



UNIVERSIDAD DE CHILE  
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS  
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

ANÁLISIS ESTÁTICO DEL SISTEMA DE MEDIOS NOTICIOSOS  
CHILENOS EN TWITTER / *STATIC ANALYSIS OF THE CHILEAN  
NEWS MEDIA SYSTEM IN TWITTER*

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS,  
MENCIÓN COMPUTACIÓN

JORGE ANDRÉS BAHAMONDE VEGA

PROFESOR GUÍA:  
BÁRBARA POBLETE LABRA

MIEMBROS DE LA COMISIÓN:

NELSON BALOIAN TATARYAN  
JORGE PÉREZ ROJAS  
ELIANA SCHEIHING GARCIA

SANTIAGO DE CHILE  
2017

Este trabajo ha sido parcialmente financiado por CONICYT-PCHA/Magíster Nacional/2015-22151202  
y por Proyecto FONDECYT 11121511 de Dra. Poblete

# Resumen

A medida que el nivel de uso de redes sociales en línea tales como Facebook y Twitter ha aumentado, los medios noticiosos tradicionales se han vuelto involucrados en ellas. Diversos diarios, canales de televisión y otros medios poseen cuentas en diferentes redes sociales, usándolas para diseminar información noticiosa. Además, se ha posibilitado la existencia de medios noticiosos completamente electrónicos, así como la participación activa de los usuarios en la difusión de las noticias. La creciente disponibilidad de datos provenientes de estas plataformas vuelve factible la posibilidad de estudiar fenómenos como la propagación y el cambio en la composición de las noticias.

La democracia necesita ciudadanos informados, así como una esfera pública inclusiva y pluralista. Los medios noticiosos juegan un rol fundamental en este ecosistema: la diversidad y el pluralismo de medios han sido llamados un pilar básico de las democracias saludables. Los estudios sobre estos conceptos usualmente tienen un foco en la propiedad de los medios y su regulación; sin embargo, el nivel en el que estos aspectos influyen la diversidad del contenido publicado no es completamente certero. De esta forma, el estudio de la diversidad de contenido producido por los medios noticiosos se vuelve una problemática importante.

Esta tesis apunta a caracterizar los medios noticiosos chilenos en base al contenido que publican en la plataforma de microblogging Twitter. Se propone una metodología para la exploración de la diversidad de contenido en medios noticiosos, y se aplica para obtener una visión panorámica de los medios chilenos. Esta metodología consiste en la definición de similitudes basadas en contenido y su aplicación al contenido publicado por medios noticiosos. Luego de estos pasos, se realiza la detección de grupos de medios similares, mediante técnicas de Minería de Datos y Recuperación de la Información. Estos grupos son caracterizados y comparados con características externas de los medios correspondientes, como su propiedad y su audiencia.

Las contribuciones de este trabajo incluyen tanto la metodología utilizada como los resultados obtenidos. Se observa una falta de diversidad en los medios noticiosos, particularmente en el caso de medios locales reportando sobre noticias de escala nacional. Además, su comportamiento se encuentra correlacionado con su propiedad, lo que sugiere que estos medios poseen una fuente común para noticias de este tipo. Se observa, también, que la audiencia de los medios se ve relacionada con el foco geográfico que los medios muestran. Este trabajo provee una visión de la diversidad de medios que complementa las metodologías tradicionales. Se presentan, además de estos resultados, visualizaciones que muestran cómo la metodología aplicada puede ayudar a los usuarios a diversificar el contenido que consumen.

# Abstract

As the usage level of online social networks such as Twitter and Facebook has risen, traditional news media have become interested in them: many newspapers, TV stations and other news media currently have accounts on different online social network platforms and use them to spread information. Furthermore, purely online news media outlets have appeared, and users can become active players in the information diffusion process. The growing availability of data derived from these online platforms makes studying phenomena such as information propagation and evolution a feasible possibility.

Democracy requires informed citizens, as well as an inclusive and pluralist public sphere. News media play a fundamental role in this ecosystem: media pluralism and diversity have been called a cornerstone of healthy democracies. Ownership and regulation aspects are usually the main focus when analyzing these concepts; however, the degree to which aspects influence content diversity is not completely certain. Therefore, studying content diversity in news media becomes an important problem.

This thesis aims to characterize Chilean news media based on the content they publish through the Twitter microblogging platform. We propose a methodology for exploring content diversity in news media, and apply it to obtain a view of the Chilean news media landscape. Our methodology consists in defining content-based similarities and applying them to content published by news media outlets. Following these steps, groups of similar news media outlets are discovered using Data Mining and Information Retrieval techniques. We characterize and compare these groups against external media features, such as audience and ownership.

Our contributions include both the methodology we used and the results we obtained. We observe a lack of diversity in news media, especially in the case of local news media reporting on national-scope topics. In addition to this, their behavior is strongly linked to ownership, suggesting a common source for national-scope news. We also see that audience seems to be related to the geographical scope news media outlets display. We believe this work provides a view of media diversity that complements traditional approaches. In addition to these results, we present visualizations that show a way in which our methodology and results can help users to diversify the content they consume.

# Agradecimientos

Este trabajo de tesis no habría sido posible sin la guía de la profesora Bárbara Poblete. De igual forma, las ideas, opiniones y aportes de Johan Bollen, Leo Ferres, Erick Elejalde y Miguel Paz, así como la ayuda de Mauricio Quezada y Vanessa Peña, dejaron una marca indeleble en este trabajo. ¡Gracias!

Gracias además, a mi familia, que no sólo incluye a aquellos con quienes tengo un lazo de sangre. Gracias a Bruno, Diggo, Dani, Kari, Johan, Seba, Hiho, Esteban, Mathias ¡y muchos otros!. Por preocuparse, por escucharme y hacerme sonreír: por sostenerme cuando ha sido necesario :)

And you...

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Research Question and Challenges . . . . .	2
1.2	Objectives . . . . .	3
1.2.1	Main objective . . . . .	3
1.2.2	Specific objectives . . . . .	3
1.3	Contributions . . . . .	3
1.4	Methodology . . . . .	4
1.5	Outline of this work . . . . .	5
<b>2</b>	<b>Background and Related Work</b>	<b>6</b>
2.1	Data Analysis Techniques . . . . .	6
2.1.1	Vector Space Model . . . . .	6
2.1.2	Clustering and community detection in graphs . . . . .	7
2.1.3	Graph Clustering Evaluation Metrics . . . . .	8
2.1.4	Topic models and event detection . . . . .	9
2.2	Content diversity and characterization in social networks . . . . .	10
2.3	Visualizations of News Media Relationships . . . . .	11
<b>3</b>	<b>Theoretical Framework</b>	<b>12</b>
3.1	Graph model . . . . .	12
3.2	Similarity measures . . . . .	13
3.2.1	Vocabulary Similarity . . . . .	13
3.2.2	Topic Similarity . . . . .	14
3.2.3	Temporal Correlation . . . . .	14
3.2.4	Ownership similarity . . . . .	17
3.2.5	Follower similarity . . . . .	18
3.3	Community discovery . . . . .	19
3.3.1	Preprocessing . . . . .	19
3.3.2	Community analysis . . . . .	19
<b>4</b>	<b>Experimental Methodology</b>	<b>21</b>
4.1	Methodology overview . . . . .	21
4.2	Dataset . . . . .	22
4.2.1	News documents source . . . . .	22
4.2.2	Outlet information sources . . . . .	24
4.3	Implementation and experiments . . . . .	24
4.3.1	Similarity measures and internal analysis . . . . .	25

4.3.2	Similarity distribution exploration . . . . .	26
4.3.3	Community discovery . . . . .	28
4.3.4	Community analysis . . . . .	28
<b>5</b>	<b>Results and Analysis</b>	<b>32</b>
5.1	Ownership structure . . . . .	33
5.2	Follower similarity . . . . .	34
5.3	Vocabulary similarity . . . . .	38
5.4	Topic similarity . . . . .	41
5.5	Temporal correlation . . . . .	44
5.5.1	Penta term set . . . . .	44
5.5.2	President term set . . . . .	47
5.6	Comparison . . . . .	49
5.7	Insights . . . . .	50
<b>6</b>	<b>Conclusions</b>	<b>51</b>
6.1	Applications . . . . .	51
6.2	Limitations . . . . .	54
6.3	Extensions and Improvements . . . . .	54
	<b>Bibliography</b>	<b>55</b>
	<b>Appendices</b>	<b>61</b>
A	User distribution for each community structure . . . . .	61
B	Similarity graph visualizations for each explored similarity measure . . . . .	64
C	Lists of stop words . . . . .	68
C.1	Stop words included from NLTK . . . . .	68
C.2	Manually included stop words . . . . .	70

# Chapter 1

## Introduction

There has been a sharp increase in the amount of both generated and available data during the last few years, in what has been described as a “data explosion”. It has been estimated that in 2007, the amount of yearly generated data exceeded the amount of available storage capacity for the first time in history [29]. By 2013, the “digital universe” comprised around 4.4 zettabytes ( $4.4 \times 10^{21}$  bytes). Nearly 22% of this data was concluded to be potentially useful after characterization and analysis; however, not even 5% of this data was actually analyzed [59].

In addition to this, the high level of usage of Twitter, Facebook and other social media platforms allows researchers to perform social and demographic studies at a previously prohibitive scale.

Twitter is a microblogging<sup>1</sup> social platform, which has been used for tracking epidemics [42], the study and geolocation of natural disasters [56,62] and the detection of controversial events [33,56], among other applications. Online-oriented studies (using techniques that allow data analysis as it is being generated) have managed to construct event-detection algorithms, as well as credibility-prediction algorithms [14].

Aside from conversation, Twitter is widely used for news propagation [37]. By 2010, the Internet was a significant news source for 61% of its users [52], and is considered as reliable as other types of media (except for newspapers) [24]. This platform also presents a high level of usage in Chile, having around 4 million users.

Chilean media has greatly developed during the last few years. A culture of openness and transparency has been fostered in the country since its return to democracy in 1990; however, this process is not obstacle-free in what pertains to pluralism and diversity in mass media [31]. Chilean news media outlets are highly centralized, as most big news media outlets are located in Santiago, the country’s capital. There is also centralization from an ownership perspective, to the point that some researchers consider the newspaper market to be a duopoly [15].

---

<sup>1</sup>This term is used due to the fact that messages published in this platform cannot exceed 140 characters.

The availability of different news media outlets has grown due to the increasing access to digital media, as well as mobile devices and social media. New, purely online media outlets have appeared, while traditional news media outlets have also taken an interest in digital and social media. They have accounts in either Facebook or Twitter in addition to their websites, and began to monitor social network platforms such as Twitter. However, the benefits of digital and social media only apply to those that have the technological and financial means to make use of them, in a country where the digital gap is still an important issue [17].

Furthermore, some events that had a high impact on the Chilean web did not manage to attract media coverage [31]. For example, the main television networks systematically failed to report on a strike in a pharmacy chain during 2010: the Chilean National Television Council determined these omissions were intentional [22]. Other cases include mobilizations on behalf of a group of jailed Mapuche natives [2] and a boycott against a gold mining project in the Andes (Pascua Lama) [31].

An imbalance or bias in production and/or consumption of media might place a lot of power in the hands of media owners, as they would be able to distort people’s opinions and essentially shape the public sphere [19]. Another issue is the possible creation of *filter bubbles* [50]. The content choices made by either producers (in the form of content personalization) or consumers can limit the latter’s exposure to different and diverse viewpoints, increasing ideological segregation [6].

All of these factors point to the need for an analysis of both production and consumption of news media. For example, through the results of a content diversity analysis, a user can see if the content she consumes is diverse or not, and take steps to improve it.

We analyze the Chilean news media landscape, applying a content-based similarity approach to tweets published by Chilean news media outlets. We look for news media outlets that produce related content based on different similarity measures; after this, we identify groups of similar news media outlets. Finally, we look for correspondences between groups identified through different similarity measures, and analyze user media consumption in relation to these groups.

We found ownership seems to influence topic diversity, as we observed a group of media outlets belonging to the same media group naturally forming a cluster when using a topic-based similarity. We also saw both vocabulary and follower similarity seem to be correlated with geographical distribution. When fixing a topic and analyzing time correlations, different behaviors are observed: for some topics ownership seems to be important to whether two media outlets will simultaneously cover the topic; for others, media scope (local versus national) seems to be a more decisive factor.

## 1.1 Research Question and Challenges

The main question addressed by this work is



## Is Chilean news media production and consumption diverse in terms of content?

For this question to be properly addressed, the concept of *content diversity* must be clearly defined. Alternatively, we can define a lack of diversity to take place when most news media outlets are similar, leading to the question “When are two media outlets similar?”.

## 1.2 Objectives

In the following, we present both this work’s main objective and its specific objectives.

### 1.2.1 Main objective

Our main objective is to evaluate the diversity of content, both in its production and consumption, in the Chilean news media landscape. This includes investigating if power structures are reproduced in the production of news content, if media ownership and content diversity are somehow related, and if users are consuming a wide array of content.

### 1.2.2 Specific objectives

- To obtain a dataset reflecting the content being produced by the Chilean news media system.
- To define what “similar news media outlets” means, in the context of an online social network platform.
- Determining similarities between news media outlets in the dataset, based on the previous definition.
- To explore the structures formed by the chosen similarity measures. For example, we can see which news media outlets are similar to many others, or what groups of similar outlets can be observed. This includes the construction and use of visualizations for exploration.
- To build a tool that helps a user to determine the degree of diversity of the news content they consume on the analyzed social network platform, and gives recommendations to improve this diversity.

## 1.3 Contributions

This work makes the following contributions:

- A methodology for studying media outlets’ content diversity based on their posts on an online social network.
- A characterization of Chilean news media outlets using this methodology and different notions of content similarity. We found interesting behaviors and correlations through these analyses.
- A simple visualization showing how this methodology can help users diversify the content they consume.

We observed that, at least in regards to national-scope topics, groups of local news media outlets tend to follow a behavior highly correlated with their ownership. This can be interpreted as these media outlets having a single source for this type of news, though further experiments are required. We found several groups of similar outlets, and that users tend to only consume from one or two of these groups. We also observed that consumption behavior seems to be correlated to the geographical focus news outlets seem to display.

We believe this work provides a complementary view of diversity in media, in addition to approaches to pluralism based on legislative, ownership and geographical aspects, some of which are used to measure pluralism in Europe and the United States [43].

## 1.4 Methodology

We performed the following steps to analyze the degree of content diversity production and consumption in Chilean news media outlets.

1. We used a dataset from Twitter, comprising around 700,000 tweets published by 84 Chilean news media outlets. We also obtained additional data such as an ownership structure from Poderopedia [28] and the list of Twitter followers of these news media outlets.
2. We defined several ways in which to measure similarity between two news media outlets. This similarity was, in most cases, derived from the content that these outlets publish on Twitter.
3. We built graph representations to study the way these similarity measures group news media outlets, where similar outlets (represented as nodes) are connected with edges.
4. We visually explored these graphs, which let us gain insight on what properties the similarity measure reflects. This also allowed us to perform procedures such as filtering these graphs for further analysis.
5. We automatically learned groups (communities) of similar news media outlets based on the defined similarity measures.
6. We analyzed these groups, both by calculating internal metrics and by contrasting them with additional data and/or analyses.

## 1.5 Outline of this work

The rest of this document is organized as follows:

- Chapter 2 introduces common data mining concepts and techniques, as well as related work on similarity analysis and visualization.
- Chapter 3 establishes the theoretical model for our work. Definitions and pipelines used to perform our analyses are explored in this chapter.
- The process of instantiating the model already defined is described in Chapter 4. Similarity measures, community detection algorithms, as well as visualization techniques used are described.
- Results and their analysis are shown in Chapter 5, both general and specific to each similarity measure we used.
- Conclusions and future work are explored in Chapter 6.

# Chapter 2

## Background and Related Work

This thesis' scope lies within Data Mining and Machine Learning. We want to analyze the level of diversity in the Chilean news media ecosystem, based on the content published in their Twitter accounts. This means we have noisy, mostly unstructured data, and we wish to identify patterns in it, extracting new knowledge.

### 2.1 Data Analysis Techniques

The following sections present some concepts, tools and techniques used throughout our experiments and analyses.

#### 2.1.1 Vector Space Model

A vector space model [46] is a way of representing documents. Each document is represented by a vector  $v = (v_1, v_2, \dots)$  where each dimension corresponds to a feature or identifier, also called *term*; the actual value of each coordinate is called its *weight*. For example, written documents can be represented by using words as terms (an approach aptly named *bag-of-words*).

Terms can be weighted in different ways. A simple way is to use term frequencies (the number of times the term appears in the document) as weights. This weighting scheme has some issues, as longer documents will have larger values in their vector components, and common terms that have not been filtered out (e.g. words that only are common in a specific context) might distort the representation. For these reasons, other weighting schemes are commonly used.

One of the most popular weighting schemes is tf-idf, which stands for term frequency-inverse document frequency, and is a statistic that can be used as a weighting scheme for bag-of-words representations. In a tf-idf scheme, two opposing factors affect a term's weight:

- **Term frequency:** as a term is more relevant to a document if it appears many times in it, it is intuitive that a term’s weight should grow with its number of appearances. The following are possible definitions for  $tf$ , for a term  $t$  and a document  $d$ :

$$tf(t, d) := f_{t,d} \tag{2.1}$$

$$tf(t, d) := \frac{f_{t,d}}{|d|} \tag{2.2}$$

$$tf(t, d) := \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} \tag{2.3}$$

where  $f_{t,d}$  is the number of times  $t$  appears in  $d$ . The second and third definitions aim to prevent a bias towards longer documents: the first does this by dividing the absolute frequency by the document’s length, while the second divides it by the maximum frequency of any term in the document. Additionally, *smoothing* terms can be added to these definitions.

- **Inverse document frequency:** even if a term appears many times in a document, it might not be very relevant if it appears in most of them. Therefore, a measure of the term’s specificity is used to lower a term’s weight. For a term  $t$  and a set of documents  $D$ , a possible definition for  $idf$  is as follows:

$$idf(t, D) := \frac{|D|}{|\{d \in D : t \in d\}|} \tag{2.4}$$

In this way, a term will get a lower score if it appears in most documents (i.e. it is not specific enough).

Finally, the  $tf$ - $idf$  term weight for a term  $t$  and a document  $d$  from a set  $D$  of documents is defined as

$$tf-idf(t, d, D) := tf(t, d) \cdot idf(t, D) \tag{2.5}$$

This balances both frequency (through the  $tf$  term) and specificity (through the  $idf$  term), as explained before.

## 2.1.2 Clustering and community detection in graphs

The process of clustering consists in finding groups (*clusters*) in a given set of objects such that elements within the same group are similar or related to each other, and different from (or unrelated to) those in different groups. The set of found clusters is called a clustering [58].

A subproblem of clustering is community detection in graphs. Many real-world graphs (such as social networks) display what is said to be a “community structure”: instead of

having an homogeneous connection density, some groups of nodes happen to be more densely connected internally than with the rest of the network: community detection is the task of finding such groups [30]. In this context, the terms cluster and community are often used interchangeably.

Many types of community detection algorithms can be distinguished [25]. Graph partitioning algorithms are a top-down approach that aim to divide the vertices in  $n$  groups, for example, by performing consecutive bisecting operations.

Hierarchical clustering algorithms, on the other hand, aim to identify a multilevel structure in the graph, identifying groups of vertices with high similarity. There are two main types of hierarchical clustering algorithms: agglomerative algorithms, which merge communities iteratively in a bottom-up approach; and divisive algorithms, which split communities by removing edges between vertices with a low similarity (which must be properly defined). Non-hierarchical, divisive approaches may remove vertices or subgraphs, aiming to separate communities instead of vertices.

Partitional clustering techniques map graph vertices to a metric space, subsequently applying more traditional clustering algorithms in this space. Finally, spectral clustering methods define a pairwise similarity measure and extract eigenvectors from its matrix representation or other matrices derived from it; clustering is then performed on the vector space defined by these eigenvectors.

Most community discovery approaches must make decisions where different heuristics can be applied: for example, when choosing which edges to remove or which communities to merge. Many of these heuristics are based on graph metrics such as modularity, clustering coefficient or clique properties.

A non-hierarchical community discovery algorithm on graphs takes a (possibly weighted) graph as an input, as well as additional parameters such as the number of communities to look for, and outputs a label for each vertex in the graph: vertices having the same label are said to belong to the same community. Hierarchical algorithms, instead, return a *dendrogram*: a tree structure where communities are progressively joined together based on their similarity. A traditional clustering can be derived from a dendrogram by, for example, specifying a target number of communities and cutting the tree at the appropriate height.

### 2.1.3 Graph Clustering Evaluation Metrics

Graph metrics can be used to assess the quality of a community structure [3]. One of such metrics, modularity, measures the degree to which the community structure displays both dense connections inside communities and sparse connections between communities, with respect to a random graph. Modularity can be defined as follows:

$$Q = Tr(e) - \|e^2\| \tag{2.6}$$

where the component  $e_{ij}$  of the matrix  $e$  is the ratio of edges that link vertices in communities  $i$  and  $j$ ,  $Tr(e)$  is the sum of this matrix’s diagonal elements and  $||x||$  stands for the sum of the elements of the matrix  $x$ . Modularity takes values on  $[-\frac{1}{2}, 1)$ , with a better community structure having a higher modularity value.

Another metric is conductance, which measures the well-connectedness of graph substructures. Two definitions of conductance can be distinguished in the context of graph clustering: internal conductance, which relates to connections within communities, and external conductance, which relates to connections between communities. The conductance  $\phi$  of a cluster  $C_i$  for a graph  $G = (V, E)$  is defined as:

$$\phi(C_i) = \frac{\sum_{u \in C_i} \sum_{v \notin C_i} e_{uv}}{\min(a(C_i), a(V \setminus C_i))} \quad (2.7)$$

where  $e_{ij}$  is the weight of the edge connecting vertices  $i$  and  $j$ , and  $a(C) = \sum_{u \in C} \sum_{v \in V} e_{uv}$ . A clustering’s conductance can be defined as the average of each cluster’s conductance [3].

Another way of defining a cluster  $C$ ’s internal conductance is by taking the subgraph induced by  $C$ , and computing the minimum  $\phi(x)$ , considering every possible vertex subset  $x$  in  $C$ . The clustering’s internal conductance is the minimum of all  $\phi_{\text{int}}(C_i)$  [39]. External conductance, on the other hand, can be defined as

$$\phi_{\text{ext}}(C) = \max \phi(C_i) \quad (2.8)$$

where the maximum is taken over the clusterings’ clusters. This is the definition used by many works, being—usually—simply called conductance.

## 2.1.4 Topic models and event detection

Topic models are algorithms for discovering themes within a collection of documents and organizing the collection according to the discovered themes [11]. One of the simplest topic models is Latent Dirichlet Allocation (LDA), which considers topics to be probabilistic distributions over a fixed vocabulary. A document is seen as having different proportions of these topic. LDA tries to infer the topic distributions and proportions from co-occurring words in the documents [12].

A task related to topic models is event detection: an unsupervised learning task that can be presented in at least two forms. In one hand, retrospective detection discovers new events in a chronologically ordered set of documents that has already been accumulated; online detection, on the other hand, tries to identify events in real time, as they emerge [66].

## 2.2 Content diversity and characterization in social networks

Media pluralism, as a concept, has been analyzed in more than one fashion: for instance, existing work distinguishes external from internal pluralism. *External* pluralism deals with the ownership structure of the media: the number of owners, media companies, editorial boards, channels and others, arguing that a “free marketplace of ideas” is needed if one wants to have a variety of opinions and ideas available. *Internal* pluralism, on the other hand, relates to the diversity of published content and the degree of representation of different opinions and viewpoints. A middle ground between these two approaches deals with users’ ease of access and their interactions with media content and services [40]. In this section, we aim to identify work related to studying internal diversity characterization in social media and, plus diversity consumption.

Benkler et al. [10] use tools such as Mediacloud [34] to examine coverage of news related to the SOPA-PIPA debate by media sources, as well as the spreading of both these stories and the languages involved. They are able to identify many actors and their roles in the debate’s evolution.

Graells-Garrido et al. [32] characterize Twitter users by scoring their most frequent n-grams (sequences of contiguous words) with respect to topics determined through LDA. Subsequently, they analyze these topics’ diversity and their relationships by examining tweets relating to multiple topics. They aim to connect users having different stances on sensitive issues but similar topical characteristics through recommendations delivered by means of an organic visualization.

An et al. [5] use a form of follower-based similarity over 80 media sources’ Twitter accounts to create a media landscape map. They use publicly available data to label 34 of these sources with a political leaning; in addition to this, they analyze both direct and indirect (through following of followers) user consumption of these media sources. In a different work [4] they estimate the political leaning of additional news media outlets through the use of another follower-based distance.

Morgan et al. [47] explore the effects of perceived ideology on both consumption and sharing of news in Twitter. They assign political leanings to 12 news outlets based on the results of a news consumption survey, and analyze consumption and sharing of these news media outlets’ content by users.

Finally, Fish and Othman [23] analyze the 50 most prominent English language sites that shared stories on the 2014 Gaza War. They use vocabulary features to build a similarity graph on which they apply community detection algorithms. They find five distinct communities which touch on different aspects of the war, giving a view of the media landscape with respect to this event.



## 2.3 Visualizations of News Media Relationships

With respect to influence mapping in networks, Ronen et al. [54] build three language influence networks, based on book translations, Wikipedia users that edit articles written in different languages and multilingual Twitter users. They use visualizations and graph metrics such as centrality to evaluate each language's influence; additionally, they compare these influence estimations with additional metrics such as the number of famous people born in countries that speak each language.

Other visualizations created to suit more specific include those shown in Lotan's article [44] on the 2014 Gaza War, social media propaganda and filter bubbles. Lotan shows two visualizations built using Gephi [7], a graph visualization tool, to detail the social media landscape after the UNRWA school bombing in 2014. These visualizations are built using friendship relationships in Twitter and co-occurrence of tags in Instagram, and show two distinctively opposite sides of the issue.

The following chapters will specify the theoretical framework we use, and the tools and metrics we choose, along with the reasoning for these choices.

# Chapter 3

## Theoretical Framework

Our aim is to measure content diversity in the Chilean news media ecosystem. This calls for a formal definition of diversity, taking into account the fact that we are using social media data.

A news media outlet is an entity that emits information about real-world events through one or multiple channels. Under this definition, even a single person can be a news media outlet: our dataset does contain individual people who consistently publish news-related content. We model an outlet as a tuple  $x = (D, M)$ , where  $D$  is a set of social media documents (such as tweets, text, images, etc.) and  $M$  is metadata for the outlet (which may include, for example, a list of the outlet’s audience or its owner). Our dataset is a set of outlets,  $\mathcal{D}$ . Instead of directly tackling the concept of diversity, we consider a setting where most news media outlets are similar as having a lack of diversity.

### 3.1 Graph model

We choose to model news media outlet relationships through graphs. A graph is a 2-tuple  $G = (V, E)$  where  $V$  is a set of finite vertices and  $E \subset V \times V$  is the set of edges that connect pairs of vertices. A *weighted* graph also considers a function  $w : E \rightarrow \mathbb{R}$  that assigns weights to edges. A graph is said to be undirected if  $E$  is a symmetrical relation.

Social network analysis considers entities as vertices and their relationships as edges. Our graph model considers *similarity* to be the key relationship to analyze: we consider weighted, undirected graphs where news media outlets are represented as vertices, while the degree of similarity between pairs of them are represented by weighted edges.

In other words, we analyze weighted, undirected graphs with a vertex for each outlet (i.e.  $V = \mathcal{D}$ ). Every pair of outlets  $x_1, x_2$  is connected by an edge  $e$  with their similarity  $s$  as its weight (i.e.  $w(e) = s(x_1, x_2)$ ). We call such a graph a *similarity graph*. Additionally, visualization of the similarity graph allows to gain insights about the network structure, as well as to communicate these insights [27].

It is important to note that there are other ways of modeling these relationships. Analyses, for example, could be performed directly over similarity values. However, graphs have merits of their own: particularly, they can easily be translated into visualizations, the usual one mapping vertices to circles and edges to straight lines connecting them. In addition to this, a graph structure might capture notions such as networks of influence better than other approaches. Although we initially consider complete graphs, filtering mechanisms can help both analysis and manual exploration.

## 3.2 Similarity measures

A similarity measure is a real-valued function that quantifies the similarity between two objects. There is no single definition for a similarity measure; for this work, we considered functions of the form

$$s : \mathcal{D} \rightarrow [0, 1] \quad (3.1)$$

with 0 representing completely unrelated objects and 1 representing fully similar objects. We also consider functions taking values on  $[-1, 1]$ , with negative values representing objects that are actually *opposite* for some interpretation.

Intuitively, we require similarity functions to be symmetric (i.e.  $s(x, y) = s(y, x)$  for any  $x$  and  $y$ ) and to take on its maximum possible value whenever an object is compared with itself (i.e.  $s(x, x) = 1$  for all  $x$ ). Finally, we say that a similarity measure  $s$  is *content-based* if it only depends on the *documents*  $D$  of an outlet (this is,  $s(x = (D, M)) = s(D)$ ). We use multiple similarity measures, leading to different analyses and conclusions.

### 3.2.1 Vocabulary Similarity

A vocabulary-based similarity measure is one of the simplest ways to define a content-based similarity measure. The intuition behind such a notion of similarity is that news media outlets might use different, distinctive vocabularies for communicating news.

Borrowing from information retrieval [46], we can see each news media outlet’s published content as a document (created by the concatenation of its tweets’ text). We can therefore define as many documents as outlets in our dataset. Each document is then represented using a vector space model with words as terms, using the tf-idf weighting scheme seen in Subsection 2.1.1. This representation allows us to use cosine similarity as a similarity measure, defined as follows [57]:

$$C(A, B) := \frac{A \cdot B}{\|A\| \|B\|} \quad (3.2)$$

where  $A$  and  $B$  are vector space model representations for two outlets. As vector components take on positive values, this similarity measure takes values on  $[0, 1]$ .

### 3.2.2 Topic Similarity

An intuitive notion of similarity between news media outlets can be based on whether they talk about the same topics. This, of course, requires defining what a topic is. For this similarity measure, we use the same selection and preprocessing criteria as with vocabulary similarity.

We use Kalyanam et al’s [38] approach for topic modeling and discovery. Topics are defined as sets of keywords, and topic discovery is performed in a non-supervised fashion, in a two-step process. Over a small timescale, frequent pairs of co-occurring words are identified as keywords. Over a bigger timescale, these pairs are coalesced into topics by connecting keyword pairs sharing a word. In a graph with keywords as vertices and edges representing keyword pairs, the resulting connected components are identified as topics. An automatic, graph-oriented stop word-detecting algorithm is run on this graph, removing words that might connect otherwise unrelated topics according to a tf-idf-inspired criterion. The algorithm assigns each word a modified tf-idf weight that gives more weight to words that are very frequent across all outlets. Then, for each topic, the top scoring words are removed until the removal does not result in the topic (seen as a connected component) splitting into two, or when it ends up with one of the original keyword pairs.

After defining events for the small timescale, we characterize media outlets by the degree with which they talk about them. As the topic discovery algorithm is based on pairs of co-occurring words as event keywords, we say a tweet talks about a given event if it includes at least one of these pairs of words.

We use the set of an outlet’s tweets during a big timescale unit as documents. The topics identified for this time period, on the other hand, are the terms, and a term occurrence is a tweet including at least one of the topic’s keyword pairs. As with vocabulary similarity, we use a vector space model with a tf-idf weighting scheme, and cosine similarity.

This procedure outputs a similarity score for each pair of outlets, for each big-scale time unit. We aggregate these scores (by taking the average) to obtain an overall similarity score for each pair of outlets.

### 3.2.3 Temporal Correlation

The previous similarity measure incorporates a temporal aspect, which can be analyzed on its own. If a certain topic (defined as a set of keywords) is predefined, we can observe its rises and falls in popularity in an outlet’s published content. We then say two outlets are similar if these rises and falls are correlated in time. If two outlets display this type of behavior, it might happen in at least two, distinct ways:

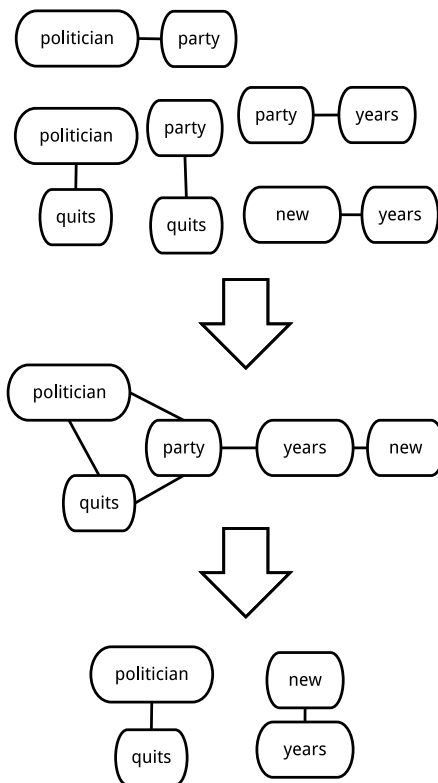


Figure 3.1: Topic discovery procedure. First, pairs of co-occurring keywords are identified. These pairs are converted into a graph; finally, stop words are detected and removed. The resulting connected components form the set of obtained topics.

- These outlets consistently display common behavior. This could be seen as a sort of imitation or influence (internal in origin).
- The outlets are covering the same singular events. They normally have different behaviors, but in the presence of an exogenous event (e.g. a volcano eruption) they cover the same topic at the same time (an external sort of similarity).

We handpick some topics of interest and choose keywords for them. We base this both on the previous analysis and our understanding of potentially controversial topics. Given a topic comprising a set of words, we say a tweet talks about the topic if it includes any of these words. This differs from topic similarity: instead of treating topics as a graph, we now see them simply as a set of words.

We define a topic’s daily popularity for an outlet,  $f_i$ , as the fraction of the outlet’s daily tweets that talk about the topic. We calculate these popularity scores for each day in the dataset and thus create a time series  $t$  for each topic and outlet:

$$t = \{f_i\}_{i=0}^n \tag{3.3}$$

We normalize these time series in two steps. First, we log-scale the time series’ values; afterwards, we replace these scaled frequencies by the differences between consecutive values.

Log-scaling is performed in order to work with relative variations, as an absolute change in a log-scaled variable is approximately proportional to a percentage change in the original variable. By using differences of scaled frequencies, we avoid being misled by topics that might show a steady and constant increase in use. We also add a small factor,  $\epsilon$ , to avoid divisions by zero:

$$t^* = \left\{ \log\left(\frac{f_i + \epsilon}{f_{i-1} + \epsilon}\right) \right\}_{i=1}^n \quad (3.4)$$

After these steps, we end up with a normalized time series for each news media outlet: we use Spearman’s rank correlation index [48] to compare them. This coefficient is a measure of statistical dependence between two variables. Therefore, if news media outlets are represented by time series, this index can be seen as a measure of the relationship between them. Intuitively, the Spearman correlation coefficient tries to describe if one time series can be warped into the other using a monotonic function.

If two random variables  $X$  and  $Y$  take on values  $(X_1, \dots, X_n)$  and  $(Y_1, \dots, Y_n)$  respectively, their Spearman correlation coefficient  $\rho$  is calculated as follows. First, values are normalized as lists of ranks  $(x_1, \dots, x_n)$  and  $(y_1, \dots, y_n)$ . A value’s rank is a real number between 1 and  $n$  indicating the value’s position when the original list of values is sorted. Ties are resolved by assigning average ranks to each tied value. After this is done for both  $X$  and  $Y$ , their Spearman rank correlation index  $\rho$  is defined as follows:

$$\rho := 1 - \frac{6 \sum_{i=1}^n d_i}{n(n^2 - 1)} \quad (3.5)$$

with  $d_i := x_i - y_i$ . Spearman’s rank correlation index takes values in  $[-1, 1]$ , with extremal values happening when one of the variables is a perfect monotonic function of the other. In other words, this similarity measures whether variables rise and fall together. We calculated Spearman’s rank correlation index for each pair of time series corresponding to the same topic.

To determine whether the resulting correlation has an external or internal nature, we calculate correlations when a single day is removed from each time series and repeat this for every day. By removing one day at a time, we obtain a distribution of correlations.

If two outlets only display common behavior because of a few days where singular events happened (and therefore an external similarity) the correlation between their time series will show variations if one of those days is removed. Therefore, singular events result in outlier values in the correlation distribution, which we measure through the concept of *kurtosis*. Kurtosis describes the degree to which a distribution has a heavy tail [65]. The standard definition of kurtosis for a random variable  $X$  is as follows [20]:

$$\text{Kurt}[X] = \frac{\mathbb{E}[(X - \mu)^4]}{(\mathbb{E}[(X - \mu)^2])^2} \quad (3.6)$$

where  $\mu$  is the distribution’s mean. Any univariate normal distribution has a kurtosis of 3, which serves as a standard for comparison. For this reason, *excess kurtosis* is defined as

$$\text{Kurt}_{\text{exc}}[X] = \text{Kurt}[X] - 3 \quad (3.7)$$

Some works use kurtosis to refer to the excess kurtosis, which this work will do from now on [49]. Distributions with negative kurtosis are called *platykurtic*, having less of a heavy tail and generating less outliers than a normal distribution; distributions with positive kurtosis, on the other hand, are called *leptokurtic*, having heavier tails than a normal distribution and producing more outliers. Finally, distributions having a kurtosis of 0 are called *mesokurtic*.

We use two additional, non content-based similarity measures, in order to improve analysis of content-based community structures.

### 3.2.4 Ownership similarity

The first non content-based similarity we explored is based on media ownership. Ownership can be modeled as a directed, acyclic graph where news media outlets and other entities (such as media groups) are represented as vertices; an edge from a vertex to another is understood as the first having full or partial ownership of the second. News media outlets typically do not own any other entities, but they are typically owned by a controller group or other sort of entity that may itself own and be owned by others.

A simple way to measure owner similarity is to consider two news media outlets to be completely similar if they share an owner in any way, and to be completely dissimilar if they do not. In other words, this method defines similarity to be 1 when comparing outlets that have the same owner, and 0 otherwise. However, the complexity of the ownership structure (in particular, owner entities being partially owned by others) suggests that a more subtle approach might be more appropriate.

Taking these observations in consideration, we use an approach related to the *lowest common ancestor* concept. In graph theory, the lowest common ancestor can be defined [9] as follows. Given a directed, acyclic graph  $G = (V, E)$  and  $x, y$  vertices in  $G$ .

- We say a vertex  $a$  is an *ancestor* of  $b$  if there exists a (directed) path from  $a$  to  $b$ .
- Let  $G_{x,y}$  be the subgraph of  $G$  induced by the set of all common ancestors of  $x$  and  $y$ ;
- Let  $\text{SLCA}(x, y)$  be the set of out-degree 0 nodes (i.e. leaves) in  $G_{x,y}$ .
- The lowest common ancestors of  $x$  and  $y$  are the elements of  $\text{SLCA}(x, y)$ .
- Given a common ancestor  $a$  of  $x$  and  $y$ , we define  $d_a(x, y) := d(a, x) + d(a, y)$ , where  $d(a, b)$  is the (potentially weighted) length of the shortest path from  $a$  to  $b$ .

Finally, we define the following similarity measure  $S$ :

$$S(x, y) := \begin{cases} \frac{2}{\min\{d_a(x, y) \mid a \in \text{SLCA}(x, y)\}} & \text{if } x \text{ and } y \text{ have a common ancestor} \\ 0 & \text{otherwise} \end{cases} \quad (3.8)$$

This is a valid similarity measure: it takes values in  $[0, 1]$  and, for example, two news media outlets belonging to the same owner will have a similarity of 1.

We note that this definition also applies for weighted graphs: if edge weights are always equal to or greater than 1, the measure still takes values in  $[0, 1]$ . In our case, we used non-unit weights when relationships involved an entity having partial ownership of another. If an entity owns a fraction  $f$  of another, we set the edge weight between them to  $\frac{1}{f}$ : in this way, as ownership fractions cannot exceed 1, edge weights are always greater than 1.

### 3.2.5 Follower similarity

Another similarity measure not based on content is obtained by focusing on the relationship users have with news media outlets, as a Twitter account's set of followers can be seen as its audience. In this approach, we represent each news media outlet by the set of its followers and therefore compute similarities for pairs of these sets.

The Jaccard similarity coefficient [36] is a first candidate to calculate similarities between sets, defined as follows:

$$J(A, B) := \frac{|A \cap B|}{|A \cup B|} \quad (3.9)$$

Another possibility is to define a similarity with a different denominator:

$$S(A, B) := \frac{|A \cap B|}{|B|} \quad (3.10)$$

Even though this function is clearly not symmetrical (as switching the roles of  $A$  and  $B$  in Equation 3.10 results in a different function), it is useful because of its interpretation: it can be seen as an estimation of  $\mathbb{P}(x \in A \mid x \in B)$ . At least two possible symmetric variations can be formulated for this function, the first of which we use as our similarity measure:

$$S(A, B) := \frac{|A \cap B|}{\max(|A|, |B|)} \quad (3.11)$$

This similarity measure can be seen as an estimation of  $\max(\mathbb{P}(x \in A \mid x \in B), \mathbb{P}(x \in B \mid x \in A))$ . This means that a high similarity value means both probabilities are high. The



other symmetric variation for Equation 3.10 is using the minimum instead of the maximum:

$$S(A, B) := \frac{|A \cap B|}{\min(|A|, |B|)} \quad (3.12)$$

However, we observed that this variation results in a great amount of outlets being highly similar to each other, muddling the analysis. Furthermore, as many small outlets end up being similar to a central, big one (in terms of followers) the resulting graph is star-like in nature, with an absence of community structures. This does not mean this measure is not interesting, but it is probably better suited for other types of analyses.

### 3.3 Community discovery

We wish to see if there are distinctive groups of similar outlets, and how they relate to content diversity. Community discovery is an useful tool for this.

Community discovery algorithms can be applied on the built graphs in order to find groups of highly connected vertices (in this case, news media outlets). As edge relationships are induced by the underlying similarity measures, the resulting communities will correspond to relatively homogeneous news media outlets.

As specified in Chapter 2, community discovery algorithms take a graph  $G = (V, E)$  (which might be weighted) as input, and output a set of communities  $C \subset \mathcal{P}(V)$ . Some algorithms assign a community to every vertex, while others allow for unlabeled vertices: in other words, community discovery algorithms output either a partition or a *partial* partition of the vertex set.

#### 3.3.1 Preprocessing

In general, using a complete graph (as those directly derived from a similarity measure) is excessive, as the distributions of similarity values tend to display an abundance of low similarity values. These usually do not hold any significant meaning, and may muddle the mathematical analysis; conversely, it can be said that similarity values far greater than the median value are the most interesting. This can be remedied by deleting all edges with a weight smaller than a certain threshold  $t$  before further analysis. This leads to a less noisy graph, which helps with both analysis and visualization.

#### 3.3.2 Community analysis

When a community structure has been determined from the similarity graph, it can be analyzed for interpretation. A first step is to characterize these communities by using information

related to the similarity measure and/or additional data, in order to see if they hold some sort of meaning.

An example is follower information. If  $F(x)$  stands for the set of users following outlet  $x$  (its audience), we define the audience coverage of a community  $c$  comprising outlets  $x_1, \dots, x_c$ , given a dataset with outlets  $y_1, \dots, y_n$  as:

$$\text{Coverage}(c) = \frac{|\bigcup_{i=1}^c F(x_i)|}{|\bigcup_{i=1}^n F(y_i)|} \quad (3.13)$$

This quantity measures how much of the audience can be reached through news media outlets from a given community. This fraction can also be expressed as a percentage.

A second step is comparison. If a community structure can be derived from additional news media outlet data (for example, from an ownership structure), it can be compared to the one obtained from the content-based similarity. If multiple similarity measures are formulated, the resulting community structures might also be compared with each other.

There are various different metrics for evaluating clusterings against a ground truth. Some of these metrics are symmetric and can be seen as a measure of the *similarity* between two clusterings. As a community structure is essentially a clustering of nodes, we apply these metrics to see whether the content-based community structures we found are related. We also include the community structures obtained from ownership and follower similarities.

The first metric we use is the Adjusted Rand Index (ARI), a metric related to the accuracy or correspondence between two clusterings, adjusted for the chance grouping of elements [35]. Another is the Adjusted Mutual Information (AMI), a similar metric based on information distance and entropy, also adjusted for chance [63].

Finally, Romano et al. [53] propose a modified concept of AMI, the Standardized Mutual Information (SMI), which is standardized with respect to variance. This results in a metric that can be transformed into statistical significance. We choose to use all three metrics, as they describe different aspects of structure correspondence.

# Chapter 4

## Experimental Methodology

In the following, we describe the way in which we apply our methodology. First of all, we describe the dataset we use, both for news documents and for additional information on news media outlets: tweets from Chilean news outlets and owner information from Poderopedia [28] (plus Twitter follower data), respectively. Next, we detail the similarity measures we use, as well as the algorithms we use for community discovery and analysis.

### 4.1 Methodology overview

Given a dataset  $\mathcal{D}$  that contains both *news documents* and *additional outlet information* and a *similarity measure*  $s$ , our methodology is as follows:

- We **compute** the similarity between every pair of news media outlets in  $\mathcal{D}$ , and obtain a distribution of similarity values. Plotting this distribution allows for visual identification of meaningful similarity values.
- We build a similarity graph for outlets from  $\mathcal{D}$  based on  $s$ , and prune it using a **threshold** deduced from the previous step. This results in a graph that can be explored and **visualized**. In particular, the thresholding procedure can be applied dynamically during visualization.
- We perform **community discovery** on this graph. This outputs a **community structure** that can also be **visualized**.
- We perform both internal and external **analyses** on these communities, to obtain an interpretation for them.
- We generate **user-oriented visualizations** based on our findings.

As we have several similarity measures, these steps are repeated for each one of them.

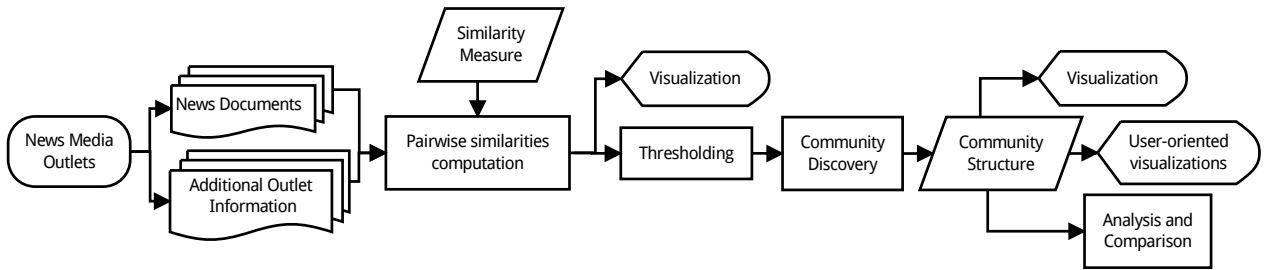


Figure 4.1: Our methodology as a flowchart. This process is repeated using different similarity measures.

## 4.2 Dataset

The dataset we use comprises, as specified before, a source of news documents and a source of additional information about news media outlets.

### 4.2.1 News documents source

Our news documents consist of messages published by Chilean news media outlets in the Twitter microblogging platform. Twitter is an online social networking service oriented to sharing short messages, as our news documents source. We now define some terminology used when talking about the platform [61].

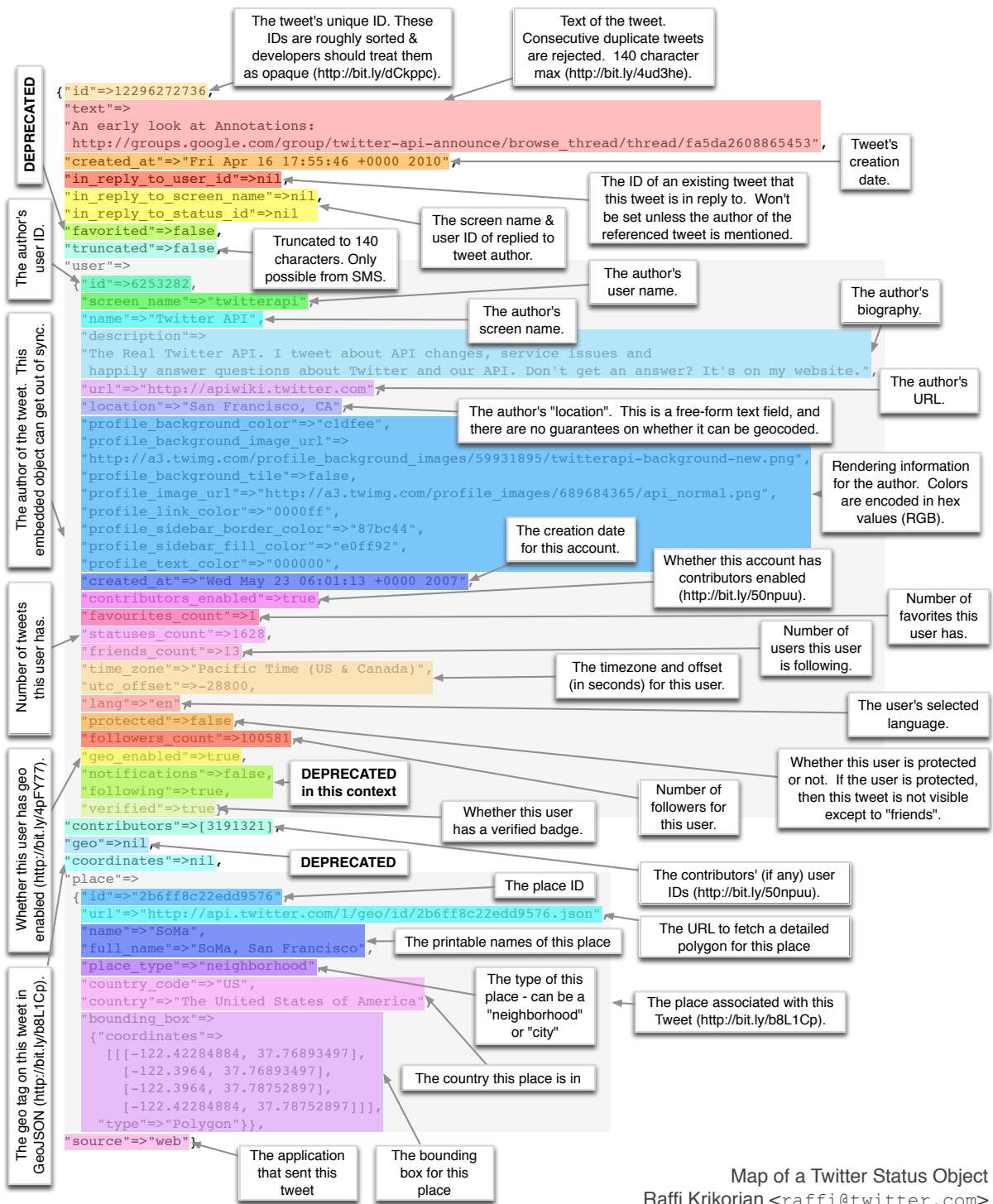
A *tweet* is a message posted to the platform by a user: it may include media such as photos, videos and links, in addition to text. A tweet’s text must not exceed 140 characters in length. A Twitter *account* allows a registered user to publish tweets, follow other users and other activities; users without an account can only read tweets.

A Twitter *@username* is a user’s identifier in the Twitter platform, and they are always preceded by the @ symbol. A tweet’s text might include a username: this is called a mention. A Twitter user can *follow* another user, which consists in subscribing to another’s Twitter account to see their tweets as soon as they post them. The first user is called a *follower*.

We use Twitter as a news document source for a number of reasons. First of all, many different news sources have Twitter accounts, including newspapers, radio stations and individual newsmen. Furthermore, Twitter provides an API<sup>1</sup> that allows, among other things, to retrieve the most recent tweets for a given Twitter username. Tweets retrieved in this way contain a great amount of metadata, as shown in Figure 4.2.

We use a preexisting dataset built by Maldonado et al. [45], which includes all tweets published by some of the most important Chilean news media outlets for a given period. Specifically, it includes 714,973 tweets from 84 news media outlets published between October 20th, 2014 through May 20th, 2015 (including retweets). Both the tweets’ text content and their metadata was obtained.

<sup>1</sup>Application programming interface.



Map of a Twitter Status Object  
 Raffi Krikorian <raffi@twitter.com>  
 18 April 2010

Figure 4.2: A tweet's anatomy when retrieved through the Twitter API, as of April 2010 [41]. Different types of information are shown in different colors. As it can be seen, they contain much more than just text and attached media. The current tweet field list can be found on Twitter's Developer page [60].



### 4.3.1 Similarity measures and internal analysis

Computing similarity values and creating a similarity graph for a given similarity measure closely follows what was detailed in the previous chapter. However, our data is inherently noisy, so additional steps must be taken in order to reduce this effect in our analyses.

#### *Vocabulary similarity*

We wish to compute an overall vocabulary for each news media outlet in our dataset. The computed vocabularies will only be meaningful if we have a reasonable amount of text for each outlet. For these reasons, we discard those outlets that average less than one tweet per day, which leaves us with 79 media outlets. Text is lowercased, and tildes and non-alphanumeric characters are removed from it, as well as multiple spaces, hashtags and hyperlinks. A set of Spanish stop words is also removed from these tweets' content, as well as context-specific stop words. These additional stop words were manually determined by looking at the most frequent words in the dataset and picking those that corresponded to generic terms.

#### *Topic similarity*

For topic similarity, we apply the same filtering and normalizing procedures used with vocabulary similarity.

As explained before, the topic discovery algorithm determines a number of keyword pairs over a small timescale, and coalesces them over a big timescale. In this work, we use hours as the small timescale, days as the big timescale and select the 6 most relevant pairs of keywords for each small scale period.

#### *Temporal correlation*

For this analysis, in addition to the selection and preprocessing applied to vocabulary and topic analysis, we choose to discard those outlets that had tweets talking about the topic for less than three days.

There are many possible similarity structures for analyzing temporal correlation, as time series can be computed for any set of terms. We include two temporal correlation analyses in this work, using the following topics:

**Penta term set:** The Penta case is an ongoing criminal case, in which employees from Chilean holding Penta, employees from Chile's national revenue system and politicians and/or their aides have been charged with tax fraud, bribery and money laundering [26, 51]. We used a term set containing only the word **penta** for this topic, as it is a very characteristic word and not probably used in many other contexts.

**President term set:** A newsworthy subject for Chilean news media outlets is, of course, the country’s current president, Michelle Bachelet. This makes it reasonable to consider her as a topic on her own: we computed time series based on the term set `{michelle, bachelet, presidenta}`.

Finally, we also include kurtosis as an additional variable when plotting similarity distributions. This results in a two-dimensional histogram instead of a one-dimensional one for this similarity measure.

### *Ownership similarity*

The main extra step in computing ownership similarity is to adapt our ownership data to the ownership model explained in Subsection 3.2.4. We consider all relationships described in Subsection 4.2.2 as equivalent. The dataset is incomplete with respect to owner shares: owner data does not always include what percentage of an entity corresponds to each owner. Furthermore, sometimes a certain percentage of ownership is known with certainty for an entity, while the rest of its owners do not have this information. In the case a portion  $r$  of ownership corresponds to  $n$  owners without specific shares known, we split it equally among these owners.

### *Follower similarity*

Follower similarity is the most straightforward. It must be noted, though, that follower data and news document data were collected at different dates, which might be inaccurate in specific situations (such as in the case of an account that was closed between the two data acquisition processes).

## 4.3.2 Similarity distribution exploration

Our first experiment consists in computing the similarity value for each pair of outlets in our dataset. We then plot a histogram of these similarity values in order to visually estimate the point where the “long tail” of the distribution begins. We use the threshold we determine in this way for the community discovery step. This way of choosing a threshold might be automated for bigger, less noisy datasets by, for example, adapting a stop word-removing algorithm based on Zipf’s law [64].

On the other hand, similarity values can be explored through visualizations, which we create using the D3.js [13] Javascript library. The most direct way to visualize the similarity graphs we compute is through a traditional circles-and-lines graph visualizations (as seen in Figure 4.4), assigning news media outlets to vertices and connecting them if their similarity reaches a certain value. We map vertex degree to vertex color through a linear white-black scale, which allows us to identify well-connected news media outlets.



Users can move the slider to select the number of edges displayed.

Users can hover over a vertex to display the username for its corresponding outlet, and drag vertices around to dynamically reposition them.

Audience size is mapped to vertex size; vertex degree is mapped to vertex color, with higher-degree vertices having darker colors.

Edges are colored according to the similarity value between the media outlets it connects; red is used for higher similarity, grey for no similarity and blue for opposite outlets, if applicable.

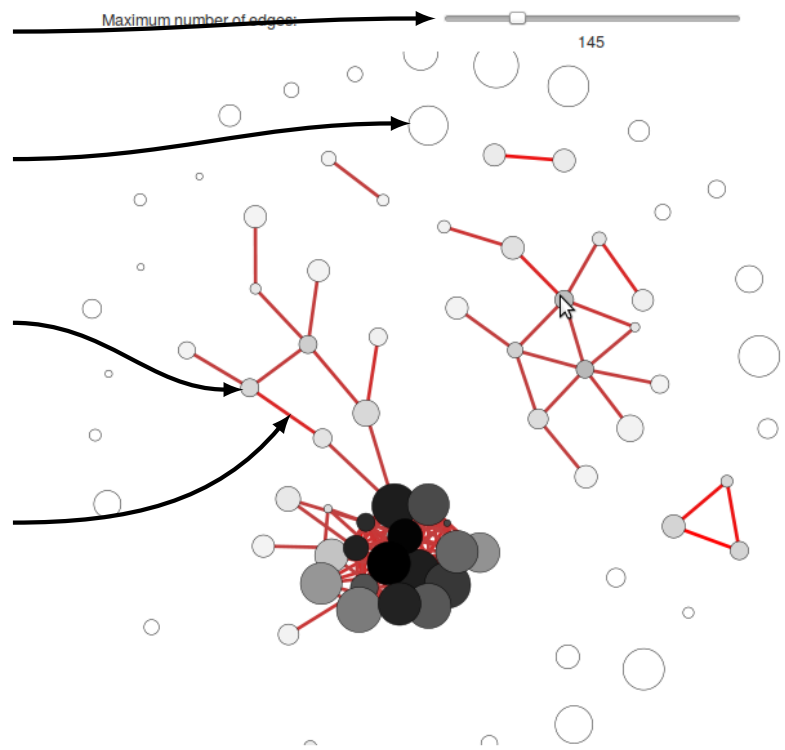


Figure 4.4: A sample of the visualizations built for analyzing the computed similarity graphs.

We map additional information to visual variables to help with interpretation: an outlet's number of followers is mapped to its corresponding vertex's radius using a logarithmic scale. We color edges between vertices according to the similarity between the corresponding media outlets: we use red for a similarity value of 1 (i.e. completely similar outlets), grey for a similarity value of 0 (i.e. unrelated outlets) and blue for a similarity value of -1 when applicable (i.e. opposite outlets). Intermediate values are colored using linear interpolation between these values.

In order to position vertices, we use a force layout algorithm provided by the D3.js Javascript library, a heuristic based on a physical model where vertices repel each other through an electrostatic-like force, while edges apply a spring-like force on the vertices they connect. The combination of these physical interactions usually results in a layout with pleasant aesthetic properties (such as symmetry), making it easier to identify groups and structures within the graph. This layout algorithm is a particular form of the *spring embedder* family of algorithms [8].

Our visualizations do not include all edges: instead, we only draw those having the  $k$  highest similarity values, and provide a slider with which the user can adjust the parameter  $k$ . We allow this parameter to take values between 0 and  $8n$ , where  $n$  stands for the number of vertices in the graph. This restriction is derived from the observation that, beyond this point, the high number of edges results in a very cluttered visualization as a single, highly connected component begins to form.

### 4.3.3 Community discovery

The second experiment is to automatically detect groups of similar outlets, based on their similarity scores. We use the igraph library [18] for community discovery, after applying the previously chosen threshold to the graph. We do not consider isolated vertices for community discovery, as many graph cluster metrics are known to be biased in the presence of this type of vertices [3]. Instead of deleting them altogether, we add them back after community discovery. For summarizing purposes, we consider isolated vertices as the set of *ungrouped elements* (which can be seen as a community itself<sup>2</sup>).

We use a hierarchical, agglomerative community discovery algorithm already implemented within igraph. It amalgamates a pair of communities on each step (starting with each vertex as a separate community), choosing those two whose merge produces the greatest modularity value for the resulting community structure. The algorithm, therefore, greedily optimizes the modularity of the discovered community structure [16].

The dendrogram returned by the algorithm is generated by the merge steps, and can be cut at a specific height to produce a community structure with a target number of communities. To choose the number of communities to aim for and analyze, we cut the dendrogram at different heights and plot modularity and external conductance for each resulting community structure. We then choose a number of communities that results in high modularity and low conductance values, as they are signs of good clustering quality (as specified in Subsection 2.1.3).

### 4.3.4 Community analysis

The first step in our analyses is to characterize the communities we find based on the respective similarity features.

#### *Internal features*

For vocabulary and topic similarity, we compute the most relevant keywords for each detected community. In the case of vocabulary similarity, each outlet has a bag-of-words representation as a normalized tf-idf vector, with each vector component corresponding to a different word. For a given community, we add its members' vectors: the components with the highest values in this resulting vector correspond to the most *relevant* words for the community.

In the case of topic similarity we have, for each day, a keyword graph where connected components are topics. We also have, for each outlet, a daily bag-of-words representation based on the keyword pairs we detect. We use a tf-idf scheme like those specified in Subsection 2.1.1, normalizing tf values according to document length (in number of tweets), as in Equation 2.2. The reason for this type of normalization is that we cannot be sure of having

---

<sup>2</sup>As noted by Bertrand Russell, though, this is a contradiction [55].

identified all topics outlets have touched upon.

This procedure gives a tf-idf score for each keyword, day and media outlet. For each outlet, we sum these scores across all days; afterwards, we add scores for members of each community. The keyword pairs with the highest scores correspond to the *relevant keyword pairs* for the community.

These keywords can provide insights into what induces the community structure. For example, geographical characteristics for a community can sometimes be inferred from its most relevant words and keywords, as they might include location names. In some cases, these aspects can also be quickly and manually inferred from outlet names, as many regional outlets have location-related names. This is useful, as not every outlet in our dataset makes use of Twitter’s geolocalization features.

### *Visualization*

After community discovery, we modify our visualizations to include community structure information. We now map community membership to vertex color, using categorical colors to clearly distinguish news media outlets belonging to different communities: this information is also displayed when the user hovers over a vertex.

In the case of vocabulary and topic similarity, we include community information in the form of relevant words and relevant keyword pairs, respectively. A community’s information is displayed when the user clicks on one of its vertices, in the form of a word cloud next to the graph, as seen in Figure 4.5. Additionally, all vertices from the community whose word cloud is being displayed are highlighted with a thick black border. See Appendix B for more examples.

### *Additional Information*

Follower information can be used to compute consumption metrics for each community structures. One way is to define the full follower set as the set of users that follow at least one of the news media outlets in our dataset. We say a user *follows* a community if it follows at least one of its members. With this, we measure the *audience coverage* for each community, defined as the percentage of the full follower set that follows the community.

Another possible metric is the number of communities each user follows. If communities do represent sets of similar news outlets, this gives us an intuition of the degree of content diversity from a consumer’s point of view, in the context of the chosen similarity measure. An ownership breakdown for each community can also be obtained, as an additional way to characterize each community.

The second step is comparison: we contrast the content-based community structures we find against those based on ownership and audience. As for indices, we use the Adjusted Rand Index, Adjusted Mutual Information and Standard Mutual Information, introduced in

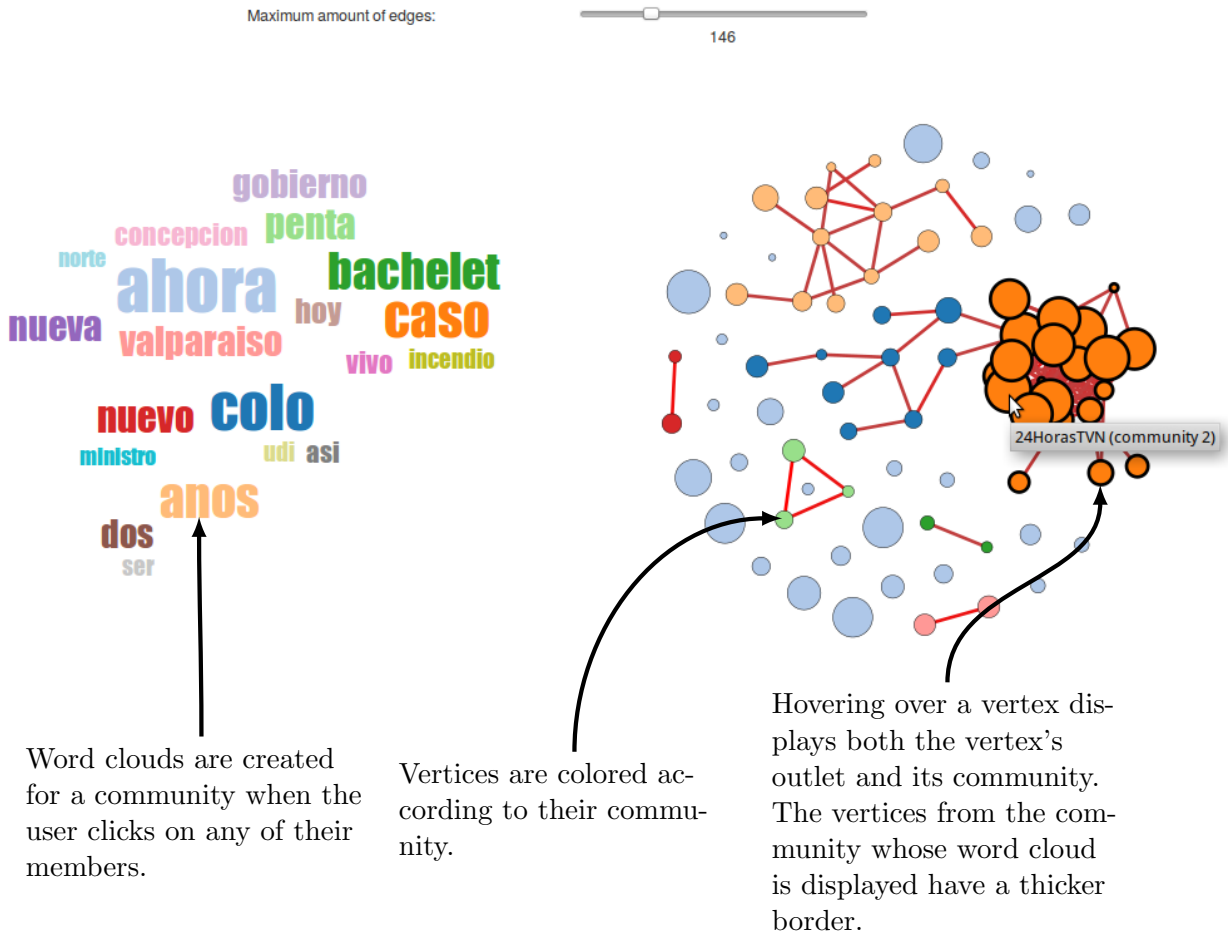


Figure 4.5: The visualization shown in Figure 4.4, modified after the community discovery procedure, in the case of vocabulary similarity.

the previous chapter. These metrics are meant to be applied on traditional clusterings on a common set of elements. Results from our experiments, however, differ in a few aspects:

- **Ungrouped elements:** Our methodology does not include isolated vertices for community discovery and therefore it is not clear whether they should be included for comparison. We consider two approaches to this issue. The first is discarding isolated vertices prior to comparison, which makes sense most of the time, as information is mostly embedded in the relationships between elements instead of being in the elements themselves. We make an exception to this in the case of ownership, as isolated vertices correspond to outlets with a known owner but different from the rest: in this case, we preserve isolated vertices as communities of size 1.
- **Differing elements:** Community structures might have different elements. For example, ownership information might be incomplete: in that case, outlets without a known owner would not be included in the community structure. When comparing two community structures, we solve this by discarding elements present in one of the structures but absent in the other.

# Chapter 5

## Results and Analysis

In the following, we detail the results we obtained. We first show general results and specify individual results for each similarity measure subsequently.

In general, we see two main drivers for content similarity: geographical distribution and ownership. Big, national-scope media with big audiences tend to group together in their own community; the remaining outlets, generally having a more local scope, group according to a mixture of their geographical and ownership features.

An interesting detail we noticed was the presence of two pairs of almost-identical news media outlets for multiple similarity measures. The first of these pairs comprises `@soyvalparaiso` and `@soyvaldiviac1`, and the second comprises `@mercurioafta` and `@estrella_antofa`. Upon further examination, it turns out the corresponding Twitter accounts publish almost exactly the same tweets. Additionally, outlets in the first pair both belong to the Soy Chile media network; all four of these outlets belong to the El Mercurio group. It might be the case that these pairs of media outlets have the same community manager in charge of their Twitter accounts.

Table 5.1 contains both parameters and metrics for thresholding and community discovery procedures for each explored similarity measure. These community structures will be detailed throughout this chapter.

We detail ownership and the follower structures before delving into content-based similarity measures, as we used both ownership breakdown and audience coverage in our analyses.

Similarity	Threshold	Communities	Modularity	Conductance
Ownership	0.2	3	0.5	0.0
Vocabulary	0.5	7	0.38	0.02
Topic	0.125	4	0.6	0.11
Temporal Correlation 1	0.35	5	0.037	0.0
Temporal Correlation 2	0.25	4	0.52	0.047
Follower	0.08	10	0.3	0.0

Table 5.1: Chosen parameters and graph clustering metrics for each explored similarity measure. Temporal Correlation 1 and 2 correspond, respectively, to the term sets `{penta}` and `{michelle, bachelet, presidenta}`. Threshold is the minimum similarity value to establish an edge between outlets; high modularity and low conductance values indicate well-defined communities.

## 5.1 Ownership structure

We group news media outlets according to their ownership: outlets having a common owner will tend to be put in the same group. The ownership similarity graph presents three well-defined connected components when all edges except for zero-valued similarities are kept. These components naturally become the detected communities, summarized in Table 5.2.

ID	Size	Media outlets	Audience coverage %
0	26	@pinguinodiario, @ladiscusioncl, @nacioncl...	89.68
1	14	@elparadiario14, @elrepuertero, @laopinon...	2.76
2	19	@Estrella_Toco, @austral_osorno, @elsurcl...	26.82
3	4	@DiarioLaHora, @lacuarta, @latercera...	28.51

Table 5.2: Ownership community structure. Audience coverage is measured in terms of percentage of unique followers with respect to the whole dataset. The community with an ID of 0 corresponds to ungrouped media outlets.

This analysis shows us a broad overview of media ownership in Chile. In Table Communities with an ID of 2 and 3 correspond to the two biggest news media groups in Chile: the El Mercurio news company and the Copesa media conglomerate, respectively, which form what has been called a newspaper duopoly. The community with an ID of 1 corresponds to a group of digital newspapers, the Mi Voz network. From this analysis, a news media outlet can be put in one of these four classes: being owned by the Copesa group, by the El Mercurio news company, by the Mi Voz group or having a different ownership, possibly unknown (to us, of course). We use these classes for describing ownership when presenting results for other similarity measures.

Visualization of these communities, as seen in Figure 5.1 shows that the Mi Voz community has outlets with smaller audience, while the El Mercurio community has both very big and small outlets. Many outlets having a big audience, such as government accounts, do not belong to any community.

It is interesting that the El Mercurio news company and the Copesa media conglomerate

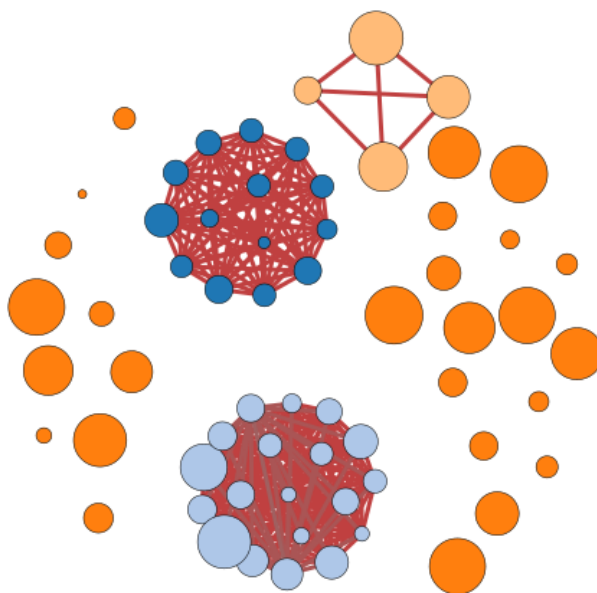


Figure 5.1: Ownership-based similarity graph. Three well-defined communities can be seen.

have similar coverage, while the Mi Voz network presents a much smaller reach. The unknown ownership community has a huge coverage, which might be explained due to the fact that it includes other types of media outlets, such as TV channels and radio stations.

## 5.2 Follower similarity

We also perform a different grouping process, where the similarity value is based on the outlets' sets of followers, in order to characterize the news media's relationship with its audience. As specified previously, we plot the distribution of similarity values in order to only preserve similarities with unusually high values. We observe that the similarity distribution has a huge peak near zero, with smaller peaks afterwards. We use a threshold that keeps all of these smaller peaks.

The computed community structure suggests that follower similarity is deeply related to geographical focus (derived from outlet names). First, the community structure has a big community (with an ID of 8) containing national-scope media outlets. The remaining communities are very small in comparison, both in number of outlets (except for community 3) and audience coverage, and exhibit a geographical nature. In fact, when we inspect these communities, we find that each of these communities only contain news media outlets targeted to inhabitants of specific administrative regions of Chile. Out of Chile's 15 administrative zones, 14 can be identified in these communities, as Table 5.3 shows.

The community with an ID of 3 is interesting, as it contains outlets targeted to six administrative regions. Two groups can be recognized: the Arica and Parinacota Region, the Antofagasta and the Atacama Regions are neighboring regions in the far north of the country, while the Los Lagos, Los Ríos and Araucanía Regions are neighboring regions on



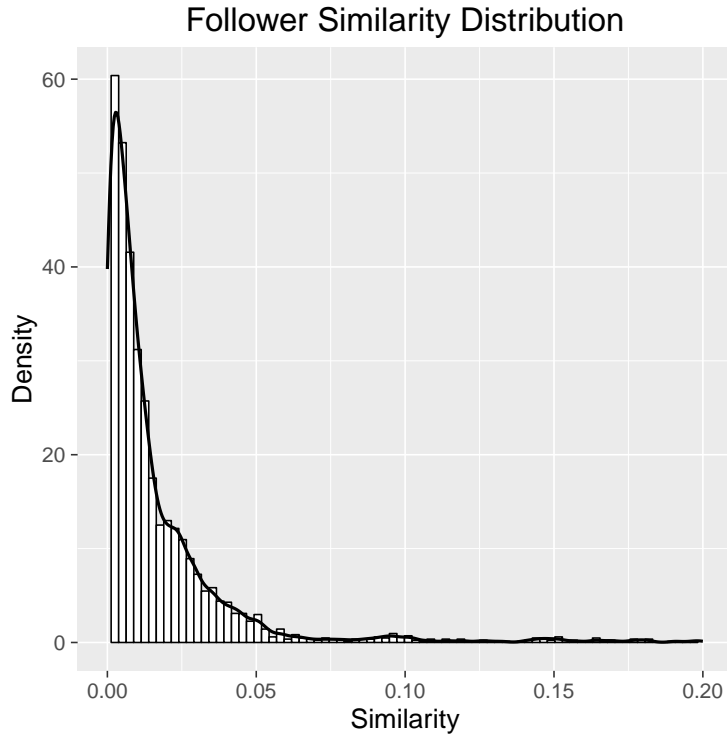


Figure 5.2: Follower similarity distribution. A histogram is shown as bars; the curve shows the estimated density distribution. Small peaks can be seen from similarity values 0.1 and higher: we decided to use a threshold such that these peaks were preserved. Due to the amount of histogram columns, the distribution is shown up to similarity values of 0.2 to avoid visual clutter.

the southern part of the country. This is evidenced when visually exploring the community structure, as seen in Figure 5.3: the orange community is loosely held together and can be interpreted as multiple communities starting to coalesce.

It is interesting to note that if the similarity threshold is raised over a value of 0.1, this grouping phenomenon disappears and the correspondence between communities and geographical zones is even higher. Therefore, this grouping is induced by a set of similarities with values near 0.1 (visible as a peak in Figure 5.2). It might be the case that the chosen similarity measure (dividing over the biggest number of followers) also distorts relationships, amplifying similarities in certain cases.

There is an administrative zone that does not get a community of its own: the Metropolitan Region (Chile’s capital region). However, due to this region containing nearly half of the country’s population, national-scope media outlets are often partially aimed towards this region.

Finally, the set of ungrouped outlets contains a mix of smaller national-scope news media outlets and media outlets aimed at various regions (including both those identified in the smaller communities and those absent).

We summarize these observation with the following interpretation: users follow news

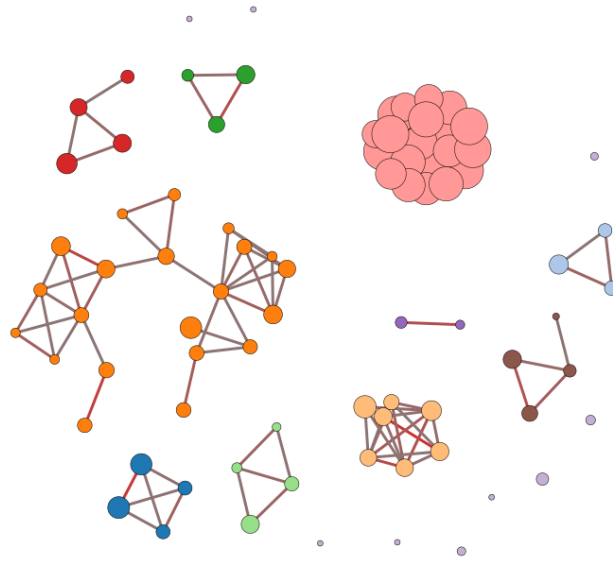


Figure 5.3: Follower-based similarity graph. Two big communities can be observed, as well as many smaller ones.

accounts depending on their geographical scope. Seeing as most users follow outlets from one or two communities, we deduce that most users are interested in national-scope news (as it is a community with a very high audience coverage); in addition to this, we speculate that users follow outlets related to their location. As we do not have user location information in our dataset, the latter statement remains to be tested.

The ownership breakdown for each of these communities can be seen in Table 5.4. As it can be seen, most of the communities have media owned by different groups. Taking into account the geographical nature of these communities, we can say that both the El Mercurio and the Mi Voz groups have regional media outlets, while the Copesa group seems to be focused on national-scope media and news media aimed at the Bío-Bío Region.

ID	Size	Media outlets	Audience coverage %	Geographical Zone(s)
0	12	ahnoticiasmega, noticiasmalleco, diariolasnotic...	1.19	-
1	4	eldia_cl, serenaycoquimbo, elobservatodo...	1.3	Coquimbo Region
2	3	diarioelcentro, el_amaule, laprensacurico	0.61	Maule Region
3	21	elrepuertero, mercurioafta, chanarcillo...	3.02	Arica and Parinacota, Antofagasta, Atacama, Los Lagos, Los Ríos and Araucanía Regions
4	7	diarioconce, cronicachillan, elsurcl...	2.07	Bio-Bio Region
5	3	elrancaguino, noticias_rgua, elrancahuaso	0.5	O'Higgins Region
6	4	laestrellaiqq, diarioelnortino,elongino...	0.47	Tarapacá Region
7	4	eo_enlinea, soyvalparaiso, elaconcagua...	1.2	Valparaíso Region
8	20	24horastvn, tv_mauricio, el_ciudadano...	99.14	National-scope media
9	2	ddivisadero, diariodeaysen	0.09	Aysén Region
10	4	elmagallanews, elpatagonicocl, pinguinodiario...	0.51	Magallanes Region

Table 5.3: Community structure summary for follower similarity. Audience coverage is measured in terms of percentage of unique followers with respect to the whole dataset. The community with an ID of 0 corresponds to ungrouped media outlets. The last column indicates the corresponding geographical zone determined by manual inspection.

Community	Other/unknown ownership %	Mi Voz %	El Mercurio %	Copesa %
0	40.0	20.0	40.0	0.0
1	66.7	33.3	0.0	0.0
2	66.7	33.3	0.0	0.0
3	0.0	37.5	62.5	0.0
4	16.7	16.7	50.0	16.7
5	50.0	50.0	0.0	0.0
6	0.0	50.0	50.0	0.0
7	50.0	25.0	25.0	0.0
8	70.6	0.0	11.8	17.6
9	100.0	0.0	0.0	0.0
10	66.7	33.3	0.0	0.0

Table 5.4: Ownership breakdown for the follower community structure. Each community’s ownership composition is specified in a row.

### 5.3 Vocabulary similarity

The first content-based similarity we explored, based on vocabulary, aims to give a global, broad view of what outlets tweet about. Again, we plot the similarity distribution to distinguish common similarity values from unusual, interesting ones.

The distribution has a peak around 0.13 (instead of around zero as a power law [1] relationship would). There might be a set of common words that were not identified as stop words, as we used a fixed stop word list. It might be interesting to see if this behavior is preserved in the presence of an automated, domain-specific stop word detecting algorithm. From this analysis we choose a threshold of 0.5 that eliminates most of the peak, while preserving two smaller ones that might hold useful information.

As for the obtained community structure, the biggest community (having an ID of 2 in Table 5.5) contains media outlets with a national scope and has a big audience coverage, with the most representative words including terms related to soccer and national-level politics. Two of the biggest cities in the country (Valparaíso and Concepción) can also be found among these words, though with a lower score (not displayed in the table). Santiago, Chile’s capital, is not among these: it might be the case that all newspapers talk about Santiago and therefore it has a low specificity.

The remaining communities correspond, given their most representative words, to broad geographical zones. This makes sense, as news often include location names to provide context to the reader. These location names would then group media outlets targeted to inhabitants of the same geographical regions. These communities are also more loosely coupled, as seen in Figure 5.5

With respect to ownership, most communities seem to contain media owned by the Mi Voz group and the El Mercurio Group, while the Copesa group only participates in the national-scope group and the ungrouped set, in a similar way to the follower-based community

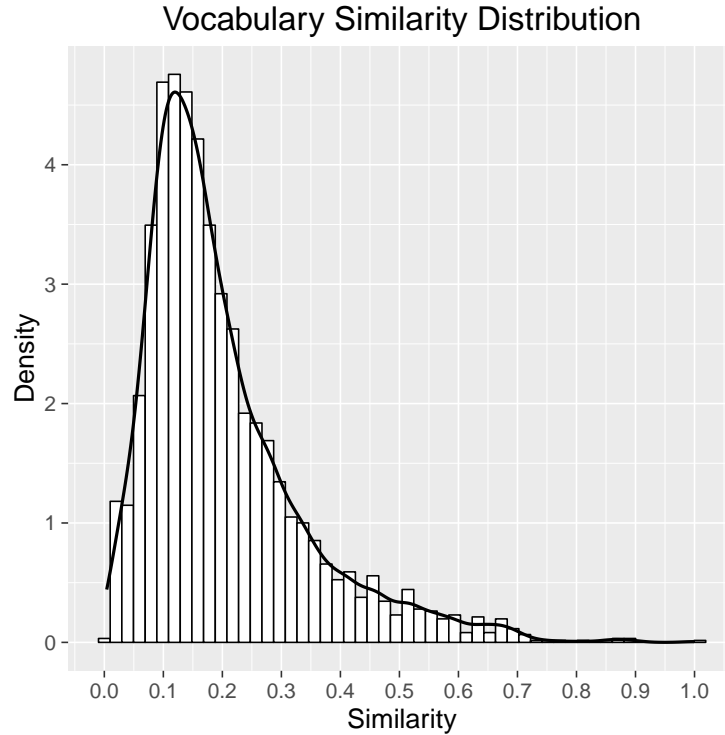


Figure 5.4: Vocabulary similarity distribution. A histogram is shown as bars; the curve shows the estimated density distribution. Interestingly, the distribution displays a peak near a similarity value of 0.13 instead of 0, possibly due to news media outlets sharing a common language.

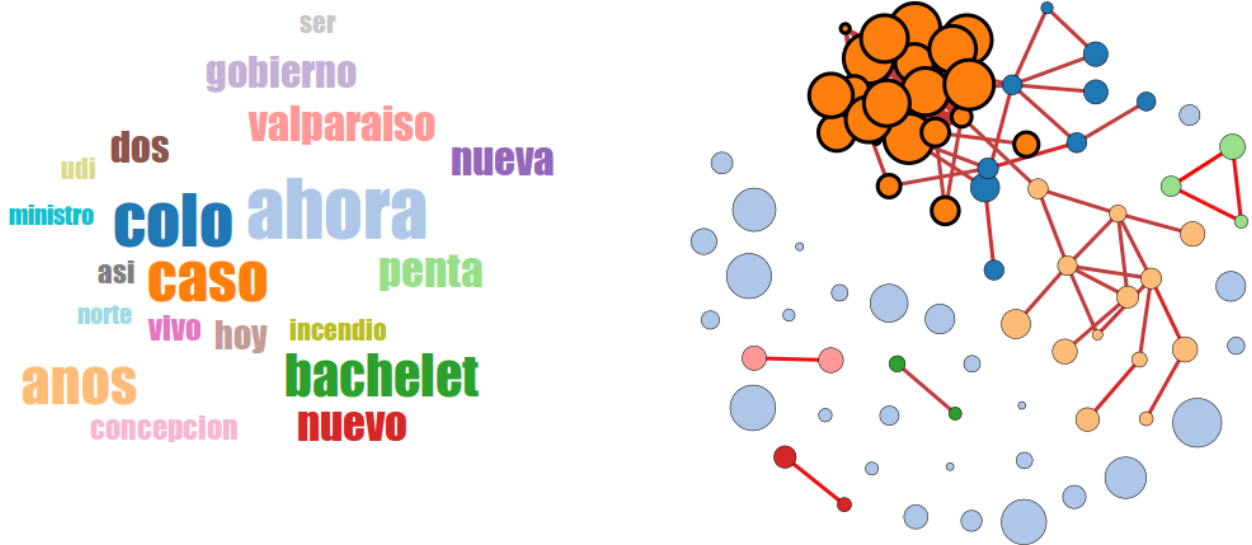


Figure 5.5: Vocabulary-based similarity graph. Three big communities can be observed, plus smaller ones. A word cloud is shown to the left, containing the most distinctive words for the biggest community.

ID	Size	Media outlets	Audience coverage %	Most relevant words
0	27	@estrella_loa, @DiarioLaHora, @lanalhue...	51.67	-
1	9	@laprensacurico, @el_amaule, @laestrellaiqq...	2.01	iquique, horoscopo, tarapaca, maule, serena, participa
2	21	@nacioncl, @24HorasTVN, @soyvaldiviacl...	91.22	ahora, colo, caso, años, bachelet, penta
3	13	@elnavegable, @elrepuertero, @AustralTemuco...	2.57	horoscopo, puerto, hoy, osorno, montt, region
4	2	@ddivisadero, @diariodeaysen	0.09	coyhaique, aysen, horoscopo, puerto, magallanes, opinion
5	3	@estrella_antofa, @Estrella_Toco, @mercurioafta	0.53	antofagasta, metales, sector, antofagastinos, londres, bolsa
6	2	@EstrelladeArica, @elmorrocotudo	0.28	arica, antofagasta, parinacota, marcos, san, hoy
7	2	@CronicaChillan, @ladiscusioncl	0.53	chillan, ñuble, arica, quillon, ñublense, san

Table 5.5: Community structure summary for vocabulary similarity. Audience coverage is measured in terms of percentage of unique followers with respect to the whole dataset. The community with an ID of 0 corresponds to ungrouped media outlets. The six most relevant words (this is, with the highest cumulative tf-idf score for the community) are shown for each community.

Community	Other/unknown %	Mi Voz %	El Mercurio %	Copesa %
0	47.1	0.0	35.3	17.6
1	14.3	71.4	14.3	0.0
2	66.7	5.6	22.2	5.6
3	16.7	58.3	25.0	0.0
4	100.0	0.0	0.0	0.0
5	0.0	0.0	100.0	0.0
6	0.0	50.0	50.0	0.0
7	50.0	0.0	50.0	0.0

Table 5.6: Ownership breakdown for the vocabulary ownership structure. Each community’s ownership composition is specified in a row.

structure.

An overall interpretation for these communities is that local news aimed toward a specific geographical region are a consistent feature of some media outlets, which sets them apart and groups them through this analysis: these features are stronger than ownership-originated similarities. An important factor for this interpretation is that this analysis is performed in a global scale, as vocabularies are computed over the whole time period.

## 5.4 Topic similarity

Topic similarity has a behavior much more akin to a power law, with a large peak around 0 and a tail toward higher values. We observe a small peak near a similarity value of 0.18, which might correspond to a community substructure. We use a threshold of 0.125 to preserve this.

As seen in Table 5.7, the computed community structure has a big community (with an ID of 2) containing national-scope media outlets. This community’s most relevant keyword pairs mainly includes political figures. There is also a small community based around one of the previously-mentioned pairs of near-identical outlets, containing outlets related to the Antofagasta region: their keywords relate to stock exchange (possibly involving the region’s mining industry).

The other two communities do not seem to display distinctive features and their most relevant keywords do not display a clear theme, even though they are very well connected, as seen in Figure 5.7. The lack of a common theme for some communities might be due to the fact that keyword relevance is computed over the full time period covered by the dataset, while keywords and similarities are computed on a daily scale. In other words, the similarity that groups these outlets might not be because of overarching topics, but because of outlets reporting on the same topics *concurrently*.

In addition to this, there is a lack of smaller, local topics, in contrast with vocabulary analysis. It is possible that these either are not detected by the algorithm we used, or are overshadowed by bigger, national-scope topics. This would alter the interpretation for this

### Topic Similarity Distribution

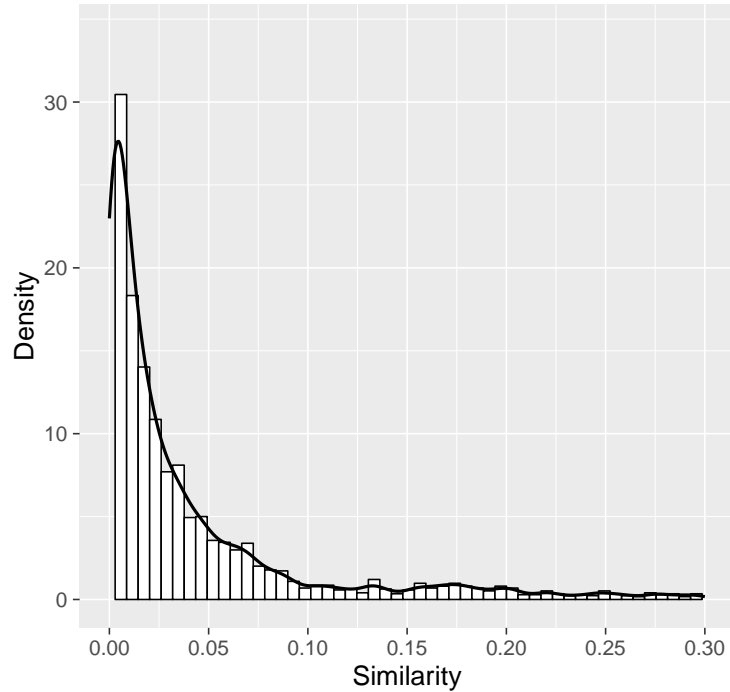


Figure 5.6: Topic similarity distribution. A histogram is shown as bars; the curve shows the estimated density distribution. A small peak around a similarity value of 0.18 can be seen, which might be an indicator of an underlying two-level community structure. Due to the amount of histogram columns, the distribution is shown up to similarity values of 0.3 to avoid visual clutter.

similarity measure: it would be focused on national topics. It might also be the case that local-oriented outlets do not cover local issues in a consistent way, and therefore these issues are not detected as topics. Changes could be done to the topic-detection algorithm, as well as relevant keyword extraction in order to get a better understanding of this behavior.

Ownership analysis provides more interesting results. As seen in Table 5.8, the community with an ID of 1 is completely owned by the Mi Voz group, while communities 3 and 4 are owned by the El Mercurio group. This suggests that ownership has a strong influence on national-scope topics discussed by media outlets, specifically on their timing. We interpret this, in combination with the nature of the most relevant topics, as evidence of owners feeding regional media with national-scope news coming from a common source.



ID	Size	Media outlets	Audience coverage %	Keyword pairs
0	29	@diarioelcentro, @eo_online, @onemichile...	26.82	-
1	14	@laopinon, @elobservatodo, @elmagallanews...	2.76	“city participa” “vida auto” “inglaterra participa” “gana campeonato” “ser expositor” “nuevos pregrado”
2	21	@PublimetroChile, @tv_mauricio, @La_Segunda...	97.43	“presidenta bachelet” “sigue señal” “caso penta” “subsecretario aleuy” “von baer” “hugo bravo”
3	12	@diariolider, @laestrellaiqq, @diarioatacama...	3.13	“revisar tendencias” “quinta vergara” “benegas ofensivo” “cariola vallejo” “desordenes vergara” “menos muertos”
4	3	@estrella_antofa, @mercurioafta, @Estrella_Toco	0.53	“sector norte” “alza cerro” “sector alto” “baja cerro” “alza bolsa”

Table 5.7: Community structure summary for topic similarity. Audience coverage is measured in terms of percentage of unique followers with respect to the whole dataset. The community with an ID of 0 corresponds to ungrouped media outlets. The six most relevant keyword pairs (this is, those with the highest cumulative scores for the community) are shown for each community. Words in each do not have an inherent order, but we manually order them if meaning is evident.

Community	Other/unknown %	Mi Voz %	El Mercurio %	Copesa %
0	81.2	0.0	6.2	12.5
1	0.0	<b>100.0</b>	0.0	0.0
2	72.2	0.0	16.7	11.1
3	0.0	0.0	<b>100.0</b>	0.0
4	0.0	0.0	<b>100.0</b>	0.0

Table 5.8: Ownership breakdown for the topic ownership structure. Each community’s ownership composition is specified in a row.

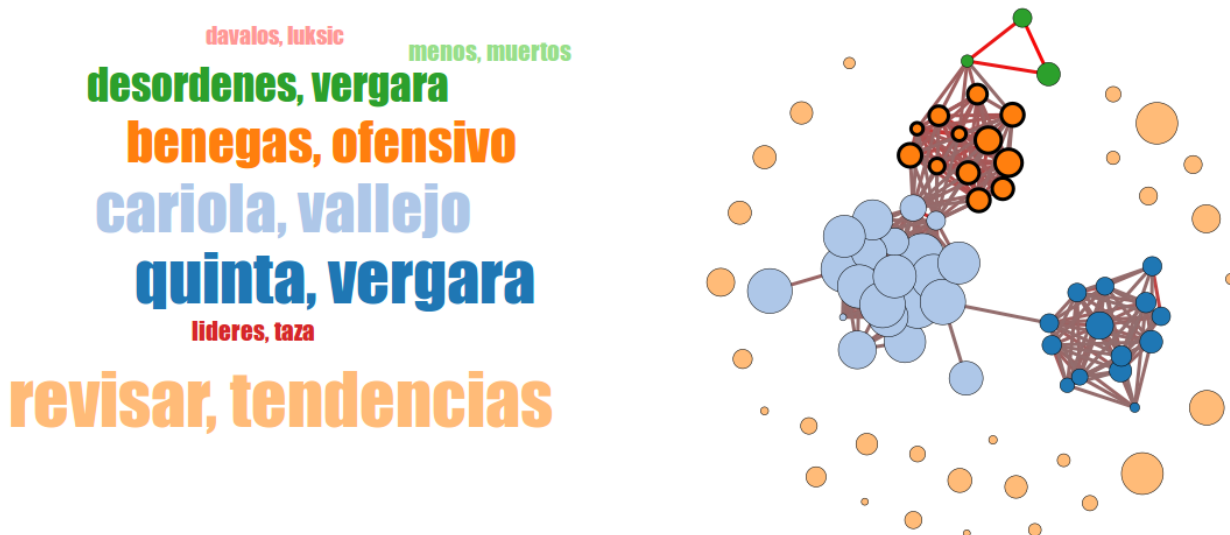


Figure 5.7: Topic-based similarity graph. Communities display a higher level of well-connectedness than for vocabulary similarity. A word cloud is shown to the left, containing the most distinctive keyword pairs for the highlighted community.

## 5.5 Temporal correlation

A similarity measure based on temporal correlation can tell us if a group of news media outlets talks about a given topic at the same time, connecting with the idea of concurrence from the previous analysis. We use two topics: the *Penta* topic, based on an ongoing criminal case, and the *President* topic, based on a newsworthy subject (Chile’s president).

### 5.5.1 Penta term set

We observe most correlation values are small, with varying kurtosis values. There is, as seen in Figure 5.8, a group that differentiates itself from the main distribution, displaying higher correlation values, with higher kurtosis as correlation increases. There is also a small peak near a correlation value of 0.4, which we choose to preserve by introducing a threshold with a value of 0.35. We note that all kurtosis values are positive, indicating an external origin for common behavior. However, the peak we preserve has a lower kurtosis than most similarities, which suggests it might be influenced by internal factors. In other words, although the observed concurrence is caused by media outlets responding to an external event, there is a group of similarities that suggests this concurrence has a partial origin in the outlets themselves.

The obtained community structure contains a single, well connected, relatively big community (with an ID of 1) and four smaller ones, as seen in Figure 5.9. We see, when taking ownership into account, that this big community is completely owned by the El Mercurio group, similar to community 3 in the previous analysis. Three of the smaller communities are almost completely owned by the Mi Voz group. This fits the previous observation on

Correlation and Kurtosis Distribution for 'President' term set

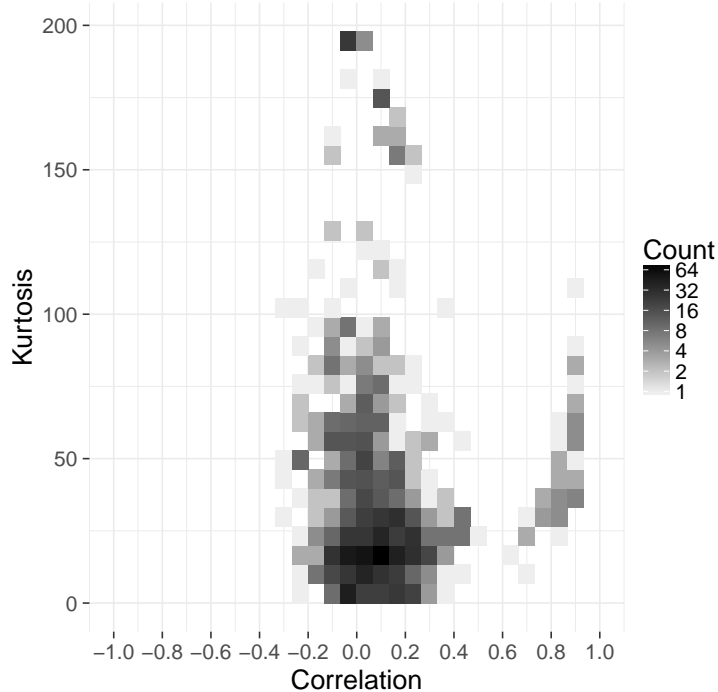


Figure 5.8: Correlation and kurtosis distribution for temporal correlation for term set {penta}. The main part of the distribution seems to be centered around 0, but a group of very high correlations can be seen, as well as a small peak around a correlation value of 0.4.

external versus internal origins for high similarity values: media outlets report at the same time on the topic because they belong to the same entity.

It is interesting that the biggest news media outlet from the El Mercurio group (El Mercurio) is not in community 1: it only contains local news media outlets, targeted to the inhabitants of specific cities and regions. Again, we observe groups of regional news media outlets grouping together according to their owners: this is consistent with the topic-based analysis, as the chosen topic has a national scope. This observation allows us to improve our interpretation for this analysis: when owned by the same entity, regional news media outlets report on national topics simultaneously. This is possibly caused by these outlets having a single, owner-managed source for national-scope news.

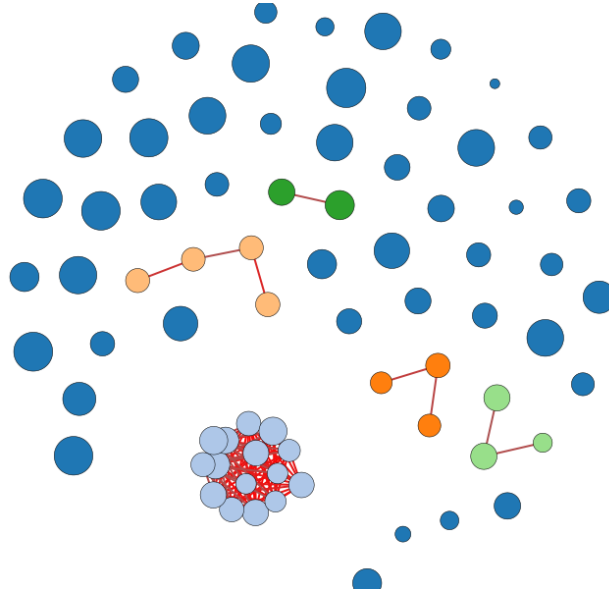


Figure 5.9: Temporal correlation-based similarity graph for the Penta term set. Small but well-defined communities can be seen.

ID	Size	Media outlets	Audience coverage %
0	46	@elrancaguino, @latercera, @biobio...	99.4
1	15	@laestrellaiqq, @estrella_loa, @Estrella.Toco...	3.83
2	3	@Elvacanudo, @elrepuertero, @elmagallanews	0.33
3	4	@elconcecuente, @elnaveghable, @el_amaule...	0.63
4	2	@ladiscusioncl, @eldia_cl	1.34
5	3	@elmartutino, @elparadiario14, @pinguinodiario	0.77

Table 5.9: Community structure summary for temporal correlation when using {penta} as the term set. Audience coverage is measured in terms of percentage of unique followers with respect to the whole dataset. The community with an ID of 0 corresponds to ungrouped media outlets.

Community	Other/unknown %	Mi Voz %	El Mercurio %	Copesa %
0	60.6	15.2	15.2	9.1
1	0.0	0.0	<b>100.0</b>	0.0
2	0.0	<b>100.0</b>	0.0	0.0
3	0.0	<b>100.0</b>	0.0	0.0
4	100.0	0.0	0.0	0.0
5	33.3	66.7	0.0	0.0

Table 5.10: Ownership breakdown for the temporal correlation’s ownership structure when using {penta} as the term set. Each community’s ownership composition is specified in a row.

## 5.5.2 President term set

The correlation distribution is, as in the previous case, centered around 0, but with a group of higher correlations. Again, this group presents lower (though positive) kurtosis values with respect to the rest of the distribution, which suggest that response to this topic is partially influenced by internal factors (such as media outlets having a common editorial line).

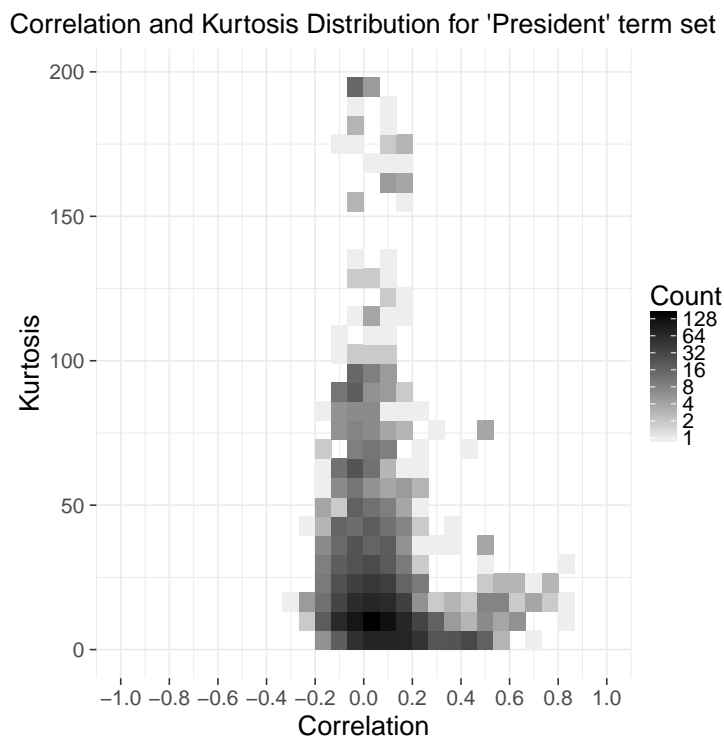


Figure 5.10: Correlation and kurtosis distribution for temporal correlation for term set {michelle, bachelet, presidenta}. The main part of the distribution seems to be centered around 0, but a group of higher correlations can be seen.

The obtained community structure has three main communities, one of them being noticeably bigger than the others, both in number of outlets and audience coverage. As seen in Figure 5.11, two of these communities are very well-connected.

Ownership analysis of this community structure shows results consistent with those for the previous term set: one of the communities is completely owned by the El Mercurio group, comprising regional news media outlets, while another is completely owned by the Mi Voz group (the less connected one, which actually broke into multiple communities in the previous analysis). The main difference with the Penta term set is the biggest community, which comprises many different news media outlets, both in ownership and geographical scope. Our interpretation is that these media outlets consider presidential issues as important, and report on these as soon as they can, becoming correlated.

ID	Size	Media outlets	Audience coverage %
0	19	@serenaycoquimbo, @DiarioConce, @mercurioafta...	28.73
1	24	@ElBoyaldia, @S_Schwartzmann, @ladiscusioncl...	97.21
2	13	@AustralTemuco, @laestrellaiqq, @estrella_loa...	3.12
3	15	@elparadiario14, @diarioelnortino, @elrepuertero...	3.0
4	2	@lun, @eo_onlinea	1.15

Table 5.11: Community structure summary for temporal correlation when using {michelle, bachelet, presidenta} as the term set. Audience coverage is measured in terms of percentage of unique followers with respect to the whole dataset. The community with an ID of 0 corresponds to ungrouped media outlets.

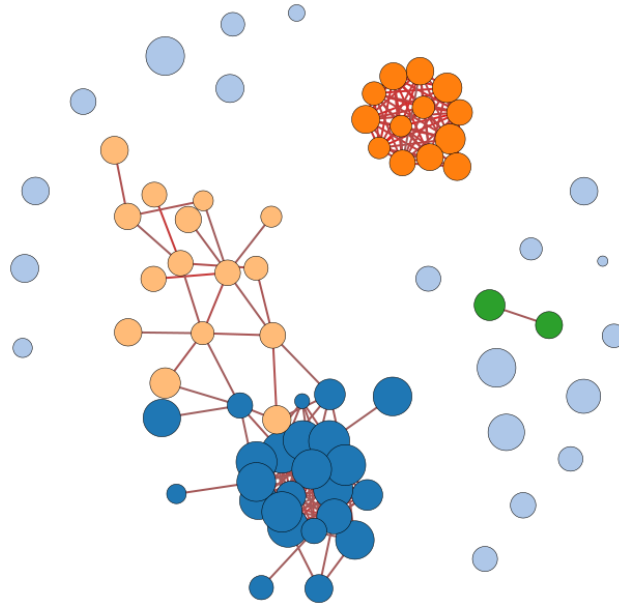


Figure 5.11: Temporal correlation-based similarity graph for the President term set. Well-defined communities can be seen.

Community	Other/unknown %	Mi Voz %	El Mercurio %	Copesa %
0	60.0	0.0	30.0	10.0
1	75.0	5.0	10.0	10.0
2	0.0	0.0	<b>100.0</b>	0.0
3	7.1	<b>92.9</b>	0.0	0.0
4	50.0	0.0	50.0	0.0

Table 5.12: Ownership breakdown for the temporal correlation’s ownership structure when using {michelle, bachelet, presidenta} as the term set. Each community’s ownership composition is specified in a row.

## 5.6 Comparison

We compare the content-based community structures we obtain with the ones we obtain from additional information in order to relate our analysis to notions of producer and consumer diversity.

The first two metrics we use, the Adjusted Rand Index (ARI) and the Adjusted Mutual Information (AMI) take values on  $[0, 1]$ : higher values corresponding to structures with a higher degree of similarity. The third metric, the Standardized Mutual Information (SMI), is not bounded, but higher values still represent a higher correspondence. The usefulness of this metric is that it can be translated to statistical significance with respect to a certain model. According to Romano et al. [53], SMI values over 4.36 (bolded in Tables 5.13 and 5.14) result in an upper bound for the p-value (under a hypergeometric null hypothesis) of 0.05.

The first comparison we perform is against the follower structure we obtained. As it can be seen in Table 5.13, vocabulary similarity has the greatest correspondence with the follower structure explained before. We interpret this as a consequence of both consumer behavior and outlet vocabularies being related to geographical scope.

Similarity	Considered Outlets	ARI	AMI	SMI
Vocabulary	50	0.3561	0.3805	<b>8.922</b>
Topics	47	0.4210	0.2946	<b>8.02</b>
Temporal Correlation 1	25	0.1590	0.1267	1.495
Temporal Correlation 2	49	0.3242	0.2053	<b>6.22</b>

Table 5.13: Comparison metrics with respect to the follower-based similarity structure. Temporal correlation instances 1 and 2 correspond, respectively, to the term sets {penta} and {michelle, bachelet, presidenta}.

On the other hand, when comparing communities against the ownership structure we have (as seen in Table 5.14), we observe that topic similarity exhibits the highest correspondence, followed by temporal correlation. This confirms our previous observations: ownership is related to topics news media outlets talk about.

Similarity	Considered Outlets	ARI	AMI	SMI
Vocabulary	46	0.0169	0.0185	0.416
Topics	47	0.4473	0.3286	<b>11.137</b>
Temporal Correlation 1	26	0.8063	0.6512	<b>8.47</b>
Temporal Correlation 2	49	0.4211	0.2575	<b>10.132</b>

Table 5.14: Comparison metrics with respect to the ownership-based similarity structure. Temporal correlation instances 1 and 2 correspond, respectively, to the term sets {penta} and {michelle, bachelet, presidenta}.

## 5.7 Insights

We finish this section with a summary of the insights we gained from these results.

- Users follow news accounts depending on their geographical scope. Most users seem to be interested in national-scope news,
- Geographical scope influences an outlet's vocabulary. Audience seems to be related to vocabulary, too, but we interpret this as a byproduct of its relation to geographical scope: we hypothesise that users are interested in news outlets that focus on their geographical locations.
- When owned by the same entity, smaller, regional news media outlets report on national topics simultaneously. This is possibly caused by these outlets having a single, owner-managed source for national-scope news. Regional media outlets seem to be heavily influenced by their owners when reporting on national-scope news, acting act as if they depend on an owner-provided source for this type of news.



# Chapter 6

## Conclusions

In this work, we presented a methodology for characterizing content diversity in news media outlets through their Twitter feeds, and its application for statically analyzing the Chilean news media system. Our work provides, therefore, a way of characterizing media diversity based on content. We believe this provides a new facet for pluralism analysis, which complements approaches focused on external pluralism.

We found a correspondence between external similarity features (such as audience and ownership) and content-based relationships. Namely, we found that news media outlets' vocabularies are related to their follower base (possibly through an underlying geographical relationship) and that the topics these outlets talk about in their feeds are related to media ownership.

We also found that repeating topics do not seem to be related to local issues. Our interpretation is that this happens because of two reasons: first, local media sharing a common owner depend on a single source for national topics, which makes them stand out in our analyses; secondly, although local media outlets provide information about events regarding their geographical locations, they do not do this in a consistent way so that the topic-detection algorithm we used can detect it.

A lack of ownership diversity is often seen as a sign of a lack of diversity in content, and our work might serve as a link between these two notions. According to other works, Chile has a newspaper duopoly [15], and our work suggests that this is reflected in content: we show that groups of news outlets having the same owner display little topic diversity.

### 6.1 Applications

A potential application for this work is the development of tools to allow news media consumers to better understand the type of content they consume, and assess its diversity. These tools can also help them find alternatives to enrich the news they receive. We built a set of visualizations that shows a way in which such a tool might be implemented. Presented as a

web form, it takes a Twitter handle as its input, plus a choice of similarity measures, which includes vocabulary-based, topic-based, ownership-based and follower-based similarities.

## Recommendation proof-of-concept

Your info:

Twitter Username

Criteria to include: Vocabulary Topics Ownership Follower

Figure 6.1: The starting point for our recommendation proof of concept. The user’s Twitter account is to be analyzed with respect to the selected criteria.

For each selected similarity metric, the user is presented with two ring graphs. The first of these shows the communities we found based on the metric, with arc lengths proportional to the number of media outlets in each community.

### Vocabulary Analysis

We found 8 groups of similar news media outlets when grouping them according to a vocabulary-based criterion. Click on a group to see some of the media outlets in it.

The following is your consumption with respect to the same criterion:



Figure 6.2: Ring graphs are used to show diversity, both in available content and in the user’s consumption.

The second graph represents the user’s consumption diversity. The full ring represents the total set of news media media outlets in our dataset the user follows, while arcs are proportional to the number of members of each community for the similarity metric. We also include the similarity graphs (with a fixed threshold) to offer a greater degree of detail.

Finally, a simple greedy algorithm is applied to recommend news outlets to the user. Given a community structure comprising sets of similar news media outlets, we say a user has a diverse consumption with respect to this community structure if it receives content from at least one outlet of each set. Intuitively, we say a user has a diverse consumption with respect to a number of community structures if they do with respect to each of them.

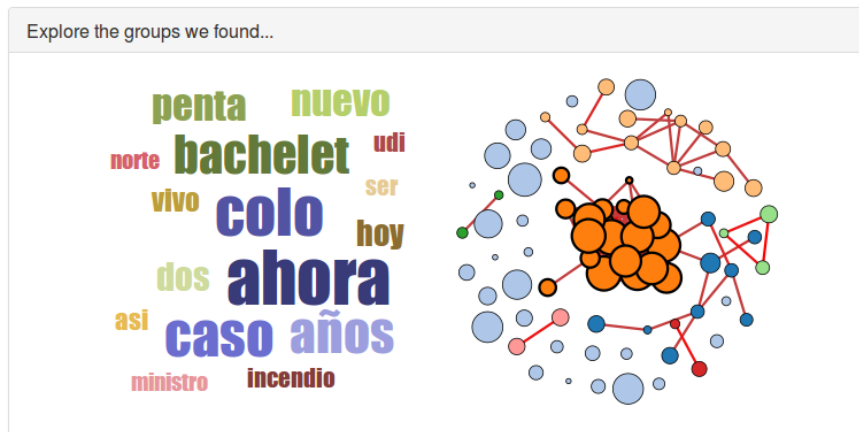









Figure 6.3: The user can optionally explore the similarity graphs for each criterion, with top keywords when applicable.

With this in mind, a recommendation is a set of news media outlets such that, when added to a user’s consumption, it becomes diverse with respect to a target set of community structures. Ideally, we would like to offer a small set of suggestions, to avoid overwhelming the user.

## Our recommendation:

You might want to follow these Twitter accounts to receive a more diverse content :

-  [@estrelladearica](#) (helps with vocabulary, topics and ownership diversity)
-  [@elboyaldia](#) (helps with vocabulary, topics and ownership diversity)
-  [@estrella\\_antofa](#) (helps with vocabulary, topics and ownership diversity)
-  [@diariodeaysen](#) (helps with vocabulary diversity)
-  [@ellanquihue](#) (helps with vocabulary, topics and ownership diversity)
-  [@latercera](#) (helps with ownership diversity)
-  [@ladiscusioncl](#) (helps with vocabulary diversity)

Click on "Try again!" again if you want to see a different set of suggestions!

Try again!

Figure 6.4: Recommendations are displayed along an explanation of which diversity criteria they help improve.

Other possible applications of our methodology include the analysis of media outlets not focused on news. An important issue, though, is to define pluralism and content diversity for works of fiction and other types of information, and formulate similarity measures that reflect these definitions. Additionally, data from sources other than Twitter could be analyzed, provided an appropriate similarity measure is defined.

Furthermore, analyses could be performed over social network users instead of restricting the process to media outlets. Finding communities of users that generate similar content might be a good way to characterize the so-called “Tweetsphere” for a given time period.

## 6.2 Limitations

One of the limitations our work has lies in the fact that our dataset does not cover all news media outlets. Many radio stations include news in their feeds, as well as government agencies. Influential Twitter users could also be included, as they play an important role in news propagation.

Additionally, even though Twitter provides a uniform interface to extract news feeds from many sources, the platform is not used in the same way by them. Some news media outlets attempt to convey news summaries through their tweets, while others attempt to entice users into visiting their websites.

## 6.3 Extensions and Improvements

Our methodology can be improved. The methodology we used includes applying a manually-obtained threshold to similarity graphs. This step might be replaced by a more formal criterion: for example, defining a threshold as the similarity distribution's mean plus a number of standard deviations.

We observed that big, national-scope media ended up together in a single community; however, there might be useful and important information within these national-scope communities. The dendrograms we obtain from community discovery algorithms contain details we flatten out when we cut them at a certain height. Preserving the full hierarchical community structure might better insights. This would bring up some challenges, such as creating appropriate visualizations for these structures and perform comparisons between them.

There is also a time scale issue. Our similarity metrics are computed either over the full time period covered by our dataset or on a daily basis. Allowing for flexibility in this respect would possibly yield interesting results, depending on the similarity metric. For example, vocabularies could be computed on a daily basis, which might produce results closer to those we obtained through topic analysis. Other interesting analyses could be at the hour scale: media outlets might choose certain times of the day to report on social media, given that different types of media seem to attract their audience at different times [21].

Other similarity measures can also be applied to our methodology, each of them conveying a different meaning and a different angle of looking at news outlets. For example, a sentiment-based similarity with respect to certain public characters might group outlets according to their political leaning. If obtainable, other audience characteristics could also be considered to help with interpretation, such as income and political affiliation or orientation. Finally, a bigger dataset, including a more varied type of news media outlets and/or having news articles instead of Twitter headlines could address some of the limitations already mentioned.

# Bibliography

- [1] Adamic, L.A. and B.A. Huberman: *Zipf's law and the internet*. Glottometrics, 3(1):143–150, 2002, ISSN 1617-8351.
- [2] Agencia de Noticias: *CNTV rechaza aplicar sanciones contra canales de TV que censuraron huelga de hambre de Presos Políticos Mapuche*. <http://www.agenciadenoticias.org/cntv-rechaza-aplicar-sanciones-contra-canales-de-tv-que-censuraron-huelga-de-hambre-de-presos-politicos-mapuche/>, December 2010. Accessed: 2016-09-14.
- [3] Almeida, H., D.O.G. Neto, W.M. Jr., and M.J. Zaki: *Is there a best quality metric for graph clusters?* In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I*, pp. 44–59, 2011. [http://dx.doi.org/10.1007/978-3-642-23780-5\\_13](http://dx.doi.org/10.1007/978-3-642-23780-5_13).
- [4] An, J., M. Cha, K. Gummadi, J. Crowcroft, and D. Quercia: *Visualizing media bias through Twitter*. In *AAAI Workshop - Technical Report*, vol. WS-12-01, pp. 2–5, Dec. 2012, ISBN 9781577355649.
- [5] An, J., M. Cha, P.K. Gummadi, and J. 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, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2825>.
- [6] Bakshy, E., S. Messing, and L. Adamic: *Exposure to ideologically diverse news and opinion on facebook*. *Science*, 2015, ISSN 0036-8075. <http://science.sciencemag.org/content/early/2015/05/06/science.aaa1160>.
- [7] Bastian, M., S. Heymann, and M. Jacomy: *Gephi: An open source software for exploring and manipulating networks*. In *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17-20, 2009*, 2009. <http://aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.
- [8] Battista, G.D., P. Eades, R. Tamassia, and I.G. Tollis: *Algorithms for drawing graphs: an annotated bibliography*. *Comput. Geom.*, 4:235–282, 1994. [http://dx.doi.org/10.1016/0925-7721\(94\)00014-X](http://dx.doi.org/10.1016/0925-7721(94)00014-X).
- [9] Bender, M.A., M. Farach-Colton, G. Pemmasani, S. Skiena, and P. Sumazin: *Lowest common ancestors in trees and directed acyclic graphs*. *J. Algorithms*, 57(2):75–94, 2005. <http://dx.doi.org/10.1016/j.jalgor.2005.08.001>.

- [10] Benkler, Y., H. Roberts, R. Faris, A. Solow-Niederman, and B. Etling: *Social mobilization and the networked public sphere: Mapping the SOPA-PIPA debate*. Political Communication, 32(4):594–624, 2015. <http://dx.doi.org/10.1080/10584609.2014.986349>.
- [11] Blei, D.M.: *Probabilistic topic models*. Commun. ACM, 55(4):77–84, 2012. <http://doi.acm.org/10.1145/2133806.2133826>.
- [12] Blei, D.M., A.Y. Ng, and M.I. Jordan: *Latent dirichlet allocation*. Journal of Machine Learning Research, 3:993–1022, 2003. <http://www.jmlr.org/papers/v3/blei03a.html>.
- [13] Bostock, M., V. Ogievetsky, and J. Heer: *D3: Data-driven documents*. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2011. <http://vis.stanford.edu/papers/d3>.
- [14] Castillo, C., M. Mendoza, and B. Poblete: *Predicting information credibility in time-sensitive social media*. Internet Research, 23(5):560–588, 2013, ISSN 1066-2243. <http://www.emeraldinsight.com/10.1108/IntR-05-2012-0095>.
- [15] Castro, C.: *Industrias de Contenidos en Latinoamérica*. Meta, 2008. [http://www.razonypalabra.org.mx/libros/libros/Gdt\\_eLAC\\_meta\\_13.pdf](http://www.razonypalabra.org.mx/libros/libros/Gdt_eLAC_meta_13.pdf).
- [16] Clauset, A., M.E.J. Newman, and C. Moore: *Finding community structure in very large networks*. Physical Review E, 70(6), Dec. 2004. <http://dx.doi.org/10.1103/PhysRevE.70.066111>.
- [17] Consultores Asociados De Marketing, Cadem: *Informe Sexta Encuesta de Accesos, Usos y Usuarios de Internet (in Spanish)*. Informe técnico., Subsecretaría de Telecomunicaciones de Chile, SUBTEL, August 2015. [http://www.subtel.gob.cl/wp-content/uploads/2015/04/Informe\\_Sexta\\_Encuesta\\_de\\_Accesos\\_Usos\\_Usuarios\\_de\\_Internet.pdf](http://www.subtel.gob.cl/wp-content/uploads/2015/04/Informe_Sexta_Encuesta_de_Accesos_Usos_Usuarios_de_Internet.pdf).
- [18] Csardi, G. and T. Nepusz: *The igraph software package for complex network research*. InterJournal, Complex Systems:1695, 2006. <http://igraph.org>.
- [19] Deane, J.: *Media, democracy and the public sphere*. In Hemer, O. and T. Tufte (eds.): *Media and glocal change: Rethinking communication for development*. Nordicom, 2005, ISBN 987-1183-26-7.
- [20] DeCarlo, L.T.: *On the meaning and use of kurtosis*. Psychological methods, 2(3):292, 1997. <http://dx.doi.org/10.1037/1082-989X.2.3.292>.
- [21] Dimmick, J., J.C. Feaster, and G.J. Hoplamazian: *News in the interstices: The niches of mobile media in space and time*. New Media & Society, 13(1):23–39, 2011. <http://dx.doi.org/10.1177/1461444810363452>.
- [22] El Mostrador: *La pugna en el CNTV por Farmacias Ahumada*. <http://www.elmostrador.cl/noticias/sin-editar/2011/01/31/la-pugna-el-el-cntv-por-farmacias-ahumada/>, January 2011. Accesed: 2016-09-14.

- [23] Fish, S. and D. Othman: *Narratives of conflict: What the 2014 Gaza War can tell us about discourse on the Internet*. In Gasser, U., J. Zittrain, R. Faris, and R. Heacock Jones (eds.): *Internet Monitor 2014: Reflections on the Digital World: Platforms, Policy, Privacy, and Public Discourse*. Berkman Center Research Publication, 2014. <http://ssrn.com/abstract=2538813>.
- [24] Flanagin, A.J. and M.J. Metzger: *Perceptions of Internet Information Credibility*, 2000, ISBN 1077-6990. ISSN 1077-6990.
- [25] Fortunato, S.: *Community detection in graphs*. CoRR, abs/0906.0612, 2009. <http://arxiv.org/abs/0906.0612>.
- [26] Franklin, J.: *Godfathers of Chilean right charged with tax fraud, bribery and money laundering*. <https://www.theguardian.com/world/2015/mar/04/chile-right-fraud-bribery-money-laundering-pinochet>, March 2015. Accessed: 2016-06-19.
- [27] Freeman, L.C.: *Visualizing social networks*. Journal of social structure, 1(1):4, 2000. <http://www.cmu.edu/joss/content/articles/volume1/Freeman.html>.
- [28] Fundación Poderomedia: *Poderopedia*. <http://www.poderopedia.org/>. Accessed: 2016-04-02.
- [29] Gantz, J.F. and S. Minton: *The Diverse and Exploding Digital Universe An Updated Forecast of Worldwide*. IDC White Paper, 36(10):20–24, 2008, ISSN 01621459.
- [30] Girvan, M. and M.E.J. Newman: *Community structure in social and biological networks*. Proceedings of the National Academy of Science, 99:7821–7826, June 2002. <http://dx.doi.org/10.1073/pnas.122653799>.
- [31] Godoy, S. and M. Gronemeyer: *Mapping Digital Media : Chile*. Techn. rep., Open Society Foundations, 2012. <https://www.opensocietyfoundations.org/sites/default/files/mapping-digital-media-chile-20121122.pdf>.
- [32] Graells-Garrido, E., M. Lalmas, and D. Quercia: *Data portraits: Connecting people of opposing views*. CoRR, abs/1311.4658, 2013. <http://arxiv.org/abs/1311.4658>.
- [33] Guzman, J. and B. Pobleto: *On-line Relevant Anomaly Detection in the Twitter Stream : An Efficient Bursty Keyword Detection Model*. In *Proceeding ODD '13 Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, pp. 31–39, 2013, ISBN 9781450323352.
- [34] Harvard Berkman Center for Internet and Society and the MIT Center for Civic Media: *Media Cloud — tools for online media analysis*. <http://mediacloud.org>. Accessed: 2016-04-07.
- [35] Hubert, L. and P. Arabie: *Comparing partitions*. Journal of Classification, 2(1):193–218, 1985, ISSN 1432-1343. <http://dx.doi.org/10.1007/BF01908075>.
- [36] Jaccard, P.: *The distribution of the flora in the alpine zone.1*. New Phytologist, 11(2):37–50, 1912, ISSN 1469-8137. <http://dx.doi.org/10.1111/j.1469-8137.1912.tb05611.x>.

- [37] Java, A., X. Song, T. Finin, and B. Tseng: *Why We Twitter : Understanding Microblogging*. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65, 2007, ISBN 1595934448.
- [38] Kalyanam, J., M. Quezada, B. Poblete, and G.R.G. Lanckriet: *Early prediction and characterization of high-impact world events using social media*. CoRR, abs/1511.01830, 2015. <http://arxiv.org/abs/1511.01830>.
- [39] Kannan, R., S. Vempala, and A. Vetta: *On clusterings: Good, bad and spectral*. J. ACM, 51(3):497–515, 2004. <http://doi.acm.org/10.1145/990308.990313>.
- [40] Klimkiewicz, B.: *Is the clash of rationalities leading nowhere? Media pluralism in European regulatory policies*. In Czepek, A., M. Hellwig, and E. Nowak (eds.): *Press Freedom and Pluralism in Europe: Concepts and Conditions*, pp. 62—64. Intellect Bristol, 2009, ISBN 9781841502434.
- [41] Krikorian, R.: *Map of a Twitter Status Object*. <http://online.wsj.com/public/resources/documents/TweetMetadata.pdf>. Accessed: 2016-04-02. Originally published at <http://mehack.com/>.
- [42] Lamos, V. and N. Cristianini: *Tracking the flu pandemic by monitoring the social web*. In *2010 2nd International Workshop on Cognitive Information Processing, CIP2010*, pp. 411–416, 2010, ISBN 9781424464593.
- [43] Lefever, K., E. Wauters, and P. Valcke: *Media pluralism in the eu-comparative analysis of measurements systems in europe and us*. [http://www.steunpuntmedia.be/wp-content/uploads/2014/03/Steunpunt-Media\\_ICRI\\_Monitoring\\_D1D2.pdf](http://www.steunpuntmedia.be/wp-content/uploads/2014/03/Steunpunt-Media_ICRI_Monitoring_D1D2.pdf), 2013. Accessed: 2016-09-14.
- [44] Lotan, G.: *Israel, Gaza, war & data - The art of personalizing propaganda*. <https://globalvoices.org/2014/08/04/israel-gaza-war-data-the-art-of-personalizing-propaganda/>, August 2014. Accessed: 2016-04-07.
- [45] Maldonado, J., V. Peña-Araya, and B. 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)*, August 2015. <http://research.microsoft.com/en-US/people/milads/taia15-5.pdf>.
- [46] Manning, C.D., P. Raghavan, and H. Schütze: *Introduction to Information Retrieval*, ch. 6 - Scoring, term weighting and the vector space model. Cambridge University Press, 2008, ISBN 9780521865715.
- [47] Morgan, J.S., C. Lampe, and M.Z. Shafiq: *Is news sharing on Twitter ideologically biased?* In *Computer Supported Cooperative Work, CSCW 2013, San Antonio, TX, USA, February 23-27, 2013*, pp. 887–896, 2013. <http://doi.acm.org/10.1145/2441776.2441877>.
- [48] Myers, J.L. and A.D. Well: *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates, New Jersey, 2003.



- [49] Nikulin, M.: *Excess coefficient*. In Michiel, H. (ed.): *Encyclopaedia of Mathematics*. Kluwer Academic Publishers, 2012, ISBN 9781402006098,1402006098. [http://www.encyclopediaofmath.org/index.php?title=Excess\\_coefficient](http://www.encyclopediaofmath.org/index.php?title=Excess_coefficient).
- [50] Pariser, E.: *The Filter Bubble: What the Internet Is Hiding from You*. The Penguin Group, 2011, ISBN 1594203008, 9781594203008.
- [51] Peralta, A.: *Chile's Penta Case Pulls Dozens into Corruption Scandal*. <https://panamapost.com/adriana-peralta/2015/01/16/chiles-penta-case-pulls-dozens-into-corruption-scandal/>, January 2015. Accessed: 2016-06-19.
- [52] Purcell, K., L. Rainie, A. Mitchell, T. Rosenstiel, and K. Olmstead: *Understanding the participatory news consumer : How internet and cell phone users have turned news into a social experience*. Pew Research Center, 2010. [http://www.pewinternet.org/files/old-media/Files/Reports/2010/PIP\\_Understanding\\_the\\_Participatory\\_News\\_Consumer.pdf](http://www.pewinternet.org/files/old-media/Files/Reports/2010/PIP_Understanding_the_Participatory_News_Consumer.pdf).
- [53] Romano, S., J. Bailey, V. Nguyen, and K. Verspoor: *Standardized mutual information for clustering comparisons: one step further in adjustment for chance*. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1143–1151, 2014.
- [54] Ronen, S., B. Gonçalves, K.Z. Hu, A. Vespignani, S. Pinker, and C.A. Hidalgo: *Links that speak: The global language network and its association with global fame*. *Proceedings of the National Academy of Sciences*, 111(52):E5616–E5622, 2014. <http://www.pnas.org/content/111/52/E5616.abstract>.
- [55] Russell, B.: *Russell (1902). Letter to Frege*. In Heijenoort, J. van (ed.): *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931*. Harvard University Press, 2002, ISBN 9780674324497.
- [56] Sakaki, T., M. Okazaki, and Y. Matsuo: *Earthquake shakes Twitter users: real-time event detection by social sensors*. WWW '10: Proceedings of the 19th international conference on World wide web, p. 851, 2010, ISSN 1605587990. <http://dl.acm.org/citation.cfm?id=1772777>.
- [57] Singhal, A.: *Modern information retrieval: A brief overview*. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001. <http://sites.computer.org/debull/A01DEC-CD.pdf>.
- [58] Tan, P., M. Steinbach, and V. Kumar: *Introduction to Data Mining*. Addison-Wesley, 2005, ISBN 0-321-32136-7.
- [59] Turner, V., J.F. Gantz, D. Reinsel, and S. Minton: *The Digital Universe of Opportunities: Rich Data and Increasing Value of the Internet of Things*. IDC White Paper, April 2014. <http://www.emc.com/leadership/digital-universe/2014iview/index.htm>.
- [60] Twitter, Inc.: *Tweets — Twitter Developers*. <https://dev.twitter.com/overview/api/tweets>. Accessed: 2016-04-02.
- [61] Twitter, Inc.: *The Twitter glossary*. <https://support.twitter.com/articles/166337>. Accessed: 2016-04-02.

- [62] Vieweg, S., A.L. Hughes, K. Starbird, and L. Palen: *Microblogging during two natural hazards events: what twitter may contribute to situational awareness*. CHI 2010: Crisis Informatics April 10–15, 2010, Atlanta, GA, USA, pp. 1079–1088, 2010. <http://dl.acm.org/citation.cfm?id=1753486>.
- [63] Vinh, N.X., J. Epps, and J. Bailey: *Information theoretic measures for clusterings comparison: Is a correction for chance necessary?* In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, pp. 1073–1080, New York, NY, USA, 2009. ACM, ISBN 978-1-60558-516-1. <http://doi.acm.org/10.1145/1553374.1553511>.
- [64] Watts, R.J., A.L. Porter, and D. Zhu: *Factor analysis optimization: Applied on natural language knowledge discovery*. In *In: Committee on Data for Science and Technology 2002: Frontiers of Scientific and Technical Data: Proceedings of the 18 th International Conference CODATA 2002*, 2002. <http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA484817>.
- [65] Westfall, P.H.: *Kurtosis as Peakedness, 1905-2014. RIP*. The American Statistician, 68(3):191–195, August 2014, ISSN 0003-1305.
- [66] Yang, Y., J.G. Carbonell, R.D. Brown, T. Pierce, B.T. Archibald, and X. Liu: *Learning approaches for detecting and tracking news events*. IEEE Intelligent Systems, 14(4):32–43, July 1999, ISSN 1541-1672. <http://dx.doi.org/10.1109/5254.784083>.

# Appendices

## A User distribution for each community structure

The following figures show user consumption for each clustering obtained. User type distribution plots specify the percentage of users that only follow media outlets from the communities we found and those who follow outlets not belonging to any community. Media consumption diversity graphs show a histogram of the number of communities users follow, both for users that only follow outlets in the communities we found and those who don't.

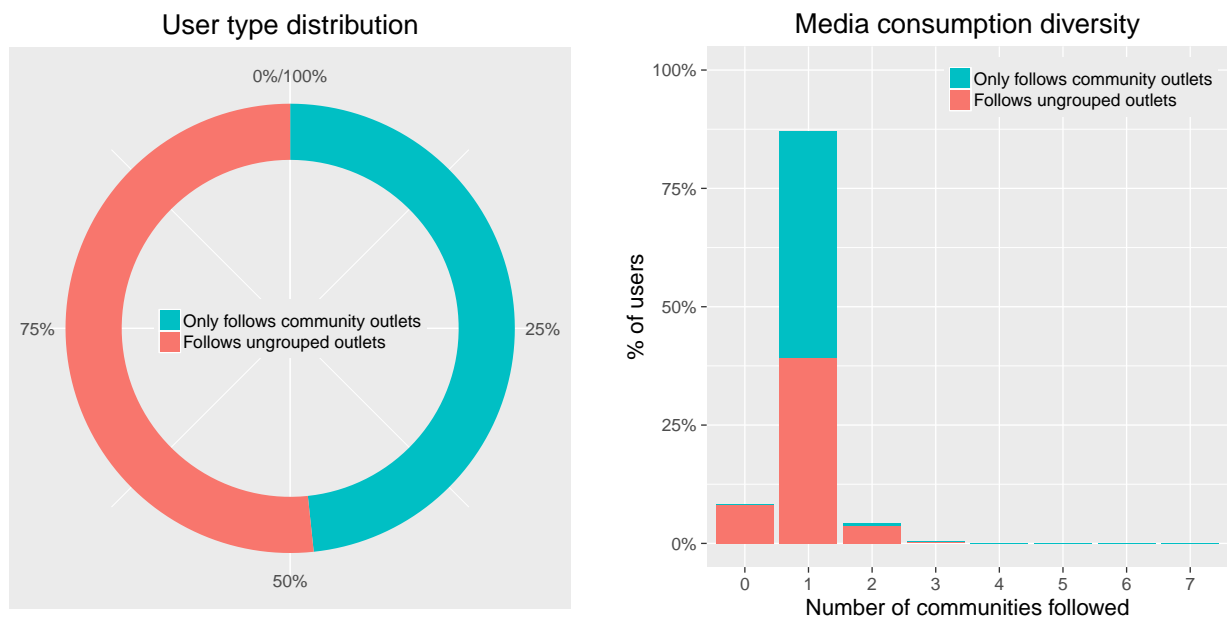


Figure A.5: Vocabulary similarity user distribution.

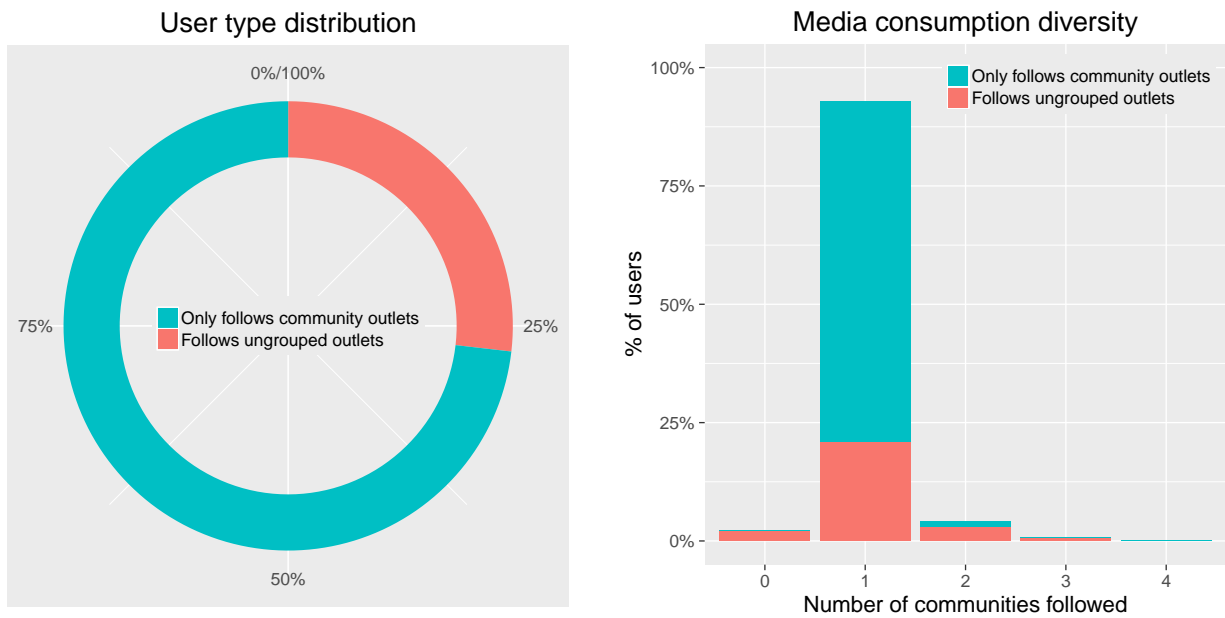


Figure A.6: Topic similarity user distribution.

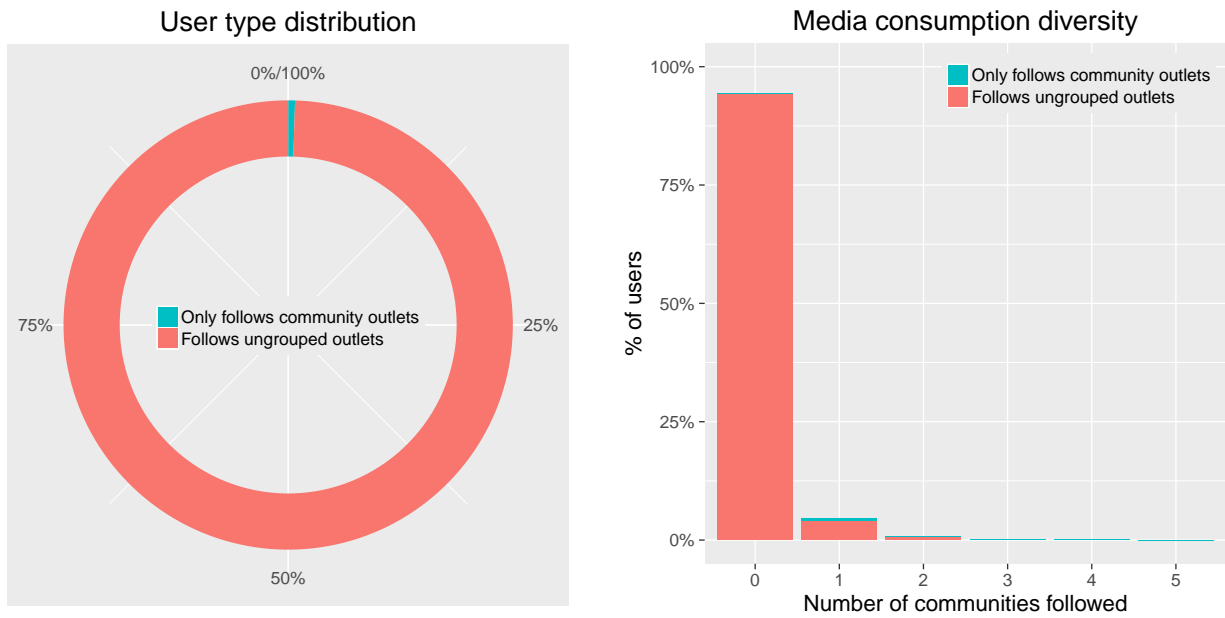


Figure A.7: Temporal Correlation user distribution for 'Penta' term set.

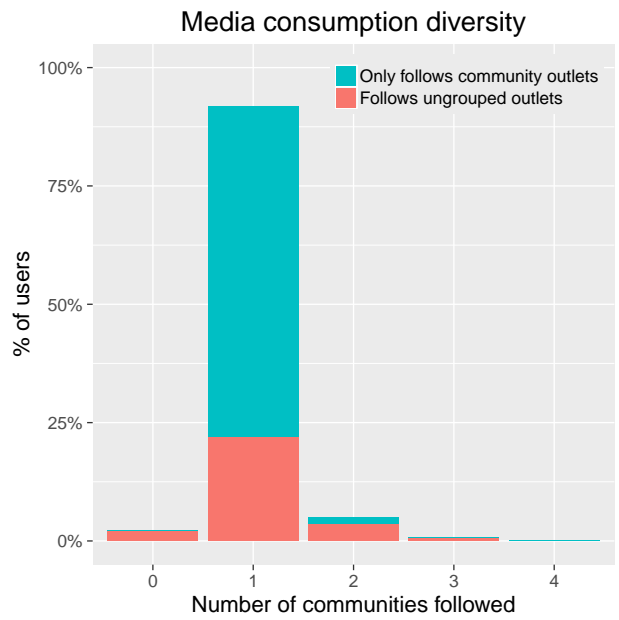
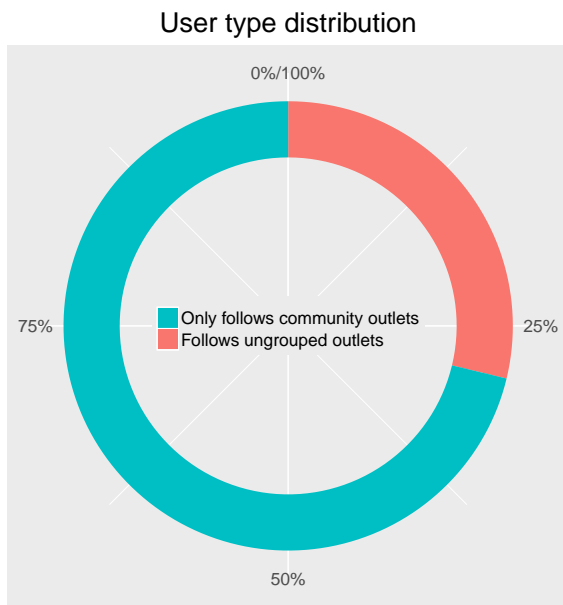


Figure A.8: Temporal Correlation user distribution for 'President' term set.

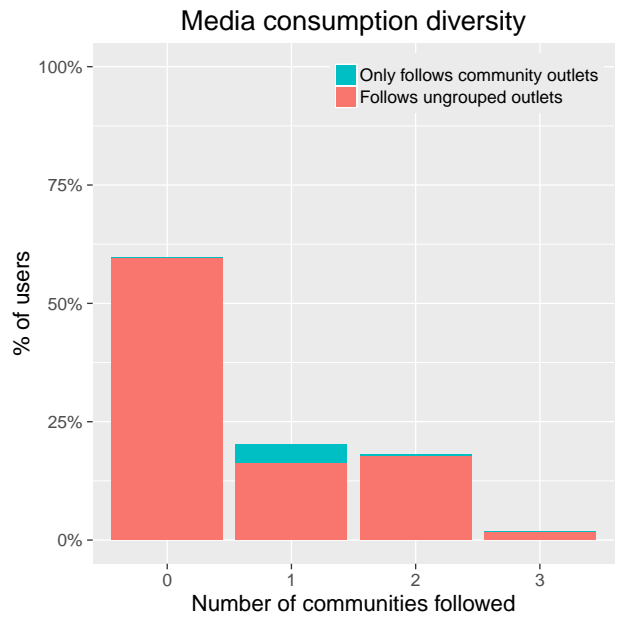
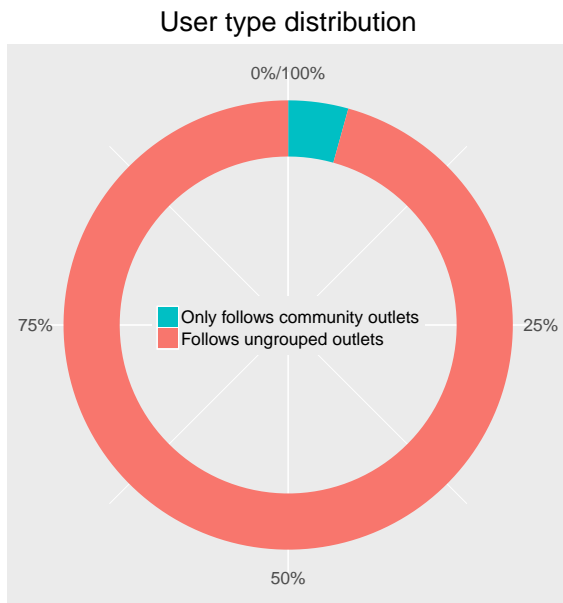


Figure A.9: Ownership similarity user distribution.

## B Similarity graph visualizations for each explored similarity measure

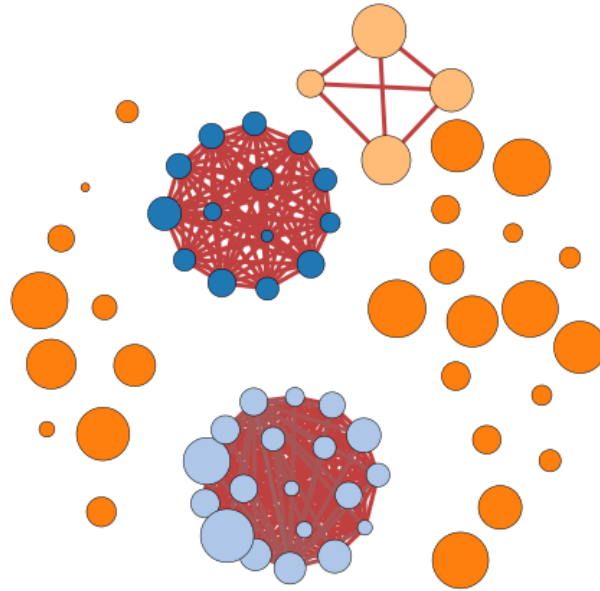


Figure B.10: Ownership-based similarity graph. Three well-defined communities can be seen.

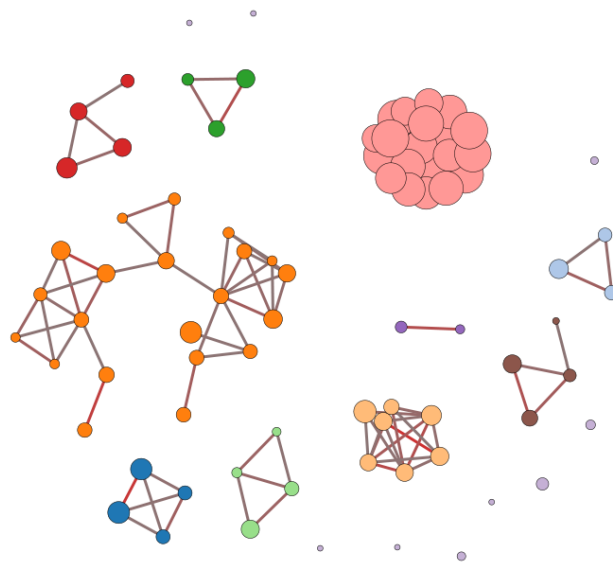


Figure B.11: Follower-based similarity graph. Two big communities can be observed, as well as many smaller ones.

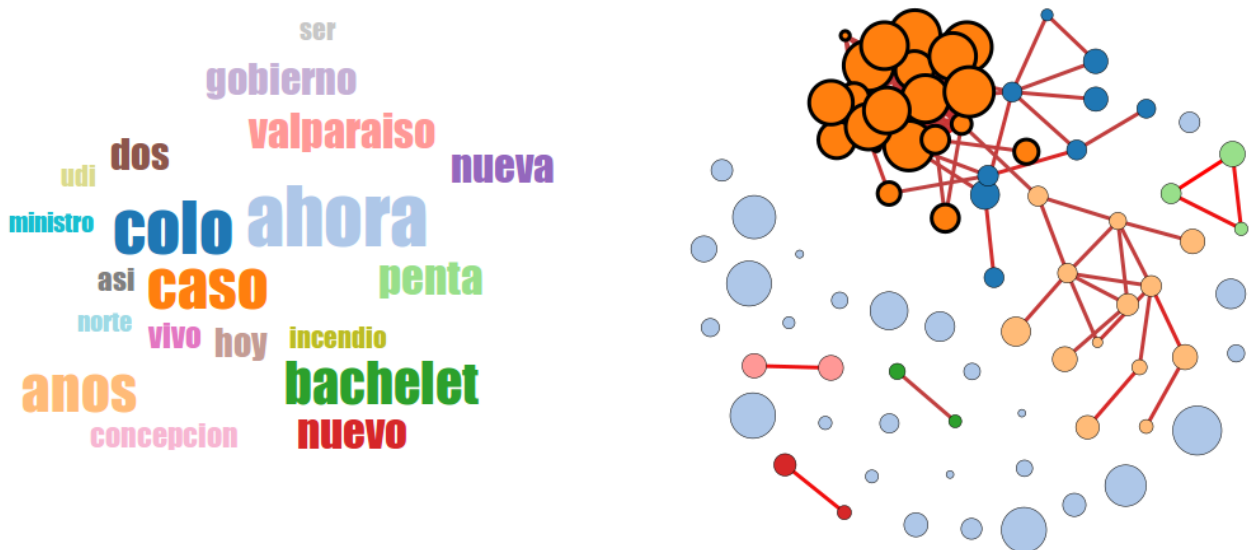


Figure B.12: Vocabulary-based similarity graph. Three big communities can be observed, plus smaller ones. A word cloud is shown to the left, containing the most distinctive words for the biggest community.

davalos, luksic      menos, muertos  
**desordenes, vergara**  
**benegas, ofensivo**  
 cariola, vallejo  
**quinta, vergara**  
 lideres, taza  
**revisar, tendencias**

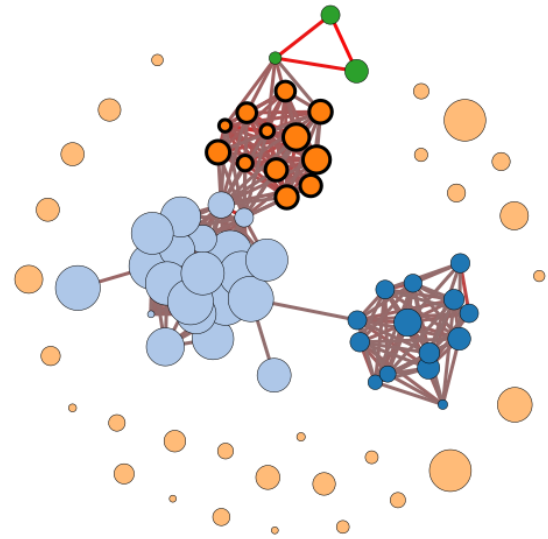


Figure B.13: Topic-based similarity graph. Communities display a higher level of well-connectedness than for vocabulary similarity. A word cloud is shown to the left, containing the most distinctive keyword pairs for the highlighted community.

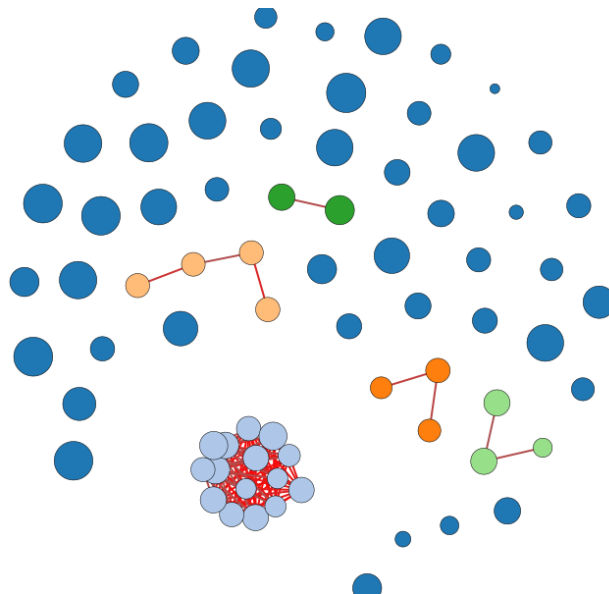


Figure B.14: Temporal correlation-based similarity graph for the Penta term set. Small but well-defined communities can be seen.



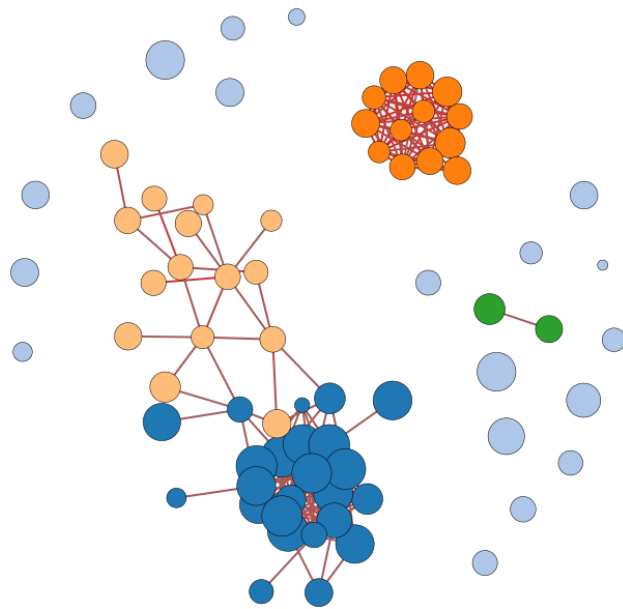


Figure B.15: Temporal correlation-based similarity graph for the President term set. Well-defined communities can be seen.

## C Lists of stop words

### C.1 Stop words included from NLTK

The NLTK Python library has access to a stop word corpus, which includes over 2400 stop words for 11 languages. As the source of these stop words is not completely clear (Martin Porter is referenced as the author in the documentation, but without a proper citation) we considered it appropriate to list them for the sake of reproducibility.

de	como	donde	algunos	algunas
la	más	quien	qué	algo
que	pero	desde	unos	nosotros
el	sus	todo	yo	mi
en	le	nos	otro	mis
y	ya	durante	otras	tú
a	o	todos	otra	te
los	este	uno	él	ti
del	sí	les	tanto	tu
se	porque	ni	esa	tus
las	esta	contra	estos	ellas
por	entre	otros	mucho	nosotras
un	cuando	ese	quienes	vosostros
para	muy	eso	nada	vosostros
con	sin	ante	muchos	os
no	sobre	ellos	cual	mío
una	también	e	poco	mía
su	me	esto	ella	míos
al	hasta	mí	estar	mías
lo	hay	antes	estas	tuyo

tuya	estéis	estuviéramos	habrás	hubiese
tuyos	estén	estuvierais	habrá	hubieses
tuyas	estaré	estuvieran	habremos	hubiésemos
suyo	estarás	estuviese	habréis	hubieseis
suya	estará	estudieses	habrán	hubiesen
suyos	estaremos	estuviésemos	habría	habiendo
suyas	estaréis	estudieseis	habrías	habido
nuestro	estarán	estudiesen	habríamos	habida
nuestra	estaría	estando	habrías	habidos
nuestros	estarías	estado	habrían	habidas
nuestras	estaríamos	estada	había	soy
vuestro	estaríais	estados	habías	eres
vuestra	estarían	estadas	habíamos	es
vuestros	estaba	estad	habíais	somos
vuestras	estabas	he	habían	sois
esos	estábamos	has	hube	son
esas	estabais	ha	hubiste	sea
estoy	estaban	hemos	hubo	seas
estás	estuve	habéis	hubimos	seamos
está	estuviste	han	hubisteis	seáis
estamos	estuvo	haya	hubieron	sean
estáis	estuvimos	hayas	hubiera	seré
están	estuvisteis	hayamos	hubieras	serás
esté	estuvieron	hayáis	hubiéramos	será
estés	estuviera	hayan	hubierais	seremos
estemos	estuvieras	habré	hubieran	seréis

serán	fuera	tengo	tendría	tuvieras
sería	fueras	tienes	tendrías	tuviéramos
serías	fuéramos	tiene	tendríamos	tuvierais
seríamos	fuerais	tenemos	tendríais	tuvieran
seríais	fuera	tenéis	tendrían	tuviese
serían	fuese	tienen	tenía	tuvieses
era	fueses	tenga	tenías	tuviésemos
eras	fuésemos	tengas	teníamos	tuvieseis
éramos	fueseis	tengamos	teníais	tuviesen
erais	fuesen	tengáis	tenían	teniendo
eran	sintiendo	tengan	tuve	tenido
fui	sentido	tendré	tuviste	tenida
fuiste	sentida	tendrás	tuvo	tenidos
fue	sentidos	tendrá	tuvimos	tenidas
fuimos	sentidas	tendremos	tuvisteis	tened
fuisteis	siente	tendréis	tuvieron	
fueron	sentid	tendrán	tuviera	

## C.2 Manually included stop words

These stop words were determined by checking the top-scoring words for vocabulary analysis and selecting words that corresponded to generic terms.

via	fotos	chile	ver	informate
dia	foto	2015	tras	registrate
uso	video	2014	galeria	