q3**, because it was the one with the lowest number of retrieved documents. This substitution increased the number of news outlets with at least one answer. There is now a stronger statistical effect for the personal issues dimension, $t(239) = 3.54, p < .001$. Interestingly, this dimension now leans to the more liberal end of the spectrum $(M = 53.57, SD = 15.63)$ (see Fig 3.7).

In Fig 3.8 we plot the scores of the ***26ers*** in the original quiz (dots) and the adapted quiz (diamonds). Note that the difference in scores between the quiz with **q3** and

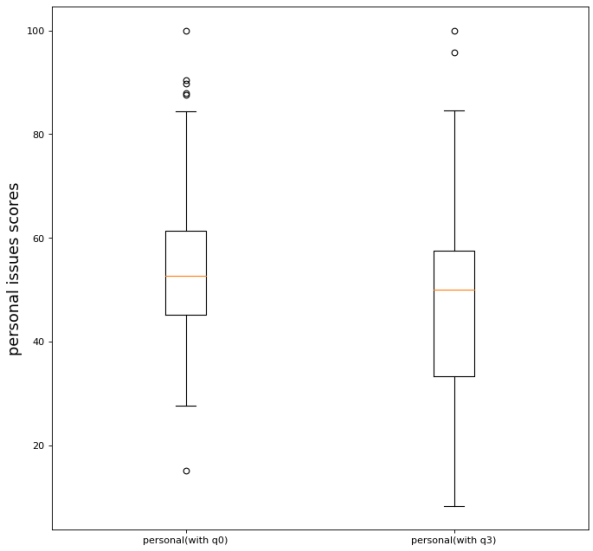


Figure 3.7: Scores before and after replacing **q3** by **q0**. These are the scores of news outlets for which we were able to answer at least one question in the corresponding dimension.

the quiz with **q0** is considerably larger (with a negative difference) for outlets in the right/conservative quadrant. This is expected and validates the model.

3.3.3 The influence of government orientation on the media landscape

As we mentioned earlier, we think that the overall behavior of the media, both in terms of their original/absolute position and the relative balance, is determined, at least in part, by the current government political alignment. This assumption is suggested by the absolute positions as shown in Figure 3.2 and supported by all the models reviewed in Chapter 2 concerned with the political-economy of the mass media. The *Propaganda Model* describes the political elite, and the government in particular, as a very influential actor. According to the PM, either by millionaire advertising contracts, controlling the sources or generating flack against opposed views, the government always tries to control the discourse. The *Media Capture* model [15] also presents the government as an important factor on the news selection and publishing process. Hallin and Mancini [66] explicitly assert that “the state always plays an important role as a source of information and primary definer of news, with enormous influence on the agenda and framing of public issues”.

In this section we investigate if our *PolQuiz* methodology is able to give evidence on the influence of the political ruling class over the mass media behavior. For our analysis we apply the political quiz to the same set of news outlets, but using a different time frame. We collected the tweets in the analogous period of the previous administration. Since the previous government (lead by Sebastian Piñera) had a different political

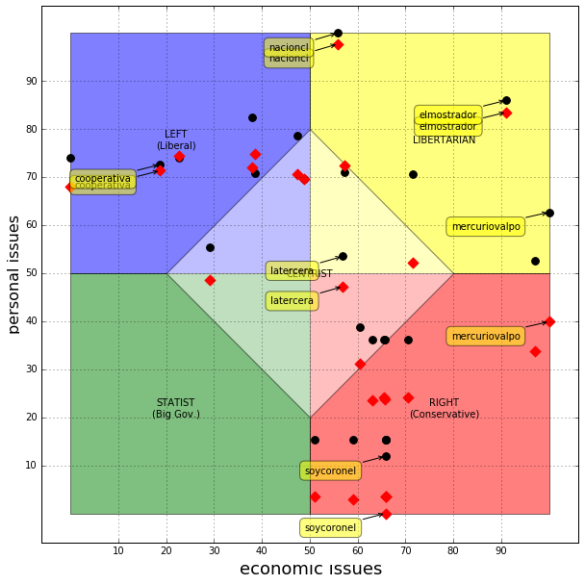


Figure 3.8: Relative position of the ***26ers***. Dots represent the scores with q3. Diamonds represent the scores with q0.

alignment (right-wing conservative), we should be able to see a shifting in the position of the outlets in between the two governments.

Using the Advanced Search from Twitter we collected 832,223 tweets, from 283 news outlets, published in the period from Oct 1st, 2010 to May 31st, 2011. This data set contains only tweet published by these outlets (no retweets are included). After applying the queries, our final dataset contains 16,176 tweets related to the issues addressed in the *PolQuiz*.

In Figure 3.9 we can see the absolute position of the 186 outlets for which we were able to answer at least one question of the *PolQuiz* using the tweets published during the previous government. In comparison with Figure 3.2 it is easy to notice that the entire context of the media as a whole is more to the center of the chart, or seem from the point of view of the current state of the media, it is more to the right and more conservative. This behavior might be due to the main topics being discussed (e.g., tax reforms vs. free high education), but ultimately this indicates that in a way or another the government is playing its role and it is dominating the discourse that prevails in the media.

We also compared the relative positions of the ***26ers*** in both governments (there are two of them that were not yet created in the first time slot). The Kendall's Tau-b correlation for the economic dimension between the two periods is $\tau_b(23) = -0.3461$ ($z = -2.37, p = .0178$). This shows that there is association between the two time periods (i.e. we can reject the null hypothesis of independence), but there are quite a few inversions in the relative order. This is because the newspapers owned by El Mercurio stayed more or less in the same position while most of the others move from being

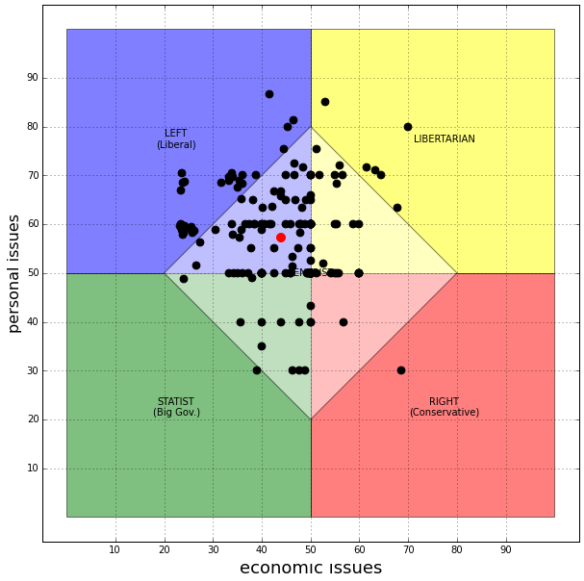


Figure 3.9: Absolute positions of the Chilean media during the Right-wing/conservative government. The chart shows the position of the outlets for which we were able to answer at least one question. The red dot shows the average position.

right to El Mercurio's to being left of their position (notice in Fig 3.10 that *mercuriovalpo* is at the left of the group). We mentioned before that *El Mercurio S.A.P* is the biggest media group in Chile, but now we can say that it is also the most stable in their editorial policy behaviour. This makes intuitively sense: their consolidated control of the market gives them more independence, and it makes them less susceptible to the government influence.

To study the individual behavior of the outlets, we also calculated the relative position of the outlets in one period with respect to their own position in the other period. For this we normalized the scores using the results from both periods combined. This allows us to see, in the overall context, the outlets transition over politically opposite regimes. Figure 3.11 shows how many points each of the ***26ers*** move within this context in the economic issues dimension. Notice that there is a tendency towards the left with the arrival of the left-wing government. Only some of the outlets owned by *El Mercurio S.A.P.* stayed in an approximately similar position or shifted to the right. *El Mercurio de Valparaiso* shows again to be one of the most important representatives for the right. Interestingly, the biggest movements to the left (in the same direction of the new government), come from outlets from which we were not able to find a clear popular perception or declare political leaning. Maybe is this flip-flopping what makes it so difficult for the public.

Figure 3.12 represents a similar individual analysis for the personal issues dimension. As in the results presented earlier, this dimension offers less information. The

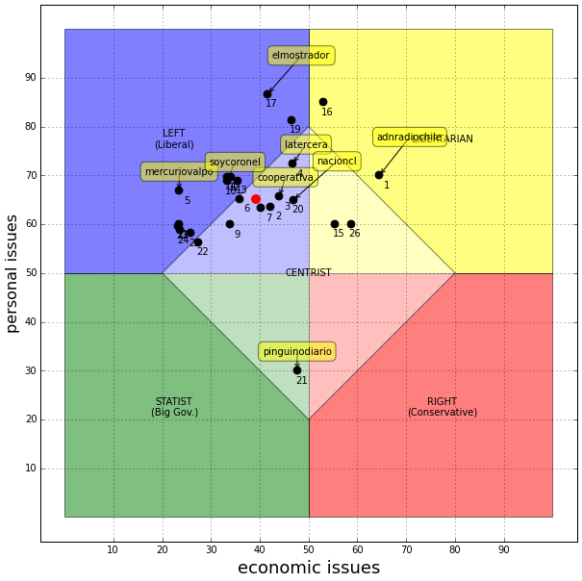


Figure 3.10: Absolute positions of the ***26ers*** during the Right-wing/conservative government. The red dot shows the average position.

direction of the movements is more divided and seems to emphasize the known position of the outlets. Of particular interest is the outlet controlled by the government (*nacioncl*). This newspaper's editorial policy seems to move (in both dimensions) to accommodate the current government. This behavior makes it a biased source of information, but a good point of reference to validate our model. Another point to notice is that, once again, the outlets without a clear perceived bias, consistently show the most significant shift in favor of the ruling side.

This empirical analysis of the behavior of the media over two politically different regimes show that, directly or indirectly, the government does successfully interfere in the news process. The case of Chile makes a good example because being only four years apart makes it harder to attribute these noticeable differences to other factors (such as a very different staff, ownership or editorial policy).

3.3.4 Investigating the nature of bias using rank difference

The *PolQuiz* showed the existence of bias in Chilean media. In this section, we investigate the nature of this bias in terms of vocabulary used and entities mentioned in the different newspapers' tweets (see Section 3.2.3). We focused on the ***26ers*** and the topic of abortion. We selected the topic of abortion, as it is one of the most polarizing issues in our dataset. Nevertheless, this is used only as an illustration: to fully understand the nature of the bias and the media landscape, the decision makers or interested parties should conduct a similar analysis on each of the questions.

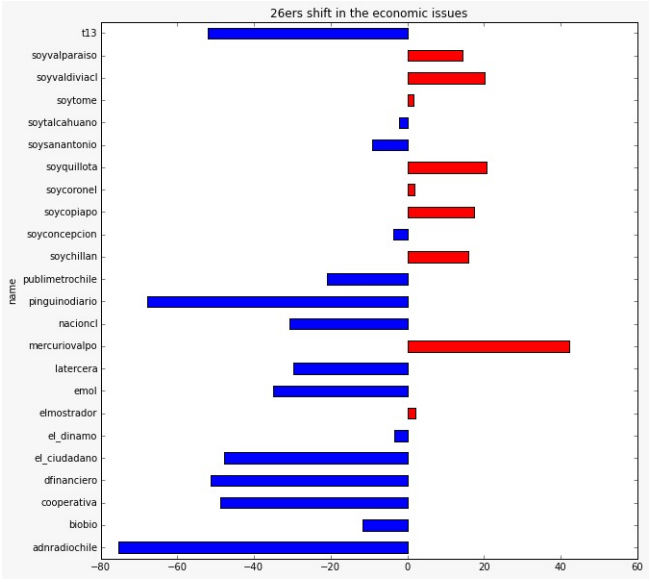


Figure 3.11: Shift in the position of the ***26ers*** for the economic dimension. The chart shows the relative shift in points from the conservative to the liberal government. A shift > 0 (red bars) means more to the right. A shift < 0 (blue bars) means more to the left left

Topic bias based on named entities

We used the Stanford's NE recognizer system [47] to extract the entities mentioned in the tweets related to the abortion issue. We compare the extracted entities against a list of politicians, public personalities and activist groups. For the list of politicians and their position in the abortion issue, we use the vote sessions in the house of representatives [25] and in the senate [139]. We manually labeled another 53 personalities and groups according to comments and events reported in the local news. The complete list L_E has 199 labeled entities. We labeled with -1 the politicians who voted against the abortion bill, and the public figures that were openly against the issue. Equivalently, we use $+1$ for politicians and personalities in favor of the subject. We assign a 0 to the entities not included in our list. We will refer to these labels as the leaning of the entities (e.g. $leaning(entity)$)

After applying the rank difference method to the NE mention counts, we calculated a score for each outlet in function of the $\tau(entity)$ and the leaning of $entity$ in the issue (for every $entity$ mentioned more than once in the news). This final score of the outlet o_i is found using the equation 3.2.

$$score(o_i) = \sum_{e \in L_E} \big(\tau(e) * leaning(e)\big) / size\_of(L_E)) \tag{3.2}$$

A low value in this score indicates that this outlet tends to mention with relatively high frequency entities with a conservative leaning and/or it tends to ignore those with

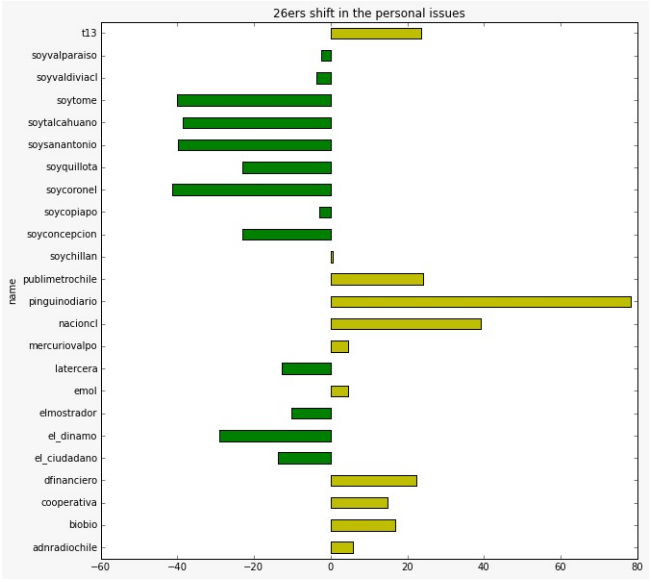


Figure 3.12: Shift in the position of the ***26ers*** for the personal dimension. The chart shows the relative shift in points from the conservative to the liberal government. A shift > 0 (yellow bars) means more liberal. A shift < 0 (green bars) means more conservative.

a more liberal view.

As expected, outlets tagged as independent, libertarian and classical-liberal have higher scores (top 10 in the ***26ers***). Interestingly, within the top 10 we also find the outlets tagged as *International*, `publimetrochile` and `adnradiochile`, which means that they behave similarly to liberal outlets under the left-liberal government in office in 2016. According to our scores, all these top-scored outlets have comparably more mentions of entities with a liberal leaning than the rest of the outlets. To our surprise, the lower values (bottom 5 in the ***26ers***) are occupied by the outlets linked to parties in the ruling coalition (Christian democracy and Left-Liberal(`nacioncl`)). Apparently these outlets focus their tweets in negative reports of the opposition. For example, when we look at the rank-difference results for `nacioncl`, within the top-20 entities, only two refer to entities with a liberal leaning ('President Michelle Bachelet' and 'Government'). To investigate more on this, we run a sentiment analysis on the most used bigrams. The results are presented in the next section.

Topic bias based on bi-grams

We again apply the rank difference method, this time using the bi-gram counts in the tweets relevant to the subject of abortion. Following the same strategy as before, we calculated a score for each outlet in function of the $\tau(bigram)$ and the sentiment calculated for $bigram$ (for every $bigram$ mentioned more than once in the news). For determining the sentiment of words and bi-grams we use the Spanish lexicon from [152].

This lexicon consists of a set of norms for valence and arousal for an extensive set of Spanish words. We found this to be one of the largest dictionaries in this language, and it includes items from a variety of frequencies, semantic categories, and parts of speech, including conjugated verbs. We weighted each word with its mean valence (we assigned the neutral value 5 for words not present in the dictionary). The weight of the bi-grams is the average of the weight of their composing words. To calculate $\tau(bigram)$ we use a formula equivalent to that shown in Equation 3.2. Accordingly, we give a similar interpretation to these scores. That is, a high value indicates that this outlet tends to convey mostly positive sentiments with the bi-grams used with relatively high frequency and/or avoid using negative sentiments when referring to the issue of abortion. For example, *elmostrador*, with the highest score, has as a frequently use bi-gram "proyecto aprobado" (tr. "project approved" - referring to the bill). This bi-gram is classified as positive by the sentiment analyzer, so it will add to the score. On the other hand, this same outlet has "injusticia gobierno" (tr. "government injustice") as a totally ignored bi-gram. Since the bi-gram is assigned a negative sentiment and the rank-difference is also negative, the bi-gram will also add to the score of the outlet, pushing it to the liberal side. Following the same reasoning, an outlet with a very low score can be understood as an outlet that uses predominantly negative words with relatively high frequency.

When we analyze the scores of the ***26ers***, we notice that *nacioncl* (controlled by the government) has the lowest score. This, together with the previous NE analysis, confirms the theory that this outlet focuses in tweeting negative reports of the opposition, at least for the abortion issue. Most of the others outlets show the expected behavior, with conservative in the lower half of the ranking (i.e. lower scores) and liberals in the higher positions.

The question that follows is if the bias that we are seeing with the *PolQuiz* and describing with the Rank Difference is perceived in the same way through the popular wisdom. We help answer this question in the next section.

3.3.5 Survey results

For the survey described in Section 3.2.4, we collected 372 answers from 54 unique Chilean users on how they perceive the bias on the topic of abortion in the different Chilean newspapers. Since this was an open and anonymous online survey, we do not have any demographic data on the users, but the IP addresses indicate we have a good representation of different regions of the country. We received between 11 and 19 answers for each of the ***26ers*** (M: 14.31, SD: 2.07). We carried out 10 Fleiss' kappa measurements; each time we selected 10 ratings at random per outlet (subject). This shows a fair agreement in the answers (M: 0.2253, SD: 0.0167). In Table 3.6 we show the ***26ers*** and their corresponding "Perceived bias" (see Section 3.2.4). The political alignment information shown in the table is again our ground-truth.

Results show that there is a perceivable difference in the language used by the

outlets in both sides of the spectrum. Note that, based on the rank difference of bi-grams, the users were able to collectively classify the outlets with over 90% precision (We are not taking into account those for which we could not find a political alignment or those that belong to international groups). Our positioning of these outlets in the adapted *PolQuiz* has also a good agreement with the direction of the Perceived bias (80%).

To evaluate the relative positions of the outlets in our *PolQuiz*, we calculated the number of inversions with respect to the ranking of the outlets in the perceived bias. The Kendall's Tau-b coefficient between the two rankings is $\tau_b(21) = 0.4203$ ($z = 2.66, p < .01$). Even though the popular perception resulting from the survey can not be seen as ground-truth for the relative positioning of the outlets, it is important to notice that our results show a good correlation with the intuition of the public. As a future work, we aim to add some other content features (e.g., leaning of the named entities) to the polarity classification of the tweets as these may help to refine the relative positioning found by our model.

To summarize, we have shown that reported political alignment is highly correlated with the *PolQuiz* results as well as with the bias, as perceived by the general audience. This implies that existing bias has a noticeable influence on how controversial issues such as abortion are reported in the media.

3.4 Conclusions

The results presented in this chapter indicate that the political orientation of the media in Chile is in line with and follows the political orientation of the government. Even though relative differences in bias or orientation between individual news outlets can be observed, public awareness of the bias of the media landscape as a whole appears to be limited: our own perception of the bias seems to be adjusted and limited by the political space defined by the news that we receive, which in turn is largely defined by governmental politics. Our model is able to discover this relative political context, which regulates the perceived bias of the media. Building upon the *PolQuiz* results, we investigated the nature of political bias and found that is exists in the chosen vocabulary and the entities covered by the newspaper. We also conducted a survey, the results of which confirm that political bias in newspapers has an impact on how controversial topics are covered and that the general audience *does* notice this bias. Our methodology does not make too many assumptions about the underlying system. The way it is designed could be applied to any Western culture. Our system can deal with any number of outlets, can compare relative quantitative positions, can show empirical evidence of consistent bias, and can partially explain the source of these tendencies.

Our methodology also contributes as empirical evidence of the media capture in modern "democratic" societies. We believe it is important to be aware of shifts, alignments and discrepancies in bias and political orientation within the government, the population and the media, as misconceptions regarding real or perceived bias may

have unexpected or negative effects.

Regardless of whether the filters exist or not, for individuals as well as for society as a whole it is important to recognize and understand media bias that is shaped by underlying general political or socioeconomic orientations. As we have shown here, these general tendencies have a clear and noticeable effect on the way concrete topics are covered and commented upon, and therefore should be investigated and published.

All the elements presented in this chapter support the predicted general behavior of the media under the Propaganda Model and, in particular, the political-elite influence in the news cycle analyzed in the first filter. As anticipated by Herman and Chomsky, we found that subjects concerning economic issues (more sensitive topics for the elite) received more coverage but with a more significant bias, which shows a stronger consent. Meanwhile, subjects in the personal dimension reflect a weaker and more heterogeneous coverage. However, the first filter suggests that media bias is primarily caused by the concentration of ownership and the influence that these owners exert on their news outlets. Our next chapter investigates whether or not editorial policies of outlets can be used to predict their ownership.

Table 3.4: Perceived bias of the *26ers* extracted from Wikipedia

Id	Name	Owner	Political alignment
1	adnradiochile	Grupo Prisa	International
2	biobio	Bo-Bo Comunicaciones	Independent
3	cooperativa	Co. Chilena de Comunicaciones	Christian democracy
4	latercera	Copesa	Classical liberalism
5	mercuriovalpo	El Mercurio	Right-wing, conservative
6	publimetrochile	Grupo metro	International
7	emol	El Mercurio	Right-wing, conservative
8	soyarauco	El Mercurio	Right-wing, conservative
9	soyconcepcion	El Mercurio	Right-wing, conservative
10	soycoronel	El Mercurio	Right-wing, conservative
11	soyquillota	El Mercurio	Right-wing, conservative
12	soysanantonio	El Mercurio	Right-wing, conservative
13	soytalcahuano	El Mercurio	Right-wing, conservative
14	soytome	El Mercurio	Right-wing, conservative
15	dfinanciero	Grupo Claro	Right-wing, conservative
16	el_ciudadano	Red de medios de los pueblos	Libertarian
17	elmostrador	La Plaza	Libertarian
18	tele13_radio	Grupo Luksic & PUC	—
19	el_dinamo	Ediciones Giro Pais	Christian democracy
20	nacioncl	Estado de Chile	Left, Liberal
21	pinguinodiario	Patagnica Publicaciones	—
22	soychillan	El Mercurio	Right-wing, conservative
23	soycopiapo	El Mercurio	Right-wing, conservative
24	soyvaldiviacl	El Mercurio	Right-wing, conservative
25	soyvalparaiso	El Mercurio	Right-wing, conservative
26	t13	Grupo Luksic & PUC	—

The list is sorted by the perceived bias. Outlets with an unclear Political Alignment (shadowed rows in the table) were left out of the analysis.

Table 3.5: Perceived bias of the *26ers* extracted from Wikipedia

Id	Name	Owner	Political alignment	Perceived bias	Personal issues
21	pinguinodiario	Patagnica Publicaciones	—	-66.67	39.18
24	soyvaldiviacl	El Mercurio	Right, conservative	-66.67	-50.49
22	soychillan	El Mercurio	Right, conservative	-57.14	-50.55
25	soyvalparaiso	El Mercurio	Right, conservative	-43.75	-51.81
8	soyarauco	El Mercurio	Right, conservative	-42.86	-51.27
12	soysanantonio	El Mercurio	Right, conservative	-30.77	-92.98
13	soytalcahuano	El Mercurio	Right, conservative	-30.77	-92.98
18	tele13_radio	Grupo Luksic & PUC	—	-28.57	52.42
9	soyconcepcion	El Mercurio	Right, conservative	-25.00	-94.09
14	soytome	El Mercurio	Right, conservative	-25.00	-92.92
7	emol	El Mercurio	Right, conservative	-25.00	-0.59
10	soycoronel	El Mercurio	Right, conservative	-23.53	-100
11	soyquillota	El Mercurio	Right, conservative	-18.18	-92.92
15	dfinanciero	Grupo Claro	Right, conservative	0.00	42.57
5	mercuriovalpo	El Mercurio	Right, conservative	21.43	-51.81
2	biobio	Bo-Bo Comunicaciones	Independent	23.53	6.91
6	publimetrochile	Grupo metro	International	25.00	47.50
17	elmostrador	La Plaza	Libertarian	26.32	70.95
19	el_dinamo	Ediciones Giro Pais	Christian democracy	29.41	-30.79
4	latercera	Copesa	Classical liberalism	33.33	-3.38
1	adnradiochile	Grupo Prisa	International	37.50	52.98
16	el_ciudadano	Red de medios de los pueblos	Libertarian	37.50	44.54
23	soycopiapo	El Mercurio	Right, conservative	38.46	-36.21
3	cooperativa	Co. Chilena de Comunicaciones	Christian democracy	57.14	46.04
26	t13	Grupo Luksic & PUC	—	57.14	48.45
20	nacioncl	Estado de Chile	Left, Liberal	63.64	100

The list is sorted by the perceived bias. Outlets with an unclear Political Alignment (shadowed rows in the table) were left out of the analysis.

Table 3.6: Results from popular survey for the ***26ers***.

Chapter 4

Power Structure in Chilean News Media

Chomsky commented on the role of a free and diverse press: “The smart way to keep people passive and obedient is to strictly limit the spectrum of acceptable opinion, but allow very lively debate within that spectrum.” [30] This is in fact the position that many advanced democracies find themselves in as the diversity of news coverage seems to shrink, whereas news coverage itself seems to continuously expand in a non-stop news cycle. Lack of diversity of viewpoints, topics, and representation of communities has been attributed to this relentless process of consolidation [177].

The market-driven consolidation of the news media industry may lead to concentration of ownership. According to the Propaganda Model [72] (PM), this concentration of ownership may have direct and indirect effects on editorial policies. For example, the editorial board of a newspaper owned by a group that also invests in agriculture may perceive a pressure to report more favorably about agricultural initiatives. Unlike other factors such as the (frequently explicitly publicized) political and historical mission of the outlet and its readership, ownership bias may thus exert a more insidious effect on editorial policies that is difficult to operationalize and quantify. Nevertheless, it may have a significant effect on the degree to which news consumers perceive the world and their ability to gather objective and effective information.

The strong presence of most news outlets in online social media platforms give the possibility of real-time distribution of news content [101]. These online environments provide an opportunity to test hypotheses with respect to the drivers affecting news diversity, such as consolidation, coverage, ownership, and network homophily.

As in previous chapters we use Twitter as our prime source for news content. The Twitter platform is particularly interesting as foundation for the study of news diversity and coverage since it is designed to constitute a large-scale social network. The ensemble of users-following-users establishes a social network where tweets travel along the edges of the network. This renders the social media platform an ideal laboratory to apply the toolkit of network science to the investigation of news diversity and coverage from a top-down (user to user to news outlet) as well as a bottom-up perspective (news outlets to their followers).

In this chapter we research the most important predictions of the first filter of the PM. This is, the influence that ownership relations have on news media content and coverage. Here we quantify the strength of the relation between news media ownership and news media content diversity in Twitter. We analyze the user accounts of news media outlets to study how their content evolves and overlaps, and whether or not these observations are linked to their known ownership structure.

As in the rest of our thesis we focus on Chilean news outlets. The Chilean media landscape is well documented due to the availability of detailed, publicly available data with respect to its ownership structure, compiled by *Poderopedia* [127], a journalist NGO that aims to understand power relationships between people, companies, and organizations.

We use the latter information to trace the existing ownership structure of Chilean media outlets which we then compare to the structural properties of their Twitter coverage and content, in particular with respect to the similarity of the content they publish in social media. To this end, we define a content similarity metric and evaluate its relation to ownership.

Prior work has focused on studying story selection similarity within different news outlets [49]. However, we extend this research by searching for indications of deeper interconnections in the *mediasphere* at the intra-country level. An *et al.* [6] modeled the outline of digital media on Twitter, analyzing media similarity based on the degree of overlap between their respective follower communities. They reported a strong tendency for members of the communities to read news from multiple sources, mostly on similar topics. Park *et al.* [121] proposed a system to identify and track events, in order to present different points of view of the same affair to readers to counteract opinion bias in news. Saez-Trumpe *et al.* [137] define a methodology to identify "selection" or "gatekeeping"-bias which consist of editorial decisions to publish certain stories and not others. They study these biases with respect to the prominence of the stories and the geographical location of the outlet. Since their work uses a data set of media from different countries, not surprisingly they find that geography might influence the selection of the stories. Our work complements prior research by searching for potential causal pathways to explain the homophilic relations between groups of news outlets. This might help to identify and characterize potential influence of the owners over the editorial policies of their respective news outlets. This should further consolidate the evidence toward the applicability of the first filter of the PM.

4.1 Methodology

Our goal is to analyze whether ownership and content are correlated in the domain of digital media news outlets. We approach this problem by studying the similarity networks and clusters that emerge from the content published on Twitter by news outlets in Chile. We contrast groups of similar news accounts with their ownership in the real-world to see if they are related according our similarity metric.

In particular, we study the similarity between pairs of news accounts from the perspective of *minhash-based topics*. In [10], we complement our analysis by comparing the results of *minhash-based topics* against alternative similarity metrics such as *vocabulary* and *keyword-based topics*. We aim to determine if there exist consistent similarity-based communities among news media outlets and if this same consistency arises in relation to ownership.

In order to achieve this, we perform independent static analyses of news media outlets for two years, 2015 and 2016. For each year, we study the communities of news outlets that are produced by using community detection over the similarity graph built for our similarity metric. In addition, we identify clusters of similar outlets with the purpose of checking consistency of the resulting similarity groups. Below, we detail our similarity metric, and community and clustering algorithms. Our data analysis started from the following:

- **Chilean News Twitter (*ds15*):** all tweets published by 84 prominent Chilean news media outlets from October 30th, 2014 through May 20th, 2015 (including retweets). This data set contains $714,973$ tweets and was created by Maldonado *et al.* [94] for their study that characterized Chilean news events.
- **Chilean News Twitter (*ds16*):** Our manually curated list of news outlets in Chile for year 2016. As we mention in previous chapters, this list derived from the Wikipedia page listing Chilean news media [173] and the independent journalistic website Poderopedia.

We joined both sets and kept all of those that had an active Twitter account; then, we downloaded all of the tweets generated by those accounts from October 25, 2015 to January 25, 2016. Overall, the *ds16* collection contained 365 news accounts and 756,864 tweets (also including retweets). Both data sets include tweet metadata, such as timestamp and user identifiers.

Standard text normalization and cleaning techniques were used to convert Tweet content in both data sets to lower case and remove stop-words, URLs, and punctuation. In addition, news outlets that posted less than one tweet per day on average were removed, leaving 79 news outlets for *ds15* and 341 in *ds16*.

As for ownership information, we manually mapped Poderopedia's influence database [127] to our lists of news media accounts on Twitter. As for grouping news media outlets according to their owners, we simply consider two outlets to belong to the same group if and only if they're owned by the same entity. There are at least two possible issues with this. First, some news media outlets are owned by multiple entities: in this case, we selected the major partner. On the other hand, there also exist ownership relationships *between* owners. In this case, we selected the entity that subsumes all others as the owner. As a result we obtained the first complete database of newspaper ownership information in Chile.

Our datasets include news outlets that belong to the two biggest news media groups in Chile: the *El Mercurio* group and the *Copesa* media conglomerate, which form what has been called in the past a newspaper duopoly[26]. We also have representatives of a group of digital newspapers, the *Mi Voz* network. Other owners with smaller number of outlets are also included, as well as a group of *unknown-to-us ownership*. We note that we are interested not only in news outlets that share owners, but also those that behave as if they did.

4.1.1 Topics detection

We use text-content based clustering of the publications of news outlets publications to identify “stories” that relate to a common event or topic [137, 49].

Minwise Hashing was originally proposed by Broder [20] for finding similar documents in the AltaVista search engine. Later, Broder et al. [22] show that to compute the similarity of two documents it suffices to keep a small number of signatures (summaries or sketches) for the sets representing each document. Finally, Broder et al. [22] presented an algorithm technique called *min-wise hashing*. Minwise hashing approaches have been successfully applied to a wide range of applications including compressing Web graphs and social networks [24, 73, 29], tracking Web spam [161], genome assembly [14].

More relevant to our investigation, minwise hashing based on n-grams has been used to obtain clusters of similar documents in the Twitter context [138]. This technique has also been compared against the cosine similarity measure [144], which is commonly found in literature to approach this or similar tasks [49, 137]. The study in [144] shows that minhash outperforms cosine similarity in most practical cases.

Therefore, we apply Minwise Hashing to our collection of news tweets. The obtained clusters are considered our news topics.

For each text we extract groups of n consecutive words (n-shingles or n-grams) [96]. Hence, each tweet is represented by the set of n-shingles that correspond to its text. Two documents are said to be similar if they have several shingles in common. To group similar tweets we search for those who share a subset of n-shingles. This way, we are not only looking for tweets with a similar set of words, but similar phrases.

The shingles (SH) model can be applied at character and word levels, but it has been shown that using long n-shingles based on characters to simulate words leads to an unacceptably high number of false positives. In contrast, using n-shingles (also called w-shingles) based on words has been used successfully in small and large documents. For instance, using a n of 2 or 3 in email documents (short documents) and $n = 4$ in large documents such as web collections [20, 21, 97] news articles or blog posts [130] provided the best results.

Since tweet are short by definition (maximum length allowed is 140 characters), we set $n = 3$. If, on top of the short length of the text, we also remove stop-words, we find that, even with a small value of n, the probability of occurrence of each n-shingle is small. We set $n = 3$ to obtain a fine grain classifier identifying specific stories, rather than broader topics for which a smaller value of n may have been chosen. Since the number of distinct n-shingles can be very large, it is possible to apply a hash function to every n-shingle. Other strategies can be employed to build summaries or sketches to reduce the space without deteriorating the effectiveness of our matching algorithms [20, 21].

In this work, we represent tweets using 3-shingles and then apply 4 keys minwise hashing over each tweet. We define clusters based on similar sets of minwise hashings. As mentioned before, we consider these clusters our set of discovered topics.

We look only for topics corresponding to tweets from *multiple* news outlets (henceforth *multi-outlet-clusters*). With these topics we can analyze how many outlets cover the same event and/or how many time two outlets coincide in their selection of stories.

4.1.2 Similarity metric between news outlets

We are interested in finding how related news outlets are using their content. In other words, we study their vocabularies and how the stories they select may indicate a connection between their editorial policy.

Similarity between two news outlets is defined by the co-occurrence of two news outlets with respect to a same topic. In particular, the value of the similarity between outlet A and B is the conditional probability $Pr(A|B)$ of the occurrence of A in a cluster given that B occurs in that same cluster. This similarity measure is directional, expressing how likely it is that a story tweeted by B is also tweeted by A [6]. In order to define a symmetric similarity measure we further specify the similarity between A and B as $sim(A, B) = max(Pr(A|B), Pr(B|A))$.

In order to identify similar news outlets, we create different *similarity graphs* based on the aforementioned similarity metric and each of the datasets. Formally, we define a generic similarity graph $G = (V, S)$ for the set of news outlets $V = \{v_1, v_2, \ldots, v_n\}$ and similarity measure $S : (v_i, v_j) \rightarrow \mathbb{R}^+$, as a graph where each pair of outlets v_i and v_j are connected by an edge of weight $S : (v_i, v_j)$. This yields a complete, weighted, and undirected graph.

Using each similarity graph, we apply graph partitioning techniques to find groups of similar news outlets. For each similarity graphs we used a hierarchical, agglomerative community discovery algorithm [115], and the normalized cut technique [143]. This methodology has been proved to be successful in similar problems [92].

4.2 Results

Figure 4.1 shows the communities (represented as boxes) of news outlets obtained from Topics collected in *ds15* (left column) and *ds16* (right column). The curves that connect both columns of boxes represent the number of outlets (or the proportion of outlets) shared by the communities at the two ends of the curve. For example, the first community in the top left from *ds15* has 16 (16/18) outlets in common with the first community in the top right from *ds16* (which is completely cover by these 16 outlets) and shares only 1 (1/18) outlet with the second and 1 (1/18) outlet with the third community in *ds16*. The figure shows that results do not vary significantly between the two datasets. In the same vein, the graph partitioning algorithms (community detection and clustering) show that the resulting communities in both cases are very similar (see Fig 4.2). This congruence supports the notion that the communities found are significant, denoting a real structure in the data .

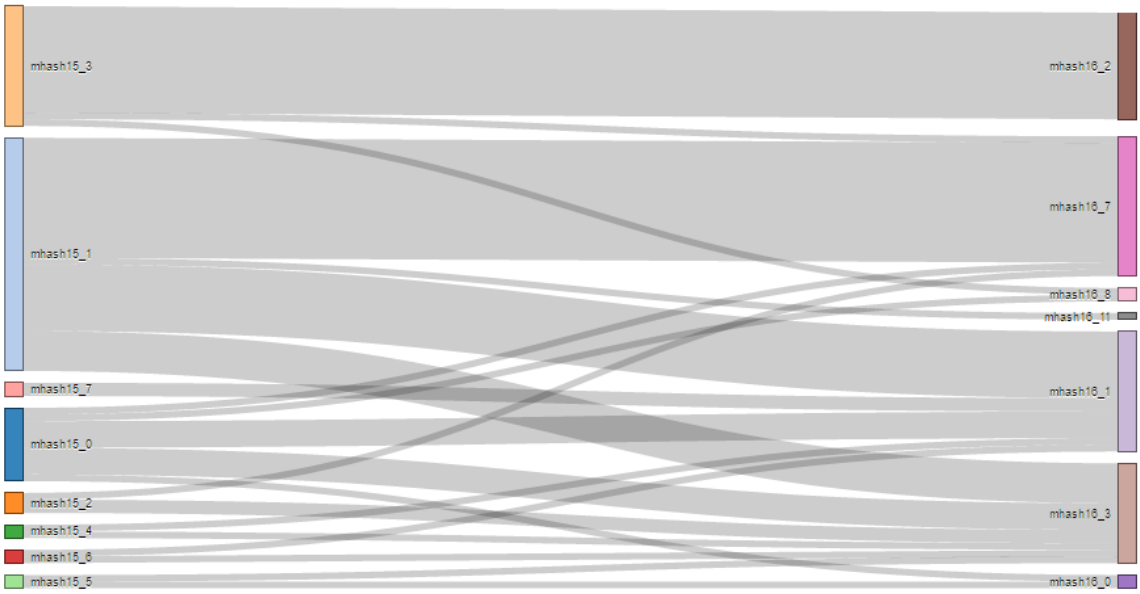


Figure 4.1: *ds15* vs. *ds16* communities on *Topic (minhash-based)* similarity.

Table 4.1 summarizes several metrics for community discovery over both data sets. For comparison, we also show equivalent results for other similarity metrics (based both in topic detection and vocabulary) applied to the same datasets [10]. The *Topic* similarity is obtained by mining frequent term-sets from the tweets posted each day and then joining these sets by word co-occurrence (within the same day). Then, a daily vector representation is computed for each outlet based on the day's topics; daily similarities between pairs of outlets are obtained as the cosine similarities of pairs of these vectors. Finally, the overall topic similarity between two outlets is defined as their average daily similarity. For the *Vocabulary* similarity each news media outlet is represented as a single document composed of all of the tweets posted by its news account during the time of our data collection. Each document is converted to its vector-space representation using a *tf-idf* weighting scheme[98]. Similarity is then computed as the cosine similarity between two vectors[147].

In Table 4.1, the column *Outlets* is the initial number of outlets in the similarity graph (see Section 4.1), while *Grouped* is the number of outlets that were included in at least one community. Column *Comm.* is the number of communities found by the algorithm, while *Mod.* and *Cond.* show, respectively, the modularity and conductance of the subgraphs formed by the *Grouped* outlets within returned communities (see Section 4.1).

The first thing to notice is that *Topic* similarity creates the lowest number of communities. Also, this similarity for the Dataset *ds15* includes almost all outlets, but for *ds16* it only grouped about a third of the total dataset. We think this is because this similarity is a coarse-grained classification that only captures the strongest signals. If we focus on *ds15*, the *topic* similarity has the highest modularity, which means it creates well defined communities. However, we have to take into account that this dataset only

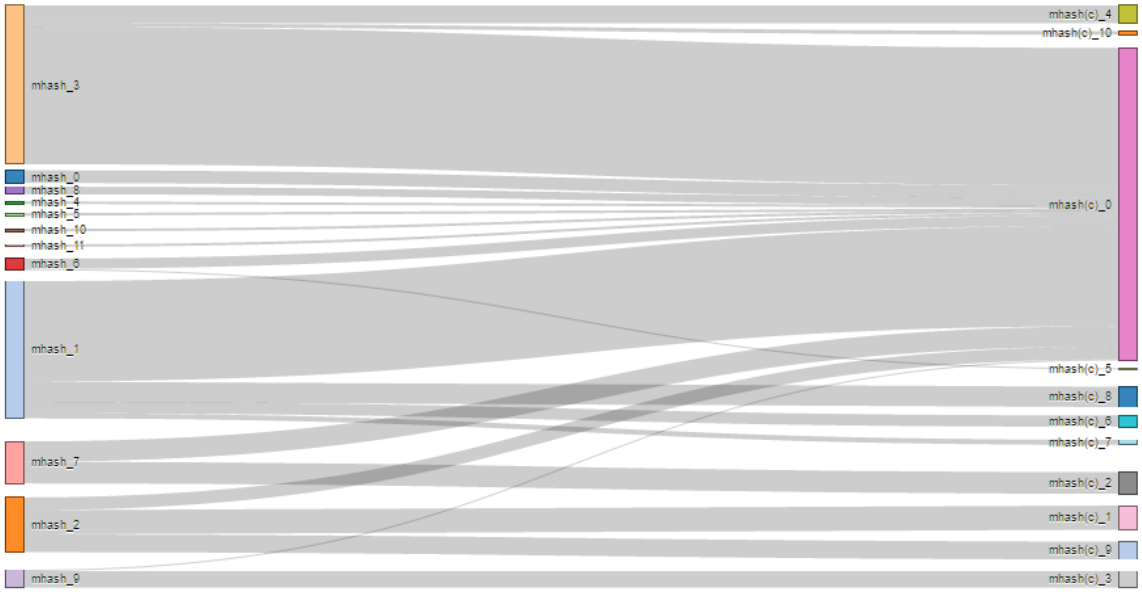


Figure 4.2: Communities vs. Clusters *Topic (minhash-based)* similarity on *ds16*.

Similarity	**Outlets**		**Grouped**		**Comm.**		**Mod.**		**Cond.**	
	ds15	*ds16*	*ds15*	*ds16*	*ds15*	*ds16*	*ds15*	*ds16*	*ds15*	*ds16*
Vocabulary	79	341	52	262	7	14	0.38	0.38	0.02	0.35
Topics	79	341	50	133	4	7	0.60	0.58	0.11	0.26
MinHash	75	365	50	355	6	11	0.40	0.74	0.01	0.04

Grouped outlets (Grouped) correspond to those belonging to a discovered community. Modularity (Mod.) and conductance (Cond.) are calculated with respect to this subgraph.

Table 4.1: Internal metrics for community structures derived from each explored similarity measure for the *ds15* and *ds16* datasets.

contains 84 outlets that comprise most of the largest, most famous newspapers of the country. When we look at the *ds16* dataset we find a more diverse set of outlets (in size and content). In *ds16*, *Topic* similarity shows similar performance if un-grouped outlets are excluded. In turn, *MinHash* seems to be more sensitive to weaker signals, creating a more fine-grained classification. We can see this in the high modularity achieved with *ds16* in spite of having included most outlets. On the other hand, *Vocabulary* similarity has the lowest performance in both datasets, which gives us the intuition that there are no particularly strong differences in vocabulary between the analyzed outlets.

4.2.1 Analysis of the communities

Using the *minhash* technique over the tweets in *ds16*, we identified 100,774 topics that contain 438,353 tweets. In the case of *ds15*, we identify 83,582 topics containing a total

of 254,650 tweets. We looked for topics that had tweets from multiple news outlets. In *ds16*, out of all the topics, 31,423 contained tweets from more than one news outlet (31.2%) and in *ds15* $17,211$ (20.6%). Using these topics, we used the co-occurrences for each pair of news outlets to calculate their similarity (see Section 4.1). We found that all news outlets co-occur at least once with some other news source, for both *ds16* and *ds15*.

Results for Minhash-based similarity have features like those seen in the keyword-based communities. These communities are easily identifiable even by visual inspection (see Fig 4.3). We observe two big communities with many different media outlets (with IDs 1 and 3), some small ones (with IDs of 4, 5, 10 and 11) which under manual inspection of the outlets name seem to have specific scopes (*e.g.*, the Linares Province, aquaculture, the Chiloe Province or the Maipú commune).

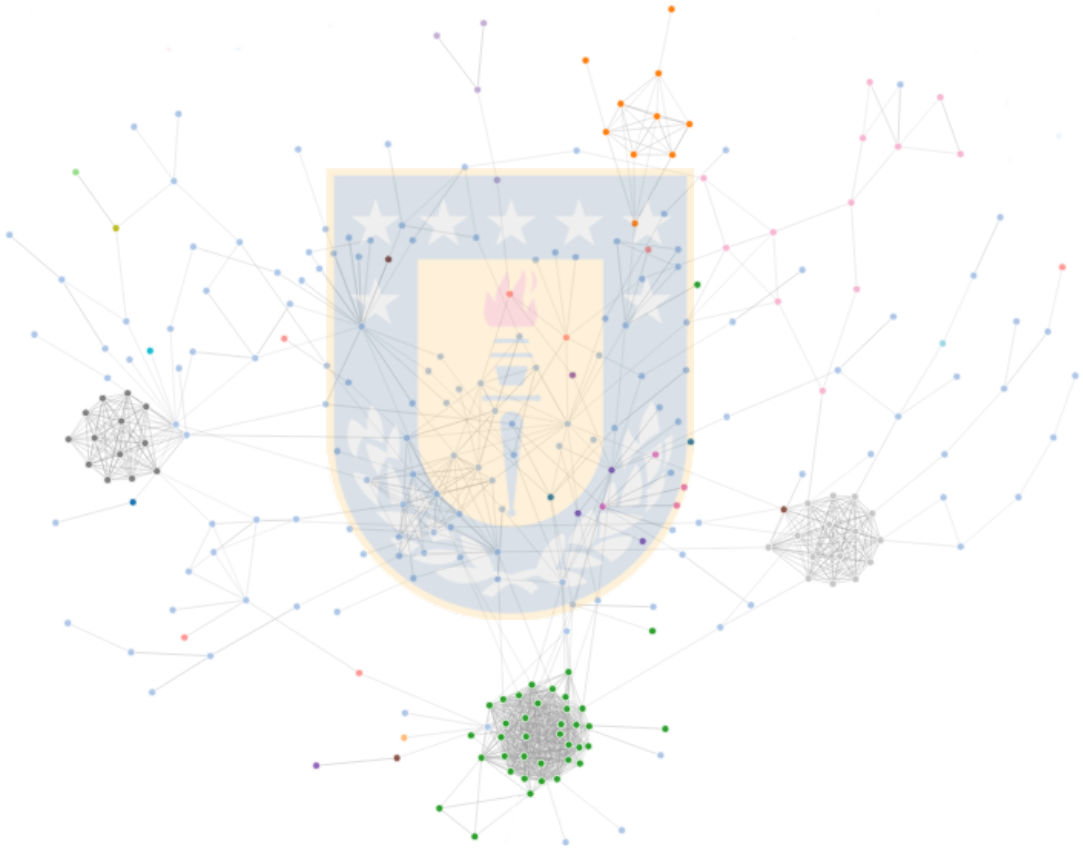


Figure 4.3: Similarity graph, using *Topic (MinHash-based)* similarity on *ds16*. Only representing edges with weight over (mean+2std). We assigned different colors to the biggest owners.

Ownership features help explain the remaining communities as shown in Table 4.2. Communities 2, 6, 7 and 9 have entities that own a big part of them. The small community with an ID of 11 does not have a clear meaning or unifying theme. Notice that

in this case the community with ID of 3 seems to represent the outlets that could not be grouped into any other commune. This community has the biggest number of outlets and the main owner has less than 10% of them.

ID	Size	Main owner(s)	Owner(s)% [#]	Unk. owner % [#]
0	10	El Mercurio	20.00 [2]	30.00 [3]
1	109	Red de Diarios Comunales	12.84 [14]	21.10 [23]
2	43	El Mercurio	83.72 [36]	11.63 [5]
3	130	Copesa	4.62 [6]	30.00 [39]
4	2	Radio Ancoa de Linares	50.00 [1]	0.00 [0]
		Comunicaciones del Sur	50.00 [1]	
5	2	Editec	50.00 [1]	0.00 [0]
		Sociedad Medios Comunicaciones	50.00 [1]	
6	9	Betazeta Networks	100.0 [9]	0.00 [0]
7	36	Asesorias e Inversiones Comunidades Ciudadanas	41.67 [15]	30.56 [10]
8	6	El Mercurio	16.67 [1]	33.33 [2]
		Copesa	16.67 [1]	
		Troya Comunicaciones	16.67 [1]	
		Servicios de Radio Difusion Pedro Felidor Roa Barrientos	16.67 [1]	
9	14	Grupo Diarios en Red	92.86 [13]	0.00 [0]
10	2	Mono Manco	50.00 [1]	0.00 [0]
		Camilo Montalban Araneda	50.00 [1]	
11	2	Sociedad Radiodifusora Primordial FM	50.00 [1]	50.00 [1]

The community with an ID of 0 corresponds to un-grouped media outlets. Entities owning over 10% of the outlets in a community are listed next to it.

Table 4.2: Ownership properties for Minhash-based communities for the *ds16* dataset.

Even though the communities in *Minhash-based* similarity graph are partially explained by ownership, the correlation is not as strong as in the clustering obtained from the normalized cut algorithm. Unlike with the other similarities, for the minhash-based similarity graph the normalized cut clustering algorithm improves results over community detection (see Table 4.3).

If we assume that the biggest cluster (with ID 0) is the one containing the outlets that do not fit in any other group (equivalent to the un-grouped outlets in the community detection), then we get clusters that are very similar to the communities that are obtained for *topic (keyword-based)* similarity.

On one hand, the clusters leave out a bigger number of outlets than the community structure. This reduces the number of clustered outlets to an amount similar to that found with *topic (keyword-based)* similarity (see Table 4.1). On the other hand, it finds a classification with a better owner separation. As we can see in Table 4.3, there are two relatively small clusters (with ID 8 and 10). Beside those two, all other clusters are

ID	Size	Main owner(s)	Owner(s)% [#]	Unk. owner % [#]
0	245	Copesa	5.71 [14]	25.71 [63]
1	14	-	-	50.00 [7]
2	19	El Mercurio	100.0 [19]	0.00 [0]
3	16	Asesorias e Inversiones Comunidades Ciudadanas	93.75 [15]	0.00 [0]
4	14	El Mercurio	100.0 [14]	0.00 [0]
5	13	Grupo Diarios en Red	100.0 [13]	0.00 [0]
6	9	Grupo Prisa	100.0 [9]	0.00 [0]
7	4	Editorial Televisa Chile	100.0 [4]	0.00 [0]
8	1	Betazeta Networks	100.0 [1]	0.00 [0]
9	14	Red de Diarios Comunales	85.71 [12]	0.00 [0]
10	3	Estado de Chile	33.33 [1]	0.00 [0]
		ITV Patagonia	33.33 [1]	
		Corporacion de Television de la Pontificia Universidad Catolica de Valparaiso	33.33 [1]	

Entities owning over 10% of the outlets in a cluster are listed next to it.

Table 4.3: Ownership properties for Minhash-based clustering for the *ds16* dataset.

heavily, if not entirely, dominated by one owner.

4.2.2 Clustering metrics

Based on these results, we hypothesize that ownership relationships are similar to the ones based on content. Given that we have the actual owners of most news outlets in our data sets, we used this as a ground truth to evaluate the performance of our methodology. To this end, we computed different clustering metrics using ownership information as class labels.

We used the Adjusted Rand Index (**ARI**) to quantify the degree of correspondence between the set of communities found by our methodology and the sets of clusters defined by the actual owners of the news outlets. **ARI** scores are normalized against chance, so scores close to 0.0 indicate random label assignments, 1.0 indicates a perfect match, and negative scores indicate a correspondence lower than what is expected for random assignments. Similarly, the Adjusted Mutual Information (**AMI**) index gives a sense of how much information we can obtain about one distribution given the other one. **AMI** scores are also adjusted with respect to the expected value (subtracting the expected value from the Mutual Information score). Again, scores close to 0.0 indicate random assignments and a 1.0 score indicates two identical assignments. The Normalized variation of the Mutual Information Index (**NMI**) also gives a greater score as the communities are closer to a perfect recreation of ownership classes. Moreover, **NMI** does not penalize if the classes are further subdivided into smaller clusters. The results of the application of these indices (given in Table 4.4) suggest non-random

clusters. Homogeneity (**Hom**) is maximized when each cluster contains members of a single class, while completeness (**Com**) measures the desirable objective of assigning all members of a class to a single cluster.

Prior to calculations, we removed outlets without ownership information and outlets with owners that only have a single outlet, since they do not add any relevant information. Additionally, there are outlets that do not belong to any of the communities we found. They could be discarded, but we might be deleting valuable information: our algorithm indicates their content is different from the others'. For this reason, we preserve each of them as a community of size 1. Though reasonable, this might distort some comparison metrics, as the correspondence of single-outlet communities is perfect if they're isolated in both the content-based and the owner-based community structures. As both the number of considered outlets and the number of communities is altered by these decisions, we specify them in Table 4.4 (columns *Outlets* and *Comm.* respectively).

Similarity	Outlets	Comm.	ARI	AMI	NMI	Hom	Com
Vocabulary	157	28	0.1834	0.3007	0.5362	0.4748	0.6056
Topic: Keywords	157	66	0.4246	0.4261	0.7313	0.8113	0.6592
Topic: Minhash	167	12	0.4301	0.4584	0.6593	0.5460	**0.7961**
Cluster Topic: Minhash	169	90	**0.5326**	**0.4652**	**0.8365**	**1.0000**	0.6997

Rows represent the different metrics used to calculate the similarity graphs. Columns represent the scores of the indices calculated using our ground-truth as reference (**ARI**: Adjusted Rand Index, **AMI**: Adjusted Mutual Information Based, **NMI**: Normalized Mutual Information Based, **Hom**: Homogeneity, **Com**: Completeness)

Table 4.4: Comparison of community structures and ownership.

Table 4.4 shows the results of these indices over the communities obtained from our similarity graph. For comparison, we again show the results obtained for the other similarity graphs. Once more, we can see that the vocabulary-based similarity does not give a good prediction on news outlets that belong to the same owner: *Vocabulary* gets the lowest score for all metrics. On the other hand, we can see that topic similarities do show higher degrees of correspondence with owner classes, which is consistent with previous observations. Keyword-based similarity communities have high homogeneity, while Minhash-based communities show very high completeness, shown in Figs 4.4 and 4.5.

We also included in Table 4.4 the results for indices over the normalized cut clustering obtained for the *topic (MinHash-based)* similarity. We follow the same procedure for clusters, i.e., we removed outlets without ownership information and outlets with owners that only have a single outlet. Also, we moved each remaining outlet in cluster ID 0 to its own individual cluster. The completeness (**Com**) is altered by the bigger number of clusters of size one created from outlets in the cluster with ID 0. The results for this variation (clusters over minhash-based similarity) are the highest for most indices, presenting this technique as a very good predictor for a common-owner relationship.

4.3 Conclusions

In this chapter we introduced an analysis of Chilean news media outlets based on the information that each news source chooses to post in Twitter and the similarity clusters of news outlets that arise from this content. The study of whether ownership influences the content produced by different news sources is an important area of research to make possible biases explicit. More importantly, however, for the purposes of this thesis, it gives a considerable first step toward proving the applicability of the first filter of the Propaganda Model.

In general, our results indicate that ownership does play an important role in news content similarity. Big, national-scope media with big audiences tend to group together in their own community; other outlets, generally with a more local scope, group according to a mix of ownership and some geographical features that we inferred (such as *Mercurio de Valparaíso*).

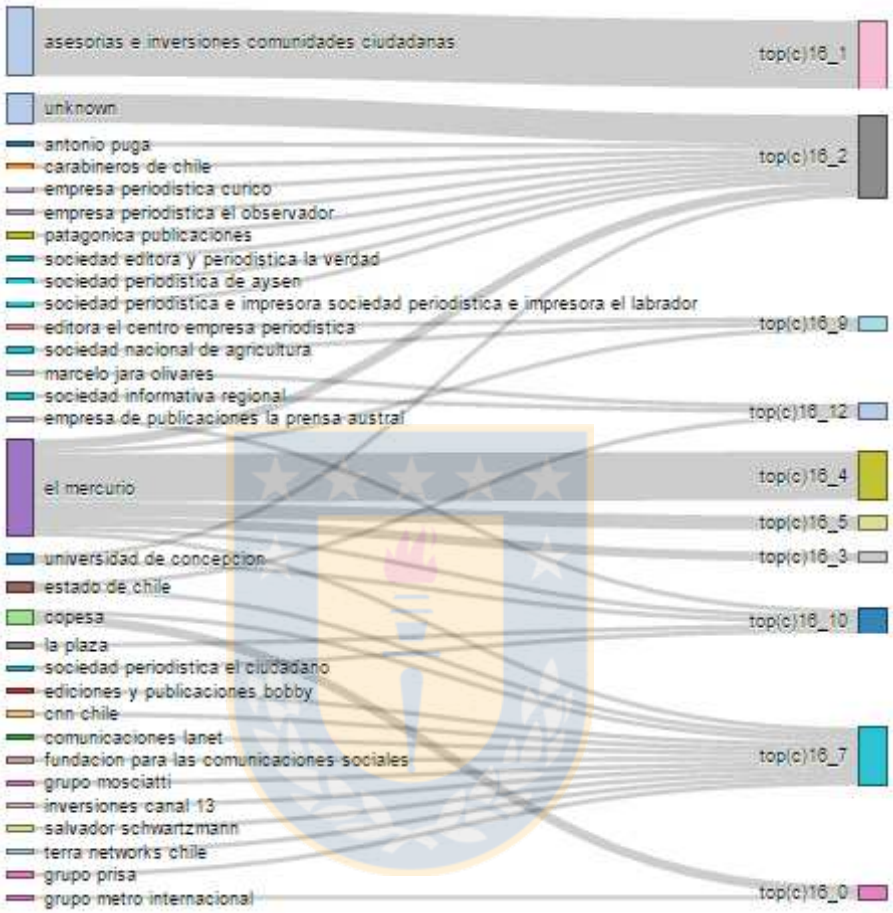
We studied several similarity metrics as well as different ways in which to identify clusters (or communities) in our data. The indices that we calculated (e.g., **ARI** and **AMI**) suggest non-random clusters. These results seem in agreement with our hypothesis, since similarity based on vocabulary seems that may be attributed to other factors (e.g. geographic zones); on the other hand, our analyses indicate a correlation between owners and their selection of topics (which can be interpreted as a common editorial policy for outlets owned by the same entity).

We show that our results are consistent over time by studying two non-overlapping in time datasets, *ds15* and *ds16*, and show that they both present consistent properties.

A limitation of our methodology is that when outlets are too specialized (e.g. magazines on automobiles, fashion, etc.), even if they belong to the same owner, they do not cluster together. This is due to the nature of the stories that they publish, since by design they do not share any significant part of their content. It is therefore difficult to conclude that there is any influence by owners in the content generation of these specific type of outlets.

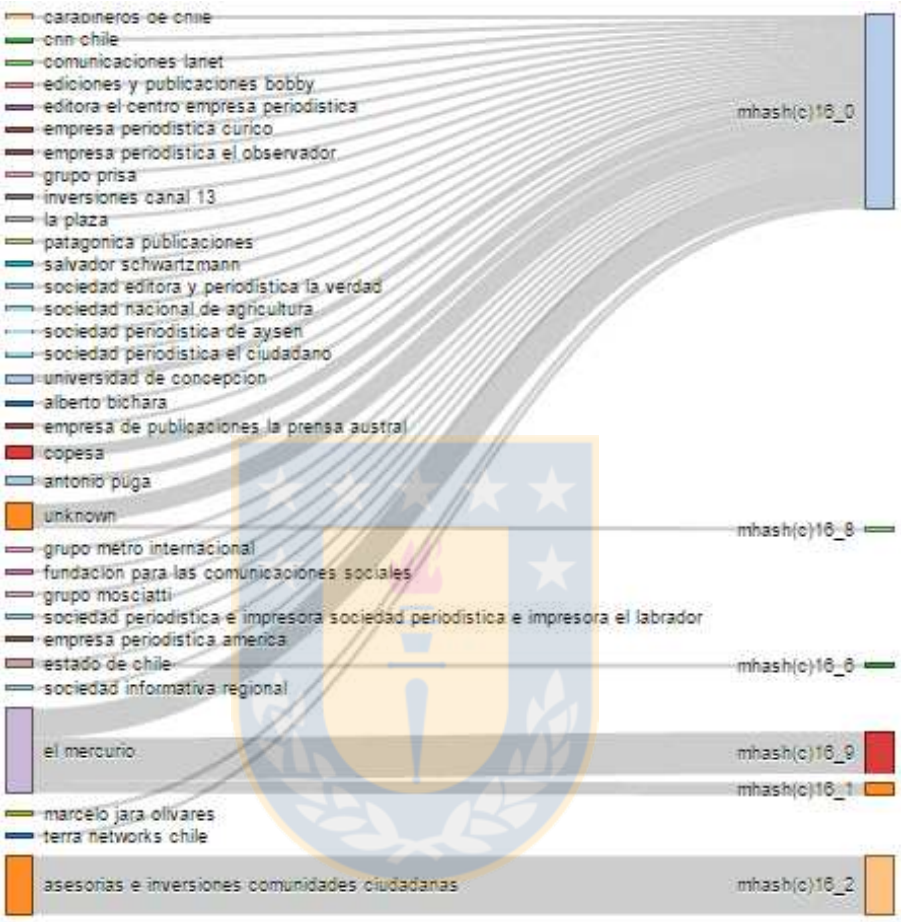
Some owners only have one news outlet. As we saw in section 4.2, our similarity measure based on topic-detection does a good job in pulling apart these special cases from the biggest groups of owners by clustering them together in a different community (our cluster with *id* 0 above). What we are after, however, is the identification of more than one outlet with a single owner such that content in those different media may be affected by a single editorial line.

We have shown that using a language-independent and fast approach we are able to easily discover “editorial homophily”. Besides the contribution for the Propaganda Model, this method can have several applications, such as help mitigate the “filter bubble effect” in people’s news media consumption by recommending more diverse news sources. It can also help identify “hidden owners”, in the sense that we can identify news sources that behave as if they had a same owner, despite not declaring so publicly. Overall, our findings could help towards promoting a media structure that is less biased towards very few groups as in the trending concentration of ownership.



Owners are displayed on the left, while communities are displayed on the right. The width of a flow connecting an owner and a community is proportional to the number of outlets in the community belonging to that owner.

Figure 4.4: Ownership vs. Topic (keyword-based) community structure.



Owners are displayed on the left, while communities are displayed on the right. The width of a flow connecting an owner and a community is proportional to the number of outlets in the community belonging to that owner.

Figure 4.5: Ownership vs. Topic (minhash-based) community structure.

Chapter 5

Diversity and Health of Online News Ecosystem

Even in developed countries with an active free press, news coverage can be dominated by only a few players. For example, we have found that approximately 90% of Twitter users who follow a Chilean news outlet, follow at least one own by one of the three biggest companies (Copesa, El Mercurio, Publimetro). This can lead to a reduction of topical and community diversity. As we saw in the previous chapter=, ownership structures might further limit coverage by implicitly or explicitly biasing editorial policies (see Chapter 4). Our analysis on the nature of the bias (see Chapter 3) shows that due to its reliance on social networking relations for news propagation, social media may be subject to a variety of social issues that may restrict news coverage and topical diversity, e.g. as a result of information bubbles [120] and social conformity bias[100].

On the other hand, some might argue that the exponential growth of new communication technologies can solve, or at least alleviate, many diversity problems. More accessible and cheaper channels of communication should provide new content producers with better opportunities and less friction to compete in a larger media market. However, this assumption has not been tested empirically. Indeed, early indicators point to high levels of bias as well as a lack of diversity in terms of topics covered and communities addressed [77, 133, 57]. This would indicate that the postulates of the Propaganda Model [72] also apply to the new channels of news propagation.

In the past, news ecosystems have primarily been modeled from a political and economic perspective. In the *Media Capture Model* proposed by Besley and Prat [15] the authors state that, for a truly democratic society, the more information we can have as voters the better. They predict that a *low* number of independent outlets will make the news media industry *more* susceptible to be fully captured by the political and economic elite. In other words, when there is enough pluralism, the media behave more independently (see section 2.3.2).

Both the *Media Capture Model* and *The Propaganda Model* warn against the negative consequences of the concentration of ownership in the mass media. Having a large share of the media industry in the hands of just a few mega-conglomerates poses the risk of the system not necessarily representing the interest of the common good, the media's original primary purpose.

A different metric of pluralism is discussed in [128]. The author defines two classes of pluralism: External Pluralism (EP) and Internal Pluralism (IP). EP requires that all political opinions have room and are represented in at least some of the suppliers of content in the media market. On the other hand, IP is achieved when every media company covers all sides of the main political issues in a society. These measures are closely

related to the concept of *political parallelism* as defined by Hallin and Mansini [66] (see section 2.3.1). EP benefits from a larger number of media outlets if the users are really free to choose. If the public favors (or are limited to) a small set of news media, then it is important to analyze market concentration; i.e., the number of companies and the percentage of the total news production that each of them represents.

Here we postulate that the news industry can be modeled as a complex system [89], an ecosystem that consists of many different interacting components, such as news outlets, their owners, reporters, news consumers, advertisers, all subject to and responding to a variety of social factors. Through their interactions among themselves and with external drivers, these components collectively shape our news ecosystems. Given these broad similarities, we hypothesize that we can apply techniques developed to study the health and diversity of biological ecosystems to online news (eco)systems.

In this chapter we use a set of ecological indicators [76] to analyze the health of the "news ecosystem" as viewed from Twitter. This analysis allow us to assess the impact of the ownership concentration process over the news diversity in the online Chilean news ecosystem. Our work shows that, by more than one metric, this news ecosystem can be considered to be in a "poor" state in terms of heterogeneity, diversity and access to varied information.

5.1 Methodology

Our objective is to measure the diversity of a news ecosystem, taking into account the variety of different news sources, the producers of news content, as well as the news consumers. We start with considering each individual news tweet as an *entity* and its corresponding news outlet as its *type*. We then apply well-known ecology indices to quantitatively measure how "healthy" – diverse – our system is. We assume that diversity of content is a desired property of any news system, see [86].

Similarly, we have shown that ownership can influence editorial policies and bias content [10] (see chapter 4). Thus, we will also relate the relationship and type of each entity with *the owner* of the publisher outlet, rather than with the news outlet itself. This is, potentially, a stronger effect, since several newspapers may publish similar content because they belong to a single ownership group. We analyze media ownership [177] using two metrics: *numerical diversity* and *source diversity*. Numerical diversity refers to the number of outlets available to the public in a given area; source diversity indicates the number of owners that actually control those outlets. The rationale for using these indicators is our expectation that having a news industry increasingly dominated by fewer and fewer companies, increases the owners' potential influence on the published content, leading to a greater probability of reallocation of attention to their interests as proposed by the first filter of the Propaganda Model.

Several studies indicate that online news distribution and consumption can be subject to considerable bias, for example through the so-called Filter Bubble effect [120] and the prevalent tendency towards homophilic connections [100, 11] in online social

networks. This may be counteracted by the fact that social media users are generally exposed to a wider number of news sources [35, 108]. In this chapter we attempt to assess, on balance, to which degree even readers who are subscribed to a high number of the available news outlets (or are exposed to their news indirectly [6]) can still be affected by significant bias due to the lack of diversity in online news ecosystems. We want to note again that, like in the Propaganda Model, we are not considering the effect of the media behavior on the general public. In this thesis, we are characterizing the media system that is exposed to that public using a variety of tools and from different angles. As we already mentioned, news consumers only have access to the corpus of published news. Thus, they only get to see the *final product* of the system of news production which has already gone through the alleged filtering process outlined in The Propaganda Model. In this case, the news ecosystem is observed through the final news items that consumers have access to.

This is less the case for online news where consumers play an active role in the distribution, formation, and modification of news, and these processes, recorded in social media data, are observable much like the news items themselves.

Ecological science has developed extensive models of the diversity of ecosystems that may generalize and apply to online news ecologies. Ecologists have used four attributes to characterize the evolution of complex systems [164]: (1) progressive integration, (2) progressive differentiation, (3) progressive mechanization, and (4) progressive centralization. The progressive integration is represented in the news ecosystem by the current dynamics in news production where small news outlets report (or redistribute) news created by bigger news agencies, or the tendency of outlets of the same owner to report on the same topics (see chapter 4). Progressive differentiation is shown in the plethora of news outlets and magazines that create their own niche attempting to take advantage of their condition (either geographic, topic wise or by exploiting certain political position [137, 6, 57]). Progressive mechanization refers to the growing number of feedback and regulation mechanisms, that social media platforms seem to be particularly susceptible to. Finally, progressive centralization can be seem in how news has been modified and adapted to the other components in the ecosystem (*e.g.* native ads[1]).

In [132], the authors provide a conceptual definition of ecosystem health. It is largely focused on three components: (1) Vigor or scope for growth, (2) Organization (given by the diversity or complexity of the system), and (3) Resilience (in function of the system capacity to counteract stressful conditions). These components are integrated in a global Health Index (HI) that can be formulated as follows:

$$HI = V \times O \times R \tag{5.1}$$

where V represents the Vigor of the system, O represents the Organization Index, and R represents the Resilience index. As HI is directly proportional to all three factors, lowering any of them will result in a lower global health index. In this work we focus

[1] https://en.wikipedia.org/wiki/Native_advertising

in the Organization component by analyzing a variety of quantitative diversity indices, namely the Shannon Diversity [142] and Simpson Diversity [146] indices. Meanwhile, the Average Taxonomic Distinctness [165] provides a notion of the similarity we can expect from the coverage of a story, even in cases where it originates from different outlets. We restrict our analysis to the "news ecosystem" form in the context of the Chilean on-line media, and specifically on Twitter.

Like in our previous analysis on the power structure, we use Minwise Hashing for topic detection (see chapter 4 section 4.1.1). This time, we set $n = 2$ to allow tweets to cluster into wider topics. Once again, we consider these clusters our set of discovered topics. Remember that we look only for topics corresponding to tweets from *multiple* news outlets (*multi-outlet-clusters*).

5.1.1 Diversity Index

A diversity index is a quantitative measure that reflects how many different *types* there are in a data set and/or takes into account how evenly the basic *entities* are distributed among those *types*. There are three basic groups of ecological diversity indices: enrichment of species, abundance of species, and proportional abundance of species [76]. The first group, enrichment of species, only measures the number of species. Indicators of the abundance of species, besides the number of species, also try to model the distribution of their abundance. The last group of indices, proportional abundance of species, represent enrichment and uniformity in the same expression. Within this last group we can find the Shannon Diversity Index and the Simpson Index. In turn, the Average Taxonomic Distinctness index approaches the problem by taking into account different dimensions of biodiversity (e.g. taxonomic, numerical and phylogenetic). Including these aspects of the diversity helps counteract some of the problems described for previous diversity indices such as measuring functional diversity [166]. We use these indices to asses diversity of the on-line news distribution.

Shannon Diversity Index (ShDI)

Shannon Diversity Index is widely used in ecology and biology to measure the diversity of species in a community [142], but has also been extrapolated to other fields. In [159] the authors used the ShDI to measure subjectivity in the selection of dates for timeline creation in news stories. For example, dates that were considered significant in the timeline of one newspaper, as opposed to those dates that were relevant for "all" news outlets. The authors use this index to highlight dates on which important events happened, but that are likely to be ignored by many news agencies, hence, indicating how subjective (or non-diversified) a date is. According to the Propaganda Model, this subjectivity in the selective reporting may be caused by direct and/or indirect intervention of the ownership in the editorial policies of the outlet.

Here we apply ShDI to topics; we use the ShDI to express the rarity or commonness of topics, as reported by different news outlets. The idea is to quantify, for each

detected topic, how common it is across the media, and whether it was covered disproportionately by just a few newspapers. This will give us an indicator to assess whether the topic was generally considered important news and covered accordingly across a wide variety of outlets or pushed as a topic by specific outlets.

The expression to calculate the ShDI value for cluster c_i is given in Equation 5.2. Using newspapers as types and tweets as entities, Eq. 5.2 quantifies the uncertainty in predicting that a newspaper will publish a tweet taken randomly from the dataset (the dataset in this case are the tweets of the corresponding cluster).

$$ShDI(c_i) = -\sum_{i=t}^{R} p_t \ln p_t \tag{5.2}$$

In Eq. 5.2, R is the number of *types* participating in the cluster and p_t is the proportion of *entities* from type t. A low index value indicates an unhealthy (or polluted) ecosystem. This index usually takes values in the range $[0..5]$ (bits/individual), interpreted as follows: (1) High status: > 4 (2) Good status: $4-3$ (3) Moderate status: $3-2$ (4) Poor status: $2-1$ (5) Bad status: $1-0$ [107].

Our previous study of power structure showed that ownership does influence, to some extent, the editorial policies of a given media outlet [10]. Thus, it is important to test whether the subjectivity of a topic varies from newspapers to owners: owners may either want to maximize readership (making the ecosystem diverse), or focus on a single message to target a specific audience ("biasing" the ecosystem). In order to find out the effect of ownership given the ShDI, we included "owners" as *types*, while retaining each tweet as an entity associated with the owner of the newspaper that published it.

Pielou Evenness Index (PEI)

In Equation 5.2, the value of the ShDI increases both when evenness increases and when the number of *types* increases. So, to be able to compare results for different topics, we also calculated the maximum achievable ShDI ($ShDI_{MAX}$) of each cluster to normalize the results. This normalization is also known as the *Pielou Evenness Index* [126]. The PEI expression for cluster c_i is of the form:

$$PEI(c_i) = \frac{ShDI(c_i)}{ShDI_{MAX}(c_i)} = \frac{ShDI(c_i)}{\ln R} \tag{5.3}$$

where R is the number of *types* (*i.e.* outlets or owners) that participate in the cluster c_i.

Simpson Index (SiDI)

The Simpson Index [146] is another index widely used in ecology. While the ShDI is based on information theory and measures the abundance of species and diversity of individuals, the Simpson index is considered a *dominance index* that assigns a higher

weight to the most common species. This means that the presence of a few individuals of some rare species will not have an important effect in the result. This renders as very appropriate to our analysis given that just a few mega-conglomerates heavily dominate the media system.

The original index gives a value λ ($0 \leq \lambda \leq 1$) that is higher for environments with low diversity, which is counter-intuitive for a diversity index. To solve this, most authors use the *Gini_Simpson Index* (see Equation 5.4), which is a variation of the Simpson Index

$$SiDI(c_i) = 1 - \lambda = 1 - \sum_{t=1}^{R} \frac{n_t(n_t - 1)}{N(N-1)} \tag{5.4}$$

In Equation 5.4, again, R represents the number of different *types* in the cluster. We use N to represent the total number of tweets in c_i and n_t is the number of tweets in c_i published by *type* t.

This index is used in [63] as the *Participation coefficient*. With this value they measure how well-distributed are the connections of a node among the communities of the graph. Defining a range of the obtained measures helps the authors classify the different roles that a node may have in a complex system network. In [56, 78], the authors also use the Simpson Index to differentiate nodes in a social network based on the interactions of people that use different languages.

We adopt a similar interpretation: in our case, we are interested in measuring how well-distributed is the coverage of a given topic that is received from the available news sources.

Average Taxonomic Distinctness (ATxDI)

Similar to the ones above, this index takes into account the species abundance, but also includes the taxonomic distance between any two types [165]. Specifically, this index represents the expected path length through the classification tree between two entities chosen at random. For us, then, the Average Taxonomic Distinctness is the average editorial "distance" (using a similarity matrix, see below) between two news sources randomly selected from two different types in the same topic. As before, types are either news outlets or owners.

For the taxonomic distance, we use a numerical taxonomy [150]. This form of classification is basically determined by observable characters of taxa (i.e., phenetic similarities). Since we already know the different classes (i.e. our *types*), this should give us an idea of the affinity of any two types. Similarity between two news outlets is then defined by the co-occurrence of two *types* with respect to a same topic. Note that the topics are extracted also from observations of homologies in words (n-grams) of our entities, so the similarity could be further rooted in these lower level aspects. We reuse the similarity matrix defined in the analysis of the power structure (see section

4.1.2). Finally, we use the agglomerative hierarchical clustering algorithm[2] to create a tree (using the arithmetic mean for the linking method - also known as the UPGMA algorithm). For the clustering algorithm we first transform the similarity matrix into a distance matrix (i.e. $dist(A, B) = 1 - sim(A, B)$). The more similar two types are the closer they will be. With the tree obtained from the clustering we can calculate the length of the path between any two types.

The Average Taxonomic Distinctness for a topic c_i is described in the following formulation:

$$ATxDI(c_i) = \frac{\sum_{j=1}^{R}\sum_{k=1}^{j-1}\omega_{jk}n_j n_k}{\sum_{j=1}^{R}\sum_{k=1}^{j-1}n_j n_k} \tag{5.5}$$

where R still represents the number of different *types* in the cluster and n_j is the number of tweets in c_i published by *type* j. The factor ω_{jk} represents the length of the path connecting types j and k in the tree. The double summation accounts for all pairs of types. Equation 5.5 comes as the result of dividing the *average taxonomic diversity* [165] by the Simpson Index. Doing so eliminates the dominating effect of the species abundance distribution.

The approach proposed by the Taxonomic Distinctness brings a different dimension to diversity. An ecosystem under environmental disturbance could display not only a reduced number of species (as shown by the Simpson and Shannon indices) but also that the remaining species could be closely related. For a news media ecosystem, this would imply that not only the stories are dominated by a few outlets, but also that the point of view of these outlets could be very similar, which could be seem as an indication of media capture.

5.1.2 Data

For the data we selected the same set of 365 outlets used in our previous analysis of the power structure (*ds16* collection - see section 4.1). Remember this dataset includes tweets generated from October 25, 2015 to January 25, 2016, for the 365 news outlets twitter accounts (containing 756,864 tweets). The text of each tweet was lower-cased and preprocessed by removing stop-words, URLs and punctuation marks.

We treat every tweet as an independent document from which we can extract a statement/headline. Headlines of online news articles have shown to be a reliable source for adequately providing a high-level overview of news events [3, 40, 158].

Since we are working with topics, we filter out the tweets from 'specialized' news outlets, and kept only 235 outlets registered as "general-interest", those covering most subjects. Specialized outlets or magazines (such as fashion or sports) are expected to give a differentiated coverage to special subjects, which could influence our results. Thus, we focused only on those topics/events that were considered of interest to the

[2]This time we use the version implemented in the *scipy.cluster.hierarchy.linkage* library

general public. For these general-interest news outlets we collected 563,262 tweets during the observed period.

5.2 Results

5.2.1 Topics

For the 235 general-interest news outlets, we were able to identify 79,753 clusters using the min-hash techniques described above (see section 4.1.1). These clusters account for 366,180 tweets (65% of the total). Notice here that we are only counting tweets that are contained in one of the clusters, and only those clusters that contain at least two tweets .

There were 56,496 clusters with just one news outlet in them (*single-outlet-clusters*), grouping 172,276 tweets. We found that, against Twitter Rules[3], many outlets tweet multiple times with the same text or a very small variation of it (this is considered *spam* by Twitter). After collapsing tweets with the exact same text into a single one, these clusters were left with 64,920 different tweets (only 37.7% of the tweets in *single-outlet-clusters* were original content). As many as 49,369 *single-outlet-clusters* were formed by one repeated tweet. Even if we do not use this information in our analysis of the indices, it is already a strong indication of the poor condition of diversity in our news ecosystem: 87% of single-outlet-clusters contain a single text repeated in multiple tweets. Already, this can be considered as a very low measure of Internal Pluralism, as discussed above.

For our analysis we searched for clusters that had tweets from more than one news outlet (*multi-outlet-cluster*). There were 23,257 *multi-outlet-clusters* (29.2% of total clusters). These contained 193,904 tweets. After removing tweets with the exact same text published by the same outlet, there were 143,092 tweets (73.8% of the total number of tweets in *multi-outlet-cluster*).

To check how effective our method of clustering was, we calculated the Jaccard Index ($JI(x, y)$) for each tweet x against every other tweet y on its cluster, assigning the mean of the JI to that tweet x. The JI_c of the cluster is the mean of the JI of the tweets it contains (see Equation 5.6).

$$JI_c(c_i) = \frac{\sum_{x \in c_i} \sum_{y \in \{c_i - x\}} JI(x, y)}{N(N-1)} \tag{5.6}$$

In the case of *multi-outlet-cluster*, for $JI_c \geq 0.8$, we had 22,025 clusters (94.7% of total multi-outlet-clusters). Even for clusters with $JI_c < 0.8$, the content of the tweets within the cluster is still very similar for most cases. The smaller value in the JI_c is mainly because of the shortness of the messages: as a result, changing just a few words would lower the value of the JI. For example, two tweets with the text[4] *"bolsa*

[3] https://support.twitter.com/articles/18311

[4] This is the text after removing the stop-words

santiago parte incremento" and *"bolsa santiago parte ganancias"* have a $JI = 0.6$.

To check for inter-cluster similarity we ran a second pass of our clustering procedure. This time using as input a bag of words for each of our initial clusters: less than 3.0% of the topics clustered in these "second-level-clusters". Showing a very low inter-cluster similarity.

In summary, we were able to identify a fair amount of topics where more than one news outlets are involved. These are events that were considered newsworthy by at least two different sources. This set of *multi-outlet-clusters* constitutes the dataset of topics for our diversity analysis.

5.2.2 Diversity

ShDI and PEI

As a reference, we first calculated the maximum achievable ShDI taking into account all the newspapers in our data set (as opposed to only those with at least one tweet in the cluster, $ShDI_{MAX}$). In other words, we find the maximum achievable ShDI if all 235 "general-interest" outlets publish on the same topic approximately the same number of tweets. We will refer to this value as $ShDI'_{OPT}$. In this case, we get a $ShDI'_{OPT}$ of 5.4337 for news outlets and 4.4426 for the owners. These values are in the range of a good/high status of diversity, which means that the Chilean media have the potential to be a healthy system.

We calculated the ShDI (Shannon Diversity Index) for each topic. For our first experiment (using the newspapers as *types*), the average ShDI among all clusters is 1.3455 and the $ShDI_{MAX}$ is 1.3484. When considering the owners as *types*, the average ShDI and average $ShDI_{MAX}$ among all clusters was 0.1408 and 0.1526 respectively. These are very low (see above), even considering just the $ShDI_{MAX}$, which means that there is a very low agreement between outlets to select the topics they publish on Twitter.

We obtained an average normalized ShDI (i.e. Pielou Evenness Index - PEI) of 0.9971 for the newspapers. The average PEI for owners stands at a low 0.1887. Looking only at the PEI value obtained for the outlets, one might conclude the system is doing well in terms of diversity, but the PEI obtained for owners indicates the critical condition of the ecosystem. These reinforce the ShDI results above, but are more telling of the concentration problem in the media industry.

Finally, the ratio $PEI' = ShDI/ShDI'_{OPT}$ shows how far the Chilean news ecosystem is from becoming this ideal system: outlets are on average 24.7% of their full potential diversity, while owners stand at an extremely low 3.1%.

Even when the indices are low for both types (i.e. outlets and owners), we can see that diversity between news outlets is at least one order of magnitude larger than between owners. This low values in the diversity means that newspapers and, in a higher degree, owners are pushing their own agenda by the introduction of topics that are newsworthy almost exclusively to them. This behavior is consistent with the hypothesis of the Propaganda Model and, more specifically its first filter.

SiDI

For the same set of topic used in the previous section, we also applied the Simpson Diversity Index (SiDI) to search for indications of concentration of the market and/or dominance of the news cycle by just a few sources.

When using the outlets as *types*, we found a very high average result, $SiDI = 0.9884$. Recall that the Simpson index only range from 0 to 1, so these values indicate that the reporting on these topics does not seem to be controlled by just a few outlets. On the contrary, it appears that each subject is being equally covered by most of the outlets that participate in it.

On the other hand, using owners as *types*, we obtain an average $SiDI = 0.1778$. This indicates that the media system in Chile shows clear symptoms of market concentration and a severe lack of diversity. Once again, the difference in the results between both evaluations *reveals the artificial illusion of diversity created by the multiplicity of outlets owned each by dominant companies*.

A very telling sign of artificial diversity created by the owners that control the market can be seen by analyzing the percentage of topics each *type* (outlets or owners) participates in. Figure 5.1 shows this statistic for the 30 outlets with the largest participation. The graph shows that these 30 outlets have a fairly balanced presence on the topics discussed. However, out of these top 30, 25 belong to the same owner (*El Mercurio S.A.P*), and only one news outlet from a different owner participates in more than 5% of the *multi-outlet-cluster* topics. Figure 5.2 shows the severity of the dominance of this one company in the selection of subjects (i.e. outlets owned by *El Mercurio* participate in more than 66% of the topics shared by more than one media Twitter account, where the closest competitor is under 10%).

ATxDI

Finally, we used the Average Taxonomic Distinctness to evaluate the expected editorial distance between *types* publishing the same topic; i.e., in the same cluster. As mentioned before, this index gives another dimension to our diversity analysis, taking into account not only how many different sources participate in a given topic, but also how similar or dissimilar these sources are.

As with previous indices, we first analyzed news outlets as the types of our entities (i.e., the tweets). For the 23,257 topics found, we obtained an average value of $ATxDI = 4.08$. Note that, given the way we constructed the similarity tree, the shorter path between any two outlets has length at least two (e.g., if two outlets are siblings). When we assign a random outlet (taken from the list of outlets) to each tweet, we get an average $ATxDI = 7.60$. If we also take into account that the average number of news outlets owned by one company is 2.64, we can see that we have low distinctness in the news ecosystem.

In the case of owners as types, we observe a lower average distinctness for the list of topics ($ATxDI = 0.71$). When comparing this result against its equivalent for

a random assignment of outlets to each tweet we obtain a mean of $ATxDI = 13.04$. Even if we limit our analysis to the 4,740 topic share by more than one owner, the average *ATxDI* is only 3.48.

5.3 Conclusion

In this chapter we applied three ecological diversity indices, which are commonly used to asses the health of biological ecosystems, namely the Shannon Diversity Index, the Gini-Simpson Index, and the Average Taxonomic Distinctness, to assess the health of the Chilean news system as an ecosystem. Our results across the different indices suggest that the Chilean on-line news ecosystem lacks diversity, in terms of coverage, topics covered, editorial policies, and ownership, possibly leading to a deterioration of an individual's access to variety in their news coverage and news sources. As we saw before, this coincides with characteristics of the media system described by the Propaganda Model.

Following the analysis proposed by Polo [128], we find low external pluralism (EP) or low diversity. Topic selection seems to be driven by subjective factors rather than objective criteria, such as the newsworthiness of events. Topics are covered by only a few news outlets and even less owners. Furthermore, we find that outlets that cover the same topics exhibit high levels of content similarity, indicating a lack of independent reporting. This suggests that many outlets are not only subject to similar editorial policies, but rely on similar content. Although our results indicate relatively high numerical diversity (many news outlets), which should in principle contribute to a healthy Chilean news ecology, we observe a significant lack of *source-driven* diversity.

We found that the health of the system is considerably more critical when we use owners instead of outlets as types: we saw between 5 and 6 times more diversity for outlets according to Average Taxonomic Distinctness (ATxDI) and Simpson Index (SiDI), and almost ten times for the Shannon Diversity Index (ShDI). The lower ownership diversity vs. outlet diversity is indicative of high levels of concentration in the Chilean news market: few owners control many outlets and may influence their editorial policies (e.g., topic selection). The fact that outlets owned by the same company systematically share the same topics/clusters indicates low internal pluralism (IP). Again, we see how a few mega-conglomerates controlling a large number of outlets presents a clear risk to news diversity, coverage, and representativeness[129, 128].

The current Chilean media ecology seems highly concentrated in terms of ownership and coverage. However, one may expect that Internal Pluralism (IP) may mitigate this issue in terms of news diversity. Our observation suggest this may be difficult to achieve due to high levels of topic concentration and indications of biased topic selection. External Pluralism, on the other hand, can be achieved, but requires policy intervention to sustain a healthy and diverse media ecology. Our analysis may provide quantitative input to such decision-making.

We developed quantitative measures of the healthiness of news (eco)systems in

general whose usefulness extends beyond their specific application to the Chilean case. As we have shown, measures for internal and external pluralism, as well as a range of diversity measures, provide detailed insights on the diversity or congruity of the media landscape. Furthermore, as proposed by the Propaganda Model and demonstrated across this thesis, apart from the newspapers and their content themselves, ownership may be an important factor to determine to what extent news diversity may be affected by economic drivers. The measures we used are internally meaningful, without the need to compare them with those obtained for different countries or regions. This allows us to draw more generalizable conclusion about news ecologies, and the factors that drive their ecological health, even from data that pertains to online news distribution in the Chilean context.

However, we need to caution about some assumptions and limitations of our approach. First, our analysis pertains strictly to content that outlets *publish on twitter*. More traditional publishing methods, such as paper-based newspapers, may exhibit lower degrees of concentration and greater ecological health. However, due to the higher barriers of entry of traditional publishing compared to online media, this is unlikely. Furthermore, the online distribution of news is growing rapidly to the degree that it may soon become the dominant medium. Hence, our analysis sheds light on a phenomenon that will become increasingly important for the health and diversity of our news ecology. Second, we do not assess within-topic coverage differences, i.e., two outlets that publish tweets in the same cluster might in principle take opposite approaches to the same topic, but our analysis will not acknowledge such differences. Instead, we assume that a systematic co-occurrence in clusters implies similar interests and points of views. Finally, our results are descriptive in nature. They do not pertain to the causal mechanisms that define low ecological diversity and readership, or to the Chilean population in general.

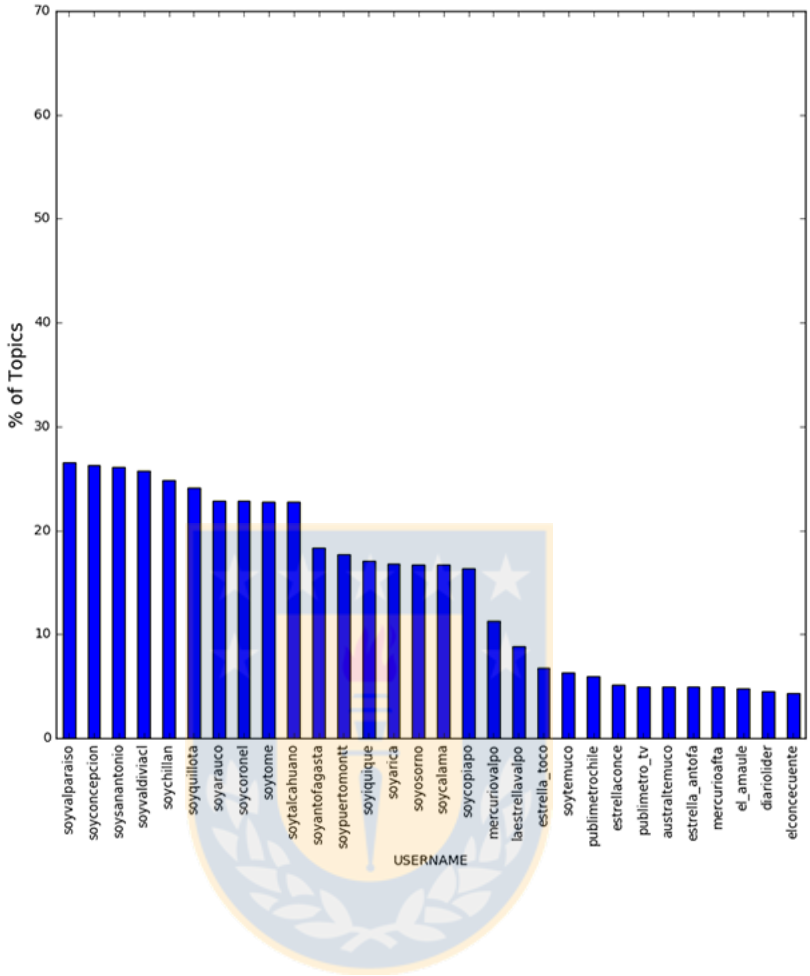


Figure 5.1: Percentage of topics where each type participates. Top 30 ranking (outlet as types).

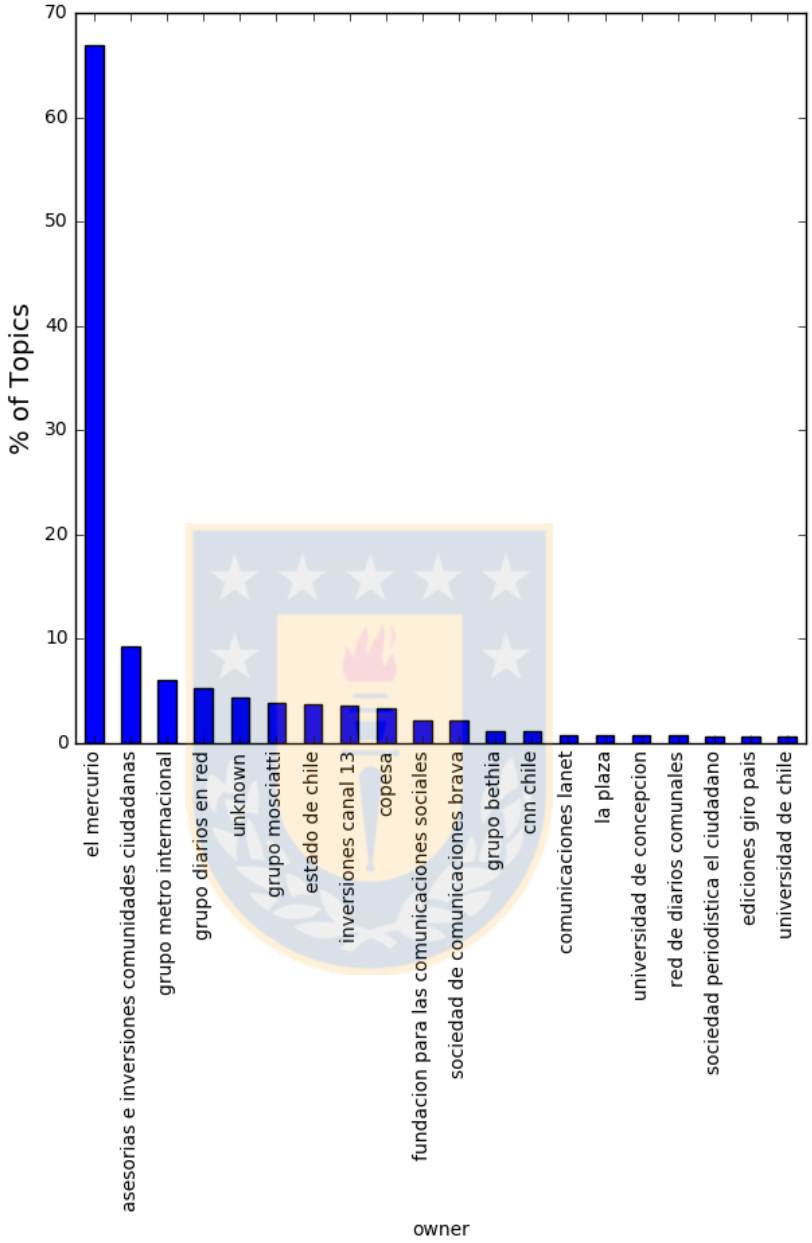


Figure 5.2: Percentage of topics where each type participates. Top 20 ranking (owners as types).

Chapter 6

Understanding news outlets audience-targeting patterns

We have seem in previous chapters that many factors influence media coverage. We could naturally assume that one of the clearest exampe would be the geographic distribution of people. That is, the "value" of information decreases as we move away from the news source. [182]. However, according to the Propaganda Model (PM) [72], there are other factors that shape the distribution of news, such as direct targeting of specific sectors of the population and/or the political ideology of the media outlet itself.

The PM states that each linguistic account of an event must pass through five filters that define what is newsworthy. The second filter, the *advertising* filter, predicts that most news outlets will try to reach a specific audience (segments of the population) with the objective of maximizing "profit", instead of actually informing. In other words, outlets will try to cater to the target demographic's expectations, rather than being "fair" in their treatment of what is news. For instance, some advertisers will prefer to take their businesses to outlets with target audiences of high purchasing power, which will eventually marginalize working-class audiences; or by political reasons, with advertisers declining to do business with outlets perceived as "ideological enemies" or indeed any media unfavorable to their private interests. As mentioned before (see section 2.4.3), Prat and Strömberg [129] better define the same concept of a media system entirely driven by profit-maximization. The authors identify the characteristics that a public needs to have so that issues of informative value for them receive the attention of the media under such media system (*Proposition 4* in [129]). In a media system driven by profit, areas of low population density, minorities, and low-income classes will be relatively under-served and underrepresented in mainstream news coverage. In general, this theoretical model predicts that a less informed audience will have a lower impact on the election of politician and hence the politician will pay less attention to these sectors of the population. This creates a loop that ends up neglecting (policy-wise) the most vulnerable segments of society just for having limited access to the media.

The mass media is one of the social forces with the strongest transformative power. Despite this, there has been comparatively little large-scale, quantitative research on the relationship between the quality and diversity of the contents the media generate, and the socioeconomic indices of a particular area of coverage. In this chapter, we try to find whether or not an outlet's coverage deviates from the purely geographic influence to a more sophisticated behavior involving the predictions of the second filter of the Propaganda Model (i.e., the weight of political and socioeconomic interest). We examine the degree to which different geographic locations in the same country are

covered by existing news outlets using our dataset of Chilean social media data. We quantify how much of this coverage can be explained by the natural geographic bias (e.g., local newspapers will give more importance to local news), and how much of the bias can be attributed to the politic/socioeconomic profile of the areas they serve.

We show empirical evidence on the relation between the number of Twitter followers of an outlet on a commune and each of the main conditions of an audience that attract media coverage according to *Proposition 4* in [129]. The affinity of these elements is not a direct proof of a causal relation. But we assume that, in general, there is a natural order of information demand and supply. News media models usually presume that readers gets some value from the news (e.g., entertainment or arguments to decide on a private action) [7, 129]. Here we consider that people following a newspaper account are interested in the "editorial line" of that newspaper. This means that, following some criteria, the news outlet is creating content that is attractive to this specific audience.

serve.

6.1 Methodology

As we mentioned earlier, we will take advantage of the formal theoretical framework provided by Prat and Strömberg theory, and we will assume the model and predictions of their *Proposition 4* as a superset of the second filter of the PM. The *Proposition 4* suggest that among the main factors that influence the mass media coverage of an event are: (a) if the matter is of interest for a large group of people (a group may be characterized by a political stand, geographic location, ethnicity, etc.), (b) if it has a significant advertising potential (e.g., it may attract readers with a higher purchasing power), (c) if it is newsworthy to a group within easy reach (i.e., it is cheap to distribute news to that group). In this chapter, we look for empirical evidence on each of these three cases at a large scale and with data available thanks to new communication technologies and the massive adoption of social networks. We use statistical models to test how much of the distribution of Twitter followers can be explained based on the geographic, political and socioeconomic features of the different areas.

6.1.1 Geographic bias

According to Zipf's *Gravity Model* [182, 181], as we move farther away from the source of a piece of news, the interest/relevance of a story drops. Given that news outlets tend to cover stories where reporters can get quickly and easily (again, to minimize the cost of the piece of news), their followers are expected to be predominantly from populations that are closer to them. Also, the size of a population at a particular place may influence how newspapers cover events originating in that area. Newspapers work on an economy of scale with a considerable first copy cost. According to the Gravity Model, we could predict the flow of information in the news media system and hence, indirectly, the proportional distribution of followers a target area i will have for a news

outlet j. For this we use the population and location of both the source of the medium and the target area. Equation 6.1 represent this relation.

$$F_i^j = \frac{P_i * P_j}{D_{ij}} \tag{6.1}$$

Here P_i is the population of i, and P_j is the population of the commune in which outlet j is located. D_{ij} represents the distance between the two communes. Then, F_i^j should give us a value that represents the expected number of followers that outlet j will have in the commune i.

Distance and population size are also essential magnitudes to describe profit in the model proposed by Prat and Strömberg (*Equation 5* in [129]). News outlets will favor in their coverage issues that may draw the attention of larger groups (e.g., big cities) and to which it is cheaper to deliver the news (e.g., at a shorter distance).

We first use the Gravity Model as defined in Equation 6.1 to identify how much of the followers' distribution of the media system can be explained just by the geographic factors of distance and population. We run the model for each news outlet so we can analyze the geographic bias behaviour for different types of media.

For this study, we manually locate each news outlets in its source commune. The location may be determined by the intended audience if the name of the commune is in the name of the outlets (e.g., *soyConcepcion* is assigned to *Concepcion* city) or by the location of its headquarters. At the intra-country level, big news media companies may have more than one headquarter, but in most cases they either work under a different name (with a more "local" name) or report directly to the central headquarters which ultimately define the editorial line. For example, *Soy Concepcion* is owned by the *El Mercurio* Group, which is also the group that owns one of the largest newspapers of the capital region (also called *El Mercurio*).

Regarding the location of followers, there is an extensive body of work that focuses on geo-tagging Twitter users [118, 1, 88]. Most of this work can be divided into two groups according to their approach: content-based and network-based. Methods based on content can be further subdivided into those that use a gazetteer [5], as in our case, to find direct references to geographic places and those based on a Language Models that try to learn a probabilistic text model [140]. The performance of the former depends heavily on the quality of the used dictionary. The latter may achieve high precision for the geo-localization of users at a country level, or even within country regions or cities [28, 136]. However, to achieve a good performance at a finer grain classification, such as commune/neighborhood level, massive corpora of social media annotation is required [88]. On the other hand, the geo-localization of users based on their network (based on the assumption that users are more likely to interact with other users that are geographically closer to them) are more accurate at a finer level [8, 87]. The problem is that crawling the connections of several million users and dealing with the corresponding graph is very time consuming and computationally intensive.

In this chapter we decided to test our hypothesis using only the users that we were able to geolocate based on their profile's *location* field. We use these as a sample

of the population. The number of follower per commune in our sample is strongly correlated with the actual population distribution ($r(343) = .61, p < .01$). So, we will use this information to represent our ground-truth in the proportions of the distribution of readers for the news outlets in our database.

Finally, for every pair of communes we use their estimated populations (obtained from the INE [74]) and GPS coordinates. We calculate the direct distance between them using the Haversine formula. We represent each outlet j as a vector F^j. The elements of F^j are the expected proportion of followers in each commune for outlet j obtained from the Gravity Model. We also create a vector T^j for each outlet with the actual number of followers on each commune i obtained from our ground-truth. Using the two vectors that represents each outlet we calculate the Pearson product-moment correlation coefficients. This coefficient will give us, for each news outlet, an idea of how much of the distribution of readers can be attributed to the geographic bias.

6.1.2 Socioeconomic bias

Another factor that theoretically influences the news coverage is the socioeconomic level of a geographic area. As we mentioned earlier, hypothetically, a valid strategy for news outlets to increase their revenue in terms of advertising, will be, for example, to target sectors of the population with a higher purchasing power. Herman and Chomsky point out in the second filter of the PM that advertising, being a fundamental source of income for news outlets, plays an important role to maintain the hegemony of the top news companies in the free market. News outlets that can secure good advertising contracts may afford lower sell prices and become more competitive. This business model breaks the natural market rules that give the final buyer's choice the power to decide. In this case, the advertisers' contracts have a significant impact on the media growth or even their survival. So, outlets are forced to comply and demonstrate to the announcer how their content may serve to its needs. The audience of a newspaper becomes its product, which can then be "sold" to the sponsors.

The second filter of the PM is in line with the predictions in the Proposition 4(b) in [129] (see above). This filter suggest that in their effort to align their content with the advertisers' interests, the media have shifted to a lighter and less controversial programming (e.g., lifestyle, fashion, sports, etc.) [124, 125] In [67], the author present some evidence on the same direction, showing, for example, media preference for more "soft news" content since this is favored by advertisers as it targets a demographic of female and young people audience. A more recent and direct example on how advertisers may influence the content of the media is the evolution from product placements to *Native Ads* [174], which makes it difficult to the reader to differentiate between news and advertisement. This type of pseudo-content provides a significant part of the outlets' revenue [50].

Being able to detect this kind of behavior in the media is of utmost importance. For example, a socioeconomic bias in the media system can be very damaging as it may exacerbate the gap between rich and poor areas. A population with limited

access to the news is less inform and, consequently, less likely to hold authorities responsible for public expenditure and providing broad public goods [133, 77]. In turn, this motivates the incumbent to prioritize and divert resources to places where they will receive more media coverage and not necessarily where they are most needed. According to Chomsky and Herman [72], these characteristics make the news media system comparable to a political scheme where votes are weighted by income.

Other aspects of a community and their links to different socioeconomic conditions have been studied. For example, the diversity in the individuals' relationships [42] or patterns in the urban mobility [148] have shown to be useful indicators for the deprivation levels of a region. However, there has been comparatively little large-scale, quantitative research on the relationship between the media coverage, and the socioeconomic indices of a particular area.

6.1.3 Political bias

Political bias is probably the most studied type of bias in the mass media [121, 57, 180, 154], and we have cover this topic extensively in previous chapters. Specifically, in chapter (see chapter 3), we analyze the nature of bias through a political quiz. Our study shows that even the political bias could have some economic factors [44]. Extra evidence of this is given in [57]. The authors estimate the bias in newspapers according to how similar is the language to that used by congressmen for which a right/left stand is known. They do not find a direct relationship between the "slant" of a newspaper and the political preference of the owners. Instead, bias in the news is found to be more correlated to the political inclinations of the readers, showing a tendency in these news outlets to align themselves with the political preferences of their target audience and hence, maximizing selling profits. We think this is an important result because, although outlets may seem to take a political stand in their editorial line, evidence suggests that this may be another strategy to generate revenue by targeting a specific group of people. For example, governmental offices at various levels assign a considerable part of their budgets to advertising. Newspapers sympathizers of the government policies may benefit from lucrative advertising contracts with the incumbent. So, outlets discrimination can be also influenced by political reasons, with advertisers declining to do business with media that are perceived as ideological enemies or generally unfavorable to their interests.

6.1.4 Regression Model

We use a regression model to study the influence of the different features that represent the dimensions in our hypothesis, namely the socioeconomic and political characteristics of the communes that may attract profit-driven media coverage. We include again the geographic factor in our model to measure its influence and to keep a reference. We use as the geographic feature only the distance from the commune to the news source, given that the actual population is to closely related to our target variable

(a function on the number of followers). For the political factor, we use the right/left-leaning of the commune (see section 6.2). Meanwhile, we estimate the socioeconomic level of an area as its expected household income. Utilizing a regressor, trained on the political, economic and geographic features, we try to model the ranking of communes for each outlet based on the share of followers from each commune.

For the model we use a random forest regressor [19] (implemented in the module `RandomForestRegressor` within the python library `scikit-learn`). This estimator is based on classifying decision trees. As mention in a previous analysis, we prefer models based on decision trees as they are less susceptible to overfitting, considering that our training sets only have as many samples as communes with a valid entry for ech news outlet.

We evaluate the model using a random shuffle cross validation that leaves 20% of the dataset for testing, and trains the regressor in the remaining 80%. Each experiment is repeated 100 times and the average score and standard deviation are reported. We measure the quality of the fit with the explained variance.

We also measure the explanatory power of each individual dimension on the media coverage. For this, we calculate the Kendall's Tau correlation of the corresponding feature against the number of followers per commune for each news outlet. The results of these measurements should give information on the marketing strategy of different outlets.

To validate the model and test that the results are not an artifact of the social media itself, we repeat the experiment of predicting the ranking of communes based on the number of followers with a different dataset. This time we use the profiles of the Twitter followers for the players that were part of the Chilean national football[1] team in the "Copa America Centenario 2016" tournament. We expect the distribution of sport-celebrity fans to be influenced by different aspects and hence, the correlation to the investigated features should be significantly lower.

6.2 Data

6.2.1 Sources, collection process and pre-processing

Once again we make use of our news outlets database. Our database contains 403 *active* accounts. An account is considered *active* if it tweets at least once a month. We enriched the information of each outlet by adding relevant information such as geographic location, scope, number of Twitter followers, etc. We use the Twitter followers as a proxy for the actual audience of a news outlet. For our analysis, we download the user's profile on Twitter for each follower of the outlets in our database. We collected the profile of 4,943,351 unique users. Each user may simultaneously follow more than one news outlet.

[1] "Soccer", in the US dialect.

To find the users that are following more than one outlet we use the identifier from each user's profile on Twitter. From these users, we use only those that have a non-empty *location* field. That brings our list down to 1,579,068 accounts (31% of the initial amount).

In [70] the authors analyze the nature of the *location* field in the Twitter profile. Given that this is an open text field, users not always enter a valid (or even geographic) information. So, a pre-processing of these data is in order if we are making any study involving geolocation of the users based on this field (as we are).

In our remaining 31%, some of the users have GPS coordinates, and others have a text description of their location. Since the text description is a free text entered by the user, it ranges from an exact postal address to a completely useless text (e.g., "The milky way"). Using a gazetteer, we could extract 996,326 users with a recognizable location, which represents the 20% of the initial amount. We tried to assign each user to a commune with a given level of confidence. For the users with a pair of GPS coordinates, we used a shape-file [2] of the communes of Chile to find the one that enclosed the point. Only 4,829 of the users had GPS coordinates. The users with a text description making explicit mention of a commune were assigned to that commune. For those who mentioned only a province or a region, we could allocate them in the city/commune capital of that region. Given that these cities have the most prominent population density in the area, we would maximize the chance to be correct when making a guess. Nevertheless, we choose to work only with users for which we have high confidence in their location, namely: those with GPS coordinates or explicit mention of a commune. Thus, our final list contains 602,810 users, which is over 10% of the total number of unique followers. We decided to use the commune as our location unit given that this is the smaller political division in Chile, but at the same time, it is big enough to create both a statistical and popularly perceived socioeconomic profile at the population level.

We obtained the total actual population of each commune and some other demographic indices from the National Institute of Statistics (INE) [74]. The demographic indices were already aggregated by commune.

We also needed information on the socioeconomic development of each zone. This kind of information is harder to obtain. The most reliable source is the national census. The problem is that censuses are very expensive and therefore are performed very infrequently (sometimes more than a decade apart - last completed valid census performed in Chile was in 2002)[2]. Instead, we use the data from the National Socio-Economic Characterization Survey (CASEN) from 2013[3] [149]. This study is conducted by the Ministry of Social Development in Chile. From the CASEN survey, we can obtain the socioeconomic indicator at the level of commune for all Chile. This data is extracted by using the expansion factor to calculate the weighted average income per home for

[2]There was another census in 2012, but it was methodologically flawed, with problems in coverage, and a supposed manipulation of some of the key indices [18].

[3]There is a CASEN survey from 2015 but the expansion factor for the communes is not complete (not even for the communes in Santiago)

each commune.

Our last dimension has to do with the political leaning of the communes. To measure the political tendencies of each geographical area, we use the results from the presidential election. The Chilean Electoral Service [141] provides detailed information district-wise on the Chilean presidential elections since 1989 (that is, since Chile's return to democracy after the dictatorship of Augusto Pinochet).

We aggregated the raw number of votes receives by each party on each commune in the past three elections. We manually annotated each political party as left-wing, right-wing or centrist according to their self-declared position. Political parallelism on the media system is seen when the media outlets are popularly perceived as leaned to one broad side in the political spectrum (not necessarily linked to a political party but rather to a political range) [66]. So, we aggregated the votes for all parties that have a similar political ideology. With this, we measure how "left-leaning" or "right-leaning" is a commune.

The data in all three dimensions was aggregated at the level of communes and normalized by calculating the z-score of each area on each feature.

6.2.2 Chilean national soccer team followers

To compare, we used sport-celebrity followers to discard that the patterns found in the media are endogenous to the social media and not to the selected newspaper accounts. We repeated the procedure describe above to download and filter the Twitter followers, this time using as the "followee", the players that integrated the Chilean national football team who participated in the "Copa America Centenario 2016" tournament.

We download the Twitter's profile of 6,568,769 unique users that follow at least one of the 21 players for which we were able to find an official Twitter account. From these, only 2,434,183 had a non-empty *location* field. We followed the same methodology for the geolocation of these user. We found 540,828 users that match a valid location in Chile. Out of the valid users, only 2,041 had a GPS set of coordinates, and 381,166 did explicit reference to a commune. This gave us a total of 383,207 unique followers that we were able to assign with high confidence to one of the 346 communes in Chile. This is comparable with the 602,810 users that we will use as our sample of followers of the news outlets. In this new dataset, the number of follower per commune is also strongly correlated with the actual population distribution of Chile ($r(344) = .66, p < .01$). So, it is comparable in size to our newspapers-followers dataset and it is representative sample of the actual Chilean population.

6.3 Results

6.3.1 Gravity Model

We represent each outlet as a vector F^j with the expected number of followers in each commune obtained from the Gravity Model. We also created a vector T^j for each outlet with the actual number of followers on each commune i obtained for our ground-truth. Using the two vector that represents each outlet we calculate the Pearson correlation coefficients. In Figure 6.1 we can see the distribution of the correlation coefficients. We can see that the coverage bias of a big number of outlets can be almost entirely explained just by the geography. Actually, half the outlets correlate over 0.7. However, there is an important number of outlets for which the geographic bias explains very little or none of their observed coverage.

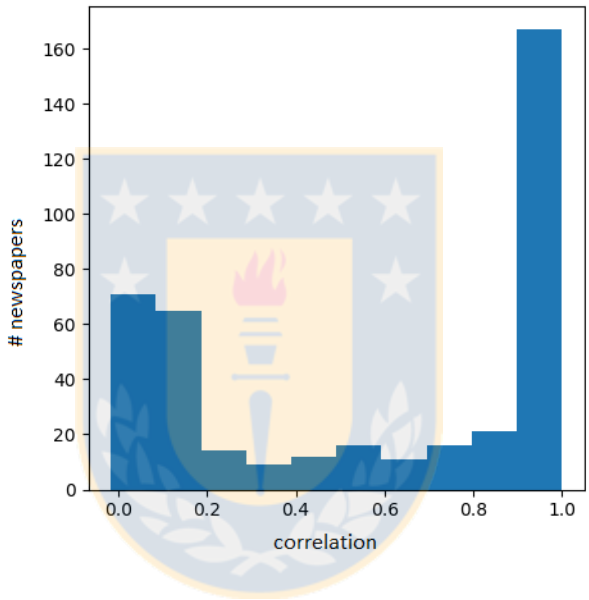


Figure 6.1: Distribution of the correlation coefficients (Gravity Model vs. Ground-Truth) for **All** news outlets

Table 6.1 shows some stats that help to better describe the characteristics of the news outlets with the lowest and highest correlation. Not surprisingly, the group that falls farther from the predicted coverage is dominated by the newspapers in the capital city (i.e., Santiago) and with a national scope. These are expected to be the ones with the most prominent political and socioeconomic bias, given that they are the most influential and the ones that dominate news production. Their leading position also ensures that they receive the biggest share in the investment of advertisers. Hence, these outlets are the most exposed to external pressures. On the other hand, news outlets with a local scope behave as described by the Gravity Model, at least in average. Figures 6.2 and 6.3 show the distribution of the correlation for the outlets with a local and national scope respectively. The figures illustrate the behavioral difference of

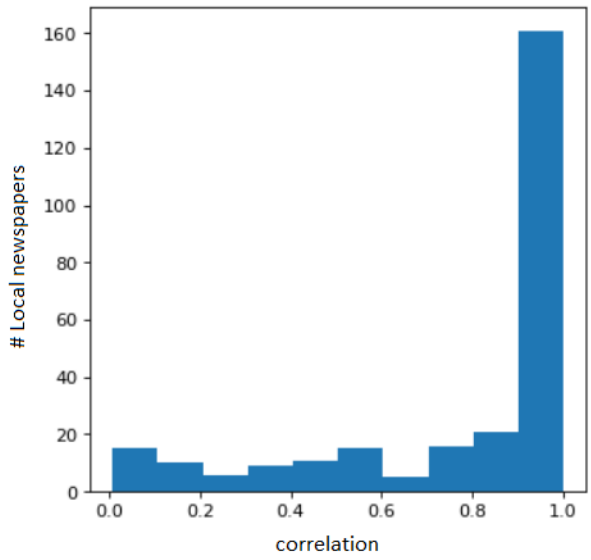


Figure 6.2: Distribution of the correlation coefficients (Gravity Model vs. Ground-Truth) for **Local** news outlets

these two classes of outlets.

Category	Total	$\rho > 0.7$	$\rho < 0.2$
Outlets	402	203	141
w/ nacional scope	133	23	94
w/ local scope	269	180	47
located in Santiago	156	29	108
w/ nacional scope & located in Santiago	126	18	93

Table 6.1: Stats about the news outlets' correlation coefficients.

From the previous results, we can conclude that geographic bias is not enough to describe the nature of the news media. If we look for example at the communes *Lo Prado* [172] and *San Miguel*[176], they have a similar population and are situated at a similar distance from the center of Santiago, where an important number of news outlets (local and national) are located. If we take only these news outlets located in the center of Santiago, the average difference in the expected number of followers between the two communes according to the Gravity Model is just over 1%. In other words, based only on geographic factors these communes should be virtually indistinguishable. But, if we look at the actual number of follower for the same set of outlets, the average difference is almost 250%, with an overwhelming dominance of followers from *San Miguel*. Moreover, a general query in the Twitter API for tweets geo-located near "Lo Prado, Chile" during August 2017, gives almost 100,000 unique users, while

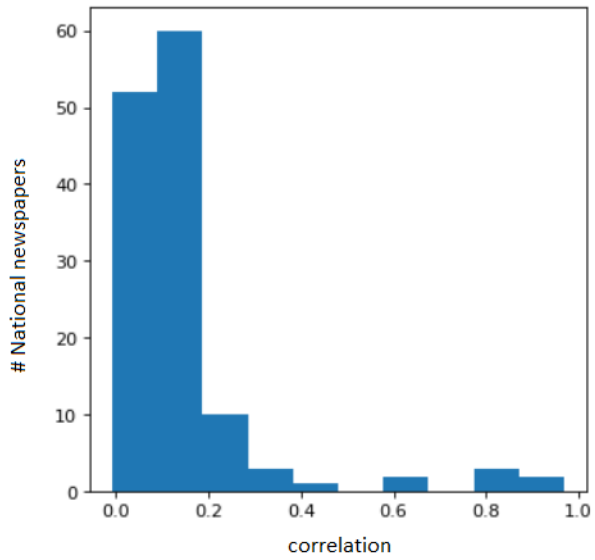


Figure 6.3: Distribution of the correlation coefficients (Gravity Model vs. Ground-Truth) for **National** news outlets

the same query for tweets near “San Miguel, Chile” throws only 61,165 unique users. Thus, the difference in news outlets’ followers cannot be thought as the result of a disparity in Twitter penetration. One possible factor that may influence this striking contrast is the gap in socioeconomic conditions and deprivation levels between the two communes. *Lo Prado*, despite being located within the capital city, is ranked in the top ten of the poorest communes of Chile [117]. On the other hand, *San Miguel*, even though it is a predominantly residential commune, it is also an important economic/industrial pole of the city. In fact, *San Miguel* ranks in the 40^{th} position out of 93 communes in the Index of Urban Quality of Life for Chile [117]. Thus, the hypothesis of our theoretical models in the political economy of the mass medial [72, 129] support the idea that the socioeconomic characteristics of a sector can make its population more or less attractive to the media.

6.3.2 Filtering the data

Given that we are analyzing geographic coverage and its relation to socioeconomic and political factors, we have to take into account the specific characteristics of Chile. From every point of view, Chile is a heavily centralized country. The previous results detailed in Table 6.1 give evidence of this. According to a study from 2013 [163], in proportion to its size, population, and economic development, Chile is the most centralized country in Latin America. The data obtained from the INE [74] gives us a total estimated population of 17.9 million people for the entire country, out of which 7.4 million (41%) are located in the Metropolitan Region (where its capital, Santiago, is located). If we also add that this is the smallest (in area) of the 15 regions that compose Chile, we

have a very dense population area. Only for its geographic demographic characteristics, Santiago is already a desirable market for the media based on *Proposition 4* of Prat and Strömberg [129]. Now, on the political side, each region in Chile is headed by an *Intendente* (equiv. Mayor), but they are appointed and respond directly to the president. Moreover, members of the House of Representatives who legislate on behalf of the different districts of the country reside in Santiago. This organization concentrates all the political power in Santiago. In the same way, according to the annual report published by the Central Bank of Chile for 2016 [12], the Metropolitan Region participated with 46% of the GDP (5x the next highest contribution). With this heavily concentrated power in all spheres, and based on our set of hypothesis, the capital of Chile checks all the right boxes to receive an extensive media coverage.

Consequently, based on our results of the Gravity Model, we will focus on the community of outlets identified as the least influenced by the geographic bias. That is, we filtered our database to keep only those news outlets (locals and national) with headquarter in the capital. The centralization of the Chilean population it is also perceived in our collection of followers: out of 15 regions, 36.9% of our geolocated followers are in the Metropolitan Region. To minimize the noise in our model, we decided to limit the study of the coverage only to the communes in Santiago. With 51 communes and a wide range of socioeconomic conditions, the Metropolitan Region offers a good case of study on its own. To further strengthen the signal, we also limited the analysis to the 25 news outlets with the highest number of followers.

6.3.3 Regression Model

To extend our model and study the influence of other factors such as the political and socioeconomic characteristics in the distribution of news media followers, we use a regression model. As mention before, our target variable is, given a news outlet and a commune, the ranking position based on the number of followers of that communes for that outlet.

We include three features in our model: `right-leaning`, representing the political dimension; `income`, representing the socioeconomic dimension; and `distance`, representing the geographic dimension. In table 6.2 we show the Pearson correlation between our three features (using the filtered data). One thing to notice is the high correlation between the expected income of an average household and the political leaning of the area where it is located. In Chile (and Latin America in general), right-conservative political parties are popularly associated with wealthy people. At least in the last few year, left-leaning parties tend to be more populists.

Using these features, our trained model is able to represent the mass media behavior with high precision. The results of the regression indicate the three predictors explained on average up to 93.9% $(SD = 0.01)$ of the variance in cross-validation. Figure 6.4 shows the learning curve for the selected model.

We were also interested in modeling the coverage behavior of each individual outlet to see how they fit with respect to these three dimensions. To do this, we used the

Feature	right-leaning	income	distance
right-leaning	1.00	0.82	-0.11
income	0.82	1.00	-0.34
distance	-0.11	-0.34	1.00

Table 6.2: Correlation between features

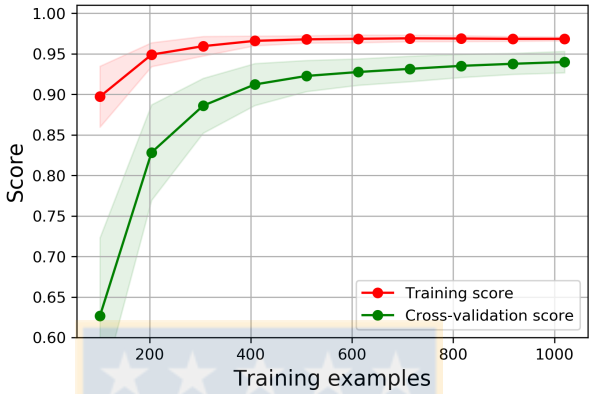


Figure 6.4: Learning curve of the Random Forest regressor model for the top 25 news outlets in Santiago. Each step represents the average of 100 iterations of shuffle split cross-validation with 20% of the data for validation.

selected features to create a regression model for each news outlet. This model is then used in the same way. That is, we predict their audience-based ranking of the communes in Santiago but using data related only to the selected outlet. The results, shown in figure 6.5, confirm that with the selected features, the regressors are able to approximate the distribution of followers very well $(M = 0.76, SD = 0.05)$.

We also studied the ranking of the communes in relation to each feature. We used Kendall's Tau (KT) correlations to have an indication of how strong is the influence of each factor in the prediction. Figures 6.6, 6.7 and 6.8 show the distribution of the KT correlation for the top 25 news outlets in Santiago with respect to the communes' political leaning, expected income and distance to the origin, respectively. Results are shown regarding their absolute values because the direction of impact is not important for our model. For example, if a news outlet favors a commune based on the area being right-leaning, for our model this is as telling as another news outlet disregarding the commune for the same reason. In both cases, the outlets are biased based on political factors. The results show that the behavior of news outlets is very similar in terms of the discriminating influence of these three dimensions in the news coverage, at least within this group.

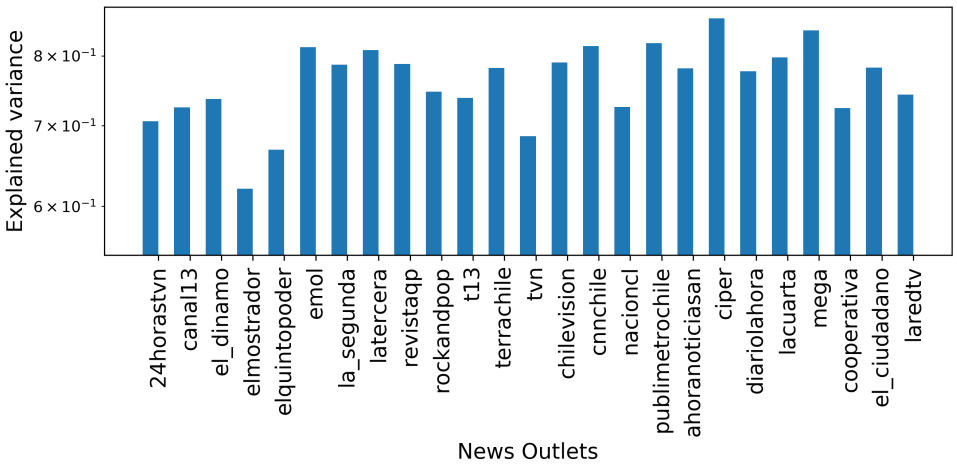


Figure 6.5: Explained variance using the regressor model for the top 25 news outlets in Santiago.

In figure 6.9 we show a comparison of the KT correlation coefficient for all three dimensions for each of the top 25 news outlets. This comparison can be used as a characterization/profiling of each outlet's coverage behavior. For example, the coverage of Radio Cooperativa (*cooperativa*) [175] seems to be driven by political and economic factors, with practically no attention to the location of the commune. This is a radio station with a national scope. According to a survey conducted in 2015, it is the second in audience in the region of Santiago [75] and the first one among people with the highest income (last quintile). Moreover, its editorial line is "popularly perceived" to be associated to the Christian Democratic Party [175, 44]. Actually, from the early 70's until the late 90's the radio was directly owned by this party (currently belongs to El Mercurio Group). This profile coincides with the characterization reflected by our model. On the other hand, El Quinto Poder (*elquintopoder*) [51] is an online news website/community where any member can contribute with its column. This newspaper follows the concept of citizen journalism popularized by sites like `http://www.ohmynews.com/`. Its editorial line and community rules explicitly prohibit any content that is aimed at a personal or institutional gain. In the same way, political opinions can only be expressed through personal profiles (rather than an organization profile). In our model, for this newspaper the influence from the political and economic factors are equated, but also the geographic dimensions is the highest within these top 25 outlets.

Just as a comparison, we repeat the analysis, this time filtering the dataset to keep only the top 25 "newspapers" in Santiago - i.e., excluding radio station, TV channels, etc. (see figure 6.10). Here, for example, it is easy to distinguish newspapers with a local scope, such as *portaldemeli* or *betazeta*. For those, the influence of the geographic factor is higher than that of the economic and even the political features. Actually, in the case of *portaldemeli* (a small commune's local digital newspaper), the economic factor is almost non-existing.

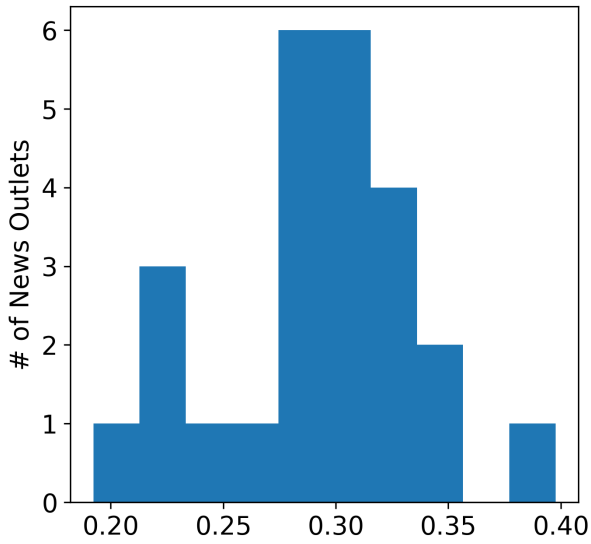


Figure 6.6: Distribution of the KT correlation for communes in the Metropolitan Region. Number of followers vs. Right-leaning.

To validate the results of our method we repeat the experiments using the information for the followers of a group of football players (see section 6.1.4). We again filtered the data to keep only football players that were born, play or live in Santiago (this condition matched six players). We also kept only the followers that were geolocated in one of the 51 communes of the capital region. The regression model trained with the three selected features, on average, is able to explain only 84% $(SD = 0.06)$ of the variance in cross validation. Although the model also gives a good fit for this data, it is clearly less explanatory than for the news outlets (over 10% lost of precision compared to the news outlets). Notice that it is very difficult, if not impossible, to find a public/popular figure for which the followers are not influenced by neither of these three factor. So, the results must be evaluated relative to each other.

Another way to differentiate the two datasets is by comparing the individual influence of each dimension. We calculated the KT correlation coefficient for all three dimensions for each of the top 6 players (see figure 6.11). We found that, compared with the news outlets, the difference in the average correlation is statically significant for all three features (right-leaning: $t = 8.31, p < .001$; income: $t = 7.93, p < .001$; distance $t = 2.39, p = .03$).

We can also see that the profile obtained for each individual player (shown in the shaded area in figure 6.9) differs from that of the news outlets. In this case, football players tend to have a comparatively stronger influence from the geographic factor and politics plays a lesser role.

The found differences between the two datasets indicate that the distribution of followers for the news outlets is not determined by the social media but is defined but the characteristics of the entities.

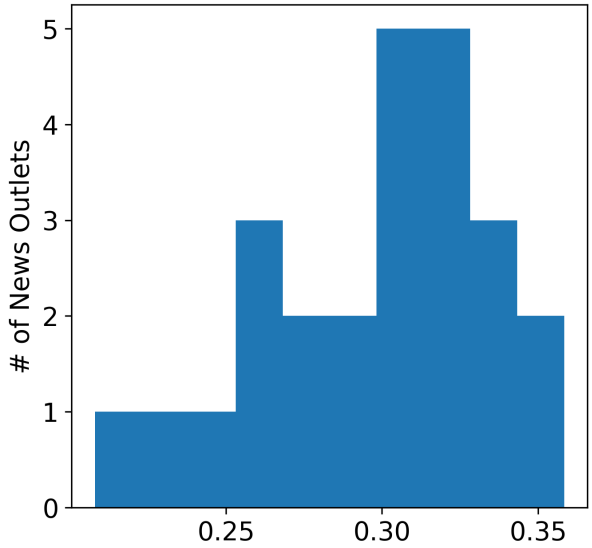


Figure 6.7: Distribution of the KT correlation for communes in the Metropolitan Region. Number of followers vs. Avg. Income.

In general, our results support the idea of a media system entirely motivated by economic interest as described by the Propaganda Model. This profit-driven media system seems to promote selective coverage that targets specific segments of the population based on the "quality" of the readership.

6.4 Conclusions

This chapter presents a method to characterize the news outlets in the media system based on the geographic, socioeconomic and political profile of their audiences. Under the assumption of a natural order of information demand and supply (i.e., readers gets some value from the news [15]), this modeling of the media can imply a conscious targeting of some specific public by catering to their preferences.

Using data from multiple sources we found that news outlets systematically prefer followers from densely populated areas with an specific socioeconomic profile. The political leaning of the commune proved to be the most discriminating feature on the prediction of the level of readership ratings. These findings give supportive evidence to the second filter of the Propaganda Model that describe the news media outlets as profit-driven companies.

Although our model seems to generalize quite well and lends evidence to the hypotheses, we recognize that our method has some limitations. First, we are restricted in the amount of users that we are able to geolocate using only the *location* field from the Twitter profile. A more sophisticated method of location could increase the number

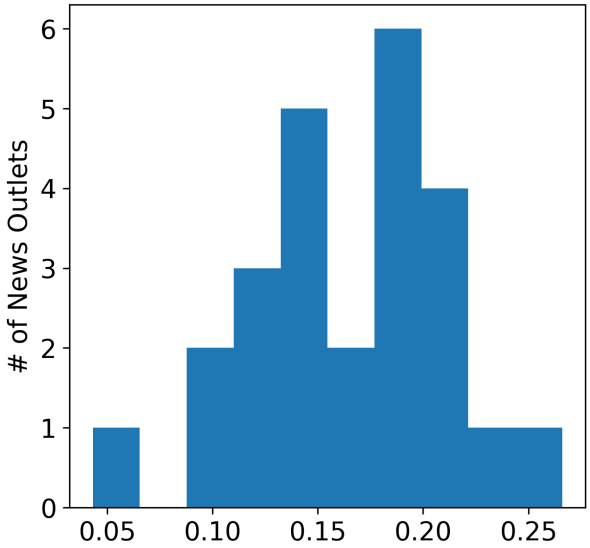


Figure 6.8: Distribution of the KT correlation for communes in the Metropolitan Region. Number of followers vs. Distance.

of valid users and maybe increase the precision of the model for other regions. A second limitation comes from the fact that the political and economic dimension seem to be closely related. This prevents us from creating a characterization of the outlets that better reflect the actual preference for a population with either a certain political profile or a socioeconomic range, but not both. The entanglement of these two dimension may be due to the reality of the studied country.

In summary, the results seem to support the hypothesis that outlets focus on reaching and acquiring an audience with a higher "quality", that can be latter sell to advertisers. This type of media system neglect areas of low population (e.g., rural communes) and high deprivation levels, causing these to be underserved and underrepresented in

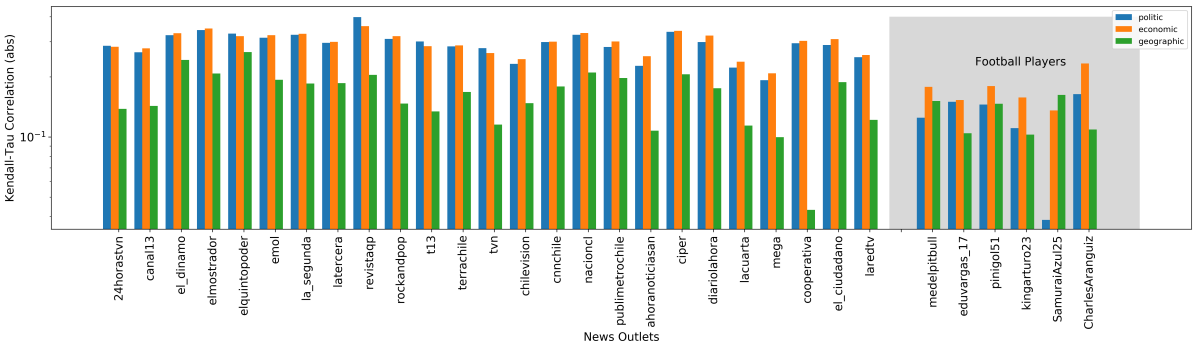


Figure 6.9: KT correlation for communes in the Metropolitan Region for top 25 news outlets. Each feature is correlated with the number of followers' ranking

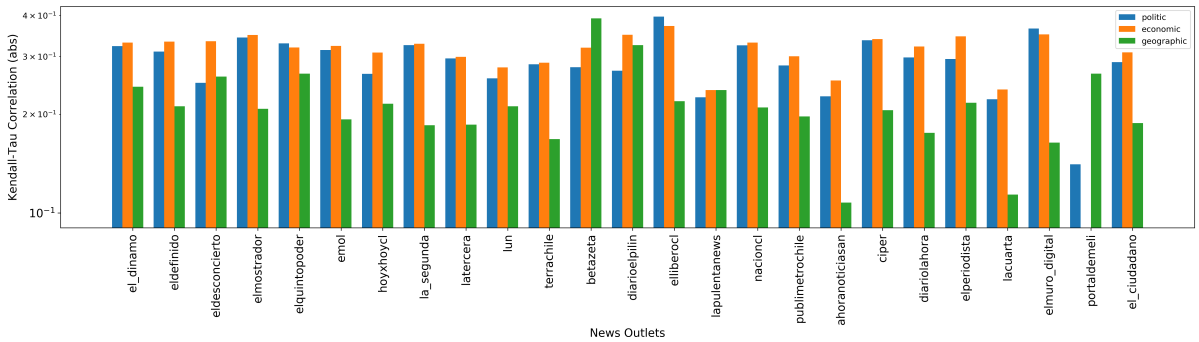


Figure 6.10: KT correlation for communes in the Metropolitan Region for top 25 **newspapers**. Each feature is correlated with the number of followers ranking

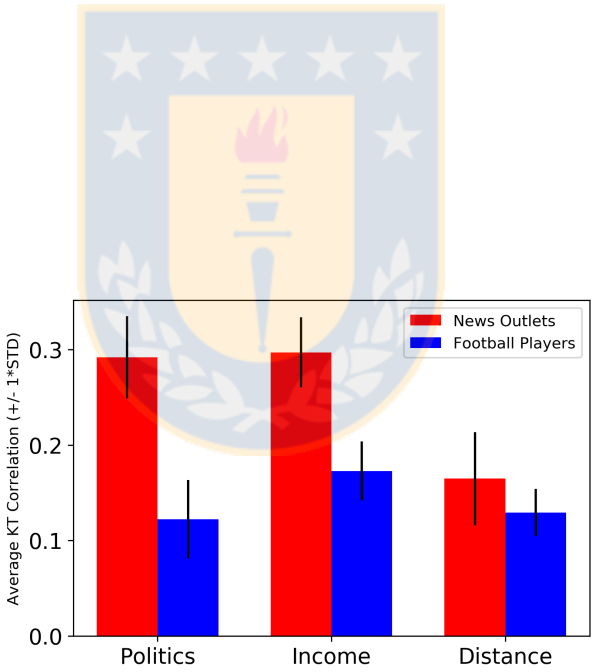


Figure 6.11: KT correlation for communes in the Metropolitan Region comparing News Outlets and Football players behaviour. Each feature is correlated with the number of followers' ranking of the corresponding dataset.

the news coverage. In turn, this creates a full cycle when public policies and politicians overlook sectors of the population that are less informed and hence, are less likely to influence the *status quo* of the political elite.

Chapter 7

Future Work

This thesis opens the door to an extensive line of research. The complexity of each filter in the Propaganda Model allows for multiple analysis, each with various dimensions and alternative points of view. We were able to cover just a small fraction of the topic. In this chapter we want to mention at least some of the extension that we think may contribute to continue expanding our capability to automatically model and predict the behaviour of the mass media and, in general, enrich the set of tools we have to analyze political economy theories using data science.

In this work we were interested in the behaviour and bias displayed by each different outlet in the media landscape, regardless of the size. This notwithstanding, we wanted to make sure we were covering the entire spectrum of the Chilean media. Our methodology is able to to characterize the outlet's position in the political and socioeconomic context of the Chilean system, even it published just a few tweets. The evident extension of a more general analysis that takes into account the weight that each news outlet contributes to the global media bias is left is an interesting path to explore in a future work.

Also, it would be interesting to compare our results to similar analysis conducted over full-text articles published by the same news outlets. As discussed during the thesis, this will require more sophisticated NLP tools and more human supervision, but it could shed some light on the similarities and differences between traditional media and social media.

The model used to analyze the nature of the media bias could benefit from adding some other content features (e.g., leaning of the named entities) to the polarity classification of the tweets as these may help to refine the relative positioning found by our model.

For that same part of our study, we are also interested to see what is the most accurate way to score the missing answers (i.e., outlets for which we did not find any tweet related to some of the questions in the quiz). Since "coverage" is a form of bias [137], perhaps the outlet is not being *neutral* by not mentioning a specific subject. Even when the decision of which stories/events are newsworthy is subjective and depends on the editorial strategy [159], there are some events that are very relevant in the national context and are covered for the majority of the media. So, a complete silence of a news outlet on such an event may be interpreted as something other than neutrality.

For example, question **q7** is related to international free trade. Taking the number of tweets and re-tweet as an indicator of important events [102], we can see in Figure 7.1 that this topic has had at least one major event during this period. This event was

the ascription of Chile to the Trans-Pacific Partnership (TPP) signed by the country on Feb 3th, 2016. Despite the magnitude of the event, only 135 out of 198 newspapers with a section on politics mentioned it. A plausible cause is that the other news outlets decided not to report about this event, in other words 'bias by omission'.

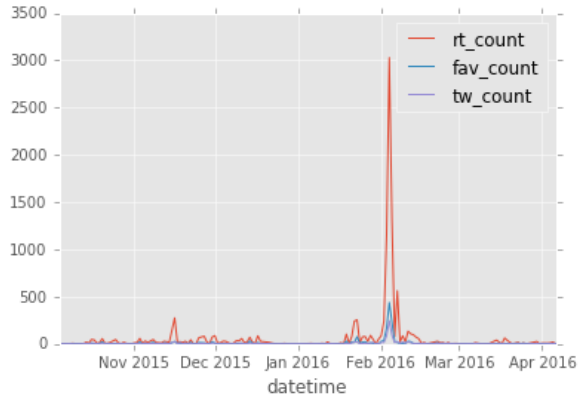


Figure 7.1: Counting tweets and re-tweets for **q7** under president Bachelet

We show that a careful selection or update of the questions of the *PolQuiz* may lead to a significant improvement in the results. If we have an inside understanding of the socio-economic environment from where the news are being collected, then we could replace the questions to capture more relevant topics. In this sense, we could benefit from advances in systems that focus on identifying controversial topics in social media [53]. On the other hand, if we do not have any intuition on the news collected, then we can accumulate the new questions so we can widen the spectrum of topics and have a better chance of capturing relevant events/discussions with our queries.

In our investigation of the second filter we studied the selective news coverage based on political and socioeconomic characteristics of the target area. We geolocated the followers of each news outlet based on their self reported location in their Twitter's profile. This limited our analysis to 10% of the followers. An examination with a bigger representation of the population, involving more sophisticated methods of geo-tagging could improve the results.

Besides the limitations in the location of the users, the choice to use Twitter followers as a proxy for the audience of the news outlet may introduce some bias and noise to our study. For example, it is difficult to determine the actual demographics of the population in the social media [27]. An alternative method to effectively define the actual audience of the news outlets (e.g., monitoring the passive and active traffic on the selected Twitter accounts or websites) could complement our method and improve the predictive capability of our model in areas with a weaker signal (e.g., beyond the Metropolitan Region). This is left for future work.

Chapter 8

Conclusions

In this work we show empirically evidence in support of a major socioeconomic theory: the Propaganda Model, put forth by Herman and Chomsky [72]. The authors of the theory try to support it by discussing a few anecdotal examples of the different treatment of the media to similar events. Here, we make a more exhaustive analysis using a computational approach. Previous work has had used machine learning and natural language processing techniques, but they have focused only on showing some leaning to a political party by a sample of the major news outlets. In our analysis, we used the footprint left by news outlets and their followers in social medias to study its behavioral patterns at a bigger scale. At this point, we have started a first attempt at the formal definition of the model and the filters. This has guided the studies in the work we report across this thesis.

Results show that the media have a measurable bias, and illustrate this by showing the favoritism of Chilean media for the ruling political parties in the country. This favoritism becomes clearer as we empirically observe a shift in the position of the mass media when there is a change in government. Even though relative differences in bias between news outlets can be observed, public awareness of the bias of the media landscape as a whole appears to be limited by the political space defined by the news that we receive as a population. As predicted by the Propaganda Model, we found that the nature of the bias is reflected in the vocabulary used and the entities mentioned by different news outlets.

We have studied the topic similarities in news reported by different outlets by clustering their tweets. We found important similarities between outlets with the same owner, showing indications of a global policy rather than individual guidelines of each editorial institution. A network analysis reveals that Chilean media is highly concentrated both in terms of ownership as well as in terms of topics covered. Moreover, we found evidence showing that newspapers and, in a higher degree, owners are pushing their own agenda by the introduction of topics that are newsworthy almost exclusively to them. This is in tune with the first filter of the Propaganda Model (ownership). Our methods can be used to determine which groups of outlets and ownership exert the greatest influence on news coverage.

Our studies on the geographic news coverage also give indications of the presence of another filter (advertising). We found that major news outlets are not interested in covering the entire territory, but just those regions with the greatest population and better socioeconomic status. The political leaning of the commune proved to be the most discriminative feature on the prediction of the level of readership ratings. Anticipating

the location of the audience based on these features is of key importance in our investigation to better understand the dynamics of news system. Our experiments on predicting the communes with the biggest share of readership show promising results for this approach.

As far as we know, this is the first time that there has been an attempt to empirically prove this political economy theory using data science. Having a more accurate method to measure and characterize the media behavior will help readers position outlets in the socioeconomic landscape, even when a (sometimes opposite) self-declared position is stated. This will empower readers to better reflect on the content provided by their news outlets of choice.

Bibliography

[1] Oluwaseun Ajao, Jun Hong, and Weiru Liu. A survey of location inference techniques on twitter. *J. Inf. Sci.*, 41(6):855–864, December 2015.

[2] Christoph Albers. Mapas de las provincias y regiones de chile. Cartografa Rulamahue. Available from: `http://www.rulamahue.cl/` [Accessed 18-October-2017], 2016.

[3] Scott L. Althaus, Jill A. Edy, and Patricia F. Phalen. Using substitutes for full-text news stories in content analysis: Which text is best? *American Journal of Political Science*, 45(3):707–723, 2001.

[4] David L. Altheide. Terrorism and the politics of fear. *Cultural Studies Critical Methodologies*, 6(4):415–439, 2006.

[5] Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. Web-a-where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 273–280, New York, NY, USA, 2004. ACM.

[6] Jisun An, Meeyoung Cha, Krishna Gummadi, and Jon Crowcroft. Media landscape in Twitter: A world of new conventions and political diversity. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, Menlo Park, CA, USA, 2011. AAAI.

[7] Simon P Anderson and John McLaren. Media Mergers and Media Bias with Rational Consumers. CEPR Discussion Papers 7768, C.E.P.R. Discussion Papers, March 2010.

[8] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 61–70, New York, NY, USA, 2010. ACM.

[9] Ben H Bagdikian. *The new media monopoly: A completely revised and updated edition with seven new chapters*. Beacon Press, 2014.

[10] Jorge Bahamonde, Johan Bollen, Erick Elejalde, Leo Ferres, and Barbara Poblete. Power structure in chilean news media. *CoRR*, abs/1710.06347, 2017.

[11] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[12] Banco Central de Chile. Cuentas nacionales de chile. PIB regional 2016. Available from: http://www.bcentral.cl/documents/20143/32019/CCNNPIB_Regional2016.pdf/90a16087-69d8-fcc6-cfe1-5f2ce741f40e [Accessed 30-December-2017], 2017.

[13] Tabe Bergman. The case for a dutch propaganda model. *International Journal of Communication*, 8(0), 2014.

[14] Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology*, 2015.

[15] Timothy Besley and Andrea Prat. Handcuffs for the grabbing hand? media capture and government accountability. *American Economic Review*, 96(3):720–736, June 2006.

[16] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583, 2009.

[17] Oliver Boyd-Barrett. Judith miller, the new york times, and the propaganda model. *Journalism Studies*, 5(4):435–449, 2004.

[18] David Bravo, Osvaldo Larra naga, Isabel Milln, Magda Ruiz, and Felipe Zamorano. Informe final comisin externa revisora del censo 2012. Technical report, PNUD, Santiago, Chile, 2013. http://www.cl.undp.org/content/dam/chile/docs/pobreza/undp_cl_pobreza_informe_censo_2013.pdf?download.

[19] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[20] Andrei Z Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings*, pages 21–29. IEEE, 1997.

[21] Andrei Z Broder. Identifying and filtering near-duplicate documents. In *Combinatorial pattern matching*, pages 1–10. Springer, 2000.

[22] Andrei Z Broder, Moses Charikar, Alan M Frieze, and Michael Mitzenmacher. Min-wise independent permutations. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 327–336. ACM, 1998.

[23] Maurice Bryson and William McDill. The political spectrum: A bi-dimensional approach. *Rampart Journal of Individualist Thought*, 4(2):19–26, 1968.

[24] Gregory Buehrer and Kumar Chellapilla. A scalable pattern mining approach to web graph compression with communities. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 95–106. ACM, 2008.

[25] Camara de Diputados de Chile. Trabajo en sala. boletn no. 9895-11. Available from: https://www.camara.cl/trabajamos/sala_votacion_detalle.aspx?prmID=23099 [Accessed 12-August-2017], 2016.

[26] Cosette Castro. Industrias de contenidos en latinoamérica. *Santiago de Chile: Cepal*, 2008.

[27] Nina Cesare, Christan Grant, and Elaine Okanyene Nsoesie. Detection of user demographics on social media: A review of methods and recommendations for best practices. *CoRR*, abs/1702.01807, 2017.

[28] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.

[29] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Michael Mitzenmacher, Alessandro Panconesi, and Prabhakar Raghavan. On compressing social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 219–228. ACM, 2009.

[30] Noam Chomsky. *The common good*. Odonian Press Distributed through Common Courage Press/LPC Group, Monroe, ME, 1998.

[31] Noam Chomsky. *Necessary illusions : thought control in democratic societies*. House of Anansi Press, Toronto, 2003.

[32] J. R. Clark, Ashley S. Harrison, and Bradley K. Hobbs. The current status of free enterprise chairs and professorships in academe. *Journal of Private Enterprise*, 26(2):15–46, Spring 2011.

[33] Raviv Cohen and Derek Ruths. Classifying political orientation on twitter: It's not easy! In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM*. The AAAI Press, 2013.

[34] M Conover, B Gonçalves, J Ratkiewicz, A Flammini, and F Menczer. Predicting the political alignment of twitter users. In *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom)*, 2011.

[35] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Political polarization on twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[36] John Corner. The model in question: A response to klaehn on herman and chomsky. *European Journal of Communication*, 18(3):367–375, 2003.

[37] David Cromwell and David Edwards. Guardians of power: The myth of the liberal media. *London: Pluto*, 2006.

[38] Alexander Dallmann, Florian Lemmerich, Daniel Zoller, and Andreas Hotho. Media bias in german online newspapers. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pages 133–137, New York, NY, USA, 2015. ACM.

[39] Elynor Davis and Darrell F. Parker. Student attitudes toward regulation, politics, and free enterprise. *Journal of Legal, Ethical and Regulatory Issues*, 7(1):155–168, 2004. Copyright - Copyright The DreamCatchers Group, LLC 2004; Document feature - ; Last updated - 2013-03-20.

[40] Daniel Dor. On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 35(5):695 – 721, 2003.

[41] Sergio Godoy E. Media ownership and concentration in chile. In *Who Owns the World's Media?*, pages 641–673. Oxford University Press, jan 2016.

[42] Nathan Eagle, Michael Macy, and Rob Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

[43] Rick Edmonds, Emily Guskin, Tom Rosenstiel, and Amy Mitchell. Newspapers: Building digital revenues proves painfully slow. *The State of the News Media. Journalism & Media. Pew Research Center*, 2012. http://assets.pewresearch.org/wp-content/uploads/sites/13/2017/05/24141622/State-of-the-News-Media-Report-2012-FINAL.pdf [Accessed 20-May-2016].

[44] Erick Elejalde, Leo Ferres, and Eelco Herder. The nature of real and perceived bias in chilean media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, pages 95–104, New York, NY, USA, 2017. ACM.

[45] Robert M. Entman. Dissent on manufacturing consent: A reply. *Journal of Communication*, 40(3):190–192, 1990.

[46] Hans Eysenck. *The Psychology of Politics*. Routledge, 1988.

[47] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[48] Ilias Flaounas, Omar Ali, Thomas Lansdall-Welfare, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. Research methods in the age of digital journalism. *Digital Journalism*, 1(1):102–116, 2013.

[49] Ilias Flaounas, Marco Turchi, Omar Ali, Nick Fyson, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. The structure of the eu mediasphere. *PLoS ONE*, 5(12):e14243, 12 2010.

[50] Paul Fletcher. Native advertising will provide a quarter of news media revenue by 2018. Forbes. Available from: https://www.forbes.com/sites/paulfletcher/2016/11/30/native-advertising-will-provide-a-quarter-of-news-media-revenue-by-2018/#75950afa2d0c [Accessed 12-August-2017], 2017.

[51] Fundacin Democracia y Desarrollo. El quinto poder. FD+D. Available from: http://www.elquintopoder.cl/ [Accessed 30-December-2017], 2010.

[52] Marco Gambaro and Riccardo Puglisi. What do ads buy? daily coverage of listed companies on the italian press. *European Journal of Political Economy*, 39(Supplement C):41 – 57, 2015.

[53] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy in social media. *CoRR*, abs/1507.05224, 2015.

[54] Venkata Rama Kiran Garimella and Ingmar Weber. Co-following on twitter. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 249–254, New York, NY, USA, 2014. ACM.

[55] Gary J. Gates. Majority of u.s. voters think media favors clinton. Gallup. Available from: http://www.gallup.com/poll/197090/majority-voters-think-media-favors-clinton.aspx [Accessed 12-August-2017], 2016.

[56] Ruth Olimpia G. Gavilanes, Diego Gomez, Denis Parra Santander, Christoph Trattner, Andreas Kaltenbrunner, and Eduardo Graells. Language, twitter and academic conferences. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, pages 159–163, New York, NY, USA, 2015. ACM.

[57] Matthew Gentzkow and Jesse M. Shapiro. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71, 2010.

[58] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[59] Jennifer Golbeck and Derek Hansen. Computing political preference among twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1105–1108, New York, NY, USA, 2011. ACM.

[60] Brian Michael Goss. *Rebooting the Herman & Chomsky propaganda model in the twenty-first century*. Peter Lang, 2013.

[61] Thomas Gryta, Keach Hagey, Dana Cimilluca, and Amol Sharma. AT&T reaches deal to buy Time Warner for $85.4 billion. *The Wall Street Journal*, 2016. https://www.wsj.com/articles/at-t-reaches-deal-to-buy-time-warner-for-more-than-80-billion-1477157084 [Accessed 12-December-2017].

[62] Y. Gu, T. Chen, Y. Sun, and B. Wang. Ideology Detection for Twitter Users with Heterogeneous Types of Links. *ArXiv e-prints*, December 2016.

[63] Roger Guimerà and Luís A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, feb 2005.

[64] Stuart Hall, Chas Critcher, Tony Jefferson, John Clarke, and Brian Roberts. *Policing the crisis: Mugging, the state and law and order*. Palgrave Macmillan, 1978.

[65] D.C. Hallin. *We Keep America on Top of the World: Television Journalism and the Public Sphere*. Communication and society. Routledge, 1994.

[66] D.C. Hallin and P. Mancini. *Comparing Media Systems: Three Models of Media and Politics*. Communication, Society and Politics. Cambridge University Press, 2004.

[67] James T. Hamilton. *All the News That's Fit to Sell: How the Market Transforms Information into News*. Princeton University Press, Princeton, New Jersey, 2004.

[68] Jesse Hearns-Branaman. A political economy of news media in the peoples republic of china. *Westminster papers in communication and Culture*, 6(2), 2009.

[69] J.O. Hearns-Branaman. *The Political Economy of News in China: Manufacturing Harmony*. Lexington Books, 2016.

[70] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 237–246, New York, NY, USA, 2011. ACM.

[71] Edward S. Herman. The propaganda model: A retrospective. `http://human-nature.com/reason/01/herman.html`, 2003.

[72] E.S. Herman and N. Chomsky. *Manufacturing consent: the political economy of the mass media*. Pantheon Books, 1988.

[73] Cecilia Hernández and Gonzalo Navarro. Compressed representations for web and social graphs. *Knowledge and information systems*, 40(2):279–313, 2014.

[74] INE. Demogrficas vitales. Instituto Nacional de Estadisticas de Chile (INE). Available from: http://www.ine.cl/canales/chile_estadistico/familias/demograficas_vitales.php [Accessed 12-December-2017], 2013.

[75] Ipsos. Ranking general de audiencia gran santiago. ipsos radio 2016. Available from: http://www.ipsos.cl/ipsosradioalaire/pagdos.htm [Accessed 30-December-2017], 2015.

[76] S E Jorgensen, F L Xu, and R Costanza. *Handbook of Ecological Indicators for Assessment of Ecosystem Health*. Applied Ecology and Environmental Management. CRC Press, 2005.

[77] Philip Keefer and Stuti Khemani. Democracy, public expenditures, and the poor: Understanding political incentives for providing public services. *The World Bank Research Observer*, 20(1):1–27, 2005.

[78] Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 243–248, New York, NY, USA, 2014. ACM.

[79] Chunyu Kit and Xiaoyue Liu. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 14(2):204–229, 2008.

[80] J. Klaehn. *Filtering the news: essays on Herman and Chomsky's propaganda model*. Black rose books. Black Rose Books, 2005.

[81] Jeffery Klaehn. A critical review and assessment of herman and chomsky's'propaganda model'. *European Journal of Communication*, 17(2):147–182, 2002.

[82] Jeffery Klaehn. Behind the invisible curtain of scholarly criticism: revisiting the propaganda model. *Journalism Studies*, 4(3):359–369, 2003.

[83] Jeffery Klaehn. Model construction: Various other epistemological concerns: A reply to john corner's commentary on the propaganda model. *European Journal of Communication*, 18(3):377–383, 2003.

[84] Jeffery Klaehn. The propaganda model: Theoretical and methodological considerations. *Westminster Papers in Communication and Culture*, 6(2), 2009.

[85] Jeffery Klaehn and Andrew Mullen. The propaganda model and sociology: understanding the media and society. *Synaesthesia: Communication Across Cultures*, 1(1):10–23, 2010.

[86] Joseph T. Klapper. *The effects of mass communication*. Free Press Glencoe, Ill, 1960.

[87] Longbo Kong, Zhi Liu, and Yan Huang. Spot: Locating social media users based on social network context. *Proc. VLDB Endow.*, 7(13):1681–1684, August 2014.

[88] G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris. Geotagging text content with language models and feature mining. *Proceedings of the IEEE*, 105(10):1971–1986, Oct 2017.

[89] James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67, 2013.

[90] Eli Lehrer. Chomsky and the media: A kept press and a manipulated people. In Peter Collier and David Horowitz, editors, *The Anti-Chomsky Reader*, pages 67–87. Encounter Books, 2004.

[91] Yu-Ru Lin, James P. Bagrow, and David Lazer. ”quantifying bias in social and mainstream media” by yu-ru lin, james p. bagrow, and david lazer with ching-man au yeung as coordinator. *SIGWEB Newsl.*, pages 5:1–5:6, July 2012.

[92] Haokai Lu, James Caverlee, and Wei Niu. Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM ’15, pages 213–222, New York, NY, USA, 2015. ACM.

[93] Aibek Makazhanov and Davood Rafiei. Predicting political preference of twitter users. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM ’13, pages 298–305, New York, NY, USA, 2013. ACM.

[94] Jazmine Maldonado, Vanessa Pea-Araya, and Barbara Poblete. Spatio and temporal characterization of chilean news events in social media. In *SIGIR 2015 Workshop on Temporal, Social and Spatially-aware Information Access (TAIA’15)*, 2015.

[95] Momin M. Malik and Jrgen Pfeffer. A macroscopic analysis of news content in twitter. *Digital Journalism*, 4(8):955–979, 2016.

[96] Udi Manber et al. Finding similar files in a large file system. In *Usenix Winter*, volume 94, pages 1–10, 1994.

[97] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150. ACM, 2007.

[98] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Scoring, term weighting and the vector space model. *Introduction to information retrieval*, 100:2–4, 2008.

[99] Robert W McChesney. *The political economy of media: Enduring issues, emerging dilemmas*. NYU Press, 2008.

[100] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[101] Marcus Messner, Maureen Linke, and Asriel Eford. Shoveling tweets: An analysis of the microblogging engagement of traditional news organizations. *International Symposium on Online Journalism. isoj.org*, 2(1):74, 2012.

[102] Felix Ming, Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets and retweets. In *Proceedings of the 7th International AAAI Conference on Web and Social Media*, Boston, Massachusetts, USA, 2013. AAAI Press.

[103] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Elisa Shearer. Pathways to news. *The Modern News Consumer. Journalism & Media. Pew Research Center*, 2016. http://www.journalism.org/2016/07/07/pathways-to-news/.

[104] Amy Mitchell, Jeffrey Gottfried, Elisa Shearer, and Kristine Lu. How americans encounter, recall and act upon digital news. *Analysis. Journalism & Media. Pew Research Center*, 2017. http://www.journalism.org/2017/02/09/how-americans-encounter-recall-and-act-upon-digital-news/.

[105] Brian Patrick Mitchell. *Eight ways to run the country: A new and revealing look at left and right*. Greenwood Publishing Group, 2007.

[106] Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonalves, Qian Zhang, and Alessandro Vespignani. The twitter of babel: Mapping world languages through microblogging platforms. *PLoS ONE*, 8(4):e61981, 04 2013.

[107] Jarle Molvær, Jon Knutzen, Jan Magnusson, Brage Rygg, Jens Skei, and Jan Sørensen. Klassifisering av miljøkvalitet i fjorder og kystfarvann. Technical report, Norsk institutt for vannforskning, Oslo, 2004.

[108] Jonathan Scott Morgan, Cliff Lampe, and Muhammad Zubair Shafiq. Is news sharing on twitter ideologically biased? In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 887–896, New York, NY, USA, 2013. ACM.

[109] Andrew Mullen. Editorial. *Westminster Papers in Communication and Culture*, 6(2):1, oct 2009.

[110] Andrew Mullen. The propaganda model after 20 years: Interview with edward s. herman and noam chomsky. *Westminster Papers in Communication and Culture.*, 6(2):12–22, 2009.

[111] Andrew Mullen. Twenty years on: the second-order prediction of the herman-chomsky propaganda model. *Media, Culture & Society*, 32(4):673–690, 2010.

[112] Andrew Mullen and Jeffery Klaehn. The hermanchomsky propaganda model: A critical approach to analyzing mass media behaviour. *Sociology Compass*, 4(4):215–229, 2010.

[113] Andy Mullen. Twenty years at the margins: The herman-chomsky propaganda model, 1988-2008. *Fifth-Estate-Online: International Journal of Radical Mass Media Criticism*, 2008.

[114] PATRICIO NAVIA and RODRIGO OSORIO. El Mercurio lies, and La Tercera lies more. political bias in newspaper headlines in chile, 1994-2010. *Bulletin of Latin American Research*, 34(4):467–485, apr 2015.

[115] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[116] DF Nolan. Classifying and analysing politico-economic systems. *The Individualist*, pages 5–11, January 1971.

[117] Arturo Orellana. *ICVU 2017: ndice de Calidad de Vida Urbana. Comunas y Ciudades de Chile*. Instituto de Estudios Urbanos y Territoriales. Pontificia Universidad Catolica de Chile, 2017.

[118] Ozer Ozdikis, Halit Oğuztüzün, and Pinar Karagoz. A survey on location estimation techniques for events detected in twitter. *Knowl. Inf. Syst.*, 52(2):291–339, August 2017.

[119] Pace News Limited. The political compass. Available from: `https://www.politicalcompass.org/test` [Accessed 12-August-2017], 2017.

[120] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The, 2011.

[121] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. A computational framework for media bias mitigation. *ACM Trans. Interact. Intell. Syst.*, 2(2):8:1–8:32, June 2012.

[122] Joan Pedro. Evaluación crítica del modelo de propaganda de herman y chomsky/a critical evaluation of herman and chomsky's propaganda model. *Revista Latina de Comunicación Social*, pages 210–223, 2009.

[123] Joan Pedro. The propaganda model in the early 21st century (part i). *International Journal of Communication*, 5:41, 2011.

[124] Pew Research Center for the People & the Press. Too much celebrity news, too little good news. *U.S. Politics & Policy. Pew Research Center*, 2007. http://www.people-press.org/2007/10/12/too-much-celebrity-news-too-little-good-news/ [Accessed 02-Jan-2018].

[125] Pew Research Center for the People & the Press. Haiti, snowstorms, economy vie for publics attention. *The State of the News Media. Journalism & Media. Pew Research Center*, 2010. http://www.people-press.org/2010/02/17/haiti-snowstorms-economy-vie-for-publics-attention/ [Accessed 02-Jan-2018].

[126] E. C. Pielou. *Ecological diversity*. Wiley New York, 1975.

[127] Poderopedia. Poderomedia Foundation. Mapa de medios. [Online: http://apps.poderopedia.org/mapademedios/index/; accessed 12-May-2017].

[128] Michele Polo. Regulation for pluralism in the media markets. *The Economic Regulation of Broadcasting Markets: Evolving Technology and the Challenges for Policy*, pages 150–188, 2005.

[129] Andrea Prat and David Strömberg. The political economy of mass media. *CEPR Discussion Paper No. DP8246*, 2011.

[130] Anand Rajaraman, Jeffrey D Ullman, Jeffrey David Ullman, and Jeffrey David Ullman. *Mining of massive datasets*, volume 77. Cambridge University Press Cambridge, 2012.

[131] Sheldon Rampton. Has the internet changed the propaganda model? *PRWatch: The Center for Media and Democracy*, 2007. http://www.prwatch.org/news/2007/05/6068/has-internet-changed-propaganda-model.

[132] D J Rapport, R Costanza, and A J McMichael. Assessing ecosystem health. *Trends in Ecology & Evolution*, 13(10):397–402, 1998.

[133] Ritva Reinikka and Jakob Svensson. Fighting corruption to improve schooling: Evidence from a newspaper campaign in uganda. *Journal of the European Economic Association*, 3(2-3):259–267, 2005.

[134] John W Robertson. The propaganda model in 2011: Stronger yet still neglected in uk higher education? *Synaesthesia: Communication Across Cultures*, 1(1):25–34, 2011.

[135] Piers Robinson. *The Propaganda Model: Still Relevant Today?*, pages 77–96. Palgrave Macmillan UK, London, 2015.

[136] KyoungMin Ryoo and Sue Moon. Inferring twitter user locations with 10 km accuracy. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 643–648, New York, NY, USA, 2014. ACM.

[137] Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. Social media news communities: Gatekeeping, coverage, and statement bias. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1679–1684, New York, NY, USA, 2013. ACM.

[138] Surendra Sedhai and Aixin Sun. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 223–232, New York, NY, USA, 2015. ACM.

[139] Senado Repblica de Chile. Sala de sesiones. sesin: 84/364. Available from: `http://www.senado.cl/appsenado/index.php?mo=sesionessala&ac=detalleVotacion&votaid=6668` [Accessed 12-August-2017], 2017.

[140] Pavel Serdyukov, Vanessa Murdock, and Roelof van Zwol. Placing flickr photos on a map. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 484–491, New York, NY, USA, 2009. ACM.

[141] SERVEL. Elecciones presidenciales 1989 al 2013 por circunscripcin electoral. Servicio Electoral de Chile. Available from: `https://www.servel.cl/elecciones-presidenciales-1989-al-2013-por-circunscripcion-electoral/` [Accessed 12-August-2017], 2017.

[142] Claude E. Shannon and Warren Weaver. *A Mathematical Theory of Communication*. University of Illinois Press, Champaign, IL, USA, 1963.

[143] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[144] Anshumali Shrivastava and Ping Li. In defense of minhash over simhash. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pages 886–894, 2014.

[145] Fred Seaton Siebert, Theodore Peterson, and Wilbur Schramm. *Four theories of the press: The authoritarian, libertarian, social responsibility, and Soviet communist concepts of what the press should be and do*. University of Illinois Press, 1956.

[146] Edward H Simpson. Measurement of diversity. *Nature*, 1949.

[147] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[148] Chris Smith, Daniele Quercia, and Licia Capra. Finger on the pulse: Identifying deprivation using transit flow analysis. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 683–692, New York, NY, USA, 2013. ACM.

[149] Observatorio Social. Encuesta de caracterizacin socioeconmica nacional (CASEN) 2013. Minesterio de Desarrollo Social (Gobierno de Chile). Available from: `http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/casen_2013.php` [Accessed 12-December-2017], 2015.

[150] Robert R. Sokal. The principles and practice of numerical taxonomy. *Taxon*, 12(5):190–199, 1963.

[151] Colin Sparks. Extending and refining the propaganda model. *Westminster Papers in Communication and Culture*, 4(2):68–84, 2007.

[152] Hans Stadthagen-Gonzalez, Constance Imbault, Miguel A. Pérez Sánchez, and Marc Brysbaert. Norms of valence and arousal for 14,031 spanish words. *Behavior Research Methods*, pages 1–13, 2016.

[153] Galen Stocking. Digital news fact sheet. *The State of the News Media. Journalism & Media. Pew Research Center*, 2017. http://www.journalism.org/fact-sheet/digital-news/.

[154] Saatviga Sudhahar, Thomas Lansdall-Welfare, Ilias Flaounas, and Nello Cristianini. Electionwatch: Detecting patterns in news coverage of us elections. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 82–86, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[155] The Advocates for Self Government. The world's smallest political quiz. Available from: `http://www.theadvocates.org/quiz/quiz.php` [Accessed 12-August-2017].

[156] The Advocates for Self Government. The world's smallest political quiz taken 23 million time online!. Facebook. Available from: `https://www.facebook.com/725411117509651/posts/1094121767305249/` [Accessed 12-August-2017], 2016.

[157] The Advocates for Self Government. About. Available from: `https://www.theadvocates.org/our-mission/` [Accessed 12-August-2017], 2017.

[158] Giang Tran, Mohammad Alrifai, and Eelco Herder. *Timeline Summarization from Relevant Headlines*, pages 245–256. Springer International Publishing, Cham, 2015.

[159] Giang Binh Tran and Eelco Herder. Detecting filter bubbles in ongoing news stories. In Alexandra I. Cristea, Judith Masthoff, Alan Said, and Nava Tintarev, editors, *UMAP Workshops*, volume 1388 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[160] Trefis Team. Twitter's surprising user growth bodes well for 2017. Forbes. Available from: `https://www.forbes.com/sites/greatspeculations/2017/04/27/twitters-surprising-user-growth-bodes-well-for-2017/#151668952e11` [Accessed 12-August-2017], 2017.

[161] Tanguy Urvoy, Emmanuel Chauveau, Pascal Filoche, and Thomas Lavergne. Tracking web spam with html style similarities. *ACM Transactions on the Web (TWEB)*, 2(1):3, 2008.

[162] Justine Zhang Cristian Danescu-Niculescu-Mizil Jure Leskovec Vlad Niculae, Caroline Suen. Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of WWW 2015*, 2015.

[163] Heinrich von Baer and Felipe Torralbo. Chile descentralizado y desarrollado: Fundamentos y propuestas para construir una poltica de estado en descentralizacin y desarrollo territorial en chile. *95 propuestas para un Chile mejor*, 2013.

[164] Ludwig Von Bertalanffy. Problems of life; an evaluation of modern biological thought. *The Yale Journal of Biology and Medicine*, 25(4), 1953.

[165] R. M. Warwick and K. R. Clarke. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, 129(1/3):301–305, 1995.

[166] R. M. Warwick and K. R. Clarke. Taxonomic distinctness and environmental assessment. *Journal of Applied Ecology*, 35(4):532–543, 1998.

[167] David Watson. In the danger: self censorship, the propaganda model, and the saving grace. In *Forum on Public Policy: A Journal of the Oxford Round Table*. Forum on Public Policy, 2008.

[168] Jeffrey N. Weatherly, Thomas V. Petros, Kimberly M. Christopherson, and Erin N. Haugen. Perceptions of political bias in the headlines of two major news organizations. *Harvard International Journal of Press/Politics*, 12(2):91–104, apr 2007.

[169] Wikipedia. El mercurio. Wikipedia, The Free Encyclopedia. Available from: `https://en.wikipedia.org/wiki/El_Mercurio` [Accessed 12-August-2017], 2017.

[170] Wikipedia. El mostrador. Wikipedia, The Free Encyclopedia. Available from: https://en.wikipedia.org/wiki/El_Mostrador [Accessed 12-August-2017], 2017.

[171] Wikipedia. La tercera. Wikipedia, The Free Encyclopedia. Available from: https://en.wikipedia.org/wiki/La_Tercera [Accessed 12-August-2017], 2017.

[172] Wikipedia. Lo prado. Wikipedia, The Free Encyclopedia. Available from: https://es.wikipedia.org/wiki/Lo_Prado [Accessed 12-August-2017], 2017.

[173] Wikipedia. Medios de comunicacin en chile. Wikipedia, The Free Encyclopedia. Available from: https://es.wikipedia.org/wiki/Medios_de_comunicaci%C3%B3n_en_Chile [Accessed 12-August-2017], 2017.

[174] Wikipedia. Native advertising. Wikipedia, The Free Encyclopedia. Available from: https://en.wikipedia.org/wiki/Native_advertising [Accessed 12-August-2017], 2017.

[175] Wikipedia. Radio cooperativa (chile). Wikipedia, The Free Encyclopedia. Available from: https://es.wikipedia.org/wiki/Radio_Cooperativa_(Chile) [Accessed 30-December-2017], 2017.

[176] Wikipedia. San miguel (chile). Wikipedia, The Free Encyclopedia. Available from: https://es.wikipedia.org/wiki/San_Miguel_(Chile) [Accessed 12-August-2017], 2017.

[177] Dwayne Winseck. The state of media ownership and media markets: Competition or concentration and why should we care? *Sociology Compass*, 2(1):34–47, jan 2008.

[178] Arjumand Younus, Muhammad Atif Qureshi, Suneel Kumar Kingrani, Muhammad Saeed, Nasir Touheed, Colm O'Riordan, and Pasi Gabriella. Investigating bias in traditional media through social media. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 643–644, New York, NY, USA, 2012. ACM.

[179] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.

[180] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the political leaning of news articles and users from user votes. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2011.

[181] George Kingsley Zipf. The P_1P_2/D hypothesis: On the intercity movement of persons. *American Sociological Review*, 11(6):677–686, 1946.

[182] George Kingsley Zipf. Some determinants of the circulation of information. *The American Journal of Psychology*, 59(3):401–421, 1946.