An Analysis of Novelty Dynamics in News Media Coverage

Ronaldo Cristiano Prati
Universidade Federal do ABC
Santo André, São Paulo, Brazil
ronaldo.prati@ufabc.edu.br

Walter Teixeira Lima Júnior
Univerisdade Federal do Amapá
Macapá, Amapá, Brazil
contato@walterlima.net

Abstract

Computer Science has affected almost all fields of human knowledge, contributing to scientific advances in many branches of Natural and Social Sciences. Journalism is one of the fields that is benefiting of the advance of computer science. Among the journalistic concepts that can be analyzed computationally is News Value. Novelty is one of the most important news value. A possible approach to get novelty elements in a story considers word frequency, through of the capacity to collect and analyze massive amounts of data. In this paper, we use the News Coverage Index dataset (NCI), maintained by the Pew Research Center, to analyze the novelty dynamics of news coverage, using the novelty signatures proposed by [12]. As a definition of novelty, we used the first appearance of a new lead newsmaker. Results show a good fit of the model to the dataset. Furthermore, an analysis by media sector and broad topic shows interesting insights for the analysis of media coverage.

1 Introduction

The Computational Science has affected almost all fields of human knowledge, contributing to scientific advances in many branches of Natural and Social Sciences. For instance, the capacity to collect and analyze massive amounts of data has transformed intensely fields such as biology and physics [7].

In Social Science, despite the difficulties to formalize computationally many scientific subjects of the human behavior, "a computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale" [7]. Unfortunately, most of the advances in this area have been progressing at a much slower pace. However, substantial barriers that might limit progress are being overcome in recent years. The emergence of a powerful new field of data analysis of Social Science has also influenced the research on a branch of it, Journalism. Journalism is an important social practice. Therefore, to find non-trivial information on content produced by journalism, it is necessary to count with the support of the current stage of technologies to advance in analytical techniques "Computation can advance journalism by drawing on innovations in topic detection, video analysis, personalization, aggregation, visualization, and sense making [10].

Among the journalistic concepts that can be analyzed computationally is News Value. News value as a concept was thought by Johan Galtung and Mari Holmboe Ruge's seminal publication in the Journal of Peace Research. In 1965, the paper suggested a range of attributes that establish news values in discursive elements contained in newspapers and broadcast news. Galtung and Ruge established the news values elements as Frequency; Threshold; Unambiguity; Meaningfulness; Consonance; Unexpectedness; Continuity; Composition; Reference to Elite Nations; Reference to Elite People; Reference to Persons; and Reference to Something Negative [3]. These factors have been the base to compose the structure of the theory of newsworthiness. The theory is based on the psychology of individual perception and explain which factors influence newsworthiness of an event [6].

News values are studied considering a range of at-

In: M. Martinez, U. Kruschwitz, G. Kazai, D. Corney, F. Hopfgartner, R. Campos and D. Albakour (eds.): Proceedings of the NewsIR'16 Workshop at ECIR, Padua, Italy, 20-March-2016, published at http://ceur-ws.org

tributes contained in discursive elements. It is also possible to verify the news value through a range of "more specific cognitive constraints that define news values (Novelty, Regency, Presupposition, Consonance, Relevance, Deviance and Negativity, Proximity) [2]. The news value named Novelty can be analyzed by words such as reveal or revelation. These words announce semantically 'unexpected aspects of an event News stories are frequently about happenings that surprise us, that are unusual or rare' [1].

The novelty can be understood by concepts as out of the ordinary, least expected, or not predicted, news values relating to the novelty, newness or unexpectedness of an event/happening [2]. The quality of being interesting enough to the public (newsworthiness) is also based on if a journalistic fact is out of the ordinary, it will have a greater effect than something that is an everyday occurrence (unexpectedness). The unexpectedness power of attraction is in the factor that "there is new information that has been uncovered and evaluations of importance can make the eliteness of a source explicit" [2]. This means that readers or viewers can know facts or different people or unusual to their quotidian, however, "this is the old man-bites-dog syndrome which needs little more explanation" [9, 2]. When a fact or term first come up, the human attention is captured, but "the fact that the novelty of a story tends to fade with time and thus the attention that people pay for it. This can be due to either habituation or competition from other new stories" [13].

As previously observed, novelty is also elaborated "mainly through using evaluative language, references to surprise/expectations and comparisons" [1]. This way of perception of novelty on the construction of journalistic contents is based on analyzes produced by reading the news. However, it is possible to get novelty elements in the story considering word frequency, through the capacity to collect and analyze massive amounts of data. Over the years, there is a massive increase in the availability of journalistic data and creation of new tools to extract the value from data that are helping to understand our lives, organizations, and societies.

In this paper, we used a recent model of novelty dynamics to analyze news coverage. The main idea is to analyze whether different news sources present different novelty dynamics. This paper is organized as follows: Section 2 presents novelty signatures that emerge in some dynamical processes. Section 3 describes the data set used in our study. Section 4 presents the results of applying the novelty signatures to the NCI dataset, and Section 5 concludes the paper.

2 Novelty in dynamical processes

Tria et. al [12] have recently analyzed novelty as new events occurring in a dynamical process evolving over time. Given a sequence of events, a novelty occurs whenever a new element first appears in a sequence. They have analyzed four different data sets: books from Gutenberg project Corpus, annotations in the social bookmarking platform Delicious, songs and singers at Last FM streaming portal and, entries appearance in English Wikipedia. The novelties in these data are, respectively, the occurrence of new words in books, the use of new annotation tags in the bookmarks, the inclusion of a new artist/song in a play list the user had never listen to and the first edition of a page in the collaborative encyclopedia.

They were able to model novelty as a simple mathematical model based on random draws sampling with replacement of an Urn [4] that increases when a novel item is observed. The model predicts statistical laws for the rate at which novelties happen (Heaps' law [5]) and for the probability distribution on the space explored (Zipf's law [14]), as well as signatures of the process by which one novelty sets the stage for another.

The first signature is based on quantifying the rate at which novelties occur in a temporally ordered sequence of elements of length N by analyzing the growth of the number $D(N)$ of distinct elements in this sequence. This relation would imply in a Heap's law, which states that the rate at which novelties occur decreases over time as t^{β}, where β is the coefficient of a power law distribution of $D(N)$ over N fitted over the data.

The second signature is related to the frequency of occurrence of different elements in the data. The frequency-rank distribution would follow an approximate Zipffian distribution (Zipf's law). In this distribution, the frequency of any element is inversely proportional to its rank in the frequency table, *i.e.*, the frequency $F(R)$ of an element at rank R is proportional to $R^{-\alpha}$, where α is the coefficient of a power law distribution of $F(R)$ over R fitted over the data.

It is well known that α and β are inversely correlated [8]. The larger the β coefficient, the higher the frequency of appearance of new elements in the sequence, thus there is a high propensity for novelty. On the other hand, the larger the α coefficient, the higher the occurrence of the most frequent elements in the sequence. The key result reported in [12] is that in the four data sets analyzed, the model was able to capture the novelty behavior in the data. An interesting research question is then whether News delivery also shows these novelty signatures. This paper is an initial attempt towards such analysis.

3 News Coverage Index dataset

In our analysis, we used the data gathered by the Pew Research Center[1]. Every week, this institution produced the News Coverage Index (NCI) by identifying and annotating the main subjects covered by the U.S. mainstream media. The dataset used this research is the most updated dataset (2013), published by Pew Research Center. Until this moment, no other similar dataset that can be used to update the data or serve to comparison.

> The NCI captured and analyzed 52 news outlets in real time to determine what was being covered and what was not in the U.S. news media. The analysis was conducted weekly, Monday - Sunday. The key variables included source, story date, big story, broad story topic, placement, format, geographic focus, story word count, duration of broadcast story and lead newsmaker. The outlets studied came from print, network TV, cable, online, and radio. They included evening and morning network news, several hours of daytime and prime time cable news each day, newspapers from around the country, the top online news sites, and radio, including headlines, the long form programs and talk [11].

By focusing on the topic of the story, the index measures by what percentage of the analyzed news hole is about that topic. Data were collected from January 2007 to May 2012. Table 1 presents the number of news stories collected per year. Note that the year 2012 has a few stories because the collection period ranges from January to May, rather than January to December.

Table 1: Number of news stories per year

Year	2007	2008	2009	2010	2011	2012
CableTV	22823	21892	18856	17087	15324	6472
NetworkTV	21320	19796	19427	13016	11858	5186
Newspaper	6559	7350	7370	5626	5190	1977
Online	6520	6539	7830	7818	7744	3242
Radio	13515	14365	15234	9067	8439	3570
All	70737	69942	68717	52614	48555	20447

The codebook includes variable names, definitions, applicable procedures and changes that were made to certain variables. For each story, it was annotated the date, source, broadcast start time (morning, noon, afternoon, evening and night, or not broadcast), duration in seconds, word counts, placement prominence, story format, big story, geographic focus (local, US national, US international, non-US international), broad story topic, media sector (cable TV, network TV, newspaper, online and radio), and lead newsmaker. The number of outlets and individual programs vary considerably within each media sector, as do the number of stories and size of the audience.

The index is a good source for analyzing, through time, how stories emerge and sink. Other possibilities include how the character or narrative focuses of the story change and how much of the broad topic's categories get more coverage, when compared to the others. However, the index does not provide information for additional possible questions, such as tone, sourcing or other matters.

The key variable chosen in this study was "lead newsmaker", a variable that "determines the person whose actions or statements constitute the main subject matter of the story". In the NCI, the derivation of the "lead newsmaker' variable used a methodology that examined the outlets daily by the coding team. The researchers establish as a definition: variable lead newsmaker determines the person whose actions or statements constitute the main subject matter of the story discussed with at least 50% of the story (in time or space).

Therefore, in our analysis, a news story is flagged as a novelty whenever the first appearance of a new lead newsmaker occurs, considering an ordered sequence of histories by date in the NCI. Obviously, this approach does not completely capture all the aspect of novelty in news coverage. It is perfectly possible (and indeed very common) that some new factor is being published by some lead newsmaker who appeared before. However, this approach does capture some aspect of novelty, in a sense that different subjects are being noticed in the media. Furthermore, the approach sheds some interesting insights, as discussed next.

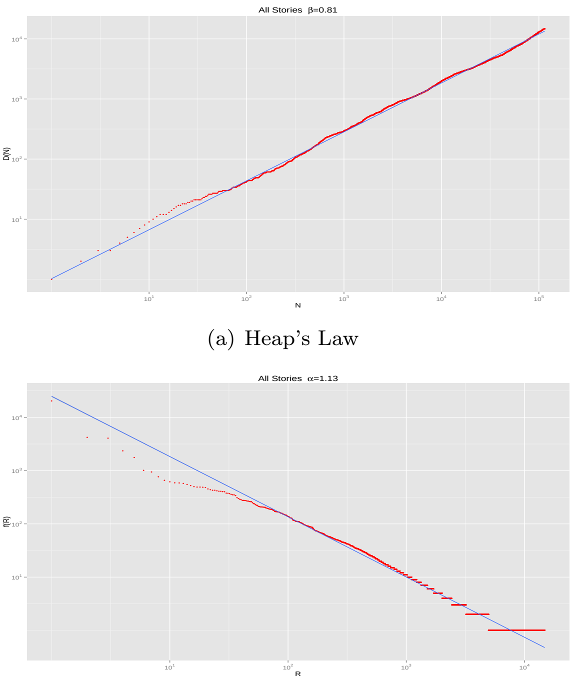
4 Results and Discussion

In this section we present the results of the novelty signatures as proposed by [12] to the NCI dataset. As the main variable used in this study was lead newsmaker, we removed from the dataset all stories where the lead maker was not identified, resulting in a total of 135,205 entries in the dataset.

Figure 1 shows the two novelty signatures for all stories in the NCI dataset, for the Heaps' law and Zipf's law, respectively. The graphs show a very good fit (the blue line in the graphs), indicating that the dynamic of novelties also follows the model proposed in [12] for the NCI dataset.

This is an interesting result per se, but we can move beyond that by conditioning the analysis by some news groups. Figure 2 does this, where we have split the analysis by the media sector (newspaper, online, radio,

[1] Formerly Project for Excellence in Journalism (PEJ).



(a) Heap's Law

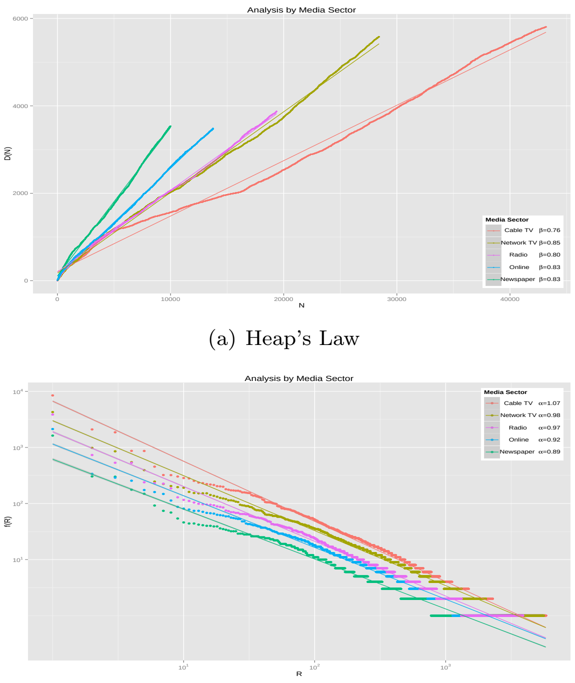
(b) Zipf's Law

Figure 1: Novelty Signatures for lead newsmaker over all stories in NCI dataset. The blue line is the best data fit.

broadcast TV and cable TV). Figure 2(a) shows how novel lead newsmakers appears in the news sequence, for each media sector collected by the NCI. The interpretation of these results is, the steeper the line, the more novelty the media sector has (according to the definition of novelty used in this paper). Surprisingly, newspapers is the media

sector with the larger ratio of lead newsmakers per story, followed by online portals, radio, network TV and cable TV. Figure 2(b) shows an orthogonal insight for this result, which shows the rank distribution of lead newsmakers for each sector. As Heaps' law and Zipfs' law are inverse correlated, the interpretation of these results are, the steeper the line, the more a media sector concentrates the coverage in a few lead makers. Cable TV repeats lead newsmakers more often than other sectors, and (proportionally) uses fewer leads newsmaker than the other media sectors. Newspapers, on the other hand, proportionally use the top ranked lead news makers less often, and have a larger number of histories with different lead makers.

We can speculate that the higher frequency of novel lead newsmakers in newspaper media is due to that this media needs competitiveness in relation to other media (digital and electronic), which are characterized by dissemination of news in real time. As the newspaper is a diary media, it always needs to have something different to present than what was published on the previous day on TV, radio and, Internet. Despite being late in relation to events in one day (it generally publishes stories from the eve), the newspaper still continues to be a source for other rival media because it intends always having something new in their pages. On the other hand, TVs have a rotating audience, and focus on a narrow range of topics. Thus, the presented histories focus in a few lead newsmakers. Radios an online media are somehow in between these two extremes.



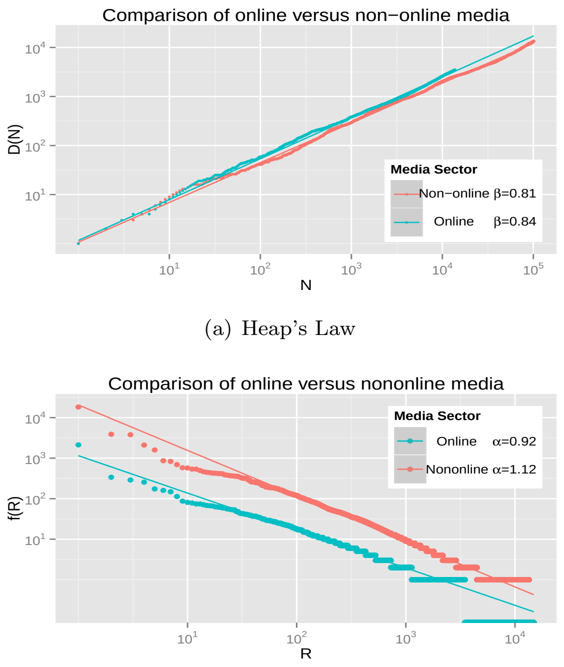
(a) Heap's Law

(b) Zipf's Law

Figure 2: Novelty Signatures for lead newsmaker in NCI dataset grouped by media sector

To gain some insight in the online versus offline scenario, we break down the analysis in online versus offline media, as shown in Figure 3. The interpretation of the graphs is the same as of 2. Figure 3(a) shows that online sector introduce more lead makers in their stories, and Figure 3(b) indicate that the same lead maker appears less often in online media. As can be seen from the graphs, online media have stronger novelty signatures. Therefore, online media have a bias towards introducing more different lead newsmakers, and a lower tendency to echo the same leading maker in future stories.

A possible reason for this is that online outlets have a high propensity to show new stories due to the difference in media consumption from the target audience. In general, the audience for online news sources is of younger people (as discussed in the previous section). These users have a less tendency to in-depth stories, fo-



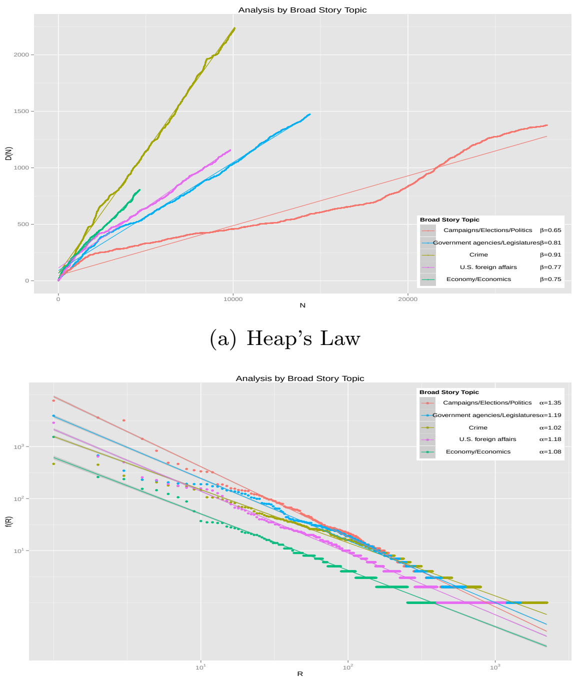
(a) Heap's Law

(b) Zipf's Law

Figure 3: Novelty Signatures for lead newsmaker in NCI dataset grouped by online versus offline media

cusing in the headlines. They also are more connected, and access the news more often, thus the necessity of novelty in the news stories.

We did a similar analysis, but conditioning on the five most frequent broad story topics. The broad story topic variable identifies which of the broad topic categories is addressed by a story. NCI has 32 broad story categories, but most of them have low frequencies. These low frequencies difficult an analysis, due to a lack of data. Figure 5 shows these results. The interpretation of the graphs is the same as of 2, except for the fact that instead of media sectors, we have topics in these graphs. Figure 4(a) shows how novel lead newsmakers appear in the news sequence, for each of the five most frequent topics collected by the NCI. In these figures, one traditional attribute of news value, Negativity (any reference that is negative), emerges as through Crime broad story topic. Crime is the topic with the largest rate of novel lead makers, followed by economy/economics, US foreign affairs, government agencies/legislatures and campaigns/elections/politics with the lowest rate. Figure 4(b) indicates that the most frequent lead makers appear proportionally less often in the news than the most frequent lead makers in campaigns/elections/politics. Furthermore, crime is the sector with the largest proportion of lead makers to appear in fewer histories.



(a) Heap's Law

(b) Zipf's Law

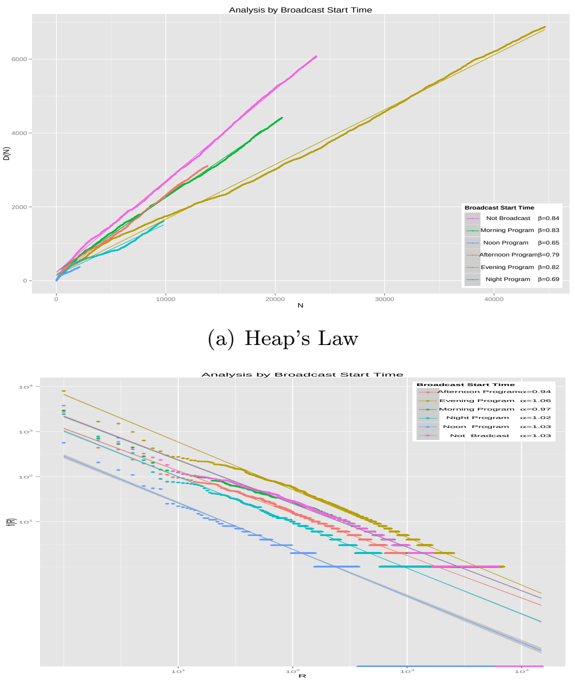
Figure 4: Novelty Signatures for lead newsmaker in NCI dataset grouped by media sector (top 5 sectors)

A similar analysis was performed by start time of the program, as shown in Figure 6. The interpretation of the graphs is the same as of 2, except for the fact that instead of media sectors, we have the program start time in these graphs. Figure 5(a) shows that, in general, morning programs introduce more often new lead makers, while night programs have few novelty lead makers. On the other hand, Figure 5(b) shows that night program cites more often the more noticed lead makers than morning programs. We believe this also is related to the target audience, which in the evening/night has a higher prevalence of elderly people, which is more interested in-depth coverage.

5 Concluding Remarks

In this paper, we examine the dynamic of novelties in the NCI dataset. We used the lead newsmaker as the main variable to define the concept of novelty in our framework. We verified a very good fit of these data to the two novelty signatures discussed in [12].

We obtained interesting and insightful insights when conditioning the analysis do media sector and broad story topic. Regarding media sector, we verified that newspapers is the sector with largest novelty, in terms of the introduction of new lead newsmakers. Furthermore, online media have a largest novelty, when compared to non-line media. In terms of story topic, crime is the sector with more novelty, also in terms of lead newsmakers.



(a) Heap's Law

(b) Zipf's Law

Figure 5: Novelty Signatures for lead newsmaker in NCI dataset grouped by starting time

We believe these patterns somehow tend to follow the interest of the public in order to get her attention. Thus, there is the necessity to provide news on topics to better reach a target audience, tailoring the audience. An interesting future work is to analyze whether these patterns would be similar in the next years, because the online young audience became a generation more mature. Would the behavior be the same, and the sectors have to adapt to the news consumption patterns of this generation or they will change their tastes, showing a similar behavior or their previous generation as accessing the in-depth stories?

This research has two obvious limitations. First, our adopted definition of novelty does not capture all aspects of novelty, as new information can be published about lead Newsmakers which already appeared in the sequence. However, we believe this definition do capture some aspects of novelty, and were able to provide some interesting insights on the topic. Furthermore, the data set has a bias towards the U.S.A. media coverage. An interesting further research direction is to broaden this research to different sources.

References

[1] M. Bednarek and H. Caple. 'value added': Language, image and news values. *Discourse, Context & Media*, 1(2–3):103–113, 2012.

[2] H. Caple and M. Bednarek. Delving into the discourse: Approaches to news values in journalism studies and beyond. Technical report, Reuters Institute for the Study of Journalism, 2013.

[3] J. Galtung and M. H. Ruge. The structure of foreign news the presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1):64–90, 1965.

[4] J. Haigh. Polya urn models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):942–942, 2009.

[5] H. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, New York, 1978.

[6] H. Kwak and J. An. Understanding news geography and major determinants of global news coverage of disasters. In *Computer+Journalism Symposium*, New York, USA, 2014.

[7] D. Lazer, A. Pentland, A. Lada, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Life in the network: the coming age of computational social science. *Science*, 323(5915):721–723, 2009.

[8] L. Lü, Z.-K. Zhang, and T. Zhou. Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5(12):e14139, 12 2010.

[9] M. Masterton. Asian journalists seek values worth preserving. *Asia Pacific Media Educator*, 1(16):41–48, 2005.

[10] B. O'Connor, D. Bamman, and N. A. Smith. Computational text analysis for social science: Model assumptions and complexity. In *Second NIPS Workshop on Comptuational Social Science and the Wisdom of Crowds*, 2011.

[11] Pew Research Center. News coverage index methodology, 2013. `http://www.journalism.org/news_index_methodology/99/`.

[12] F. Tria, V. Loreto, V. D. P. Servedio, and S. H. Strogatz. The dynamics of correlated novelties. *Sci. Rep.*, 4, 2014.

[13] F. Wu and B. A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104(45):17599–17601, 2007.

[14] G. K. Zipf. The psycho-biology of language. *Language*, 12(3):196–210, 1935.