Quantifying the ecological diversity and health of online news

Erick Elejalde[a,*], Leo Ferres[b,c,*], Eelco Herder[d], Johan Bollen[e]

[a] *Department of Computer Science, University of Concepcion, Concepción, Chile*
[b] *Data Science Institute, Faculty of Engineering, Universidad del Desarrollo, Chile*
[c] *Telefónica R&D, Santiago, Chile*
[d] *Institute for Computing and Information Sciences, Radboud Universiteit, Nijmegen, Netherlands*
[e] *School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*

Abstract

Even in developed countries with an active free press, news coverage can be dominated by only a few players. This can lead to a reduction of topical and community diversity. Ownership structures might further limit coverage by implicitly or explicitly biasing editorial policies. In this paper, we apply ecological diversity measures to quantify the health of the Chilean online news ecology using extensive ownership and social media data. Results indicate that high levels concentration characterizes the Chilean media landscape in terms of ownership and topical coverage. Our methods reveal which groups of outlets and ownership exert the greatest influence on news coverage and can be generalized to any nation's news system.

Keywords: news ecosystem, diversity indices, media ownership, topic selection
2010 MSC: 68-11, 68-35

*Corresponding author
Email addresses: eelejalde@udec.cl (Erick Elejalde), lferres@udd.cl (Leo Ferres)

1. Introduction

News is increasingly aggregated by and consumed through social media[1] such as Reddit, Facebook and Twitter[2]. Due to its reliance on social networking relations for news propagation, social media may be subject to a variety of social issues that may restrict news coverage and topical diversity, e.g. as a result of information bubbles [1] and social conformity bias [2].

On the other hand, some might argue that the exponential growth of new communication technologies can solve, or at least alleviate, many diversity problems. More accessible and cheaper channels of communication should provide new content producers with better opportunities and less friction to compete in a larger media market. However, this assumption has not been tested empirically. Indeed, early indicators point to high levels of bias as well as a lack of diversity in terms of topics covered and communities addressed [3, 4, 5].

In the past, news ecosystems have primarily been modeled from a political and economical perspective. Theoretical models focusing on the political economy of the mass media show that, for a truly democratic society, the more information we can have as voters the better. For example, in [6], Besley and Prat proposed the *Media Capture Model* which predicts that a *low* number of independent outlets will make the news media industry *more* susceptible to be fully captured by the political and economic elite. In other words, when there is enough pluralism, the media behave more independently. Separately, the Propaganda Model [7] tried to explain the behaviour of the media by focusing on how a number of external factors filter what is finally published by the media. Each linguistic account of an event must pass through a number of filters that define what is newsworthy thereby shaping discourse and interpretation.

Both the *Media Capture Model* and *The Propaganda Model* [7] warn against the negative consequences of the concentration of ownership in the mass media. Having a large share of the media industry in the hands of just a few

[1] `http://www.journalism.org/2016/07/07/pathways-to-news/`

[2] `https://goo.gl/wofqQ7`

mega-conglomerates poses the risk of the system not necessarily representing the interest of the common good, the media's original primary purpose. For an in-depth survey of the political economy of the mass media, see [8].

A different metric of pluralism is discussed in [9]. The author defines two classes of pluralism: External Pluralism (EP) and Internal Pluralism (IP). EP requires that all political opinions have room and are represented in at least some of the suppliers of content in the media market. On the other hand, IP is achieved when every media company covers all sides of the main political issues in a society. EP benefits from a larger number of media outlets if the users are really free to choose. If the public favors (or are limited to) a small set of news media, then it is important to analyze market concentration; i.e., the number of companies and the percentage of the total news production that each of them represents.

Here we postulate that the news industry can be modeled as a complex system [10], an ecosystem that consists of many different interacting components, such as news outlets, their owners, reporters, news consumers, advertisers, all subject to and responding to a variety of social factors. Through their interactions among themselves and with external drivers, these components collectively shape our news ecosystems. Given these broad similarities, we hypothesize that we can apply techniques developed to study the health and diversity of biological ecosystems to online news (eco)systems. Other works that exploit parallels between information systems and ecology have proven to be fruitful [11].

This paper uses a set of ecological indicators [12] to analyze the health of the "news ecosystem" as viewed from Twitter, an online social network specifically designed for the social propagation of information and increasingly used as a platform for the dissemination of news, fake or not.

Our analysis relies on information about Chilean news outlets since they have established a significant social media presence with a high number of Chilean users [13]. The Chilean media landscape is furthermore well documented due to the availability of detailed, publicly available data with respect to its own-

ership structure, compiled by *Poderopedia*[3], a journalist NGO that aims to understand power relationships between people, companies, and organizations. Finally, Chile is also on a par with several other important news systems of the world such as the US and Canada, part of southern Europe, Australia and South Africa, as evidenced by the Reporters Without Borders' 2017 World Press Freedom Index[4]. The case of Chilean news thus provides an interesting addition to previous media studies that are mostly focused mostly on Northern hemissphere countries, with a strong inclination to the United States and Western Europe.

Our work shows that, by more than one metric, the online Chilean news ecosystem can be considered to be in a "poor" state in terms of heterogeneity, diversity and access to varied information.

2. Background and preliminaries

Our objective is to measure the diversity of a news ecosystem, taking into account the variety of different news sources, the producers of news content, as well as the news consumers. We start with considering each individual news tweet as an *entity* and its corresponding news outlet as its *type*. We then apply well-known ecology indices to quantitatively measure how "healthy" – diverse – our system is. We assume that diversity of content is a desired property of any news system, see [14].

Similarly, it has been proposed that ownership can influence editorial policies and bias content [15, 16]. Thus, we will also relate the relationship and type of each entity with *the owner* of the publisher outlet, rather than with the news outlet itself. This is, potentially, a stronger effect, since several newspapers may publish similar content because they belong to a single ownership group. We analyze media ownership [17] using two metrics: *numerical diversity* and *source diversity*. Numerical diversity refers to the number of outlets available to the public in a given area; source diversity indicates the number of owners

[3] `http://www.poderopedia.org/`

[4] `https://rsf.org/en/ranking/2017`

that actually control those outlets. The rationale for using these indicators is our expectation that having a news industry increasingly dominated by fewer and fewer companies, increases the owners' potential influence on the published content, leading to a greater probability of reallocation of attention to their interests.

Several studies indicate that online news distribution and consumption can be subject to considerable bias, for example through the so-called Filter Bubble effect [1] and the prevalent tendency towards homophilic connections [2, 18] in online social networks. This may be counteracted by the fact that social media users are generally exposed to a wider number of news sources [19, 20]. Our work attempts to assess, on balance, to which degree even readers who are subscribed to a high number of the available news outlets (or are exposed to their news indirectly [21]) can still be affected by significant bias due to the lack of diversity in online news ecosystems.

It is important to note that news consumers only have access to the corpus of published news (and the associated commentaries, blogs, tweets, Facebook posts, /ldots). Thus, they only get to see the *final product* of the system of news production which has already gone through the alleged filtering process outlined in The Propaganda Model [7]. In this case, the news ecosystem is observed through the final news items that consumers have access to.

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

Ecological science has developed extensive models of the diversity of ecosystems that may generalize and apply to online news ecologies. Ecologists have used four attributes to characterize the evolution of complex systems [22]: (1) progressive integration, (2) progressive differentiation, (3) progressive mechanization, and (4) progressive centralization. The progressive integration is represented in the news ecosystem by the current dynamics in news production where small news outlets report (or redistribute) news created by bigger news agencies. Progressive differentiation is shown in the plethora of news outlets

and magazines that create their own niche attempting to take advantage of their condition (either geographic, topic wise or by exploiting certain political position [23, 21, 5]). Progressive mechanization refers to the growing number of feedback and regulation mechanisms, that social media platforms seem to be particularly susceptible to. Finally, progressive centralization can be seem in how news has been modified and adapted to the other components in the ecosystem (*e.g.* native ads[5]).

In [24], the authors provide a conceptual definition of ecosystem health. It is largely focused on three components: (1) Vigor or scope for growth, (2) Organization (given by the diversity or complexity of the system), and (3) Resilience (in function of the system capacity to counteract stressful conditions). These components are integrated in a Health Index (HI) that can be formulated as:

$$HI = V \times O \times R \tag{1}$$

where V represents the Vigor of the system, O represents the Organization Index, and R represents the Resilience index. As HI is directly proportional to all three factors, lowering any of them will result in a lower global health index. In this work we focus in the Organization component by analyzing a variety of quantitative diversity indices, namely the Shannon Diversity [25] and Simpson Diversity [26] indices. Meanwhile, the Average Taxonomic Distinctness [27] provides a notion of the similarity we can expect from the coverage of a story, even in cases where it originates from different outlets. We restrict our analysis to the "news ecosystem" form in the context of the Chilean on-line media, and specifically on Twitter.

2.1. Topic detection

We first use text-content based clustering of the publications of news outlets publications to identify "stories" that relate to a common event or topic [23, 28].

Minwise Hashing was originally proposed by Broder [29] for finding similar documents in the AltaVista search engine. Later, Broder et al. [30] show that

[5] `https://en.wikipedia.org/wiki/Native_advertising`

to compute the similarity of two documents it suffices to keep a small number of signatures (summaries or sketches) for the sets representing each document. Finally, Broder et al. [30] presented an algorithm technique called *min-wise hashing*. Minwise hashing approaches have been successfully applied to a wide range of applications including compressing Web graphs and social networks [31, 32, 33], tracking Web spam [34], genome assembly [35].

More relevant to our investigation, minwise hashing based on n-grams has been used to obtain clusters of similar documents in the Twitter context [36]. This technique has also been compared against the cosine similarity measure [37], which is commonly found in literature to approach this or similar tasks [28, 23]. The study in [37] shows that minhash outperforms cosine similarity in most practical cases.

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

For each text we extract groups of n consecutive words (n-shingles or n-grams) [38]. Hence, each tweet is represented by the set of n-shingles that correspond to its text. Two documents are said to be similar if they have several shingles in common. To group similar tweets we search for those who share a subset of n-shingles. This way, we are not only looking for tweets with a similar set of words, but similar phrases.

The shingles (SH) model can be applied at character and word levels, but it has been shown that using long n-shingles based on characters to simulate words leads to an unacceptably high number of false positives. In contrast, using n-shingles (also called w-shingles) based on words has been used successfully in small and large documents. For instance, using a n of 2 or 3 in email documents (short documents) and $n = 4$ in large documents such as web collections [29, 39, 40] news articles or blog posts [41] provided the best results.

Since tweet are short by definition (maximum length allowed is 140 characters), we set $n = 2$. If, on top of the short length of the text, we also remove stop-words, we find that, even with a small value of n, the probability of occurrence of each n-shingle is small. Since the number of distinct n-shingles can

be very large, it is possible to apply a hash function to every n-shingle. Other strategies can be employed to build summaries or sketches to reduce the space without deteriorating the effectiveness of our matching algorithms [29, 39].

In this work, we represent tweets using 2-shingles and then apply 4 keys minwise hashing over each tweet. We define clusters based on similar sets of minwise hashings. As mentioned before, we consider these clusters our set of discovered topics.

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

2.2. Diversity Index

A diversity index is a quantitative measure that reflects how many different *types* there are in a data set and/or takes into account how evenly the basic *entities* are distributed among those *types*. There are three basic groups of ecological diversity indices: enrichment of species, abundance of species, and proportional abundance of species [12]. The first group, enrichment of species, only measures the number of species. Indicators of the abundance of species, besides the number of species, also try to model the distribution of their abundance. The last group of indices, proportional abundance of species, represent enrichment and uniformity in the same expression. Within this last group we can find the Shannon Diversity Index and the Simpson Index. In turn, the Average Taxonomic Distinctness index approaches the problem by taking into account different dimensions of biodiversity (e.g. taxonomic, numerical and phylogenetic). Including these aspects of the diversity helps counteract some of the problems described for previous diversity indices such as measuring functional diversity [42]. We use these indices to asses diversity of the on-line news distribution.

2.2.1. Shannon Diversity Index (ShDI)

Shannon Diversity Index is widely used in ecology and biology to measure the diversity of species in a community [25], but has also been extrapolated to other fields. In [43] the authors used the ShDI to measure subjectivity in the selection of dates for timeline creation in news stories. For example, dates that were considered significant in the timeline of one newspaper, as opposed to those dates that were relevant for "all" news outlets. The authors use this index to highlight dates on which important events happened, but that are likely to be ignored by many news agencies, hence, indicating how subjective (or non-diversified) a date is.

Here we apply ShDI to topics; we use the ShDI to express the rarity or commonness of topics, as reported by different news outlets. The idea is to quantify, for each detected topic, how common it is across the media, and whether it was covered disproportionately by just a few newspapers. This will give us an indicator to assess whether the topic was generally considered important news and covered accordingly across a wide variety of outlets or pushed as a topic by specific outlets.

The expression to calculate the ShDI value for cluster c_i is given in Equation 2. Using newspapers as types and tweets as entities, Eq. 2 quantifies the uncertainty in predicting that a newspaper will publish a tweet taken randomly from the dataset (the dataset in this case are the tweets of the corresponding cluster).

$$ShDI(c_i) = -\sum_{i=t}^{R} p_t \ln p_t \tag{2}$$

In Eq. 2, R is the number of *types* participating in the cluster and p_t is the proportion of *entities* from type t. A low index value indicates an unhealthy (or polluted) ecosystem. This index usually takes values in the range [0..5] (nats/individual), interpreted as follows: (1) High status: > 4 (2) Good status: $4-3$ (3) Moderate status: $3-2$ (4) Poor status: $2-1$ (5) Bad status: $1-0$ [44].

In a previous study we showed that ownership does influence, to some extent, the editorial policies of a given media outlet [15]. Thus, it is important to test

whether the subjectivity of a topic varies from newspapers to owners: owners may either want to maximize readership (making the ecosystem diverse), or focus on a single message to target a specific audience ("biasing" the ecosystem). In order to find out the effect of ownership given the ShDI, we included "owners" as *types*, while retaining each tweet as an entity associated with the owner of the newspaper that published it.

2.2.2. Pielou Evenness Index (PEI)

In Equation 2, the value of the ShDI increases both when evenness increases and when the number of *types* increases. So, to be able to compare results for different topics, we also calculated the maximum achievable ShDI ($ShDI_{MAX}$) of each cluster to normalize the results. This normalization is also known as the *Pielou Evenness Index* [45]. The PEI expression for cluster c_i is of the form:

$$PEI(c_i) = \frac{ShDI(c_i)}{ShDI_{MAX}(c_i)} = \frac{ShDI(c_i)}{\ln R} \tag{3}$$

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

2.2.3. Simpson Index (SiDI)

The Simpson Index [26] is another index widely used in ecology. While the ShDI is based on information theory and measures the abundance of species and diversity of individuals, the Simpson index is considered a *dominance index* that assigns a higher weight to the most common species. This means that the presence of a few individuals of some rare species will not have an important effect in the result.

The original index gives a value λ ($0 \leq \lambda \leq 1$) that is higher for environments with low diversity, which is counter-intuitive for a diversity index. To solve this, most authors use the *Gini_Simpson Index* (see Equation 4), which is a variation of the Simpson Index

$$SiDI(c_i) = 1 - \lambda = 1 - \sum_{t=1}^{R} \frac{n_t(n_t - 1)}{N(N-1)} \tag{4}$$

In Equation 4, again, R represents the number of different *types* in the cluster. We use N to represent the total number of tweets in c_i and n_t is the number of tweets in c_i published by *type* t.

This index is used in [46] as the *Participation coefficient*. With this value they measure how well-distributed are the connections of a node among the communities of the graph. Defining a range of the obtained measures helps the authors classify the different roles that a node may have in a complex system network. In [47, 48], the authors also use the Simpson Index to differentiate nodes in a social network based on the interactions of people that use different languages.

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

2.2.4. Average Taxonomic Distinctness (ATxDI)

Similar to the ones above, this index takes into account the species abundance, but also includes the taxonomic distance between any two types [27]. Specifically, this index represents the expected path length through the classification tree between two entities chosen at random. For us, then, the Average Taxonomic Distinctness is the average editorial "distance" (using a similarity matrix, see below) between two news sources randomly selected from two different types in the same topic. As before, types are either news outlets or owners.

For the taxonomic distance, we use a numerical taxonomy [49]. This form of classification is basically determined by observable characters of taxa (i.e., phenetic similarities). Since we already know the different classes (i.e. our *types*), this should give us an idea of the affinity of any two types. Similarity between two news outlets is then defined by the co-occurrence of two *types* with respect to a same topic. Note that the topics are extracted also from observations of homologies in words (n-grams) of our entities, so the similarity could be further rooted in these lower level aspects. In particular, the value of the similarity between outlet A and B is defined as the conditional probability

$Pr(A|B)$ of the occurrence of A in a cluster given that B occurs in that same cluster. This similarity measure is directional, expressing how likely it is that a story tweeted by B is also tweeted by A [21]. In order to define a symmetric similarity measure, we further specify the similarity between A and B as $sim(A, B) = max(Pr(A|B), Pr(B|A))$.

Finally, we use the agglomerative hierarchical clustering algorithm[6] to create a tree (using the arithmetic mean for the linking method - also known as the UPGMA algorithm). For the clustering algorithm we first transform the similarity matrix into a distance matrix (i.e. $dist(A, B) = 1 - sim(A, B)$). The more similar two types are the closer they will be. With the tree obtained from the clustering we can calculate the length of the path between any two types.

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

$$ATxDI(c_i) = \frac{\sum_{j=1}^{R} \sum_{k=1}^{j-1} \omega_{jk} n_j n_k}{\sum_{j=1}^{R} \sum_{k=1}^{j-1} n_j n_k} \tag{5}$$

where R still represents the number of different *types* in the cluster and n_j is the number of tweets in c_i published by *type* j. The factor ω_{jk} represents the length of the path connecting types j and k in the tree. The double summation accounts for all pairs of types. Equation 5 comes as the result of dividing the *average taxonomic diversity* [27] by the Simpson Index. Doing so eliminates the dominating effect of the species abundance distribution.

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

[6] We use the version implemented in the *scipy.cluster.hierarchy.linkage* library

3. Data

Twitter is now widely used by the overwhelming majority of news outlets, allowing the automated collection of their news streams through its open API (Application Programming Interface), as well as data pertaining to the individual Twitter users that subscribe to these feeds, how news items travel from one user to the next, as well as the social networking connections between both the news outlets and their consumers. In addition, the Twitter API can be queried for user profiles, followers, and tweeting history. In fact, Chile ranks among the top-10 countries on the average number of Twitter users per 1000 individuals [13]. This makes it possible to explore the behavior and interactions of personal and institutional accounts, developing and testing social theories at a previously unseen scale.

We treat every tweet as an independent document from which we can extract a statement/headline. Headlines of online news articles have shown to be a reliable source for adequately providing a high-level overview of news events [50, 51, 52].

To create our database of outlets, we used different sources, with Poderopedia's "influence" database as our baseline[7], manually adding other news outlets in Chile. We built a database of 365 Chilean news outlets with an active presence on Twitter. Then we proceeded to get the tweets generated from October 25, 2015 to January 25, 2016, for all the 365 news outlets twitter accounts. This dataset contains 756,864 tweets. Both the tweets' text content and their metadata was obtained. The text of each tweet was lower-cased and preprocessed by removing stop-words, URLs and punctuation marks.

Since we are working with topics, we filter out the tweets from 'specialized' news outlets, and kept only 235 outlets registered as "general-interest", those covering most subjects. Specialized outlets or magazines (such as fashion or sports) are expected to give a differentiated coverage to special subjects, which

[7] `http://apps.poderopedia.org/mapademedios/index/`

could influence our results. Thus, we focused only on those topics/events that were considered of interest to the general public. For these general-interest news outlets we collected 563,262 tweets during the observed period.

As for ownership information, we also relied on Poderopedia's influence database. Poderopedia aims to identify and understand relationships between people, companies and organizations. Part of this effort involves collecting data on ownership relationships among media companies. We complemented Poderopedia's database with manually added information, based on our own research. As far as we know, this is the most complete database of newspaper ownership information in Chile, which we make available at `https://github.com/eelejalde/news_ecosystem/`.

4. Results

4.1. Topics

For the 235 general-interest news outlets, we were able to identify 79,753 clusters using the min-hash techniques described above (See Section 2.1). These clusters account for 366,180 tweets (65% of the total). Notice here that we are only counting tweets that are contained in one of the clusters, and only those clusters that contain at least two tweets .

There were 56,496 clusters with just one news outlet in them (*single-outlet-clusters*), grouping 172,276 tweets. We found that, against Twitter Rules[8], many outlets tweet multiple times with the same text or a very small variation of it (this is considered *spam* by Twitter). After collapsing tweets with the exact same text into a single one, these clusters were left with 64,920 different tweets (only 37.7% of the tweets in *single-outlet-clusters* were original content). As many as 49,369 *single-outlet-clusters* were formed by one repeated tweet. Even if we do not use this information in our analysis, it is already a strong indication of the poor condition of diversity in our news ecosystem: 87% of single-outlet-

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

clusters contain a single text repeated in multiple tweets. Already, this can be considered as a very low measure of Internal Pluralism, as discussed above.

Our analysis does not take into account who reported the issue first. Single-outlet-clusters represent stories that were ignored by the rest of the media. We work under the assumption that if a story is newsworthy, it will be retweeted/reported by other outlets, regardless of who had the exclusive and which outlet was the first one to break the news. Moreover, had we included single-outlet-clusters, that would have only strengthened our point since these are topics that were considered newsworthy by this one single outlet (lowering the average diversity per topic).

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

To check how effective our method of clustering was, we calculated the Jaccard Index ($JI(x, y)$) for each tweet x against every other tweet y on its cluster, assigning the mean of the JI to that tweet x. The JI_c of the cluster is the mean of the JI of the tweets it contains (see Equation 6).

$$JI_c(c_i) = \frac{\sum_{x \in c_i} \sum_{y \in \{c_i - x\}} JI(x, y)}{N(N-1)} \tag{6}$$

In the case of *multi-outlet-cluster*, for $JI_c \geq 0.8$, we had 22,025 clusters (94.7% of total multi-outlet-clusters). Even for clusters with $JI_c < 0.8$, the content of the tweets within the cluster is still very similar for most cases. The smaller value in the JI_c is mainly because of the shortness of the messages: as a result, changing just a few words would lower the value of the JI. For example, two tweets with the text[9] *"bolsa santiago parte incremento"* and *"bolsa santiago parte ganancias"* have a $JI = 0.6$.

To check for inter-cluster similarity, we ran a second pass of our clustering

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

procedure. This time using as input a bag of words for each of our initial clusters: less than 3.0% of the topics clustered in these "second-level-clusters", showing a very low inter-cluster similarity.

We performed an additional analysis to clarify this point. If our topic clustering produces correct results, each cluster should be about a distinct topic. To test this assumption, we calculated all pairwise topic similarities between clusters on the basis of the named entities that their tweets contain, e.g. "Santiago", "Traffic", etc.. We extract the named entities from the tweets in each cluster using the Stanford's Named Entity (NE) recognizer system [53]. We then create a TF.IDF vector for each cluster based on the frequencies of its NEs, representing the distribution of topics that the cluster is about. If our clustering produces good results, the cosine distance between the TF.IDF vectors of most pairs of clusters should be high, i.e. the clusters are indeed about distinct topics. Fig. 4.1 shows that the distribution of cosine distances between our clusters is indeed very much skewed to the right (high distance values). In other words, very few pairs of clusters have low distance (high similarity) whereas the overwhelming majority (note y-axis log scale) have very low similarities (high distance). This confirms, that our clusters of tweets discovered based on their minwise hashing similarity represent different news topics.

Note that all Chilean sources are only in Spanish, so language itself is not a factor in clustering. Still, in [15] we have found evidence of a low correlation between vocabulary and ownership. However, in [54] we have found that outlets write more frequently about entities of their own political leaning and mostly in their favor. In summary, we find little evidence of a relationship between vocabulary and ownership, but there is some evidence of a relation between terminology and political leaning. In the latter case, the political leaning of the outlet may or may not be influenced by ownership structure [5] (we do not make any assumptions in this respect). Regardless of the factors that determine the topic selection process, if two or more news outlets agree on the importance and newsworthiness of an issue, and report it accordingly, we assume the system will gain in diversity by receiving different points of view.

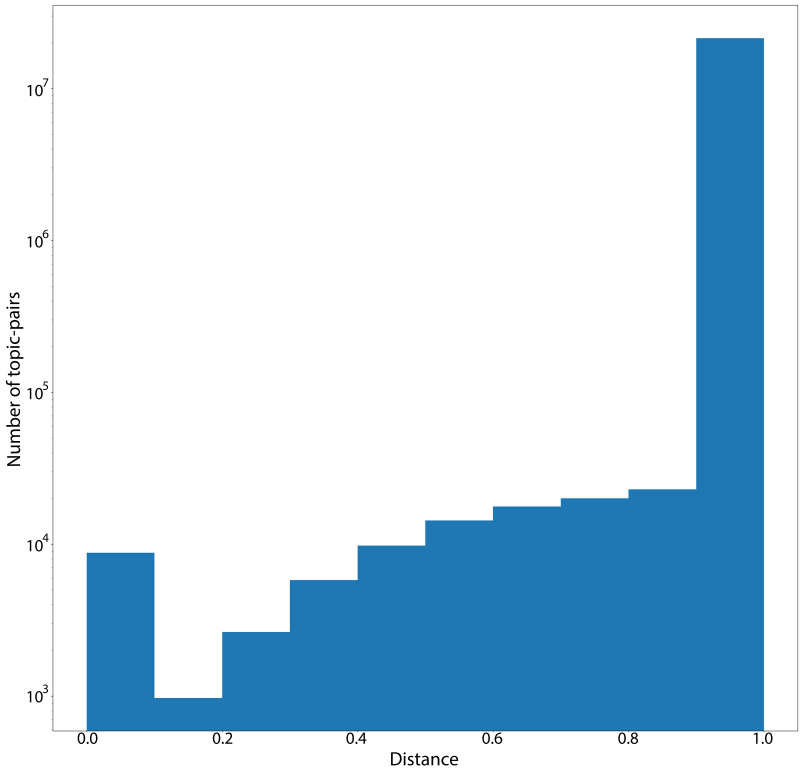

Figure 1: Pairwise clusters TF.IDF. cosine distance distribution.

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

4.2. Diversity

4.2.1. ShDI and PEI

We calculated the maximum achievable ShDI taking into account all the newspapers in our data set (as opposed to only those with at least one tweet in the cluster, $ShDI_{MAX}$). In other words, we find the maximum achievable ShDI if all 235 "general-interest" outlets publish on the same topic approximately the same number of tweets. We will refer to this value as $ShDI'_{OPT}$. In this case,

we get a $ShDI'_{OPT}$ of 5.4337 for news outlets and 4.4426 for the owners. These values are in the range of a good/high status of diversity, which means that the Chilean media have the potential to be a healthy system.

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

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

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

Even when the indices are low for both types (i.e. outlets and owners), we can see that diversity between news outlets is at least one order of magnitude larger than between owners.

4.2.2. SiDI

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

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

covered by most of the outlets that participate in it.

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

A very telling sign of artificial diversity created by the owners that control the market can be seen by analyzing the percentage of topics each *type* (outlets or owners) participates in. Figure 4.2.2 shows this statistic for the 30 outlets with the largest participation. The graph shows that these 30 outlets have a fairly balanced presence on the topics discussed. However, out of these top 30, 25 belong to the same owner (*El Mercurio S.A.P*), and only one news outlet from a different owner participates in more than 5% of the *multi-outlet-cluster* topics. Figure 4.2.2 shows the severity of the dominance of this one company in the selection of subjects (i.e. outlets owned by *El Mercurio* participate in more than 66% of the topics shared by more than one media Twitter account, where the closest competitor is under 10%).

4.2.3. ATxDI

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

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

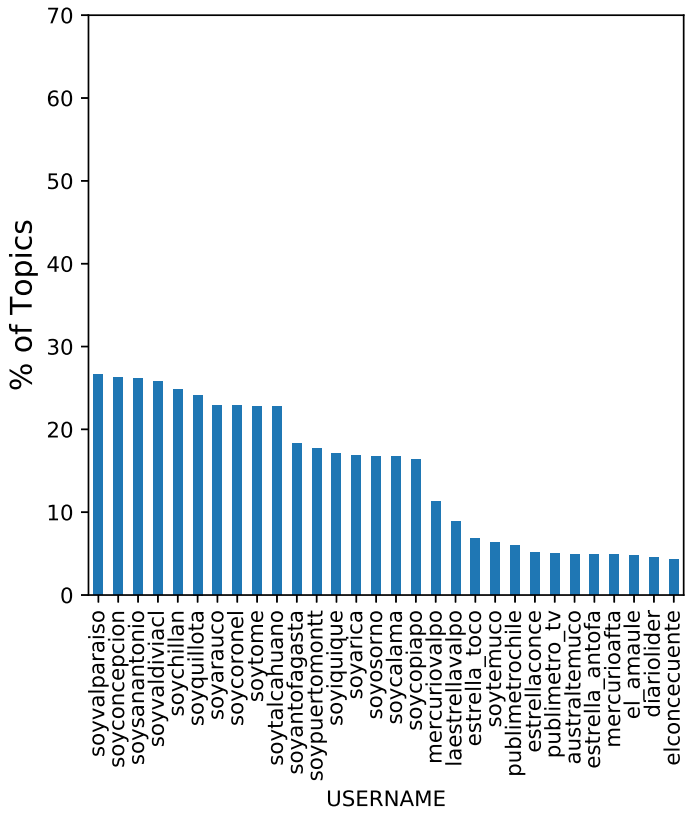

Figure 2: Percentage of topics where each type participates. Top 30 ranking (outlet as types).

into account that the average number of news outlets owned by one company is 2.64, we can see that we have low distinctness in the news ecosystem.

In the case of owners as types, we observe a lower average distinctness for the list of topics ($ATxDI = 0.71$). When comparing this result against its equivalent for a random assignment of outlets to each tweet we obtain a mean of $ATxDI = 13.04$. Even if we limit our analysis to the 4,740 topic share by more than one owner, the average $ATxDI$ is only 3.48.

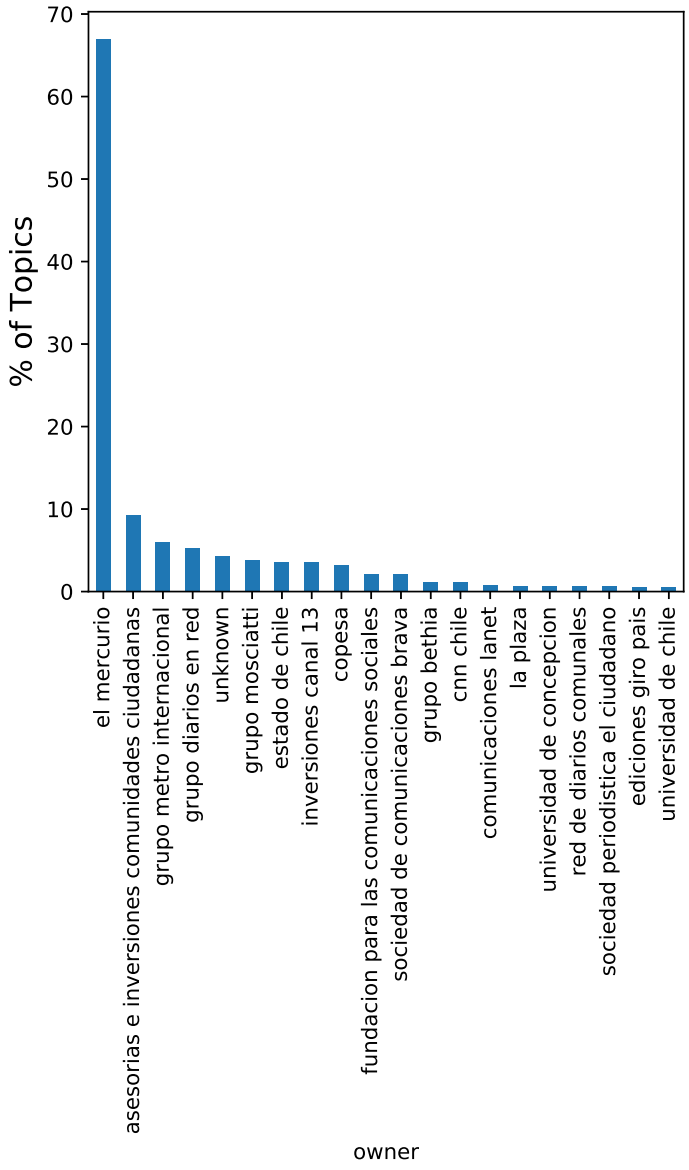

Figure 3: Percentage of topics where each type participates. Top 20 ranking (owners as types).

5. Discussion and conclusion

We applied three ecological diversity indices, which are commonly used to asses the health of biological ecosystems, namely the Shannon Diversity Index,

the Gini-Simpson Index, and the Average Taxonomic Distinctness, to assess the health of the Chilean news system as an ecosystem. Our results across the different indices suggest that the Chilean on-line news ecosystem lacks diversity, in terms of coverage, topics covered, editorial policies, and ownership, possibly leading to a deterioration of an individual's access to variety in their news coverage and news sources.

Following the analysis proposed by Polo [9], we find low external pluralism (EP) or low diversity. Topic selection seems to be driven by subjective factors rather than objective criteria, such as the newsworthiness of events. Topics are covered by only a few news outlets and even less owners. Furthermore, we find that outlets that cover the same topics exhibit high levels of content similarity, indicating a lack of independent reporting. This suggests that many outlets are not only subject to similar editorial policies, but rely on similar content. Although our results indicate relatively high numerical diversity (many news outlets), which should in principle contribute to a healthy Chilean news ecology, we observe a significant lack of *source-driven* diversity.

We found that the health of the system is considerably more critical when we use owners instead of outlets as types: we saw between 5 and 6 times more diversity for outlets according to Average Taxonomic Distinctness (ATxDI) and Simpson Index (SiDI), and almost ten times for the Shannon Diversity Index (ShDI). The lower ownership diversity vs. outlet diversity is indicative of high levels of concentration in the Chilean news market: few owners control many outlets and may influence their editorial policies (e.g., topic selection). The fact that outlets owned by the same company systematically share the same topics/clusters indicates low internal pluralism (IP). A few mega-conglomerates controlling a large number of outlets presents a clear risk to news diversity, coverage, and representativeness [8, 9].

We do not establish a cause-effect relationship between ownership and the editorial policies of their outlets. However, the final effect on the diversity that we are measuring remains relevant. When a news company buys a news outlet, it has two options: to close it so he can eliminate the competition, or to leave

it open and use it in its favor (either because it is already leaning in his favored direction, or by trying to subvert its editorial staff). Since we analyze only active accounts (i.e., they decided to leave them open) and we found a huge gap in diversity between outlets and owners, it is safe to conclude that the artificial diversity created by the multiplicity of outlets own by each company is related to the ownership structure.

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

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

However, we need to caution about some assumptions and limitations of our approach. First, our analysis pertains strictly to content that outlets *publish on twitter*. More traditional publishing methods, such as paper newspapers, may exhibit lower degrees of concentration and greater ecological health. However,

due to the higher barriers to entry of traditional publishing compared to online media, this is unlikely. Furthermore, the online distribution of news is growing rapidly to the degree that it may soon become the dominant *modus operandi*. Hence, our analysis sheds light on a phenomenon that will become increasingly important for the health and diversity of our news ecology. Second, we do not assess within-topic coverage differences, i.e., two outlets that publish tweets in the same cluster might in principle take opposite approaches to the same topic, but our analysis will not acknowledge such differences. Instead, we assume that a systematic co-occurrence in clusters implies similar interests and points of views. Third, since all news outlets in our database are Chileans, it would be interesting to analyze the penetration of international news outlets in the Chilean news system. Readers, particularly online readers, may complement their views by consuming content from more geographically remote sources, perhaps contributing to a healthier ecosystem. Still, there is some evidence which reports that: "Interest in international news varies by geographic region. Europeans are most likely to say they follow international news closely (median of 65%), while people in Latin America express the lowest level of interest in this type of news (35%) [55]. We leave this for future work. Finally, our results are descriptive in nature. They do not pertain to the causal mechanisms that define connect low ecological diversity and readership, or on the Chilean population in general.

All in all, we believe that the methods discussed in this paper can help shed light on the health status of the news systems of the world in a convenient way, by neutral parties.

Acknowledgment

This work was supported in part by the doctoral scholarships of CONICYT-PCHA No. 63130228. The second author received funding from CORFO 13CEE2-21592 (2013-21592- 985 1-INNOVA PRODUCCION2013-21592-1).

References

[1] E. Pariser, The Filter Bubble: What the Internet Is Hiding from You, Penguin Group , The, 2011.

[2] M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks, Annual review of sociology 27 (1) (2001) 415–444.

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

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

[5] M. Gentzkow, J. M. Shapiro, What drives media slant? evidence from u.s. daily newspapers, Econometrica 78 (1) (2010) 35–71.

[6] T. Besley, A. Prat, Handcuffs for the grabbing hand? media capture and government accountability, American Economic Review 96 (3) (2006) 720–736.

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

[8] A. Prat, D. Strömberg, The political economy of mass media, CEPR Discussion Paper No. DP8246.

[9] M. Polo, Regulation for pluralism in the media markets, The Economic Regulation of Broadcasting Markets: Evolving Technology and the Challenges for Policy (2005) 150–188.

[10] J. Ladyman, J. Lambert, K. Wiesner, What is a complex system?, European Journal for Philosophy of Science 3 (1) (2013) 33–67.

[11] P. Pirolli, S. Card, Information foraging., Psychological Review 106 (4) (1999) 643–675. doi:10.1037/0033-295X.106.4.643.
URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.106.4.643

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

[13] D. Mocanu, A. Baronchelli, N. Perra, B. Gonalves, Q. Zhang, A. Vespignani, The twitter of babel: Mapping world languages through microblogging platforms, PLoS ONE 8 (4) (2013) e61981.

[14] J. T. Klapper, The effects of mass communication, Free Press Glencoe, Ill, 1960.

[15] J. Bahamonde, J. Bollen, E. Elejalde, L. Ferres, B. Poblete, Power structure in chilean news media, CoRR abs/1710.06347.

[16] G. Murdock, Large corporations and the control of the communications industries, Culture, society and the media (1982) 118–150.

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

[18] E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on facebook, Science 348 (6239) (2015) 1130–1132.

[19] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonalves, F. Menczer, A. Flammini, Political polarization on twitter, in: ICWSM, 2011.

[20] J. S. Morgan, C. Lampe, M. Z. Shafiq, Is news sharing on twitter ideologically biased?, in: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, CSCW '13, ACM, New York, NY, USA, 2013, pp. 887–896.

[21] J. An, M. Cha, K. Gummadi, 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, AAAI, Menlo Park, CA, USA, 2011.

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

[23] D. Saez-Trumper, C. Castillo, M. Lalmas, Social media news communities: Gatekeeping, coverage, and statement bias, in: Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management, CIKM '13, ACM, New York, NY, USA, 2013, pp. 1679–1684.

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

[25] C. E. Shannon, W. Weaver, A Mathematical Theory of Communication, University of Illinois Press, Champaign, IL, USA, 1963.

[26] E. H. Simpson, Measurement of diversity., Nature.

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

[28] I. Flaounas, M. Turchi, O. Ali, N. Fyson, T. De Bie, N. Mosdell, J. Lewis, N. Cristianini, The structure of the eu mediasphere, PLoS ONE 5 (12) (2010) e14243.

[29] A. Z. Broder, On the resemblance and containment of documents, in: Compression and Complexity of Sequences 1997. Proceedings, IEEE, 1997, pp. 21–29.

[30] A. Z. Broder, M. Charikar, A. M. Frieze, M. Mitzenmacher, Min-wise independent permutations, in: Proceedings of the thirtieth annual ACM symposium on Theory of computing, ACM, 1998, pp. 327–336.

[31] G. Buehrer, K. Chellapilla, A scalable pattern mining approach to web graph compression with communities, in: Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM, 2008, pp. 95–106.

[32] C. Hernández, G. Navarro, Compressed representations for web and social graphs, Knowledge and information systems 40 (2) (2014) 279–313.

[33] F. Chierichetti, R. Kumar, S. Lattanzi, M. Mitzenmacher, A. Panconesi, P. Raghavan, On compressing social networks, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2009, pp. 219–228.

[34] T. Urvoy, E. Chauveau, P. Filoche, T. Lavergne, Tracking web spam with html style similarities, ACM Transactions on the Web (TWEB) 2 (1) (2008) 3.

[35] K. Berlin, S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin, A. M. Phillippy, Assembling large genomes with single-molecule sequencing and locality-sensitive hashing, Nature biotechnology.

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

[37] A. Shrivastava, P. Li, In defense of minhash over simhash, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014, 2014, pp. 886–894.

[38] U. Manber, et al., Finding similar files in a large file system., in: Usenix Winter, Vol. 94, 1994, pp. 1–10.

[39] A. Z. Broder, Identifying and filtering near-duplicate documents, in: Combinatorial pattern matching, Springer, 2000, pp. 1–10.

[40] G. S. Manku, A. Jain, A. Das Sarma, Detecting near-duplicates for web crawling, in: Proceedings of the 16th international conference on World Wide Web, ACM, 2007, pp. 141–150.

[41] A. Rajaraman, J. D. Ullman, J. D. Ullman, J. D. Ullman, Mining of massive datasets, Vol. 77, Cambridge University Press Cambridge, 2012.

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

[43] G. B. Tran, E. Herder, Detecting filter bubbles in ongoing news stories., in: A. I. Cristea, J. Masthoff, A. Said, N. Tintarev (Eds.), UMAP Workshops, Vol. 1388 of CEUR Workshop Proceedings, CEUR-WS.org, 2015.

[44] J. Molvær, J. Knutzen, J. Magnusson, B. Rygg, J. Skei, J. Sørensen, Klassifisering av miljøkvalitet i fjorder og kystfarvann, Tech. rep., Norsk institutt for vannforskning, Oslo (2004).

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

[46] R. Guimerà, L. A. Nunes Amaral, Functional cartography of complex metabolic networks, Nature (7028) (2005) 895–900.

[47] R. O. G. Gavilanes, D. Gomez, D. Parra Santander, C. Trattner, A. Kaltenbrunner, E. Graells, Language, twitter and academic conferences, in: Proceedings of the 26th ACM Conference on Hypertext & Social Media, HT '15, ACM, New York, NY, USA, 2015, pp. 159–163.

[48] S. Kim, I. Weber, L. Wei, A. Oh, Sociolinguistic analysis of twitter in multilingual societies, in: Proceedings of the 25th ACM Conference on Hypertext and Social Media, HT '14, ACM, New York, NY, USA, 2014, pp. 243–248.

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

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

[51] D. Dor, On newspaper headlines as relevance optimizers, Journal of Pragmatics 35 (5) (2003) 695 – 721.

[52] G. Tran, M. Alrifai, E. Herder, Timeline Summarization from Relevant Headlines, Springer International Publishing, Cham, 2015, pp. 245–256.

[53] J. R. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by gibbs sampling, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, 2005, pp. 363–370.

[54] E. Elejalde, L. Ferres, E. Herder, On the nature of real and perceived bias in the mainstream media, PLOS ONE 13 (3) (2018) 1–28.

[55] A. Mitchell, K. Simmons, K. E. Matsa, L. Silver, Publics around the world follow national and local news more closely than international, `https://goo.gl/T1sDLB`, accessed: 2018-04-19 (2018).